národní
úložiště
šedé
literatury

Error Estimates for Approximate Optimization by the Extended Ritz Method

Kůrková, Věra
2002

Dostupný z http://www.nusl.cz/ntk/nusl-34100

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 17.07.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní nusl.cz .

Institute of Computer Science
Academy of Sciences of the Czech Republic

Error estimates for approximate optimization by the extended Ritz method

Věra Kůrková and Marcello Sanguineti

Technical report No. 882

October 2002

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: (+4202) 6605 1111, fax: (+4202) 86 58 57 89,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Error estimates for approximate optimization by the extended Ritz method[1]

Věra Kůrková[2] and Marcello Sanguineti[3]

Technical report No. 882

October 2002

Abstract:

An alternative to the classical Ritz method of approximate optimization is investigated. In the extended Ritz method, sets of admissible solutions are approximated by their intersections with linear combinations of n-tuples from a given set. This approximation scheme, called variable-basis approximation, includes functions computable by trigonometric polynomials with free frequencies, neural networks, free-node splines, and many other nonlinear approximating families. Estimates of rates of approximate optimization by the extended Ritz method are derived. For problems with argminima, upper bounds on rates of convergence of approximate infima and argminima to a global infimum and argminimum are expressed in terms of the "degree" n of variable-basis functions, of the modulus of continuity of the functional to be minimized, of the modulus of Tychonov well-posedness of the problem, and of certain norms tailored to the type of variable basis. Classes of high-dimensional optimization problems are described, for which rates of approximate optimization do not exhibit the curse of dimensionality with respect to the number of variables of admissible solutions. The results are applied to convex best approximation problems and kernel methods in machine learning.

Keywords:
approximate optimization, extended Ritz method, rates of convergence of approximate infima and argminima, high-dimensional optimization problems, curse of dimensionality, convex best approximation, learning from data.

[1]Collaboration between V. K. and M. S. was supported by the Scientific Agreement Italy-Czech Republic, Area MC 6, Project 22: "Functional Optimization and Nonlinear Approximation by Neural Networks." V. K. was partially supported by GA ČR Grant 201/02/0428 and M. S. by a CNR - Agenzia 2000 Grant, Project "New Algorithms and Methodologies for the Approximate Solution of Nonlinear Functional Optimization Problems in a Stochastic Environment."

[2]Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, P.O. Box 5 – 182 07, Prague 8, Czech Republic – vera@cs.cas.cz.

[3]Department of Communications, Computer, and System Sciences (DIST), University of Genova Via Opera Pia 13, 16145 Genova, Italy – marcello@dist.unige.it.

1 Introduction

In many high-dimensional optimization problems (e.g., routing in communications networks, closed-loop optimal control, inventory problems, optimal management of water resources, large-scale traffic networks, etc. [10], [14], [26], [46]), theoretically optimal admissible solutions cannot be found analytically or, even when they can be found, they may not be implementable. However, optimal solutions can be approximated by suboptimal ones that are implementable. The classical Ritz method [23] considers a sequence of approximate solutions achievable over intersections of the original set of admissible solutions with a nested family of linear subspaces of increasing dimensionality.

Although linear approximation methods have many convenient properties, their practical applications are limited by the "curse of dimensionality" [11], i.e., an exponential growth with the number of variables of the dimension of a linear subspace needed for a given accuracy of optimization. Experimental results confirm that the Ritz method is often unable to deal efficiently with high-dimensional optimization tasks [46]. However, a systematic theoretical study of rates of approximation for the Ritz method have not yet been made. The estimates available in the literature [4], [42], [22], [43] either are formulated in the case of only one variable or do not explicitly state their dependence on the number of variables of admissible solutions.

In [46], an alternative to the classical Ritz method, called *the extended Ritz method*, was introduced. According to this method, instead of linear subspaces of increasing dimensionality, a nested family of so-called variable-basis functions is used to approximate admissible functions. The variable-basis approximation scheme includes a variety of nonlinear approximators such as free-node splines, trigonometric polynomials with free frequencies, and feedforward neural networks [27], [32]. The introduction of the extended Ritz method was motivated by successful applications of feedforward neural networks for the approximate solution of high-dimensional optimization problems [2], [3], [6], [7], [8], [12], [13], [34], [35], [36], [37], [38], [41], [45], [46]. When such networks are used as a variable basis, the extended Ritz method reduces the original optimization task to a nonlinear programming problem in which the optimal values of the network parameters can be determined by means of a suitable descent algorithm such as the backpropagation one [40].

In this paper, we investigate the extended Ritz method theoretically. We derive upper bounds on the speed of convergence of approximate infima and argminima over nested families of variable-basis functions to the global infima and argminima. The upper bounds are formulated in terms of the "degree" n of variable-basis functions, norms tailored to the type of basis, the modulus of continuity of the functional to be minimized and the modulus of well-posedness of the problem. By inspection of these bounds we obtain a description of high-dimensional optimization problems for which the extended Ritz method does not exhibit the curse of dimensionality. As our estimates are not merely asymptotic, they allow one to estimate the quality of approximate solutions achievable over admissible sets that are implementable.

We illustrate our results on two examples. The first is the convex best approximation problem and the second is learning from data, modeled as the minimization of the so-called regularized empirical error functional.

The paper is organized as follows. Section 2 introduces basic concepts and results from optimization theory that we use throughout the paper. Section 3 describes the variable-basis approximation scheme and the extended Ritz method. Section 4 contains our main results on the speed of convergence of the extended Ritz method, and Section 5 states their refinements for convex optimization problems. Sections 6 and 7 apply our estimates to convex best approximation problems and kernel methods in machine learning, resp. Section 8 provides a brief discussion.

2 Preliminaries

By a normed linear space $(X, \|.\|)$ we mean a *real normed linear space*. $\mathcal{R}$ denotes the set of real numbers and $\mathcal{R}_+$ the set of positive reals. For a positive integer d, $\Omega \subseteq \mathcal{R}^d$ and

$p \in [1,\infty)$, $(\mathcal{L}_p(\Omega), \|.\|_p)$ denotes the space of measurable, real-valued functions on Ω such that $\int_\Omega |f(x)|^p\,dx < \infty$ endowed with the $\mathcal{L}_p$-norm.

A ball, a sphere, resp., of radius r centered at $h \in X$ is denoted by $B_r(h, \|.\|) = \{f \in X : \|f - h\| \leq r\}$, $S_r(h, \|.\|) = \{f \in X : \|f - h\| = r\}$. We write shortly $B_r(\|.\|) = B_r(0, \|.\|)$ and $B_r(h) = B_r(h, \|.\|)$, $B_r = B_r(0)$ when it is clear which norm is used; similarly for spheres. Sequences (of real numbers, sets or elements of normed linear spaces) are denoted by $\{x_n\}$ instead of $\{x_n : n \in \mathcal{N}_+\}$, where $\mathcal{N}_+$ is the set of positive integers.

A functional $\Phi : X \to (-\infty, +\infty]$ is called *proper* if it is not identically equal to $+\infty$. The set $dom\,\Phi = \{f \in X : \Phi(f) < +\infty\}$ is called the *domain of* Φ.

Φ is *continuous* at $f \in dom\,\Phi$ if for all $\varepsilon > 0$ there exists $\eta > 0$ such that for every $g \in dom\,\Phi$, $\|f - g\| < \eta$ implies $|\Phi(f) - \Phi(g)| < \varepsilon$ and the *modulus of continuity* of Φ at f is the function $\alpha_f : \mathcal{R}_+ \to \mathcal{R}_+$ defined as $\omega_f(t) = \sup\{|\Phi(f) - \Phi(g)| : f, g \in dom\,\Phi, \|f - g\| \leq t\}$. We write merely α instead of α_f when f is clear from the context. Φ is *Lipschitz continuous* on M with a Lipschitz constant c if for all $f, g \in M$, $|\Phi(f) - \Phi(g)\| \leq c\|f - g\|$.

A functional Φ is *convex* on a convex set $M \subseteq X$ if for all $h, g \in M$ and all $\lambda \in [0,1]$, $\Phi(\lambda h + (1-\lambda)g) \leq \lambda\Phi(h) + (1-\lambda)\Phi(g)$. Φ is *uniformly convex* on a convex set $M \subseteq X$ if there exists a non-negative function $\delta : \mathcal{R}_+ \to \mathcal{R}_+$, such that $\delta(0) = 0$, $\delta(t_0) > 0$ for some $t_0 > 0$ and for all $h, g \in M$ and all $\lambda \in [0,1]$, $\Phi(\lambda h + (1-\lambda)g) \leq \lambda\Phi(h) + (1-\lambda)\Phi(g) - \lambda(1-\lambda)\delta(\|h - g\|)$. Any such function δ is called a *modulus of convexity* of Φ [44]. The functional Φ is called *strictly uniformly convex* on M if $\delta(t) > 0$ for all $t \in \mathcal{R}_+$.

Using standard notation [19], we denote by (M, Φ) the problem of infimizing a functional Φ over a subset M of X. M is called a set of *admissible solutions* or *admissible set*. When both M and Φ are convex, (M, Φ) is called a *convex optimization problem*.

A sequence $\{g_n\}$ of elements of M is called Φ-*minimizing over* M if $\lim_{n\to\infty} \Phi(g_n) = \inf_{g\in M} \Phi(g)$. By the definition of infimum, for any problem (M, Φ) with M non-empty, there always exists a minimizing sequence. We denote by $argmin\,(M, \Phi) = \{g^o \in M : \Phi(g^o) = \inf_{g\in M} \Phi(g)\}$ the set of *argminima* of the problem (M, Φ) and for $\varepsilon > 0$, we denote by $argmin_\varepsilon(M, \Phi) = \{g^\varepsilon \in M : \Phi(g^\varepsilon) < \inf_{g\in M} \Phi(g) + \varepsilon\}$ the set of its ε-*near argminima*.

