

## ERROR ESTIMATES FOR APPROXIMATE OPTIMIZATION BY THE EXTENDED RITZ METHOD\*

VĚRA KŮRKOVÁ<sup>†</sup> AND MARCELLO SANGUINETI<sup>‡</sup>

**Abstract.** An alternative to the classical Ritz method for approximate optimization is investigated. In the extended Ritz method, sets of admissible solutions are approximated by their intersections with sets of linear combinations of all  $n$ -tuples of functions from a given basis. This alternative scheme, called variable-basis approximation, includes functions computable by trigonometric polynomials with free frequencies, free-node splines, neural networks, and other nonlinear approximating families. Estimates of rates of approximate optimization by the extended Ritz method are derived. Upper bounds on rates of convergence of suboptimal solutions to the optimal one are expressed in terms of the degree  $n$  of variable-basis functions, the modulus of continuity of the functional to be minimized, the modulus of Tikhonov well-posedness of the problem, and certain norms tailored to the type of basis. The results are applied to convex best approximation and to kernel methods in machine learning.

**Key words.** functional optimization, rates of convergence of suboptimal solutions, (extended) Ritz method, curse of dimensionality, convex best approximation problems, learning from data by kernel methods

**AMS subject classifications.** 49M15, 90B99, 90C90, 41A25, 68T05, 41A50

**DOI.** 10.1137/S1052623403426507

**1. Introduction.** In many high-dimensional optimization problems (routing in communications networks, stochastic optimal control, management of water resources, large-scale traffic networks [13, 24, 46, 81], etc.), optimal solutions cannot be found analytically or, even when they can be found, they may not be computable efficiently by numerical methods. However, in some cases, optimal solutions can be approximated by suboptimal ones. In the classical Ritz method [37, Chapter 8], such an approximation is accomplished by a sequence of solutions over intersections of the original admissible set with a nested family of linear subspaces of increasing dimensionality.

Although linear approximation methods have many convenient properties, their practical applications are often limited by the “curse of dimensionality” [14], i.e., an exponential growth, as a function of the number of variables, of the dimension a linear subspace would need to achieve a desired accuracy of approximation of the optimal solution. Experimental results indicate that the Ritz method is often unable to deal efficiently with high-dimensional optimization tasks [81]. Theoretical results estimating rates of convergence of the Ritz method for the case of admissible solutions dependent on only one variable were derived in [6, 19, 27, 36, 41, 73], but we have not found in the literature any explicit estimate for the multivariable case.

Since the late 1980s, neural networks became a successful alternative to linear

---

\*Received by the editors April 20, 2003; accepted for publication (in revised form) March 11, 2004; published electronically February 19, 2005. Collaboration between the authors was supported by the Scientific Agreement Italy-Czech Republic, Area MC 6, Project 22.

<http://www.siam.org/journals/siopt/15-2/42650.html>

<sup>†</sup>Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Prague 8 Czech Republic (vera@cs.cas.cz). This author was partially supported by GA ČR grant 201/02/0428.

<sup>‡</sup>Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy (marcello@dist.unige.it). This author was partially supported by a PRIN grant from the Italian Ministry of University and Research (Project “New Techniques for the Identification and Adaptive Control of Industrial Systems”).

methods for approximate solutions of high-dimensional optimization problems (see, e.g., [18, 23, 47, 61, 62, 74]). Also, a new branch of nonlinear approximation theory investigating approximation capabilities of neural networks was developed [11, 12, 21, 28, 38, 45, 50, 51, 52, 53, 54, 55, 56, 57, 58]. In a series of papers [3, 5, 8, 9, 10, 64, 65, 66, 80, 81], a new method of approximate optimization was developed, called in [81] the *extended Ritz method*. In these papers, approximate solutions were used that were obtained over restrictions of sets of admissible solutions to linear combinations of all  $n$ -tuples of functions with varying “free” parameters, instead of linear combinations of first  $n$  functions from a basis with fixed ordering as in the classical Ritz method. In the extended Ritz method, a nested family of linear subspaces of increasing dimensionality, which in the Ritz method approximates the set of admissible solutions, is replaced by a nested family of nonlinear approximating sets called *variable-basis functions*. The variable-basis approximation scheme includes a variety of nonlinear approximators such as free-nodes splines [31, Chapter 13], polynomials with free frequencies and phases [32], radial-basis-function networks with variable variances and centers [38], and feedforward neural networks [48, 56].

For bases formed by functions computable by neural-network units or, more generally, for bases consisting of functions parameterized by vectors from finite-dimensional Euclidean spaces, the extended Ritz method reduces the original optimization task to the problem of finding optimal values of finitely many parameters. This is a nonlinear programming problem, for which various algorithms are available [1, 4, 16, 39, 40, 76, 79].

The extended Ritz method with such bases was successfully tested on a variety of problems with admissible solutions dependent on a large number of variables: stochastic optimal control [64, 65, 66, 80] and optimal estimation of state variables [3] in nonlinear dynamic systems with a large number of state variables, team optimal control [8], optimal control of freeway traffic [81], routing in large-scale communication networks [9, 10], optimal fault diagnosis [5], etc. In these applications, admissible sets of variable-basis functions were used, for which the degree  $n$  necessary to guarantee a fixed approximation accuracy grows only polynomially with the number of variables of admissible solutions. Numerical comparisons with the classical Ritz method showing advantages of the extended Ritz method were made in [81].

Motivated by these experimental results, we investigate the extended Ritz method theoretically. We derive upper bounds on the speed of convergence of suboptimal solutions over nested families of variable-basis functions of increasing degree  $n$  to the optimal solution over the whole admissible set. The upper bounds depend on the degree of the variable-basis functions, a norm tailored to the type of the basis, the modulus of continuity of the functional to be minimized, and the modulus of Tichonov well-posedness of the problem. As our bounds are not merely asymptotic, they enable one to estimate the quality of suboptimal solutions achievable over admissible sets for any degree  $n$  (in particular, for  $n$  small enough to allow an efficient implementation of such suboptimal solutions).

By inspection of the derived estimates, we obtain some insights into optimization problems for which the extended Ritz method performs well. The critical term in our bounds is of the form  $1/\sqrt{n}$  multiplied by a certain norm of the optimal solution. Such a norm is tailored to the basis used in the extended Ritz method. To keep this norm small with increasing number of variables of admissible solutions, one has to increase a certain type of regularity related to smoothness [12, 21, 50, 54].

We illustrate our results on two examples. In the first one, we apply them to the problem of convex best approximation, and in the second one, to learning from data

modeled as minimization of a regularized empirical error functional over a reproducing kernel Hilbert space.

The paper is organized as follows. Section 2 introduces basic concepts and some results from optimization theory, which are used throughout the paper. Section 3 describes the variable-basis approximation scheme and the extended Ritz method. Section 4 contains our main results on rates of convergence of the extended Ritz method, and section 5 contains their interpretation in the special case of convex problems. Sections 6 and 7 apply the derived estimates to convex best approximation and to kernel methods in machine learning, respectively. Section 8 contains a brief discussion. For the reader's convenience, we include an appendix containing some tools from nonlinear approximation theory that are used in the paper.

**2. Preliminaries.** By a normed linear space  $(X, \|\cdot\|)$ , we mean a *real normed linear space*. We write only  $X$  when it is clear which norm is used.  $\mathcal{R}$  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  $p \in [1, \infty)$ ,  $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$  denotes the space of measurable, real-valued functions on  $\Omega$  such that  $\int_{\Omega} |f(x)|^p dx < \infty$  endowed with the  $\mathcal{L}_p$ -norm.

A ball and a sphere of radius  $r$  centered at  $h \in X$  are denoted by  $B_r(h, \|\cdot\|) = \{f \in X : \|f - h\| \leq r\}$  and  $S_r(h, \|\cdot\|) = \{f \in X : \|f - h\| = r\}$ , respectively. We write shortly  $B_r(\|\cdot\|) = B_r(0, \|\cdot\|)$  and merely  $B_r(h) = B_r(h, \|\cdot\|)$ ,  $B_r = B_r(\|\cdot\|)$  when it is clear which norm is used; similarly for spheres.

A Banach space  $X$  is called *uniformly convex* if for any  $\varepsilon \in (0, 2]$ , there exists  $\delta > 0$  such that if  $\|f\| = \|g\| = 1$  and  $\|(f + g)/2\| > 1 - \delta$ , then  $\|f - g\| < \varepsilon$  (i.e., whenever the midpoint of the line segment joining two points on the unit sphere approaches the sphere, then the endpoints of the segment must approach each other).