The following proposition summarizes well-known elementary properties of uniformly convex functionals.

Proposition 2.1 *Let $(X, \|.\|)$ be a normed linear space, $M \subseteq X$ convex, Φ be a uniformly convex functional on M with a modulus of convexity δ. Then*
(i) if Ψ is convex on M, then $\Phi + \Psi$ is uniformly convex on M with a modulus of convexity δ;
(ii) if $\Phi : X \to \mathcal{R}$, then for every $f \in X$ the translated functional $\Phi(\cdot - f)$ is uniformly convex on $M - f$ with a modulus of convexity δ;
(iii) if $g^o \in argmin(M, \Phi)$ then for every $g \in M$, $\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$;
(iv) if $(X, \|.\|)$ is a Hilbert space, then the functional $\|.\|^2 : X \to \mathcal{R}$ is uniformly convex with modulus of convexity $\delta(t) = t^2$.

Proof. (i) and (ii) follow directly from the definitions.
(iii) By the definition of uniform convexity, for every $\lambda \in [0,1]$ we have $\lambda(1-\lambda)\delta(\|g - g^o\|) \leq \lambda\Phi(g) - (1-\lambda)\Phi(g^o) - \Phi(\lambda g + (1-\lambda)g^o)$. As $\Phi(g^o) \leq \Phi(\lambda g + (1-\lambda)g^o)$, we get $\lambda(1-\lambda)\delta(\|g - g^o\|) \leq \lambda\Phi(g) + (1-\lambda)\Phi(g^o) - \Phi(g^o) = \lambda\,(\Phi(g) - \Phi(g^o))$. Hence $(1-\lambda)\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$. Taking the infimum over λ, we obtain $\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$.
(iv) It is easy to check that for every $h, g \in X$ and $\lambda \in [0,1]$, we have $\|\lambda h + (1-\lambda)g\|^2 \leq \lambda\|h\|^2 + (1-\lambda)\|g\|^2 - \lambda(1-\lambda)\|h - g\|^2$. □

The problem (M, Φ) is *Tychonov well-posed* if it has a unique minimum to which every minimizing sequence converges [19, p. 1]. The *modulus of Tychonov well-posedness of* (M, Φ) *at an argminimum* g^o is a function $\xi_{g^o} : \mathcal{R}_+ \to \mathcal{R}_+$ such that for every $t \in \mathcal{R}_+$, $\xi_{g^o}(t) = \inf_{g\in M\cap S_t(g^o)} \Phi(g) - \Phi(g^o)$. Note that the modulus of Tychonov well-posedness is defined for any problem that has an argminimum even when such a problem is not Tychonov well-posed.

The *linear span* of M is $span\,M = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R},\ g_i \in M,\ n \in \mathcal{N}_+\}$. The *topological interior* of M is $int\,M = \{g \in M : (\exists\,\varepsilon > 0)\,(B_\varepsilon(g) \subset M)\}$ and its *closure* is $cl\,M = \{f \in X :$

$(\forall \varepsilon > 0)\,(B_\varepsilon(f) \cap M) \neq \emptyset)\}$. If $clM = Y$, then M is said to be *dense* in Y. The *diameter* of M is defined as $diam\, M = \sup\{\|f - g\| :\ f, g \in M\}$.

The *Minkowski functional* of $M \subseteq X$ is the functional $p_M : X \to [0, +\infty]$ defined for every $f \in X$ as $p_M(f) = \inf\{\lambda \in \mathcal{R}_+ :\ f/\lambda \in M\}$. M is called *absorbing* if $dom\, p_M = X$. For every M, p_M is positively homogeneous and for M convex, p_M is convex. The following proposition states elementary properties of Minkowski functionals of convex sets containing zero, which will be used in our proofs.

Proposition 2.2 *Let $(X, \|.\|)$ be a normed linear space, $M \subseteq X$ and $r_0 = sup\{r > 0 : B_r(\|.\|) \subseteq M\}$. Then the following hold:*
(i) if M is convex with $0 \in M$, then $M \subseteq \{f \in X : p_M(f) \leq 1\}$;
(ii) if M is convex with $0 \in M$, then $\{f \in X : p_M(f) \leq 1\} \subseteq M$;
(iii) if M is closed and convex with $0 \in M$, then $M = \{f \in X : p_M(f) \leq 1\}$;
(iv) if $0 \in int\, M$, then $dom\, p_M = X$;
(v) if $0 \in int\, M$ and $r_0 < \infty$, then for every $f \in X$, $p_M(f) \leq \|f\|/r_0$;
(vi) if M is convex and $0 \in intM$, then p_M is Lipschitz on X with constant $c = 1/r_0$ if $r_0 < \infty$ and $c = 0$ if $r_0 = \infty$.

Proof. (i) By the definition of p_M, $f \in M$ implies $p_M(f) \leq 1$ and so $M \subseteq \{f \in X : p_M(f) \leq 1\}$.

(ii) Let $f \in X$ be such that $p_M(f) < 1$. By the definition of p_M, there exists $\lambda \leq 1$ such that $f/\lambda \in M$. As M is convex and $0 \in M$, $f = \lambda\,(f/\lambda) + (1 - \lambda)\,0 \in M$.

(iii) By (i) and (ii), it is sufficient to check that for every $f \in X$ with $p_M(f) = 1$, $f \in M$. By the definition of p_M, there exists a sequence $\{\lambda_i\}$ such that $\lim_{i\to\infty} \lambda_i = 1$ and for every i, $f/\lambda_i \in M$. As M is closed and $f = \lim_{i\to\infty}(f/\lambda_i)$, we have $f \in M$.

(iv) and (v) As $0 \in int\, M$, there exists $r > 0$ such that $B_r(0) \subseteq M$. So for every $f \in B_r(0)$, $p_M(f) \leq 1$. Let $g \in X$. Then, $p_M(g) = p_M\,(r\,\|g\|\,(g/r\,\|g\|))$ and by the positive homogeneity of p_M, $p_M(g) = (\|g\|/r)\,p_M\,(r\,(g/\|g\|))$. As $\|r\,g/\|g\|\| = r$, we have $r\,g/\|g\| \in B_r(0)$ and so $p_M(g) = (\|g\|/r)\,p_M\,(r\,(g/\|g\|)) \leq \|g\|/r \leq \|g\|/r_0 < \infty$.

(vi) When M is convex, p_M is also convex. By the convexity and positive homogeneity of p_M, we have $(1/2)p_M(f) = p_M\,((1/2)f) = p_M\,((1/2)g + (1/2)(f - g)) \leq (1/2)p_M(g) + (1/2)p_M(f - g)$. Thus, $p_M(f) - p_M(g) \leq p_M(f - g) \leq \|f - g\|/r_0$. By exchanging the roles of f and g, we obtain the inequality $-\|f - g\| \leq p_M(f) - p_M(g)$. Hence $|p_M(f) - p_M(g)| \leq \|f - g\|/r_0$. □

3 Variable-basis approximation and the extended Ritz method

The classical *Ritz method* [23, p. 192] for approximate optimization replaces the problem (M, Φ) with a sequence of problems

$$\big\{\big(M \cap X_n, \Phi\big)\big\},$$

where, for each n, X_n is an n-dimensional subspace of X. Under suitable conditions on Φ, M, and $\{X_n\}$ (such as continuity of Φ, compactness of M, and density of $\bigcup_{n\in\mathcal{N}_+} M \cap X_n$ in X), for every n there exists an argminimum g_n of the approximate problem $(M \cap X_n, \Phi)$, the sequence $\{g_n\}$ converges to some $g^o \in M$, and $\lim_{n\to\infty} \Phi(g_n) = \Phi(g^o)$.

Typically, the subspaces X_n are generated by the first n elements of a subset of X with a fixed linear ordering. So this approximation scheme can be called *fixed-basis approximation* in contrast to *variable-basis approximation*, which uses nonlinear approximating sets formed by linear combinations of at most n elements of a given subset G of X. Such sets are denoted by $span_n\, G = \{\sum_{i=1}^n w_i g_i :\ w_i \in \mathcal{R},\ g_i \in G\}$. The variable-basis approximation scheme includes splines with free nodes, trigonometric polynomials with free frequencies, and feedforward neural networks [27], [32].

An alternative to the classical Ritz method consists in approximating an admissible set by its intersections with a nested sequence of the form $\{span_n G\}$. For G formed by a parameterized

family $G = \{g_a : a \in A\}$ where $A \subseteq \mathcal{R}^p$, this approach was introduced in a series of papers (see [2], [3], [6], [7], [36], [37], [38], [45], [46] and the references therein) and formalized in [46] as the *extended Ritz method*. Here we use the term extended Ritz method for optimization over the intersection of an admissible set with a nested sequence of the form $\{span_n G\}$, for a general set G. This includes the important class of admissible sets computable by neural networks with one hidden layer containing n computational units computing functions from the set G (for example, G can be formed by functions computable by perceptrons, radial-basis units, etc. [29], [30].

Sets $span_n G$ are not convex and so when the classical Ritz method is replaced with the extended one, the existence of argminima over approximate admissible sets might be lost. However, argminima can be replaced with ε_n-near argminima and a sequence of ε_n-near argminima might converge to a global argminimum much faster than in the case of the classical Ritz method. Indeed, the union of subspaces spanned by all n-tuples of elements of a set G is "much larger" than a single n-dimensional subspace generated by the first n elements of G and so the functional to be minimized might achieve in such unions of subspaces values that are closer to the global argminimum.