Sequences (of elements of linear spaces or sets) are denoted by  $\{x_n\}$  instead of  $\{x_n : n \in \mathcal{N}_+\}$ , where  $\mathcal{N}_+$  is the set of positive integers.

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

$\Phi$  is *continuous* at  $f \in \text{dom } \Phi$  if for all  $\varepsilon > 0$ , there exists  $\eta > 0$  such that for every  $g \in \text{dom } \Phi$ ,  $\|f - g\| < \eta$  implies  $|\Phi(f) - \Phi(g)| < \varepsilon$ . The function  $\alpha_f : \mathcal{R}_+ \rightarrow \mathcal{R}_+$  defined as  $\alpha_f(t) = \sup\{|\Phi(f) - \Phi(g)| : f, g \in \text{dom } \Phi, \|f - g\| \leq t\}$  is the *modulus of continuity* of  $\Phi$  at  $f$ . We write merely  $\alpha$  instead of  $\alpha_f$  when  $f$  is clear from the context.  $\Phi$  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  $\Phi$  is *convex* on a convex set  $M \subseteq X$  if for all  $h, g \in M$  and all  $\lambda \in [0, 1]$ ,  $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g)$ .  $\Phi$  is *uniformly convex* on a convex set  $M \subseteq X$  if there exists a nonnegative function  $\delta : \mathcal{R}_+ \rightarrow \mathcal{R}_+$  such that  $\delta(0) = 0$ , for all  $t > 0$  it satisfies  $\delta(t) > 0$ , and for all  $h, g \in M$  and all  $\lambda \in [0, 1]$  it satisfies  $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g) - \lambda(1 - \lambda)\delta(\|h - g\|)$ . Any such function  $\delta$  is called a *modulus of convexity* of  $\Phi$  (see, e.g., [59]).<sup>1</sup>

Using standard notation [34], we denote by

$$(M, \Phi)$$

the problem of minimization of a functional  $\Phi$  over a subset  $M$  of  $X$ .  $M$  is called a set of *admissible solutions* or *admissible set*. When both  $M$  and  $\Phi$  are convex,  $(M, \Phi)$  is called a *convex optimization problem*.

<sup>1</sup>The terminology is not unified: some authors use the term “strictly uniformly convex” instead of “uniformly convex,” and they reserve the latter term for the case where  $\delta : [0, +\infty) \rightarrow [0, +\infty)$  merely satisfies  $\delta(0) = 0$  and for some  $t_0 > 0$ ,  $\delta(t_0) > 0$  (see, e.g., [78], [33, p. 10]).

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

The following proposition summarizes elementary properties of uniformly convex functionals.

**PROPOSITION 2.1.** *Let  $(X, \|\cdot\|)$  be a normed linear space,  $M \subseteq X$  be convex, and  $\Phi$  be a uniformly convex functional on  $M$  with a modulus of convexity  $\delta$ . Then the following hold:*

- (i) *if  $\Psi$  is convex on  $M$ , then  $\Phi + \Psi$  is uniformly convex on  $M$  with a modulus of convexity  $\delta$ ;*
- (ii) *if  $\Phi : X \rightarrow \mathcal{R}$ , then for every  $f \in X$  the translated functional  $\Phi(\cdot - f)$  is uniformly convex on  $M - f$  with a modulus of convexity  $\delta$ ;*
- (iii) *if  $g^\circ \in \operatorname{argmin}(M, \Phi)$ , then for every  $g \in M$ ,  $\delta(\|g - g^\circ\|) \leq \Phi(g) - \Phi(g^\circ)$ ;*
- (iv) *if  $(X, \|\cdot\|)$  is a Hilbert space, then the functional  $\|\cdot\|^2 : X \rightarrow \mathcal{R}$  is uniformly convex with a modulus of convexity  $\delta(t) = t^2$ .*

*Proof.* (i) and (ii) follow directly from the definitions.

(iii) By the definition of uniformly convex functional, for every  $\lambda \in [0, 1]$  we have  $\lambda(1-\lambda)\delta(\|g - g^\circ\|) \leq \lambda\Phi(g) + (1-\lambda)\Phi(g^\circ) - \Phi(\lambda g + (1-\lambda)g^\circ)$ . As  $\Phi(g^\circ) \leq \Phi(\lambda g + (1-\lambda)g^\circ)$ , we get  $\lambda(1-\lambda)\delta(\|g - g^\circ\|) \leq \lambda\Phi(g) + (1-\lambda)\Phi(g^\circ) - \Phi(g^\circ) = \lambda(\Phi(g) - \Phi(g^\circ))$ . Hence  $(1-\lambda)\delta(\|g - g^\circ\|) \leq \Phi(g) - \Phi(g^\circ)$ . Taking the infimum over  $\lambda$ , we obtain  $\delta(\|g - g^\circ\|) \leq \Phi(g) - \Phi(g^\circ)$ .

(iv) It is easy to check that for every  $h, g \in X$  and  $\lambda \in [0, 1]$ , we have  $\|\lambda h + (1-\lambda)g\|^2 \leq \lambda\|h\|^2 + (1-\lambda)\|g\|^2 - \lambda(1-\lambda)\|h - g\|^2$ .  $\square$

The problem  $(M, \Phi)$  is *Tikhonov well-posed* if it has a unique minimum to which every minimizing sequence converges [34, p. 1]. The *modulus of Tikhonov well-posedness* of  $(M, \Phi)$  at a minimum point  $g^\circ$  is a function  $\xi_{g^\circ} : \mathcal{R}_+ \rightarrow \mathcal{R}_+$  such that for every  $t \in \mathcal{R}_+$ ,  $\xi_{g^\circ}(t) = \inf_{g \in M \cap S_t(g^\circ)} \Phi(g) - \Phi(g^\circ)$ .

The *linear span* of  $M$  is  $\operatorname{span} M = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R}, g_i \in M, n \in \mathcal{N}_+\}$ . The *convex hull* of  $M$  is  $\operatorname{conv} M = \{\sum_{i=1}^n w_i g_i : w_i \in [0, 1], \sum_{i=1}^n w_i = 1, g_i \in M, n \in \mathcal{N}_+\}$ . The *topological interior* of  $M$  is  $\operatorname{int} M = \{g \in M : (\exists \varepsilon > 0) (B_\varepsilon(g) \subset M)\}$  and its *closure* is  $\operatorname{cl} M = \{f \in X : (\forall \varepsilon > 0) (B_\varepsilon(f) \cap M) \neq \emptyset\}$ . If  $\operatorname{cl} M = Y$  for  $Y \subseteq X$ , then  $M$  is said to be *dense* in  $Y$ . The *diameter* of  $M$  is defined as  $\operatorname{diam} M = \sup\{\|f - g\| : f, g \in M\}$ .

For a subset  $M$  of a normed linear space, its *affine hull* is defined as  $\operatorname{aff} M = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R}, \sum_{i=1}^n w_i = 1, g_i \in M, n \in \mathcal{N}_+\}$ . An element  $g \in X$  is called a *relative interior point* of  $M \subseteq X$  if it is an interior point of  $M$  in the topological sense with respect to the topology induced on  $\operatorname{aff} M$ . The set of all relative interior points of  $M$  is called the *relative interior* of  $M$  and denoted by  $\operatorname{ri} M$ . Thus,  $\operatorname{ri} M = \{g \in \operatorname{aff} M : \exists \varepsilon > 0, B_\varepsilon(g) \cap \operatorname{aff} M \subseteq M\}$ . Note that  $\operatorname{ri} M$  is the interior of  $M$  as a subset of its affine hull, instead of as a subset of the whole space  $X$ .

The *Minkowski functional* of  $M \subseteq X$  is the functional  $p_M : X \rightarrow [0, +\infty]$  defined for every  $f \in X$  as

$$p_M(f) = \inf\{\lambda \in \mathcal{R}_+ : f/\lambda \in M\}.$$

$M$  is called *absorbing* if  $\operatorname{dom} p_M = X$ . For every  $M$ ,  $p_M$  is positively homogeneous, and if  $M$  is convex, then  $p_M$  is convex, too. The following proposition states elemen-

tary properties of Minkowski functionals of convex sets containing zero, which will be used in our proofs.