To estimate rates of convergence of approximate infima and argminima for the extended Ritz method, we take advantage of a result from nonlinear approximation theory by Maurey [39], Jones [25], and Barron [9]. Here we use its reformulation in terms of a norm tailored to a given basis G. Such a norm, called G-variation and denoted by $\|.\|_G$, was introduced in [28] for a subset G of a normed linear space $(X, \|.\|)$ as the Minkowski functional of the set $cl\, conv\, (G \cup -G)$. Thus,

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

G-variation is a norm on the subspace $\{f \in X : \|f\|_G < \infty\} \subseteq X$ satisfying $\|.\| \leq s_G \|.\|_G$. When G is an orthonormal basis of a separable Hilbert space, G-variation is equal to the l_1-*norm with respect to* G, defined for every $f \in X$ as $\|f\|_{1,G} = \sum_{g \in G} |f \cdot g|$ [33], [31]. Besides being a generalization of the notion of l_1-norm, G-variation is also a generalization of the concept of total variation studied in integration theory [9].

The next theorem is a reformulation in terms of G-variation of estimates derived by Maurey [39], Jones [25], and Barron [9] for Hilbert spaces and of their extension by Darken et al. [17] to $\mathcal{L}_p$-spaces, $p \in (1, \infty)$. We shall refer to Theorem 3.1 (i) as MJB theorem or MJB bound. For $t > 0$, we define

$$G(t) = \{wg : g \in G, w \in \mathcal{R}, |w| \leq t\}.$$

Theorem 3.1 *Let $(X, \|.\|)$ be a normed linear space, G its bounded subset and $s_G = \sup_{g \in G} \|g\|$. For every $f \in X$ and every positive integer n, the following hold:*

(i) if $(X, \|.\|)$ is a Hilbert space, then

$$\|f - span_n G\| \leq \|f - conv_n\, G(\|f\|_G)\| \leq \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}}.$$

(ii) if $\Omega \subset \mathcal{R}^d$ is compact and $(X, \|.\|) = (\mathcal{L}_p(\Omega), \|.\|_p)$, $p \in (1, \infty)$, then

$$\|f - span_n\, G\| \leq \|f - conv_n\, G(\|f\|_G)\| \leq \frac{2^{1/\bar{p}+1} s_G \|f\|_G}{n^{1/\bar{q}}},$$

where $q = p/(p-1)$, $\bar{p} = \min(p, q)$, and $\bar{q} = \max(p, q)$.

Proof. (i) See [28] and [30].
(ii) By [17, Theorem 5], for every $S \subseteq X$, every $f \in cl\, conv\, S$, every $r > 0$ such that $S \subseteq B_r(f, \|.\|)$, every $\varepsilon > 0$, and every $n \in \mathcal{N}_+$, there exists $f_n \in conv_n\, S$ such that $\|f - f_n\| \leq \frac{2^{1/\bar{p}} r + \varepsilon}{n^{1/\bar{q}}}$. Setting $S = \{w\, g : g \in G, \, |w| \leq \|f\|_G\}$ and $r = 2\, s_G \|f\|_G$, we get for every $g \in S$ and $f \in X$, $\|g - f\| \leq \|g\| + \|f\| \leq s_G \|f\|_G + s_G \|f\|_G = r$, as for every $f \in X$, $\|f\| \leq s_G \|f\|_G$. Thus $S \subseteq B_r(f, \|.\|)$ and so we can apply [17, Theorem 5] to obtain for every $f \in X$, $\varepsilon > 0$ and

$n \in \mathcal{N}_+$, $f_n \in conv_n S \subseteq span_n G$ such that $\|f - f_n\| \leq \frac{2^{\frac{1}{\bar{p}}+1} s_G \|f\|_G + \varepsilon}{n^{1/\bar{q}}}$. Hence $\|f - span_n G\| \leq \frac{2^{\frac{1}{\bar{p}}+1} s_G \|f\|_G}{n^{1/\bar{q}}}$. □

As for any number d of variables of the functions in X, the bounds from Theorem 3.1 (i), (ii) are of the orders of $\mathcal{O}(n^{-1/2})$ and $\mathcal{O}(n^{-1/\bar{q}})$, resp., some authors called them "dimension-independent." However, this is misleading as both s_G and balls of fixed radii in G-variation depend on d (for properties of balls in G-variation, see [9], [33], [32], and [30]).

4 Rates of approximate optimization over variable-basis functions

In this section, we investigate approximate optimization over variable-basis functions of a problem (M, Φ) that has an argminimum. This assumption is satisfied by various convex problems in reflexive Banach spaces (e.g, the minimization of a lower semicontinuous uniformly convex functional over a closed convex admissible set [16] or the minimization of a convex lower semicontinous proper functional over a closed convex bounded set [20, p. 35]). Such problems are often derived by regularization [19, p. 29] from problems that do not have an argminimum. So the following results apply to a wide class of regularized problems.

Let g^o be an argminimum of a problem (M, Φ) to which the extended Ritz method based on the approximation of M by sets $M \cup span_n G$ is applied. As the existence of argminima of approximate problems $(M \cap span_n G, \Phi)$ is not guaranteed, we can only consider ε_n-near argminima. To estimate the speed of convergence of these ε_n-argminima to the global argminimum g^o, we take advantage of MJB theorem. But we cannot apply it directly as MJB bound estimates the distance of g^o from $span_n G$, not from $M \cap span_n G$. The following technical lemma, which extends a result from [43], allows us to construct an auxiliary sequence of elements of $M \cap span_n G$, to which MJB bound can be applied (see Figure 4.1).

Lemma 4.1 *Let A and M be subsets of a normed linear space $(X, \|.\|)$, M be closed and convex, $0 \in M$, and $\lambda A \subseteq A$, for all $\lambda \in [0, 1)$. Then for every $g \in M$ and every $f \in A$ with $p_M(f) < +\infty$, there exists $h \in M \cap A$ such that*
(i) $\|h - g\| \leq \|f - g\| + \|g\| \left|p_M(f) - p_M(g)\right|$;
(ii) if $0 \in int\, M$, then $\|h - g\| \leq (1 + c\,\|g\|)\,\|f - g\|$, where c is the Lipschitz constant of p_M on X.

Proof. (i) When $f \in A \cap cl\, M$, the estimate holds trivially with $h = f$. If $f \in A - cl\, M$, then $f \neq 0$ and so we can set $h = \frac{p_M(g)}{p_M(f)} f$. Hence $p_M(h) = p_M(g) \leq 1$, and by Proposition 2.2 (ii), $h \in M$. As $f \notin M$ again by Proposition 2.2 (ii), we have $p_M(f) > 1$. Thus $h = \frac{p_M(g)}{p_M(f)} f$ with $\frac{p_M(g)}{p_M(f)} < 1$ and $f \in A$, which implies $h \in A$. Hence $h \in A \cap M$ and $\|h - g\| = \left\|\frac{p_M(g)}{p_M(f)} f - g\right\| = \left\|\frac{p_M(g)}{p_M(f)}(f - g) - \left(1 - \frac{p_M(g)}{p_M(f)}\right) g\right\| \leq \left|\frac{p_M(g)}{p_M(f)}\right| \|f - g\| + \left|1 - \frac{p_M(g)}{p_M(f)}\right| \|g\| < \|f - g\| + \left|\frac{p_M(f) - p_M(g)}{p_M(f)}\right| \|g\| < \|f - g\| + |p_M(f) - p_M(g)|\, \|g\|$.

(ii) If $0 \in int\, M$, then, by Proposition 2.2 (v), p_M is Lipschitz continuous on X. Denoting by c its Lipschitz constant, we have $\left|p_M(f) - p_M(g)\right| \leq c\,\|f - g\|$. So $\|h - g\| \leq \|f - g\| + \|g\| \left|p_M(f) - p_M(g)\right|$ implies $\|h - g\| \leq (1 + c\,\|g\|)\,\|f - g\|$. □

Under suitable assumptions on M (which are verified, e.g., by any ball $B_r(\|.\|)$), Lemma 4.1 allows us to construct an auxiliary sequence $h_n^\varepsilon \in M \cap span_n G$ satisfying $\|g^o - h_n^\varepsilon\| \leq C\|g^o - span_n G\| + \varepsilon$, for a constant C dependent only on $\|g^o\|$ and on the Lipschitz constant of p_M ($C = 1 + c\|g^o\|$). Combining this inequality with MJB theorem, we derive the following estimates of rates of approximate optimization in terms of G-variation of an argminimum g^o of the problem (M, Φ) and the modulus of continuity of Φ at g^o.