PROPOSITION 2.2. *Let  $(X, \|\cdot\|)$  be a normed linear space,  $M$  be a subset of  $X$  containing 0, and  $r_0 = \sup\{r > 0 : B_r(\|\cdot\|) \subseteq M\}$ . Then the following hold:*

- (i) *if  $M$  is convex, then  $M \subseteq \{f \in X : p_M(f) \leq 1\}$ ;*
- (ii) *if  $M$  is convex, then  $\{f \in X : p_M(f) < 1\} \subseteq M$ ;*
- (iii) *if  $M$  is closed and convex, then  $M = \{f \in X : p_M(f) \leq 1\}$ ;*
- (iv) *if  $0 \in \text{int } M$ , then  $\text{dom } p_M = X$ ;*
- (v) *if  $0 \in \text{int } M$  and  $r_0 < \infty$ , then for every  $f \in \text{dom } p_M$ ,  $p_M(f) \leq \|f\|/r_0$ ;*
- (vi) *if  $M$  is convex and  $0 \in \text{int } M$ , then  $p_M$  is Lipschitz on  $X$  with the Lipschitz 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  $\lambda < 1$  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  $\lim_{i \rightarrow \infty} \lambda_i = 1$  and for every  $i$ ,  $f/\lambda_i \in M$ . As  $M$  is closed and  $f = \lim_{i \rightarrow \infty} (f/\lambda_i)$ , we have  $f \in M$ .

(iv) As  $0 \in \text{int } M$ , there exists  $r > 0$  such that  $B_r(\|\cdot\|) \subseteq M$ . So for every  $f \in B_r(\|\cdot\|)$ ,  $p_M(f) \leq 1$ . 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  $p_M(g) = (\|g\|/r)p_M(r(g/\|g\|)) \leq \|g\|/r < \infty$ .

(v) By the definition of  $p_M$ , for every  $r > 0$  such that  $B_r(\|\cdot\|) \subseteq M$  and every  $f \in \text{dom } p_M$  we have  $p_M(f) \leq \|f\|/r$ . Hence, by the definition of  $r_0$ ,  $p_M(f) \leq \|f\|/r_0$ .

(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  $|p_M(f) - p_M(g)| \leq \|f-g\|/r_0$ .  $\square$

**3. Variable-basis approximation and the extended Ritz method.** The classical *Ritz method* [37, p. 192] for approximate optimization replaces the problem  $(M, \Phi)$  with a sequence of problems

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

where for each  $n$ ,  $X_n$  is an  $n$ -dimensional subspace of  $X$ . Under suitable conditions on  $\Phi$ ,  $M$ , and  $\{X_n\}$ , for every  $n$  there exists a minimum point  $g_n$  of the approximate problem  $(M \cap X_n, \Phi)$ , the sequence  $\{g_n\}$  converges to some  $g^o \in M$ , and  $\lim_{n \rightarrow \infty} \Phi(g_n) = \Phi(g^o)$ .

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  $\text{span}_n G = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R}, g_i \in G\}$ . We call  $n$  the *degree of the variable-basis functions* in  $\text{span}_n G$ . The variable-basis approximation scheme includes free-node splines [31, Chapter 13], polynomials with free frequencies and phases [32], radial-basis-function networks with variable variances and centers [38], feedforward neural networks [48, 56], and so on.

In an alternative to the classical Ritz method, the problem  $(M, \Phi)$  is approximated by a sequence of problems

$$\{M \cap \text{span}_n G, \Phi\}.$$

For  $G$  formed by parameterized families of the form  $G = \{g_a : a \in A, g_a \in G\}$  with  $A \subseteq \mathcal{R}^p$ , this method was applied to a variety of optimization tasks considered in a series of papers [3, 5, 8, 9, 10, 64, 65, 66, 80] and in [81], where it was called the *extended Ritz method*. Here we use this term even more generally for an approximate optimization by the sequence of problems  $\{M \cap \text{span}_n G\}$ , where  $G$  is any subset of  $X$ .

Sets  $\text{span}_n G$  are not convex and so, when the classical Ritz method is replaced with the extended one, minimum points over approximate admissible sets might not exist. However, the requirement of achieving a minimum point can be relaxed to a merely  $\varepsilon_n$ -near minimum, for which we shall formulate our estimates.

Typically, a basis is formed by functions parameterized by vectors from a finite-dimensional Euclidean space. For such bases, minimization over  $M \cap \text{span}_n G$  reduces to a finite-dimensional nonlinear programming problem. Such a problem can be solved by algorithms based on gradient descent with stochastic perturbations [17, pp. 38–40, 103–104], genetic algorithms [39], simulated annealing [1], global stochastic optimization based on Monte Carlo [79] or quasi-Monte Carlo [77, Chapter 4] methods, etc. When a basis is formed by functions computable by neural-network units, various learning algorithms can be applied (see, e.g., [4, 16, 40, 76] and the references therein). In [5, 9, 10, 65, 66, 81], applications of some of these algorithms to the extended Ritz method are described and illustrated by numerical results showing the algorithms' effectiveness in a variety of cases.

A sequence of  $\varepsilon_n$ -near minimum points of  $\Phi$  over  $M \cap \text{span}_n G$  might converge to a minimum point of  $\Phi$  over the whole  $M$  much faster than minima over  $M \cap X_n$  in 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 over such unions of subspaces values that are closer to the infimum over the whole  $M$ .

To estimate rates of convergence of approximate solutions that can be obtained by the extended Ritz method, we take advantage of a result from nonlinear approximation theory by Maurey (reported in [68, Lemma 2, p.V.2]), Jones [45, p. 611], and Barron [12, p. 934, Lemma 1]. Here we use a reformulation of this result in terms of a norm tailored to a given basis  $G$ . Such a norm, called *G-variation* and denoted by  $\|\cdot\|_G$ , was introduced in [51] as an extension of the concept of variation with respect to half-spaces [11]. For a subset  $G$  of a normed linear space  $(X, \|\cdot\|)$ , *G-variation* is defined as the Minkowski functional of the set  $\text{cl conv}(G \cup -G)$ :

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

So *G-variation* of  $f$  measures how much the set  $G$  should be dilated to contain  $f$  in the closure of its symmetric convex hull. *G-variation* is a norm on the subspace  $\{f \in X : \|f\|_G < \infty\} \subseteq X$  and

$$(3.1) \quad \|\cdot\| \leq s_G \|\cdot\|_G.$$

Indeed, if for  $b > 0$ ,  $f/b \in \text{cl conv}(G \cup -G)$ , then  $f/b = \lim_{\varepsilon \rightarrow 0} h_\varepsilon$ , where  $h_\varepsilon \in \text{conv}(G \cup -G)$ , and so  $\|h_\varepsilon\| \leq s_G$ . Thus,  $\|f\| \leq s_G b$ . Hence, by the definition of  $\|f\|_G$  we have  $\|f\| \leq s_G \|f\|_G$ .

When  $G$  is an orthonormal basis of a separable Hilbert space,  $G$ -variation is equal to the  $l_1$ -norm with respect to  $G$ , which is defined for every  $f \in X$  as  $\|f\|_{1,G} = \sum_{g \in G} |f \cdot g|$  [55, 58]. 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.

The next theorem is a reformulation in terms of  $G$ -variation of the estimates derived for Hilbert spaces by Maurey, Jones, and Barron and of an extension of these estimates to  $\mathcal{L}_p$ -spaces,  $p \in (1, \infty)$ , derived by Darken et al. [28, Theorem 5]. For the proof, see the appendix.

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

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

$$(3.2) \quad \|f - \text{span}_n G\| \leq \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}};$$

(ii) *if  $(X, \|\cdot\|) = (\mathcal{L}_p(\Omega), \|\cdot\|_p)$ , where  $p \in (1, \infty)$  and  $\Omega \subseteq \mathcal{R}^d$  is open, then*

$$(3.3) \quad \|f - \text{span}_n G\| \leq \frac{2^{1/\bar{p}} s_G \|f\|_G}{n^{1/\bar{q}}},$$

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

In contrast to some estimates of rates of linear (i.e., fixed-basis) approximation in spaces  $X$  of  $d$ -variable functions [67, pp. 232–233], where the denominator is of the form  $n^{c/d}$  for some  $c > 0$ , in the bounds from Theorem 3.1, the denominator is  $n^{1/2}$ , independently of the number  $d$  of variables. However, for both fixed-basis and variable-basis approximation the numerators depend on  $d$  (see the discussion in section 8).

**4. Rates of approximate optimization over variable-basis functions.** In this section, we investigate approximate solutions by  $\{(M \cap \text{span}_n G, \Phi)\}$  of a problem  $(M, \Phi)$  that has a minimum point. The existence of such a point is guaranteed for various convex problems in reflexive Banach spaces [26, 35, 59, 70]. Many problems that do not have minimum points can be transformed into problems with minimum points by regularization [34, p. 29]. So the following results apply to a wide class of regularized problems.

Let  $g^\circ$  be a minimum point of the problem  $(M, \Phi)$  to which the extended Ritz method based on an approximation of  $M$  by sets  $M \cap \text{span}_n G$  is applied. As the existence of minimum points of approximating problems  $(M \cap \text{span}_n G, \Phi)$  is not guaranteed, we consider  $\varepsilon_n$ -near minimum points. To estimate the speed of convergence of these  $\varepsilon_n$ -near minimum points to the minimum point  $g^\circ$  of  $\Phi$  over the whole  $M$ , we take advantage of Theorem 3.1. As this theorem estimates the distance of  $g^\circ$  from  $\text{span}_n G$  but not from  $M \cap \text{span}_n G$ , we construct an auxiliary sequence of elements of  $M \cap \text{span}_n G$  using the following technical lemma. It extends [75, Lemma 3], proven there for finite-dimensional subspaces of a linear space, to subsets satisfying a kind of restricted homogeneity condition. Our lemma applies to a closed convex admissible set  $M$  containing zero. In the case when zero is in the interior of  $M$ , it gives an estimate in terms of the Lipschitz constant of the Minkowski functional of  $M$ . When  $M$  is a ball  $B_r(\|\cdot\|)$ , such a Lipschitz constant is equal to  $1/r$ .

LEMMA 4.1. *Let  $A$  and  $M$  be subsets of a normed linear space  $(X, \|\cdot\|)$ ,  $M$  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)  $\|h - g\| \leq \|f - g\| + \|g\| |p_M(f) - p_M(g)|$ ;
- (ii) *if  $0 \in \text{int } M$ , then  $\|h - g\| \leq (1 + c \|g\|) \|f - g\|$ , where  $c$  is the Lipschitz constant of  $p_M$  on  $X$ .*

*Proof.* (See Figure 4.1.) (i) When  $f \in A \cap M$ , the estimate holds trivially with  $h = f$ . If  $f \in A - \text{cl } M$ , 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  $f \notin M$ , by Proposition 2.2(iii) 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  $\|h - g\| = \|\frac{p_M(g)}{p_M(f)} f - g\| = \|\frac{p_M(g)}{p_M(f)}(f - g) - (1 - \frac{p_M(g)}{p_M(f)})g\| \leq \|\frac{p_M(g)}{p_M(f)}\| \|f - g\| + \|1 - \frac{p_M(g)}{p_M(f)}\| \|g\| < \|f - g\| + \|\frac{p_M(f) - p_M(g)}{p_M(f)}\| \|g\| < \|f - g\| + |p_M(f) - p_M(g)| \|g\|$ .

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

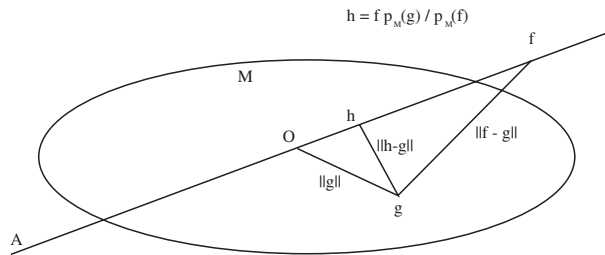


FIG. 4.1. *The construction used in the proof of Lemma 4.1.*

As we shall employ Lemma 4.1(ii) in the proof of the next theorem estimating rates of approximate optimization by the extended Ritz method, we need to assume that  $0 \in \text{int } M$ . Although this condition is restrictive, it still allows important applications. For example, when  $M$  is the whole ambient space  $X$ , one can apply the next theorem to Tikhonov’s regularization [15, pp. 68–78] (see the application in section 7), and when  $M$  is a ball of some radius  $r$  in the norm  $\|\cdot\|$ , one can apply it to Ivanov’s regularization [15, pp. 68–78]. Also, the case where  $M$  is a subspace of a Hilbert space  $X$  can be treated using the next theorem by replacing the ambient space  $X$  with  $M$  (since  $M$ , as a closed subspace of  $X$ , is a Hilbert space).

THEOREM 4.2. *Let  $(X, \|\cdot\|)$  be a Hilbert space,  $M$  and  $G$  be its subsets,  $G$  be bounded,  $s_G = \sup_{g \in G} \|g\|$ ,  $M$  be closed and convex, and  $0 \in \text{int } M$ . Let  $\Phi : X \rightarrow (-\infty, +\infty]$  be a proper functional,  $g^o \in \text{argmin}(M, \Phi)$ ,  $\Phi$  be continuous at  $g^o$  with a modulus of continuity  $\alpha$ ,  $\{\varepsilon_n\}$  be a sequence of positive reals, and  $\{g_n\}$  be such that  $g_n \in \text{argmin}_{\varepsilon_n}(M \cap \text{span}_n G, \Phi)$ . Then  $p_M$  is Lipschitz on  $X$ , and if  $c$  is its Lipschitz constant, then the following estimates hold for every positive integer  $n$ :*

- (i)

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha \left( (1 + c \|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right);$$

- (ii) if  $\|g^o\|_G < \infty$  and  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ , then  $\{g_n\}$  is a  $\Phi$ -minimizing sequence over  $M$  and

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

- (iii) if  $\xi$  is the modulus of Tikhonov well-posedness of  $(M, \Phi)$  at  $g^o$ , then

$$\xi(\|g_n - g^o\|) \leq \alpha \left( (1 + c\|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n;$$

- (iv) if  $\Phi$  is uniformly convex on  $M$  with a modulus of convexity  $\delta$ , then

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

The proof of Theorem 4.2 is based on Theorem 3.1(i) combined with the following idea: Lemma 4.1(ii) allows us to construct an auxiliary sequence  $h_n^\varepsilon \in M \cap \text{span}_n G$  satisfying  $\|g^o - h_n^\varepsilon\| \leq C\|g^o - \text{span}_n G\| + \varepsilon$  for a constant  $C = 1 + c\|g^o\|$  dependent only on  $\|g^o\|$  and on the Lipschitz constant  $c$  of  $p_M$ .

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

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

$$(4.1) \quad \|h_n^\varepsilon - g^o\| \leq (1 + c\|g^o\|) \|f_n^\varepsilon - g^o\| \leq (1 + c\|g^o\|)(\|g^o - \text{span}_n G\| + \varepsilon).$$

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

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha((1 + c\|g^o\|)\|g^o - \text{span}_n G\| + \varepsilon).$$

By Theorem 3.1(i) we have

$$(4.2) \quad \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) \leq \alpha \left( (1 + c\|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} + \varepsilon \right),$$

and by infimizing (4.2) over  $\varepsilon$ , we obtain

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

which completes the proof of (i).

(ii) By the definition of  $\varepsilon_n$ -minimum point,  $\Phi(g_n) - \Phi(g^o) \leq \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n$ . So by item (i) we have

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

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

(iii) By the definitions of  $\varepsilon_n$ -argmin and of the modulus of Tikhonov well-posedness of  $(M, \Phi)$  at  $g^\circ$  and by item (i), we have

$$\begin{aligned} \xi(\|g_n - g^\circ\|) &= \inf_{g \in M \cap S_{\|g_n - g^\circ\|}(g^\circ)} \Phi(g) - \Phi(g^\circ) \leq \Phi(g_n) - \Phi(g^\circ) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) \\ &\quad - \Phi(g^\circ) + \varepsilon_n \leq \alpha \left( (1 + c\|g^\circ\|) \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right) + \varepsilon_n. \end{aligned}$$

(iv) By the definition of  $\varepsilon_n$ -argmin, Proposition 2.1(iii), and item (i), we have

$$\begin{aligned} \delta(\|g_n - g^\circ\|) &\leq \Phi(g_n) - \Phi(g^\circ) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) + \varepsilon_n \\ &\leq \alpha \left( (1 + c\|g^\circ\|) \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right) + \varepsilon_n. \quad \square \end{aligned}$$

Theorem 4.2 shows that for  $\|g^\circ\|_G$  finite, the  $\varepsilon_n$ -near minimum points  $\{g_n\}$  form a  $\Phi$ -minimizing sequence over  $M$  and the speed of convergence of  $\{\Phi(g_n)\}$  to  $\Phi(g^\circ)$  is bounded from above by

$$\alpha \left( (1 + c\|g^\circ\|) \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right) + \varepsilon_n.$$

When minimization is performed over the whole space  $X$ , the Lipschitz constant of the Minkowski functional  $p_M = p_X$  is equal to zero; thus, Theorem 4.2 gives an upper bound

$$\alpha \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right) + \varepsilon_n,$$

which depends on the modulus of continuity  $\alpha$  of  $\Phi$ , on the  $G$ -variation norm of  $g^\circ$ , and on the ambient space norm of  $g^\circ$ .