Theorem 4.2 *Let $(X, \|.\|)$ be a Hilbert space, M and G be its subsets, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed, convex, and $0 \in int\, M$. Let $\Phi : X \to (-\infty, +\infty]$ be a functional,*

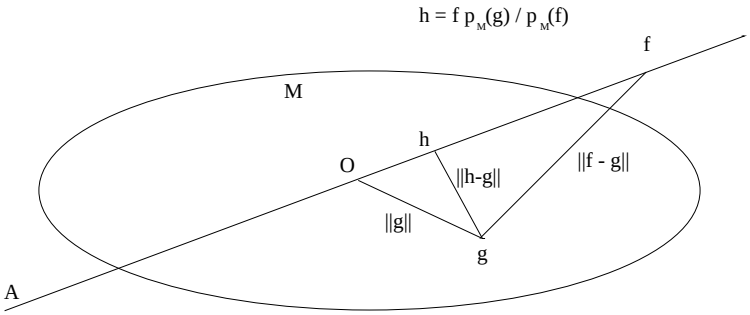

Figure 4.1:

$g^o \in argmin\,(M,\Phi)$, Φ be continuous at g^o with a modulus of continuity α, and $\{\varepsilon_n\}$ be a sequence of positive reals such that $g_n \in argmin_{\varepsilon_n}(M \cap span_n\, G, \Phi)$. Then p_M is Lipschitz on X and if c is its Lipschitz constant, the following hold for every integer n:

(i) $\inf_{g\in M\cap span_n\, G} \Phi(g) - \Phi(g^o) \le \alpha\left(\big(1+c\|g^o\|\big)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right)$;

(ii) if $\|g^o\|_G < \infty$ *and* $\lim_{n\to\infty}\varepsilon_n = 0$, *then* $\{g_n\}$ *is a* Φ*-minimizing sequence over* M *and*
$\Phi(g_n) - \Phi(g^o) \le \alpha\left(\big(1+c\|g^o\|\big)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right) + \varepsilon_n$;

(iii) if ξ *is the modulus of Tychonov well-posedness of* (M,Φ) *at* g^o, *then*
$\xi(\|g_n - g^o\|) \le \alpha\left(\big(1+c\|g^o\|\big)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right) + \varepsilon_n$;

(iv) if Φ *is uniformly convex on* M *with a modulus of convexity* δ, *then*
$\delta(\|g_n - g^o\|) \le \alpha\left(\big(1+c\|g^o\|\big)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right) + \varepsilon_n$.

Proof. (i) As $0 \in int\, M$, by Proposition 2.2 (iv) and (v), $dom\, p_M = X$ and p_M is Lipschitz on X.

For every n and every $\varepsilon > 0$, choose an ε-near best approximation f_n^ε of g^o in $span_n G$, i.e., $\|g^o - f_n^\varepsilon\| < \|g^o - span_n\, G\| + \varepsilon$. As M is closed, convex, $0 \in M$, and $f_n^\varepsilon \in dom\, p_M = X$, applying Lemma 4.1 (ii) with $f = f_n^\varepsilon$, $g = g^o$, and $A = span_n\, G$, we obtain $h_n^\varepsilon \in M \cap span_n\, G$ satisfying

$$\|h_n^\varepsilon - g^o\| \le \big(1+c\|g^o\|\big)\,\|f_n^\varepsilon - g^o\| \le (1+c\|g^o\|)(\|g^o - span_n G\| + \varepsilon). \tag{4.1}$$

As $h_n^\varepsilon \in M \cap span_n G$, we have $\inf_{g\in M\cap span_n\, G}\Phi(g) - \Phi(g^o) \le \Phi(h_n^\varepsilon) - \Phi(g^o)$. Estimating the right-hand side of this inequality in terms of the modulus of continuity α of Φ at g^o we obtain $\inf_{g\in M\cap span_n\, G}\Phi(g) - \Phi(g^o) \le \alpha\big(\|h_n^\varepsilon - g^o\|\big)$. Combining this estimate with inequality (4.1), we get

$$\inf_{g\in M\cap span_n\, G}\Phi(g) - \Phi(g^o) \le \alpha\big((1+c\|g^o\|)\|g^o - span_n G\| + \varepsilon\big).$$

By Theorem 3.1 (i), we have

$$\inf_{g\in M\cap span_n\, G}\Phi(g) - \Phi(g^o) \le \alpha\left((1+c\|g^o\|)\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} + \varepsilon\right). \tag{4.2}$$

By infimizing (4.2) over ε, we obtain

$$\inf_{g\in M\cap span_n\, G}\Phi(g)-\Phi(g^o)\le\alpha\left(\left(1+c\|g^o\|\right)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right).$$

(ii) By the definition of ε_n-argminimum, $\Phi(g_n)-\Phi(g^o)\le\inf_{g\in M\cap span_n\, G}\Phi(g)-\Phi(g^o)+\varepsilon_n$. So by item (i) we have

$$\Phi(g_n)-\Phi(g^o)\le\alpha\left(\left(1+c\|g^o\|\right)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right)+\varepsilon_n. \tag{4.3}$$

If $\lim_{n\to\infty}\varepsilon_n=0$ and $\|g^o\|_G$ is finite, then the right-hand side of (4.2) converges to zero and so $\{g_n\}$ is Φ-minimizing.

(iii) By the definitions of ε_n-argmin and of the modulus of Tychonov well-posedness of (M,Φ) at g^o, and by item (i), we have $\xi(\|g_n-g^o\|)=\inf_{g\in M\cap S_{\|g_n-g^o\|}(g^o)}\Phi(g)-\Phi(g^o)\le\Phi(g_n)-\Phi(g^o)<\inf_{g\in M\cap span_n\, G}\Phi(g)-\Phi(g^o)+\varepsilon_n\le\alpha\left(\left(1+c\|g^o\|\right)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right)+\varepsilon_n$.

(iv) By the definition of ε_n-argmin, Proposition 2.1 (iii) and item (i), we have $\delta(\|g_n-g^o\|)\le\Phi(g_n)-\Phi(g^o)<\inf_{g\in M\cap span_n\, G}\Phi(g)-\Phi(g^o)+\varepsilon_n\le\alpha\left(\left(1+c\|g^o\|\right)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right)+\varepsilon_n$.
□

Theorem 4.2 shows that for $\|g^o\|_G$ finite, the approximate argminima $\{g_n\}$ form a Φ-minimizing sequence and the speed of convergence of $\{\Phi(g_n)\}$ to the global minimum $\Phi(g^o)$ is bounded from above by $\alpha\left(\frac{(1+c\|g^o\|)s_G\|g^o\|_G}{\sqrt{n}}\right)$.

When minimization is performed over the whole space, the Lipschitz constant of the Minkowski functional $p_M=p_X$ is equal to zero; thus, we obtain from Theorem 4.2 an upper bound $\alpha\left(\frac{s_G\|g^o\|_G}{\sqrt{n}}\right)$, which is expressed in terms of the modulus of continuity of Φ and of G-variation of g^o. Similarly, when an admissible set is a ball $B_r(\|.\|)$ the Lipschitz constant is $1/r$ and we get a bound $\alpha\left(\left(1+\frac{\|g^o\|}{r}\right)\frac{s_G\|g^o\|_G}{\sqrt{n}}\right)$.

Inspection of these upper bounds enables one to describe classes of high-dimensional optimization problems that can be approximately solved up to any degree of accuracy by the extended Ritz method without incurring the curse of dimensionality, i.e., the number n of basis functions required for a satisfactory approximate optimization does not grow exponentially with the number of variables of admissible solutions. For α invertible, this is guaranteed when $\frac{(1+c\|g^o\|)s_G\|g^o\|_G}{\alpha^{-1}(\eta)}$ does not grow exponentially with the number of variables of g^o.

Estimates similar to the ones stated in Theorem 4.2 for Hilbert spaces can be obtained for $\mathcal{L}_p$-spaces, $p\in(1,\infty)$, when in the proof of Theorem 4.2 the estimate from Theorem 3.1 (ii) is used instead of the estimate from Theorem 3.1 (i).

Theorem 4.3 *Let $\Omega\subset\mathcal{R}^d$ be compact, M and G be subsets of $(\mathcal{L}_p(\Omega),\|.\|_p)$, $p\in(1,\infty)$, G be bounded, $s_G=\sup_{g\in G}\|g\|$, M be closed, convex, $0\in int\, M$, and $q=p/(p-1)$, $\bar p=\min(p,q)$, $\bar q=\max(p,q)$. Let $\Phi:X\to(-\infty,+\infty]$ be a functional, $g^o\in argmin\,(M,\Phi)$, Φ be continuous at g^o with a modulus of continuity α, and $\{\varepsilon_n\}$ be a sequence of positive reals such that $g_n\in argmin_{\varepsilon_n}(M\cap span_n\, G,\Phi)$. Then p_M is Lipschitz on X and if c is its Lipschitz constant, the following hold for every integer n:*
(i) $\inf_{g\in M\cap span_n\, G}\Phi(g)-\Phi(g^o)\le\alpha\left(\left(1+c\|g^o\|\right)\frac{2^{1/\bar p+1}s_G\|g^o\|_G}{n^{1/\bar q}}\right)$;
(ii) if $\|g^o\|_G<\infty$ *and* $\lim_{n\to\infty}\varepsilon_n=0$, *then* $\{g_n\}$ *is a* Φ*-minimizing sequence over* M *and* $\Phi(g_n)-\Phi(g^o)\le\alpha\left(\left(1+c\|g^o\|\right)\frac{2^{1/\bar p+1}s_G\|g^o\|_G}{n^{1/\bar q}}\right)+\varepsilon_n$;
(iii) if ξ *is the modulus of Tychonov well-posedness of* (M,Φ) *at* g^o, *then* $\xi(\|g_n-g^o\|)\le\alpha\left(\left(1+c\|g^o\|\right)\frac{2^{1/\bar p+1}s_G\|g^o\|_G}{n^{1/\bar q}}\right)+\varepsilon_n$;
(iv) if Φ *is uniformly convex with a modulus of convexity* δ, *then* $\delta(\|g_n-g^o\|)\le\alpha\left(\left(1+c\|g^o\|\right)\frac{2^{1/\bar p+1}s_G\|g^o\|_G}{n^{1/\bar q}}\right)+\varepsilon_n$.

5 Asymptotic estimates for convex problems

For convex problems such that the functional to be minimized is bounded in a neighborhood of an argminimum, under the additional assumption of density of $M \cap span\, G$ in M Theorem 4.2 can be simplified. As such a simplification is based on local properties of the modulus of continuity of the functional, it gives only asymptotic estimates. For $f, g : \mathcal{N}_+ \to \mathcal{N}_+$ we write $g(n) \leq \mathcal{O}(f(n))$ when there exists $a > 0$ such that, for all but finitely many $n \in \mathcal{N}_+$, $g(n) \leq a\, f(n)$.