When the admissible set is a ball  $B_r(\|\cdot\|)$ , the Lipschitz constant is  $1/r$  and we get an upper bound

$$\alpha \left( \left( 1 + \frac{\|g^\circ\|}{r} \right) \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right) + \varepsilon_n.$$

As the estimates derived from Theorem 4.2 are not merely asymptotic, they can be applied to any degree  $n$  of variable-basis functions. Moreover, for admissible sets of  $d$ -variable functions the estimates hold for any number  $d$  of variables. Inspection of the upper bounds from Theorem 4.2 allows one to describe problems for which the rates of approximate optimization do not exhibit the curse of dimensionality (i.e., the degree  $n$  of variable-basis functions required for a satisfactory approximate optimization does not grow exponentially with the number  $d$  of variables of admissible solutions). A sufficient property of such problems is that the  $G$ -variation of their minimum point  $g^\circ$  does not depend exponentially on the number  $d$  of variables. Examples of classes of functions with small variations with respect to some bases used in neurocomputing were given in [12, 58] (see also the discussion in section 8).

The next theorem is an extension of Theorem 4.2 to  $\mathcal{L}_p$ -spaces with  $p \in (1, \infty)$ . Its proof proceeds similarly as the proof of Theorem 4.2, but instead of the upper bound (i) from Theorem 3.1, it uses (ii) from the same theorem. The same remarks about the assumption  $0 \in \text{int } M$  and the replacement of  $X$  with  $M$  as those preceding Theorem 4.2 apply here, as any closed subspace of a reflexive Banach space is a reflexive Banach space [22, Proposition III.17].

**THEOREM 4.3.** *Let  $\Omega \subseteq \mathcal{R}^d$  be open,  $M$  and  $G$  be subsets of  $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$ ,  $p \in (1, \infty)$ ,  $G$  be bounded,  $s_G = \sup_{g \in G} \|g\|$ ,  $M$  be closed and convex,  $0 \in \text{int } M$ ,  $q = p/(p-1)$ ,  $\bar{p} = \min(p, q)$ , and  $\bar{q} = \max(p, q)$ . Let  $\Phi : X \rightarrow (-\infty, +\infty]$  be a proper functional,  $g^\circ \in \text{argmin}(M, \Phi)$ ,  $\Phi$  be continuous at  $g^\circ$  with a modulus of continuity  $\alpha$ , and  $\{\varepsilon_n\}$  be a sequence of positive reals such that  $g_n \in \text{argmin}_{\varepsilon_n}(M \cap \text{span}_n G, \Phi)$ . Then  $p_M$  is Lipschitz on  $X$ , and if  $c$  is its Lipschitz constant, then the following estimates hold for every positive integer  $n$ :*

(i)

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) \leq \alpha \left( (1 + c\|g^\circ\|) \frac{2^{1/\bar{p}} s_G \|g^\circ\|_G}{n^{1/\bar{q}}} \right);$$

(ii) *if  $\|g^\circ\|_G < \infty$  and  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ , then  $\{g_n\}$  is a  $\Phi$ -minimizing sequence over  $M$  and*

$$\Phi(g_n) - \Phi(g^\circ) \leq \alpha \left( (1 + c\|g^\circ\|) \frac{2^{1/\bar{p}} s_G \|g^\circ\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n;$$

(iii) *if  $\xi$  is the modulus of Tikhonov well-posedness of  $(M, \Phi)$  at  $g^\circ$ , then*

$$\xi(\|g_n - g^\circ\|) \leq \alpha \left( (1 + c\|g^\circ\|) \frac{2^{1/\bar{p}} s_G \|g^\circ\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n;$$

(iv) *if  $\Phi$  is uniformly convex on  $M$  with a modulus of convexity  $\delta$ , then*

$$\delta(\|g_n - g^\circ\|) \leq \alpha \left( (1 + c\|g^\circ\|) \frac{2^{1/\bar{p}} s_G \|g^\circ\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n.$$

In the calculus of variations, the notion of a *direct method* [37, p. 192] is used to refer to a method for solving an optimization problem  $(M, \Phi)$  by obtaining its minimum point  $g^\circ$  as a limit of a  $\Phi$ -minimizing sequence  $\{g_n\} \subseteq M$  satisfying  $\lim_{n \rightarrow \infty} \Phi(g_n) = \Phi(g^\circ)$ . Using this notion, we can rephrase our results as conditions on  $(M, \Phi)$  under which the extended Ritz method has some of the properties of a direct method. By Theorems 4.2 and 4.3 for  $\|g^\circ\|_G$  finite, any sequence  $\{g_n\}$  of  $\varepsilon_n$ -minimum points of  $(M \cap \text{span}_n G, \Phi)$  is  $\Phi$ -minimizing and  $\Phi(g^\circ) = \lim_{n \rightarrow \infty} \Phi(g_n) = \Phi(\lim_{n \rightarrow \infty} g_n)$ . The convergence of  $\{g_n\}$  to  $g^\circ$  is not always guaranteed (it depends on the behavior of the modulus of Tikhonov well-posedness of  $(M, \Phi)$  at  $g^\circ$ ). However, when applied to convex best approximation problems (see section 6) and to learning from data by kernel methods (see section 7), the extended Ritz method is a direct method.

**5. Asymptotic estimates for convex problems.** For convex problems with the functional to be minimized bounded in a neighborhood of a minimum point, under an additional assumption of density of  $M \cap \text{span } G$  in  $M$ , the upper bounds from Theorem 4.2 can be simplified. But the simplified bounds are only asymptotic as their derivation takes advantage of the local behavior of the functional in a neighborhood

of a minimum point. For  $f, g : \mathcal{N}_+ \rightarrow \mathcal{N}_+$  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  $(X, \|\cdot\|)$  be a Hilbert space,  $M$  and  $G$  be its subsets,  $G$  be bounded,  $s_G = \sup_{g \in G} \|g\|$ ,  $M$  be closed and convex,  $0 \in \text{int } M$ , and  $M \cap \text{span } G$  dense in  $M$ . Let  $\Phi : X \rightarrow (-\infty, +\infty]$  be a proper convex functional,  $g^\circ \in \text{argmin}(M, \Phi)$  be such that  $\Phi$  is bounded in its neighborhood,  $\{\varepsilon_n\}$  be a sequence of positive reals such that  $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$  and  $g_n \in \text{argmin}_{\varepsilon_n}(M \cap \text{span}_n G, \Phi)$ . Then the following estimates hold for every positive integer  $n$ :*

(i)

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) \leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right);$$

(ii) *if  $\|g^\circ\|_G < \infty$ , then  $\{g_n\}$  is a  $\Phi$ -minimizing sequence over  $M$  and*

$$\Phi(g_n) - \Phi(g^\circ) \leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right);$$

(iii) *if  $\xi$  is the modulus of Tikhonov well-posedness of  $(M, \Phi)$  at  $g^\circ$ , then*

$$\xi(\|g_n - g^\circ\|) \leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right);$$

(iv) *if  $\Phi$  is uniformly convex on  $M$  with a modulus of convexity  $\delta$ , then*

$$\delta(\|g_n - g^\circ\|) \leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right).$$

*Proof.* (i) Let  $\nu > 0$  be such that  $\Phi$  is bounded on  $B_\nu(g^\circ, \|\cdot\|)$ . As  $B_\nu(g^\circ, \|\cdot\|) \subseteq \text{dom } \Phi$ , we have  $g^\circ \in \text{int dom } \Phi$ . Since  $\Phi$  is a proper convex functional bounded on  $B_\nu(g^\circ, \|\cdot\|)$ ,  $\Phi$  is locally Lipschitz on  $B_\nu(g^\circ, \|\cdot\|)$  [35, Corollary 2.4, p. 12]. Let  $\eta \leq \nu$  be such that  $\Phi$  is Lipschitz continuous with the constant  $c_1$  on  $B_\eta(g^\circ, \|\cdot\|)$ .

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

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

$$(5.1) \quad \|h_n^\varepsilon - g^\circ\| \leq (1 + c\|g^\circ\|) \|f_n^\varepsilon - g^\circ\| < \eta.$$

So  $h_n^\varepsilon$  is in the ball  $B_\eta(g^\circ, \|\cdot\|)$ , on which  $\Phi$  is Lipschitz continuous with the constant  $c_1$ . So we have

$$(5.2) \quad \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) \leq \Phi(h_n^\varepsilon) - \Phi(g^\circ) \leq c_1 \|h_n^\varepsilon - g^\circ\|.$$

From (5.1) and (5.2) we obtain

$$(5.3) \quad \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) \leq C \|f_n^\varepsilon - g^\circ\|,$$

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

$$(5.4) \quad \|g^\circ - f_n^\varepsilon\| \leq \|g^\circ - \text{span}_n G\| + \varepsilon \leq \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} + \varepsilon.$$

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

$$\inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) \leq C \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}}.$$

(ii) As  $g_n \in \text{argmin}_{\varepsilon_n} (M \cap \text{span}_n G)$ , we have  $\Phi(g_n) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) + \varepsilon_n$ . Combining this inequality with the one from item (i) and with  $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ , we obtain

$$\Phi(g_n) - \Phi(g^\circ) \leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right).$$

(iii) By the definitions of  $\varepsilon_n$ -argmin and of the modulus of Tikhonov well-posedness of  $(M, \Phi)$  at  $g^\circ$  and by item (i), for every  $n \geq n_0$  we have

$$\begin{aligned} \xi(\|g_n - g^\circ\|) &= \inf_{g \in M \cap S_{\|g_n - g^\circ\|}(g^\circ)} \Phi(g) - \Phi(g^\circ) \\ &\leq \Phi(g_n) - \Phi(g^\circ) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) + \varepsilon_n \\ &\leq C \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} + \varepsilon_n. \end{aligned}$$

As  $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ , we obtain

$$\xi(\|g_n - g^\circ\|) \leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right).$$

(iv) By the definition of  $\varepsilon_n$ -argmin, item (i), and Proposition 2.1(iii), for all  $n \geq n_0$  we get

$$\begin{aligned} \delta(\|g_n - g^\circ\|) &\leq \Phi(g_n) - \Phi(g^\circ) < \inf_{g \in M \cap \text{span}_n G} \Phi(g) - \Phi(g^\circ) + \varepsilon_n \\ &\leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right) + \varepsilon_n. \end{aligned}$$

As  $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ , we obtain

$$\delta(\|g_n - g^\circ\|) \leq \mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right). \quad \square$$

Inspection of the proof of Theorem 5.1 shows that there exists  $n_0 \in \mathcal{N}_+$  such that the expression

$$\mathcal{O} \left( \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} \right)$$

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  $\Phi$  in a neighborhood of  $g^o$ . The proof also shows that for any sequences  $\{\varepsilon_n\}$  and  $\{g_n\}$  such that  $g_n \in \operatorname{argmin}_{\varepsilon_n}(M, \Phi)$ , 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 the estimates derived in section 4 is an 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* [34, p. 40]. We recall that  $M$  is a *Chebyshev set* if each  $f \in X$  has a unique best approximation in  $M$  (i.e., there exists a unique  $g^o \in M$  such that  $\|f - g^o\| = \|f - M\|$ ) [29, p. 21].

In [43], the classical Ritz method was used to solve approximately the problem  $(M, e_f)$  with  $M$  a closed separable subspace of  $X$ , but rates of convergence were not estimated. For  $X$  finite-dimensional, other methods of approximate solution of the best approximation problem have also been studied, and, for some of them, estimates of rates of convergence have been derived (e.g., [44, pp. 118–122]). For  $X$  infinite-dimensional, a method of approximation of best approximation for which estimates of rates of convergence are available is Dijkstra’s algorithm [29, p. 207] applied to a special class of admissible sets  $M$  of the form  $\bigcup_{i=1}^r M_i$ , where  $M_i$  are closed affine sets and  $r$  is finite [29, p. 201].

Here, taking advantage of the upper bounds from section 4, we estimate rates of convergence of approximate solutions of the problem  $(M, e_f)$ , where  $M$  is closed and convex, that are obtained by the extended Ritz method. Applying Theorem 4.2 to the best approximation problems  $(M, e_f)$  and  $(M, e_f^2)$ , we derive the following upper bounds.

**THEOREM 6.1.** *Let  $(X, \|\cdot\|)$  be a Hilbert space,  $M$  and  $G$  be its subsets,  $G$  be bounded,  $s_G = \sup_{g \in G} \|g\|$ ,  $M$  be closed and convex,  $0 \in \operatorname{int} M$ , and  $f \in X$ . Then  $p_M$  is Lipschitz on  $X$ , and if  $c$  is its Lipschitz constant, then there exists a unique minimum point  $g^o$  of  $(M, e_f)$ , and the following estimates hold for every positive integer  $n$ :*

(i)

$$\begin{aligned} \inf_{g \in M \cap \operatorname{span}_n G} e_f(g) - e_f(g^o) &= \|f - M \cap \operatorname{span}_n G\| - \|f - M\| \\ &\leq (1 + c\|g^o\|) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}}; \end{aligned}$$

- (ii) if  $M$  is bounded,  $\{\varepsilon_n\}$  is a sequence of positive reals, and for every  $n$ ,  $g_n \in \operatorname{argmin}_{\varepsilon_n} (M \cap \operatorname{span}_n G, e_f^2)$ , then

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

*Proof.* As every closed convex subset of a Hilbert space is Chebyshev [29, p. 35], the problem  $(M, e_f)$  has a unique minimum point.

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) to  $(M, e_f)$  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 a minimum point of  $(M, e_f^2)$  and by Proposition 2.1(iv), the functional  $\|\cdot\|^2$  is uniformly convex with a modulus of convexity  $\delta(t) = t^2$ .

By the triangle inequality, for every  $h, g \in X$  we have  $|e_f^2(h) - e_f^2(g)| = |(\|f - h\| - \|f - g\|)(\|f - h\| + \|f - g\|)| \leq 2 \operatorname{diam} M \|h - g\|$ , and so  $\alpha(t) = 2t \operatorname{diam} M$  is an upper bound on the modulus of continuity of  $e_f^2$ . Thus, by Theorem 4.2(iv) we get

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

Combining Theorem 4.3 with estimates of moduli of convexity of  $\mathcal{L}_p$ -spaces,  $p \in (1, \infty)$ , we obtain the following upper bounds.

**THEOREM 6.2.** *Let  $\Omega \subseteq \mathcal{R}^d$  be open,  $M$  and  $G$  be subsets of  $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$ ,  $p \in (1, \infty)$ ,  $G$  be bounded,  $s_G = \sup_{g \in G} \|g\|$ ,  $M$  be closed and convex,  $0 \in \operatorname{int} M$ ,  $f \in X$ ,  $q = p/(p-1)$ ,  $\bar{p} = \min(p, q)$ ,  $\bar{q} = \max(p, q)$ , and  $\alpha_p, \alpha_q$  be moduli of continuity of  $e_f^p$  and  $e_f^q$ , respectively, at  $f$ . Then  $p_M$  is Lipschitz on  $X$ , and if  $c$  is its Lipschitz constant, then there exists a unique minimum point  $g^o$  of  $(M, e_f)$ , and the following estimates hold for every positive integer  $n$ :*

- (i)  $\inf_{g \in M \cap \operatorname{span}_n G} e_f(g) - e_f(g^o) \leq (1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}}$ ;
- (ii) if  $\{\varepsilon_n\}$  is a sequence of positive reals,  $p \in (1, 2]$ , and  $g_n \in \operatorname{argmin}_{\varepsilon_n} (M \cap \operatorname{span}_n G, e_f^q)$ , then

$$\|g_n - g^o\|^q \leq 2^{q-2} \alpha_q \left( (1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n;$$

- (iii) if  $\{\varepsilon_n\}$  is a sequence of positive reals,  $p \geq 2$ , and  $g_n \in \operatorname{argmin}_{\varepsilon_n} (M \cap \operatorname{span}_n G, e_f^p)$ , then

$$\|g_n - g^o\|^p \leq 2^{p-2} \alpha_p \left( (1 + c\|g^o\|) \frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}} \right) + \varepsilon_n.$$

*Proof.* Since for all  $p \in (1, \infty)$  the space  $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$  is uniformly convex [2, 2.29] and every convex best approximation problem in a uniformly convex space is Tikhonov well-posed [34, p. 40], there exists a unique  $g^o \in M$  such that  $\|f - g^o\|_p = \|f - M\|_p$ . Hence, the problem  $(M, e_f)$  has a unique minimum point.