Theorem 5.1 *Let $(X, \|.\|)$ be a Hilbert space, M and G be its subsets, G be bounded, $s_G = \sup_{g\in G} \|g\|$, M be closed, convex, $0 \in int\, M$, and $M \cap span\, G$ dense in M. Let $\Phi : X \to (-\infty, +\infty]$ be a proper convex functional, $g^o \in argmin\,(M, \Phi)$ such that Φ is bounded in its neighborhood, $\{\varepsilon_n\}$ be a sequence of positive reals such that $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ and $g_n \in argmin_{\varepsilon_n}(M \cap span_n\, G, \Phi)$. Then the following hold:*

(i) $\inf_{g\in M\cap span_n\, G} \Phi(g) - \Phi(g^o) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$;

(ii) if $\|g^o\|_G < \infty$, *then* $\{g_n\}$ *is a* Φ*-minimizing sequence over* M *and*
$\Phi(g_n) - \Phi(g^o) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$;

(iii) if ξ *is the modulus of Tychonov well-posedness of* (M, Φ) *at* g^o, *then*
$\xi(\|g_n - g^o\|) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$;

(iv) if Φ *is uniformly convex with a modulus of convexity* δ, *then*
$\delta(\|g_n - g^o\|) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$.

Proof. (i) Let $\nu > 0$ be such that Φ is bounded on $B_\nu(g^o, \|.\|)$. As $B_\nu(g^o, \|.\|) \subseteq dom\, \Phi$, we have $g^o \in int\, dom\, \Phi$. Since Φ is a proper convex functional bounded on $B_\nu(g^o, \|.\|)$, Φ is locally Lipschitz on $B_\nu(g^o, \|.\|)$ [20, Corollary 2.4, p. 12]. Let $\eta \leq \nu$ be such that Φ is Lipschitz continuous with constant c_1 on $B_\eta(g^o, \|.\|)$.

As $M \cap span\, G$ is dense in M, $\lim_{n\to\infty} \|g^o - span_n G\| = 0$ and so there exist $\varepsilon_0 > 0$ and $n_0 \in \mathcal{N}_+$ such that $\|g^o - span_{n_0} G\| + \varepsilon_0 \leq \frac{\eta}{1+c\|g^o\|}$. For every $n \geq n_0$ and $\varepsilon \leq \varepsilon_0$, choose $f_n^\varepsilon \in span_n\, G$ such that $\|g^o - f_n^\varepsilon\| \leq \|g^o - span_n\, G\| + \varepsilon$.

As M is closed, convex, $0 \in int\, M$, and $dom\, p_M = X$, we can apply Lemma 4.1 (ii) with $f = f_n^\varepsilon$, $g = g^o$, and $A = span_n\, G$ to obtain $h_n^\varepsilon \in M \cap span_n\, G$ satisfying

$$\|h_n^\varepsilon - g^o\| \leq \big(1 + c\|g^o\|\big)\, \|g_n^\varepsilon - g^o\| < \eta. \tag{5.1}$$

So h_n^ε is in the ball $B_\eta(g^o, \|.\|)$, on which Φ is Lipschitz continuous with the constant c_1. So we have

$$\inf_{g\in M\cap span_n\, G} \Phi(g) - \Phi(g^o) \leq \Phi(h_n^\varepsilon) - \Phi(g^o) \leq c_1\, \|h_n^\varepsilon - g^o\|. \tag{5.2}$$

From (5.1) and (5.2) we obtain

$$\inf_{g\in M\cap span_n\, G} \Phi(g) - \Phi(g^o) \leq C\, \|f_n^\varepsilon - g^o\|\,, \tag{5.3}$$

where $C = c_1\,(1 + c\,\|g^o\|)$. By Theorem 3.1 (i) we get

$$\|g^o - f_n^\varepsilon\| \leq \|g^o - span_n\, G\| + \varepsilon \leq \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} + \varepsilon. \tag{5.4}$$

Infimizing over ε, we obtain from (5.3) and (5.4) for all $n \geq n_0$

$$\inf_{g\in M\cap span_n\, G} \Phi(g) - \Phi(g^o) \leq C\, \sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}.$$

(ii) As $g_n \in argmin(M \cap span_n G)$, we have $\Phi(g_n) < \inf_{g \in M \cap span_n G} \Phi(g) + \varepsilon_n$. Combining this inequality with the one from item (i) and $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ we obtain $\Phi(g_n) - \Phi(g^o) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$.

(iii) By the definitions of ε_n-argmin and of the modulus of Tychonov well-posedness of (M, Φ) at g^o and by item (i), we have for every $n \geq n_0$, $\xi(\|g_n - g^o\|) = \inf_{g \in M \cap S_{\|g_n - g^o\|}(g^o)} \Phi(g) - \Phi(g^o) \leq \Phi(g_n) - \Phi(g^o) < \inf_{g \in M \cap span_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n \leq C\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} + \varepsilon_n$. As $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$, we obtain $\xi(\|g_n - g^o\|) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$.

(iv) By the definition of ε_n-argmin and Propositions 2.1 (iii) and 5.1 (i), we get for all $n \geq n_0$, $\delta(\|g_n - g^o\|) \leq \Phi(g_n) - \Phi(g^o) < \inf_{g \in M \cap span_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right) + \varepsilon_n$. As $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$, we obtain $\delta(\|g_n - g^o\|) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$. □

Inspection of the proof of Theorem 5.1 shows that the expression $\mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$ can be written for $n \geq n_o$ as $C\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}$, where $C = c_1(1 + c\|g^o\|)$, c is the Lipschitz constant of p_M, and c_1 is the Lipschitz constant of Φ in a neighborhood of g^o. The proof also shows that for any sequences $\{\varepsilon_n\}$ of positive reals and $\{g_n\}$ such that $g_n \in argmin_{\varepsilon_n}(M, \Phi)$, the statements of Theorem 5.1 (ii), (iii) and (iv) hold with the bounds replaced with $\mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right) + \varepsilon_n$.

Applying Theorem 3.1 (ii) instead of Theorem 3.1(i) and following steps analogous to those in the proof of Theorem 5.1, one can obtain for $\mathcal{L}_p$-spaces estimates similar to those stated in Theorem 5.1 for Hilbert spaces (the condition $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ has to be replaced with $\varepsilon_n \leq \mathcal{O}(n^{1/\bar{q}})$, where $q = p/(p-1)$ and $\bar{q} = \max(p, q)$).

6 Application to convex best approximation problems

The simplest example illustrating results reported in Section 4 is the application of the extended Ritz method to convex best approximation problems.

For any $f \in X$, let e_f denote the functional defined as the distance from f, i.e., $e_f(g) = \|g - f\|$ for any $g \in X$. When M is a closed convex subset of X, (M, e_f) is called a *convex best approximation problem* [19, p. 40].

Applying Theorem 4.2 to the problems (M, e_f) and (M, e_f^2), we obtain the following estimates of rates of approximate optimization.

Theorem 6.1 *Let M and G be subsets of a Hilbert space $(X, \|.\|)$, G be bounded, $s_G = \sup_{g \in G}\|g\|$, M closed, convex, $0 \in int\, M$, and $f \in X$. Then there exists a unique argminimum g^o of (M, e_f) such that the following hold:*

(i) for every positive integer n, $\inf_{g \in M \cap span_n G} e_f(g) - e_f(g^o) \leq (1 + c\|g^o\|)\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}$;

(ii) if M is bounded, $\{\varepsilon_n\}$ is a sequence of positive reals, and for every n, $g_n \in argmin_{\varepsilon_n}(M \cap span_n G, e_f^2)$, then

$$\|g_n - g^o\|^2 \leq 2\, diam\, M\left((1 + c\|g^o\|)\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)^2 + \varepsilon_n.$$

Proof. As every closed convex subset of a Hilbert space is Chebyshev [18, p. 35]) (i.e., there exists a unique $g^o \in M$ such that $\|f - g^o\| = \|f - M\|$), the problem (M, e_f) has a unique argminimum.

By the triangle inequality, for every $h, g \in X$ we have $|e_f(h) - e_f(g)| \leq \|h - g\|$. So e_f is uniformly continuous on X and its modulus of continuity is $\alpha(t) = t$. Hence, applying Theorem 4.2 (i) we obtain (i).

To derive (ii), we apply Theorem 4.2 (iv) to the functional e_f^2. As $\|f - g^o\|^2 = \inf_{g \in M} \|f - g\|^2$, g^o is an argminimum of (M, e_f^2). By Proposition 2.1 (iii), the functional $\|.\|^2$ is strictly uniformly convex with a modulus of convexity $\delta(t) = t^2$.

By the triangle inequality, for every $h, g \in X$ we have $|e_f^2(h) - e_f^2(g)| = (\|f - h\| - \|f - g\|)(\|f - h\| + \|f - g\|) \leq 2\, diam\, M \|h - g\|$, and so $\alpha(t) = 2\, t\, diam\, M$ is an upper bound on the modulus of continuity of e_f^2. Thus, applying Theorem 4.2 (iv) we get $\|g_n - g^o\|^2 \leq 2\, diam\, M \left((1 + c\|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right)^2 + \varepsilon_n$. □

For convex best approximation problems in $\mathcal{L}_p$ spaces, $p \in (1, \infty)$, analogous estimates can be obtained combining Theorem 4.3 with estimates of the modulus of convexity based on geometrical properties of such spaces.