(i) 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  $\alpha(t) = t$ . Thus, by Theorem 4.3(i) we obtain (i).

(ii) When  $p \in (1, 2]$ , the estimate follows from Theorem 4.3(iv) applied to the functional  $e_f^q$  with  $q = p/(p - 1)$ , combined with Proposition A.3(i).

(iii) When  $p \geq 2$ , the estimate follows from Theorem 4.3(iv) applied to the functional  $e_f^p$ , combined with Proposition A.3(ii).  $\square$

So Theorems 6.1(i) and 6.2(i) extend Theorem 3.1 on approximation by  $\text{span}_n G$  to approximation by  $M \cap \text{span}_n G$ , where  $M$  is closed, convex, with zero in its interior, in particular when  $M = B_r(\|\cdot\|)$  for some  $r > 0$ .

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

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

$$\|f - M \cap \text{span}_n G\| \leq (1 + c\|g^\circ\|) \sqrt{\frac{(s_G \|g^\circ\|_G)^2 - \|g^\circ\|^2}{n}} + \|f - g^\circ\|;$$

(ii) *if  $(X, \|\cdot\|) = (\mathcal{L}_p(\Omega), \|\cdot\|_p)$ , where  $p \in (1, \infty)$ ,  $\Omega \subseteq \mathcal{R}^d$  is open,  $q = p/(p - 1)$ ,  $\bar{p} = \min(p, q)$ , and  $\bar{q} = \max(p, q)$ , then*

$$\|f - M \cap \text{span}_n G\| \leq (1 + c\|g^\circ\|) \frac{2^{1/\bar{p}} s_G \|g^\circ\|_G}{n^{1/\bar{q}}} + \|f - g^\circ\|.$$

Note that for  $M = X$ , Corollary 6.3 gives the same estimate as Theorem 3.1, since in such a case the Lipschitz constant of  $p_M$  is equal to 0 and  $g^\circ = f$ .

**7. Application to learning from data.** Learning from a sample  $\{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \dots, m\}$  of empirical data can be modeled as minimization of the empirical error functional (also called the empirical risk functional), defined as

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

However, the empirical error does not take into account any global properties of the input/output mapping from which the sample was chosen. Such properties can be expressed through *regularization*, which replaces the functional  $\mathcal{E}$  with  $\mathcal{E}_{\gamma, \Psi} = \mathcal{E} + \gamma \Psi$ , where  $\Psi$  is a suitable functional called *stabilizer* and  $\gamma$  is a positive real number called *regularization parameter*. The stabilizer penalizes the solutions with some undesired properties such as high-frequency oscillations, while the regularization parameter allows us to obtain a tradeoff between fitting to empirical data and fitting to properties of solutions expressed by the stabilizer.

An important class of stabilizers are squares of norms on reproducing kernel Hilbert spaces. A *reproducing kernel Hilbert space* (RKHS)  $(\mathcal{H}_K(\Omega), \|\cdot\|_K)$  is a Hilbert space of functions defined on a set  $\Omega$  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  $K : \Omega \times \Omega \rightarrow \mathcal{R}$ , called *kernel*, such that for any  $f \in \mathcal{H}_K(\Omega)$  and any  $x \in \Omega$ ,  $\mathcal{F}_x(f) = \langle f, K(x, \cdot) \rangle_K$  [7] (recall that a mapping  $K : \Omega \times \Omega \rightarrow \mathcal{R}$  is *positive semidefinite* on  $\Omega$  if for all positive integers  $m$ , all  $(a_1, \dots, a_m) \in \mathcal{R}^m$ , and all  $(x_1, \dots, x_m) \in \Omega^m$ ,  $\sum_{i,j=1}^m a_i a_j K(x_i, x_j) \geq 0$ ).

By the Cauchy–Schwarz inequality, for every  $f \in \mathcal{H}_K(\Omega)$  and every  $x \in \Omega$  we have  $|f(x)| = |\langle f, K(x, \cdot) \rangle_K| \leq \|f\|_K \sqrt{K(x, x)} \leq c_K \|f\|_K$ , where  $c_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$ .

Thus for every set  $\Omega$  and every kernel  $K$  on  $\Omega \times \Omega$  we have

$$(7.1) \quad \sup_{x \in \Omega} |f(x)| \leq c_K \|f\|_K.$$

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

$$(7.2) \quad \mathcal{E}_{\gamma, K}(f) = \frac{1}{m} \sum_{i=1}^m |f(x_i) - y_i|^2 + \gamma \|f\|_K^2.$$

The representer theorem (see, e.g., [25, p. 42], [69, pp. 538–539]) states that the problem  $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma, K})$  has a unique minimum point  $g^o$  of the form

$$(7.3) \quad g^o(x) = \sum_{i=1}^m a_i K(x, x_i).$$

It even gives a formula for computing the parameters  $a = (a_1, \dots, a_m)$  as the unique solution of the well-posed system of linear equations

$$(7.4) \quad (\mathcal{K}[x] + \gamma m \mathcal{I})a = y,$$

where  $y = (y_1, \dots, y_m)$ ,  $\mathcal{K}[x]$  is the  $m \times m$  matrix defined as  $\mathcal{K}[x]_{ij} = K(x_i, x_j)$ , and  $\mathcal{I}$  is the identity matrix [25, 69].

Thus, to compute the coefficients  $(a_1, \dots, a_m)$  of the linear combination it is necessary to solve the inverse problem (7.4), which may be ill-conditioned. To guarantee for a given  $m$  a small condition number [63, p. 33] of the matrix  $\mathcal{K}[x] + \gamma m \mathcal{I}$ , the regularization parameter  $\gamma$  must be “large” [57]. On the other hand, a “large”  $\gamma$  does not allow good interpolation of the empirical data. This limits the applicability of algorithms for computing the solution of the problem  $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma, K})$  given by the representer theorem.

It has been argued in [38, p. 219] that the “regularization principles lead to approximation schemes that are equivalent to networks with one layer of hidden units.” Indeed, the unique minimum point of the problem  $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma, K})$  is in the set  $\text{span}_m G_K$ , 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 estimates how quickly approximate solutions achievable by networks with  $n$  hidden units converge to the minimum point described by the representer theorem. We first state in the next proposition basic properties of the functional  $\mathcal{E}_{\gamma, K}$ .

**PROPOSITION 7.1.** *Let  $\Omega$  be a nonempty set,  $K : \Omega \times \Omega \rightarrow \mathcal{R}$  be a kernel,  $c_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$ ,  $\gamma > 0$ ,  $m$  be a positive integer,  $\{(x_1, y_1), \dots, (x_m, y_m)\} \subset (\Omega \times \mathcal{R})^m$ ,  $\mathcal{E}_{\gamma, K}(f) = \frac{1}{m} \sum_{i=1}^m |f(x_i) - y_i|^2 + \gamma \|f\|_K^2$ , and  $y_{\max} = \max\{|y_i| : i = 1, \dots, m\}$ . Then*

- (i) *the functional  $\mathcal{E}_{\gamma, K}$  is uniformly convex on  $\mathcal{H}_K(\Omega)$  with a modulus of convexity  $\delta(t) = \gamma t^2$ ;*

- (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  $a_1 = 2(\|f\|_K c_K^2 + y_{\max} c_K + \gamma \|f\|_K)$  and  $a_2 = c_K^2 + \gamma$ ;
- (iii) when  $M \subset \mathcal{H}_K(\Omega)$  is closed, convex, and bounded, or when  $M = \mathcal{H}_K(\Omega)$ , the problem  $(M, \mathcal{E}_{\gamma,K})$  has a unique minimum point  $g^\circ$ ;
- (iv) for any minimum point  $g^\circ$  of  $(M, \mathcal{E}_{\gamma,K})$  and any  $f \in M$ ,  $\|f - g^\circ\|_K^2 \leq \frac{|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g^\circ)|}{\gamma}$ .