Theorem 6.2 *Let $\Omega \subset \mathcal{R}^d$ be compact, M and G be subsets of $(\mathcal{L}_p(\Omega), \|.\|_p)$, $p \in (1, \infty)$, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed, convex, $0 \in int\, M$, $f \in X$, $q = p/(p-1)$, $\bar{p} = \min(p, q)$, $\bar{q} = \max(p, q)$, and α_p, α_q be moduli of continuity of e_f^p, e_f^q, resp., at f. Then there exists a unique argminimum g^o of (M, e_f) such that the following hold:*

(i) for every positive integer n, $\inf_{g \in M \cap span_n\, G} e_f(g) - e_f(g^o) \leq \left(1 + c\|g^o\|\right) \frac{2^{1/\bar{p}+1} s_G \|g^o\|_G}{n^{1/\bar{q}}}$;

(ii) $f\, M$ is bounded and $\{\varepsilon_n\}$ is a sequence of positive reals, then for every n,

if $p \geq 2$ and $g_n \in argmin_{\varepsilon_n}(M \cap span_n\, G, e_f^p)$, we have
$\|g_n - g^o\|^p \leq 2^{p-2}\, \alpha_p \left(\left((1 + c\|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right)^2 \right) + \varepsilon_n;$
if $1 < p \leq 2$ and $g_n \in argmin_{\varepsilon_n}(M \cap span_n\, G, e_f^q)$, we have
$\|g_n - g^o\|^q \leq 2^{q-2}\, \alpha_q \left(\left((1 + c\|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right)^2 \right) + \varepsilon_n.$

Proof. As for all $p \in (1, \infty)$, $(\mathcal{L}_p(\Omega), \|.\|_p)$ is uniformly convex [1, 2.29] and every convex best approximation problem in a uniformly convex space is Tychonov well-posed [19, p. 40], there exists a unique $g^o \in M$ such that $\|f - g^o\|_p = \|f - M\|_p$ and so the problem (M, e_f) has a unique argminimum.

By the triangle inequality, for every $h, g \in X$ we have $|e_f(h) - e_f(g)| \leq \|h - g\|_p$. So e_f is uniformly continuous on X and its modulus of continuity is $\alpha(t) = t$. Hence applying Theorem 4.3 (i) we obtain (i).

To derive (ii), we apply Theorem 4.3 (iv) to the functional e_f^p when $p \geq 2$, wheraes when $p \in (1, 2]$, we use the functional e_f^q with $q = p/(p-1)$.

A modulus of convexity for e_f^p and e_f^q can be estimated by means of Clarkson's inequalities [1, 2.28], which state that for every $f, g \in (\mathcal{L}_p(\Omega), \|.\|_p)$,

$$\text{if } p \geq 2, \text{ then } \left\| \frac{f+g}{2} \right\|_p^p \leq \frac{1}{2}\|f\|_p^p + \frac{1}{2}\|g\|_p^p - \frac{1}{2^p}\|f - g\|_p^p \tag{6.1}$$

$$\text{if } 1 < p \leq 2, \text{ then } \left\| \frac{f+g}{2} \right\|_p^q \leq \frac{1}{2}\|f\|_p^q + \frac{1}{2}\|g\|_p^q - \frac{1}{2^p}\|f - g\|_p^q, \tag{6.2}$$

where $q = p/(p-1)$.

Inequality (6.1) implies that for $p \geq 2$ the functional e_f^p is uniformly convex with a modulus of convexity $\delta(t) = \frac{t^p}{2^{p-2}}$. Let us now consider the case $1 < p \leq 2$. For every $1 \leq r < \infty$ and

$a, b \geq 0$, we have $(a+b)^r \leq 2^{r-1}(a^r + b^r)$ [1, 2.24], which, combined with inequality (6.2), implies that for $1 < p \leq 2$, the functional e_f^q is uniformly convex with a modulus of convexity $\delta(t) = \frac{t^q}{2^{q-2}}$. □

Theorems 6.1 (i) and 6.2 (i) imply an extension of MJB theorem (Theorem 3.1) on approximation by $span_n G$ to a theorem on approximation by $M \cap span_n G$ provided that M satisfies the assumptions of Theorem 6.1 (e.g., when M is a ball, i.e., $M = B_r(\|.\|)$).

Corollary 6.3 *Let M and G be subsets of a normed space $(X, \|.\|)$, G be bounded, $s_G = \sup_{g \in G} \|g\|$, M be closed, convex, $0 \in int\, M$, $f \in M$, $g^o = argmin(M, e_f)$. Then p_M is Lipschitz and if c is its Lipschitz constant, the following hold for every positive integer n:*

(i) If $(X, \|.\|)$ is a Hilbert space, then

$$\|f - M \cap span_n\, G\| \leq (1 + c\|g^o\|)\sqrt{\tfrac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}} + \|f - g^o\|.$$

(i) If $(X, \|.\|) = (\mathcal{L}_p(\Omega), \|.\|_p), p \in (1, \infty)$, $\Omega \subset \mathcal{R}^d$ compact, $q = p/(p-1)$, $\bar{p} = \min(p, q)$, and $\bar{q} = \max(p, q)$, then

$$\|f - M \cap span_n\, G\| \leq (1 + c\|g^o\|)\big(1 + c\|g^o\|\big)\tfrac{2^{1/\bar{p}+1} s_G \|g^o\|_G}{n^{1/\bar{q}}} + \|f - g^o\|.$$

If $M = X$, the Lipschitz constant of p_M on X is equal to 0, $g^o = f$, and so $\|f - g^o\| = 0$. Thus, in this case Corollary 6.3 gives the same estimate as MJB theorem (Theorem 3.1).

7 Application to learning from data

Application of the results presented in Section 4 allows us to obtain an approximate version of the Representer Theorem from machine learning theory [15, Proposition 8], [21, p. 18]. Learning from data can be modeled as the minimization of the *empirical error functional* (also called empirical risk functional) defined as

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

where $\{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R},\, i = 1, \ldots, m\}$ is a sample of empirical data (set of input/output pairs).

However, the empirical error only depends on the particular sample of data $\{(x_i, y_i) : i = 1, \ldots, m\}$ and does not take into account any global properties of the input/output mapping from which the sample was chosen. Such properties can be expressed by means of *regularization*, which replaces the functional $\mathcal{E}$ with $\mathcal{E}_{\gamma,\Psi} = \mathcal{E} + \gamma\,\Psi$, where Ψ is a suitable functional called *stabilizer* and γ is a positive real number called *regularization parameter*. Typically, the stabilizer models some desired property of the solution (e.g., smoothness), whereas the regularization parameter is used to one express a trade-off between fitting to a sample of empirical data and fitting to the global shape of the input/output mapping.

An important class of stabilizers are squares of norms of reproducing kernel Hilbert spaces. A *reproducing kernel Hilbert space* (RKHS) $(\mathcal{H}_K(\Omega), \|.\|_K)$ is a Hilbert space of functions defined on a set Ω such that for every $x \in \Omega$, the evaluation functional $\mathcal{F}_x$, defined for any $f \in \mathcal{H}_K(\Omega)$ as $\mathcal{F}_x(f) = f(x)$, is bounded. For any RKHS there exists a unique symmetric, positive semidefinite mapping $K : \Omega \times \Omega \to \mathcal{R}$, called *kernel*, such that for any $f \in \mathcal{H}_K(\Omega)$ and any $x \in \Omega$, $\mathcal{F}(x) = \langle f, K(x, .)\rangle_K$ [5] (a mapping $K : \Omega \times \Omega \to \mathcal{R}$ is *positive semidefinite* on Ω if for all positive integers m, all $(a_1, \ldots, a_m) \in \mathcal{R}^m$, and all $(x_1, \ldots, x_m) \in \Omega^m$, $\sum_{i,j=1}^m a_i a_j K(x_i, x_j) \geq 0$). A kernel $K : \Omega \times \Omega \to \mathcal{R}$ is called a *Mercer kernel* if Ω is compact and K is symmetric, continuous and positive definite.

With $\|.\|_K^2$ as a stabilizer, the regularized functional obtained from $\mathcal{E}$ has the form

$$\mathcal{E}_{\gamma,K}(f) = \frac{1}{m}\sum_{i=1}^{m} |f(x_i) - y_i|^2 + \gamma \|f\|_K^2.$$

For a Mercer kernel, the Representer Theorem [15, p. 42] states that the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ has a unique argminimum g^o of the form $g^o(x) = \sum_1^m a_i K(x, x_i)$. It even gives a formula for computing the parameters $a = (a_1, \ldots, a_m)$ as the solution of the well-posed system of linear equations $\mathcal{K}(x) + \gamma \mathcal{I} a = y$, where $y = (y_1, \ldots, y_m)$, $\mathcal{K}$ is the $m \times m$ matrix defined as $\mathcal{K}(x)_{ij} = K(x_i, x_j)$, and $\mathcal{I}$ is the identity matrix [24].

It has been argued in [24, p. 219] that the "regularization principles lead to approximation schemes that are equivalent to networks with one layer of hidden units." Indeed, the unique argminimum is in the set $span_m G_K$, where $G_K = \{K(x, \cdot) : x \in \Omega\}$. Functions from this set can be computed by neural networks with m hidden units. In particular, for the Gaussian kernel they can be computed by radial-basis-function networks with Gaussian units.

A drawback of this elegant result is that the number of network hidden units needed to compute the function minimizing $\mathcal{E}_{\gamma,K}$ is equal to the size of the sample of input/output data. For large data sets, such networks might not be implementable. Moreover, in typical applications of neural networks, a number of hidden units much smaller than the number of data is chosen before learning.

Using Theorem 4.2, we derive an approximate version of the Representer Theorem. It allows us to estimate how quickly approximate solutions achievable by networks with n hidden units converge to the global argminimum described by the Representer Theorem. We first state basic properties of the functional $\mathcal{E}_{\gamma,K}$.

Proposition 7.1 *Let m and d be positive integers, Ω be a compact subset of $\mathcal{R}^d$, $K : \Omega \times \Omega$ be a Mercer kernel, $\gamma > 0$ and $\{(x_1, y_1), \ldots, (x_m, y_m)\} \subset (\Omega \times \mathcal{R})^m$. Then*
(i) $\mathcal{E}_{\gamma,K}$ is strictly uniformly convex on $\mathcal{H}_K(\Omega)$ with a modulus of convexity $\delta(t) = t^2$;
(ii) at every $f \in \mathcal{H}_K(\Omega)$, $\mathcal{E}_{\gamma,K}$ is continuous with a modulus of continuity bounded from above by $\alpha(t) = a_2 t^2 + a_1 t$, where $a_1 = 2\big(m\|f\|_K c_K + mb\sqrt{c_K} + \gamma\|f\|_K\big)$, $a_2 = m c_K + \gamma$ and $b = \max\{|y_i| : i = 1, \ldots, m\}$;
(iii) for $M \subset \mathcal{H}_K(\Omega)$ closed, convex, and bounded or for $M = \mathcal{H}_K(\Omega)$, the problem $(M, \mathcal{E}_{\gamma,K})$ has a unique argminimum g^o;
(iv) for $M \subset \mathcal{H}_K(\Omega)$ closed, convex, and bounded or for $M = \mathcal{H}_K(\Omega)$, any $g^o \in argmin(M, \mathcal{E}_{\gamma,K})$ and $f \in M$, $\|f - g^o\|_K^2 \le |\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g^o)|$.

Proof. (i) It is easy to show that $\mathcal{E}$ is convex, so (i) follows from Proposition 2.1 (i) and (iv).

(ii) Set $b = \max\{|y_i| : i = 1, \ldots, m\}$. Let $f \in \mathcal{H}(\Omega)$, $t > 0$ and $g \in \mathcal{H}_K$ be such that $\|f - g\|_K < t$. Using the inequality $\|.\|_{\mathcal{C}} \le \sqrt{c_K}\|.\|_K$, we obtain

$$\begin{aligned}
|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| &= |\sum_{i=1}^{m}\big((f(x_i) - y_i)^2 - (g(x_i) - y_i)^2\big) + \gamma\left(\|f\|_K^2 - \|g\|_K^2\right)| \\
&\le |\sum_{i=1}^{m}\big(f(x_i) - g(x_i)\big)\big(f(x_i) + g(x_i) - 2y_i\big)| \\
&+ \gamma\left((\|f\|_K - \|g\|_K)(\|f\|_K + \|g\|_K)\right) \\
&\le m\,\|f - g\|_{\mathcal{C}}\big(\|f + g\|_{\mathcal{C}} + 2b\big) + \gamma\,\|f - g\|_K\left(\|f\|_K + \|g\|_K\right) \\
&\le m\,t\,\sqrt{c_K}\left(\sqrt{c_K}\|f + g\|_K + 2b\right) + \gamma\left(\|f\|_K + \|g\|_K\right)t\,.
\end{aligned}$$

As $\|g\|_K < \|f\|_K + t$, we get

$$\begin{aligned}
|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| &< mt\sqrt{c_K}\left(2\|f\|_K\sqrt{c_K} + t\sqrt{c_K} + 2b\right) + \gamma t\left(2\|f\|_K + t\right) \\
&= t^2\left(m\,c_K + \gamma\right) + 2t(m\|f\|_K\,c_K + mb\sqrt{c_K} + \gamma\|f\|_K)\,.
\end{aligned}$$

Thus, $\|f-g\|_K < t$ implies $|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| < \alpha(t) = a_2t^2 + a_1t$, where $a_2 = m\,c_K + \gamma$ and $a_1 = 2\big(m\|f\|_K\,c_K + mb\sqrt{c_K} + \gamma\|f\|_K\big)$.

(iii) As a convex lower semicontinuous functional on a reflexive space attains its minimum on every convex, closed and bounded set [16, pp. 7, 14], by (i) and (ii) there exists an argminimum of $(M, \mathcal{E}_{\gamma,K})$ for every M closed, convex, and bounded. For $M = \mathcal{H}_K(\Omega)$, the existence of a unique argminimum is proven in [15, Proposition 7].

(iv) follows from (i) and Proposition 2.1 (iii). □

So the modulus of continuity of $\mathcal{E}_{\gamma,K}$ at any $f \in \mathcal{H}_K(\Omega)$ is bounded from above by the quadratic function $a_2t^2 + a_1t$. Note that a_2 does not depend on f as it depends only on m, c_K and γ, whereas a_1 depends also on $\|f\|_K$ and b. The larger the regularization parameter γ, the larger the coefficients of this quadratic function.

Applying Theorem 4.2 to the approximate solution of the optimization problem $(\mathcal{H}(\Omega), \mathcal{E}_{\gamma,K})$, we obtain the following estimates.

Theorem 7.2 *Let $\Omega \subset \mathcal{R}^d$ be compact, $K : \Omega \times \Omega \to \mathcal{R}$ be a Mercer kernel, $(\mathcal{H}_K(\Omega), \|.\|_K)$ be the RKHS defined by K, $G_K = \{K(x,.) : x \in \Omega\}$, $s_K = \sup_{x\in\Omega}\sqrt{K(x,x)}$, $(x_i, \ldots, x_m) \in \Omega^m$, $(y_m, \ldots, y_m) \in \mathcal{R}^m$, $\mathcal{E} : \mathcal{H}_K(\Omega) \to \mathcal{R}_+$ be the empirical error functional $\mathcal{E}(f) = \frac{1}{m}\sum_{i=1}^m |f(x_i) - y_i|^2$, $\gamma > 0$, $g^o(x) = \sum_1^m w_iK(x,x_i)$ be the unique argminimum of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ given by the Representer Theorem, and $\{\varepsilon_n\}$ be a sequence of positive reals such that $g_n \in argmin_{\varepsilon_n}(span_n\,G_K, \mathcal{E}_{\gamma,K})$. Then for every positive integer n, the following hold:*

(i) $\inf_{g\in span_n\,G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \le \alpha\left(\sqrt{\frac{(s_K\|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}}\right)$;

(ii) if $\|g^o\|_G < \infty$ *and* $\lim_{n\to\infty}\varepsilon_n = 0$, *then* $\{g_n\}$ *is an* $\mathcal{E}_{\gamma,K}$*-minimizing sequence over* $\mathcal{H}_K(\Omega)$ *and*

$\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_{\gamma,K}(g^o) \le \alpha\left(\sqrt{\frac{(s_K\|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}}\right) + \varepsilon_n$;

(iii) $\|g_n - g^o\|_K^2 \le \alpha\left(\sqrt{\frac{(s_K\|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}}\right) + \varepsilon_n$;

(iv) $\|g_n - g^o\|_{\mathcal{C}}^2 \le \sqrt{c_K}\left(\alpha\left(\sqrt{\frac{(s_K\|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}}\right) + \varepsilon_n\right)$,

where $\alpha(t) = a_2t^2 + a_1t$, $a_1 = 2\big(m\|g^o\|_K\,c_K + mb\sqrt{c_K} + \gamma\|g^o\|_K\big)$, $a_2 = m\,c_K + \gamma$, $c_K = \sup_{x,y\in\Omega}|K(x,y)|$, *and* $b = \max\{|y_i| : i = 1, \ldots, m\}$.

Proof. The statements (i) and (ii) follow from Theorem 4.2 applied to $(X, \|.\|) = (\mathcal{H}_K(\Omega), \|.\|_K) = M$, $c = 0$ (the Minkowski functional of $\mathcal{H}_K(\Omega)$ is equal to zero), $\Phi(f) = \mathcal{E}_{\gamma,K}(f)$ and $G = G_K$. As for every $x \in \Omega$, $\|K(x,.)\|_K = \sqrt{\langle K(x,.), K(x,.)\rangle_K} = \sqrt{K(x,x)}$, we have $\sup_{x\in\Omega}\|K(x,.)\|_K = s_K$. By Proposition 7.1 (ii), $\mathcal{E}_{\gamma,K}$ is continuous at g^o with a modulus of continuity $\alpha(t) = a_2t^2 + a_1t$, where $a_2 = mc_K + \gamma$ and $a_1 = 2(mc_K\|g^o\|_K + mb\sqrt{c_K} + \gamma\|g^o\|_K)$.

(iii) follows from (ii) and Proposition 7.1 (iii).

(iv) follows immediately from (iii) and the inequality $\|.\|_{\mathcal{C}} \le \sqrt{c_K}\|.\|_K$ [15, p. 36]. □

As the estimates from Theorem 7.2 are not merely asymptotic, they can be applied to networks with any number of hidden units that is smaller than the number of data. Moreover, the estimates hold for any number of variables of the functions in $\mathcal{H}_K(\Omega)$. Thus inspection of these estimates enables us to describe problems for which the rates of approximate optimization guaranteed by Theorem 7.2 do not incur the curse of dimensionality. This holds when for a desired accuracy η, the quantity $\frac{s_K\|g^o\|_{G_K}}{\alpha^{-1}(\eta)}$ does not depend exponentially on the number d of variables.

8 Discussion

In the calculus of variations, the term *direct methods* [23, p. 192] is used to refer to methods of solution of optimization problems (M, Φ) based on the construction of Φ-minimizing sequences

$\{g_n\} \subseteq M$ converging to some $g \in M$ and satisfying $\lim_{n\to\infty} \Phi(g_n) = \Phi(g)$.

Using this terminology, we can rephrase our results as conditions on (M, Φ) guaranteeing some of the features of direct methods. By Theorems 4.2 and 4.3, for $\|g^o\|_G$ finite any sequence $\{g_n\}$ of ε_n-argminima of $(M \cap span_n G, \Phi)$ is Φ-minimizing and $\Phi(g^o) = \lim_{n\to\infty} \Phi(g_n) = \Phi(\lim_{n\to\infty} g_n)$. The convergence of $\{g_n\}$ to g^o is not always guaranteed (it depends on the behavior of the modulus of Tychonov well-posedness of (M, Φ) at g^o); however, it occurs in both applications presented in Sections 6 and 7. Thus when applied to convex best approximation problems and to learning from data, the extended Ritz method is a direct method. Its speed of convergence depends on the G-variation of the argminimum g^o, which can be investigated using methods described in [9] and [30].

Acknowledgments

The authors thank R. Zoppoli (University of Genova) and T. Parisini (University of Trieste) for stimulating their interest in the theoretical investigation of approximate optimization by neural networks. They are also grateful to R. Zoppoli, A. Vogt (Georgetown University), and P. C. Kainen (Georgetown University) for fruitful comments and discussions. M. Sanguineti wishes to acknowledge that a large amount of experimental results on approximate optimization by neural networks was obtained by joint research with A. Alessandri (National Research Council of Italy - Genoa), M. Baglietto (University of Genoa), C. Cervellera (National Research Council of Italy - Genoa), T. Parisini, and R. Zoppoli. This research motivated the theoretical investigations presented in this paper.

Bibliography

[1] Adams, R. A.: *Sobolev Spaces*, Academic Press, New York, 1976.

[2] Alessandri, A., Baglietto, M., Parisini, T., and Zoppoli, R.: A neural state estimator with bounded errors for nonlinear systems, *IEEE Transactions on Automatic Control*, vol. 44, pp. 2028-2042, 1999.

[3] Alessandri, A., Parisini, T., Sanguineti, M., and Zoppoli R.: Neural strategies for nonlinear optimal filtering, *Proc. IEEE Int. Conf. on Systems Engineering*, pp. 44-49, Kobe (Japan), 1992.

[4] Alt, W., On the approximation of infinite optimization problems with an application to optimal control problems, *Appl. Math. Optim.*, vol. 12, pp. 15-27, 1984.

[5] Aronszajn, N.: Theory of reproducing kernels, *Transactions of the American Mathematical Society*, vol. 68, pp. 337-404, 1950.

[6] Baglietto, M., Parisini, T., and Zoppoli, R.: Numerical solutions to the Witsenhausen counterexample by approximating networks, *IEEE Trans. on Automatic Control*, vol. 46, pp. 1471–1477, 2001.

[7] Baglietto, M., Parisini, T., and R. Zoppoli, Distributed-information neural control: the case of dynamic routing in traffic networks, *IEEE Trans. on Neural Networks*, vol. 12, pp. 485–502, 2001.

[8] Baglietto, M., Sanguineti, M., and Zoppoli, R.: Facing the curse of dimensionality by the extended Ritz method in stochastic functional optimization: dynamic routing in traffic networks, *High Performance Algorithms and Software for Nonlinear Optimization*. G. Di Pillo and A. Murli, Editors, Kluwer Academic Publishers, pp. 22-55, 2003.

[9] Barron, A. R.: Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. on Information Theory*, vol. 39, pp. 930-945, 1993.

[10] Beard, R. W. and McLain, T. W.: Successive Galerkin approximation algorithms for nonlinear optimal and robust control, *Int. J. Contr.*, vol. 71, pp. 717–743, 1998.

[11] Bellman, R.: *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, 1957.

[12] Bertsekas, D. P. and Tsitsiklis, J. N.: *Neuro-Dynamic Programming*, Athena Scientific, Belmont, Massachusetts, 1996.

[13] Chen, F.C. and Khalil, H.: Adaptive control of a class of nonlinear discrete-time systems using multilayer neural networks, *IEEE Trans. on Automatic Control* 1995, vol. 40, pp. 791-801.

[14] Chen, V. C. P., Ruppert, D., and Shoemaker C. A.: Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming, *Operations Research*, vol. 47, pp. 38-53, 1999.

[15] Cucker, F. and Smale, S.: On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, vol. 39, pp. 1-49, 2001.

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

[17] Darken, C., Donahue, M., Gurvits, L., and Sontag, E.: Rate of approximation results motivated by robust neural network learning, *Proc. Sixth Annual ACM Conference on Computational Learning Theory*. The Association for Computing Machinery, New York, pp. 303-309, 1993.

[18] Deutch, F.: *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.

[19] Dontchev, A. L. and Zolezzi, T.: *Well-Posed Optimization Problems*, Lecture Notes in Math., vol. 1543, Springer-Verlag, Berlin Heidelberg, 1993.

[20] Ekeland, I. and Temam, R.: *Convex Analysis and Variational Problems*, North-Holland Publishing Company, Amsterdam Oxford, and American Elsevier Publishing Company, Inc., New York, 1976.

[21] Evgeniou, T., Pontil, M., and Poggio, T.: Regularization networks and support vector machines, *Advances in Computational Mathematics*, vol. 13, pp. 1-50, 2000.

[22] Felgenhauer, U.: On Ritz type discretizations for optimal control problems, *Proc. 18th IFIP-ICZ Conference, Res. Notes in Math.*, Chapman-Hall, 1999, vol. 386, pp. 91-99.

[23] Gelfand, I. M. and Fomin, S. V.: *Calculus of Variations*, Prentice Hall, Englewood Cliffs, N. J., 1963.

[24] Girosi, F., Jones, M., and Poggio, T.: Regularization theory and neural networks architectures, *Neural Computation*, vol. 7, pp. 219-269, 1995.

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

[26] Johnson, S. A., Stedinger, J. R., Shoemaker C., Li, Y., and Tejada-Guibert, J.: Numerical solution of continuous-state dynamic programs using linear and spline interpolation, *Operations Research*, vol. 41, pp. 484-500, 1993.

[27] Kainen, P. C., Kůrková, V., and Sanguineti, M.: Minimization of error functionals over variable-basis functions, *SIAM Journal on Optimization*, to appear.

[28] Kůrková, V.: Dimension-independent rates of approximation by neural networks. In *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality* (K. Warwick, M. Kárný, Eds.). Birkhauser, Boston, pp. 261-270, 1997.

[29] Kůrková, V.: Neural networks as universal approximators. In *The Handbook of Brain Theory and Neural Networks* (M. Arbib, Ed.), Cambridge, MIT Press, 2002, pp. 1180–1183.

[30] Kůrková, V.: High-dimensional approximation and optimization by neural networks, in *Learning Theory and Practice* (J. Stuykens, Ed.), IOS Press, 2003 (to appear), Chapter 4 (pp. 69–88).

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

[32] Kůrková, V. and Sanguineti, M.: Comparison of worst case errors in linear and neural network approximation, *IEEE Trans. on Information Theory*, vol. 48, pp. 264-275, 2002.

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

[34] Narendra, K. S. and Parthasarathi, K.: Identification and control of dynamical systems using neural networks, *IEEE Trans. on Neural Networks*, vol. 4, pp. 4-26, 1990.

[35] Narendra, K.S. and Mukhopadhyay, S.: Adaptive control using neural networks and approximate models. *IEEE Transactions on Neural Networks*, vol. 8, pp. 475-485, 1997.

[36] Parisini, T., Sanguineti, M., and Zoppoli, R.: Nonlinear stabilization by receding-horizon neural regulators, *Int. J. of Control*, vol. 70, pp. 341–362, 1998.

[37] Parisini, T. and Zoppoli, R.: Neural networks for feedback feedforward nonlinear control systems, *IEEE Trans. on Neural Networks*, vol. 5, pp. 436-449, 1994.

[38] Parisini, T. and Zoppoli, R.: Neural approximations for multistage optimal control of nonlinear stochastic systems, *IEEE Trans. on Automatic Control*, vol. 41, pp. 889–895, 1996.

[39] Pisier, G.: Remarques sur un résultat non-publié de B. Maurey. *Séminaire d'Analyse Fonctionnelle*, vol. I, no. 12. École Polytechnique, Centre de Mathématiques, Palaiseau, 1980-81.

[40] Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning internal representation by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. I: Foundations* (D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Eds.), MIT, Cambridge, MA, 1986, pp. 318–362.

[41] Sejnowski, T. J. and Rosenberg, C. R.: Parallel networks that learn to pronounce English text, *Complex Systems*, vol. 1, pp. 145–168, 1987.

[42] Sirisena, H. R. and Chou, F. S.: Convergence of the control parametrization Ritz method for nonlinear optimal control problems, *J. of Optimization Theory and Applications*, vol. 29, pp. 369–382, 1979.

[43] Tjuhtin, V. B.: An error estimate for approximate solutions in one-sided variational problems, *Vestnik Leningrad Univ. Math.*, vol. 14, pp. 247-254, 1982.

[44] Vladimirov, A. A., Nesterov, Yu. E., and Chekanov, Yu. N.: On uniformly convex functionals, *Vestnik Moskovskogo Universiteta. Seriya 15 - Vychislitel'naya Matematika i Kibernetika*, No. 3, pp. 12-23, 1979. (English translation: *Moscow University Computational Mathematics and Cybernetics*, pp. 10-21, 1979).

[45] Zoppoli, R. and Parisini, T.: 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.), Birkhäuser, pp. 193-210, 1992.

[46] Zoppoli, R., Sanguineti, M., and Parisini, T.: Approximating networks and extended Ritz method for the solution of functional optimization problems, *J. of Optimization Theory and Applications*, vol. 112, pp. 403-440, 2002.