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

(ii) Let  $f \in \mathcal{H}_K(\Omega)$ ,  $t > 0$ , and  $g \in \mathcal{H}_K(\Omega)$  be such that  $\|f - g\|_K < t$ . Then

$$\begin{aligned} |\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| &= \left| \frac{1}{m} \sum_{i=1}^m ((f(x_i) - y_i)^2 - (g(x_i) - y_i)^2) + \gamma (\|f\|_K^2 - \|g\|_K^2) \right| \\ &\leq \left| \frac{1}{m} \sum_{i=1}^m (f(x_i) - g(x_i)) (f(x_i) + g(x_i) - 2y_i) \right| + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K) \\ &\leq \sup_{x \in \Omega} |f(x) - g(x)| \left| \sup_{x \in \Omega} |f(x) + g(x)| + 2y_{\max} \right| + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K). \end{aligned}$$

Thus, by (7.1) we have  $|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| \leq c_K \|f - g\|_K |c_K \|f + g\|_K + 2y_{\max}| + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K)$ . As  $\|g\|_K < \|f\|_K + t$ , we obtain

$$\begin{aligned} |\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| &< t c_K \left| 2\|f\|_K c_K + t c_K + 2y_{\max} \right| + \gamma t (2\|f\|_K + t) \\ &\leq t c_K \left( 2\|f\|_K c_K + t c_K + 2y_{\max} \right) + \gamma t (2\|f\|_K + t) \\ &= t^2 (c_K^2 + \gamma) + 2t \left( \|f\|_K c_K^2 + y_{\max} c_K + \gamma \|f\|_K \right). \end{aligned}$$

Hence,  $\|f - g\|_K < t$  implies  $|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g)| < \alpha(t) = a_2 t^2 + a_1 t$ , where  $a_2 = c_K^2 + \gamma$  and  $a_1 = 2(\|f\|_K c_K^2 + y_{\max} c_K + \gamma \|f\|_K)$ .

(iii) When  $M \subset \mathcal{H}_K(\Omega)$  is closed, convex, and bounded, the existence of a unique minimum point of  $(M, \mathcal{E}_{\gamma,K})$  follows from (i) and [70, Theorem 5]. When  $M = \mathcal{H}_K(\Omega)$ , it follows from the representer theorem [69, pp. 538–539].

(iv) follows from (i) and Proposition 2.1 (iii).  $\square$

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  $a_2 t^2 + a_1 t$ . Note that  $a_2$  depends on  $m$ ,  $c_K$ , and  $\gamma$ , while  $a_1$  depends, in addition to these values, also on  $\|f\|_K$  and  $y_{\max}$ .

Applying Proposition 7.1 and Theorem 4.2 to the problem  $(\mathcal{H}(\Omega), \mathcal{E}_{\gamma,K})$ , we obtain the following estimates, which hold for every positive integer  $n$  (but are only useful for  $n < m$ , as the minimum point  $g^\circ$  is in  $\text{span}_m G_K$ ).

**THEOREM 7.2.** *Let  $\Omega$  be a nonempty set,  $K : \Omega \times \Omega \rightarrow \mathcal{R}$  be a kernel,  $c_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$ ,  $(\mathcal{H}_K(\Omega), \|\cdot\|_K)$  be the RKHS defined by  $K$ ,  $G_K = \{K(x, \cdot) : x \in \Omega\}$ ,  $(x_1, \dots, x_m) \in \Omega^m$ ,  $(y_1, \dots, y_m) \in \mathcal{R}^m$ ,  $y_{\max} = \max\{|y_i| : i = 1, \dots, m\}$ ,  $\gamma > 0$ ,  $\mathcal{E}_{\gamma,K}(f) = \frac{1}{m} \sum_{i=1}^m |f(x_i) - y_i|^2 + \gamma \|f\|_K^2$ ,  $g^\circ(x) = \sum_{i=1}^m w_i K(x, x_i)$  be the unique minimum point 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  $g_n \in \text{argmin}_{\varepsilon_n} (\text{span}_n G_K, \mathcal{E}_{\gamma,K})$ . Then the following estimates hold for every positive integer  $n$ :*

(i)

$$\inf_{g \in \text{span}_n G_K} \mathcal{E}_{\gamma, K}(g) - \mathcal{E}_{\gamma, K}(g^o) \leq \alpha \left( \sqrt{\frac{(c_K \|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}} \right);$$

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

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

(iii)

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

(iv)

$$\gamma \sup_{x \in \Omega} |g_n(x) - g^o(x)|^2 \leq c_K \left( \alpha \left( \sqrt{\frac{(c_K \|g^o\|_{G_K})^2 - \|g^o\|_K^2}{n}} \right) + \varepsilon_n \right),$$

where  $\alpha(t) = a_2 t^2 + a_1 t$ ,  $a_1 = 2(\|g^o\|_K c_K^2 + y_{\max} c_K + \gamma \|g^o\|_K)$  and  $a_2 = c_K^2 + \gamma$ .

*Proof.* Statements (i) and (ii) follow from Theorem 4.2 with  $M = \mathcal{H}_K(\Omega)$ ,  $c = 0$  (as in this case  $p_M$  is the constant functional equal to zero),  $\Phi(f) = \mathcal{E}_{\gamma, K}(f)$ ,  $G = G_K$ , and  $s_G = \sup_{x \in \Omega} \|K(x, \cdot)\|_K = \sup_{x \in \Omega} \sqrt{\langle K(x, \cdot), K(x, \cdot) \rangle_K} = \sup_{x \in \Omega} \sqrt{K(x, x)} = c_K$ . By Proposition 7.1(i) and (ii),  $\mathcal{E}_{\gamma, K}$  is uniformly convex with a modulus of convexity  $\delta(t) = \gamma t^2$  and is continuous at  $g^o$  with a modulus of continuity  $\alpha(t) = a_2 t^2 + a_1 t$ , so (iii) follows from Theorem 4.2(iv). Finally, item (iv) follows from (iii) and inequality (7.1).  $\square$

For this application of Theorem 4.2, an explicit formula (7.3) describing the minimum point  $g^o$  is given by the representer theorem. Taking advantage of this formula, estimates of  $\|g^o\|_{G_K}$  and  $\|g^o\|$  in terms of the properties of the sample  $\{(x_i, y_i), i = 1, \dots, m\}$  of data, the kernel  $K$ , and the regularization parameter  $\gamma$  were derived in [57].

**8. Discussion.** We have derived upper bounds on rates of approximate optimization by the extended Ritz method for problems  $(M, \Phi)$  having a minimum point, where  $\Phi$  is continuous and  $M$  is closed, convex, and containing 0 in its interior. The bounds can be applied to a variety of problems with sets of admissible solutions equal to the ambient space, to its subspaces (restating the problems for the subspaces), and to balls of some radii in the ambient norm. Such admissible sets occur, for example, in Tikhonov's and Ivanov's regularizations.

The critical term in the bounds is of the form  $n^{-1/2}$  multiplied by the variation norm of the minimum point. To take advantage of these bounds, one needs some insights into the behavior of the variation norm tailored to the basis used for the extended Ritz method. Various methods based on integral representations (such as the Fourier transform [12, 21, 45, 58] and the Radon transform [49, 54]) were proposed to estimate the variation norm. For a survey of properties of  $G$ -variation see [53].

The role of variation norms in variable-basis approximation can be clarified by a comparison with the role played by Sobolev norms in linear approximation. Rates of linear approximation of order  $\mathcal{O}(n^{-1/2})$  for functions of  $d$  variables can be achieved when the approximation is restricted to functions from balls in Sobolev norms of degree  $s = d/2$  [67, pp. 232–233]. Similarly, in variable-basis approximation rates bounded from above by  $rn^{-1/2}$  can be obtained by restricting the approximation to balls of radii  $r$  in  $G$ -variation. Note that the “ $\mathcal{O}$ ” notation in estimates of rates of linear approximation of functions from balls in Sobolev norms hides “constants” that may depend on  $d$  [67, pp. 232–241]. Moreover with  $d$  increasing, balls in Sobolev spaces of degree  $s = d/2$  “shrink,” since some  $d$ -variable functions in the unit balls in the Sobolev norms  $\|\cdot\|_{d/2,p}$  with “large”  $((d + 1)/2)$ th derivatives cannot be extended to  $(d + 1)$ -variable functions from the unit balls in the Sobolev norms  $\|\cdot\|_{(d+1)/2,p}$ . In contrast to the linear approximation case, in variable-basis approximation with certain types of bases (such as those generated by neural-network computational units [53]) there exist families of sets of  $d$ -variable functions that can be approximated with rates  $n^{-1/2}$  and do not shrink as  $d$  increases.

**Appendix A.** For the reader’s convenience, here we state and prove some results from nonlinear approximation theory, which are used in section 4.

The following theorem states the Maurey–Jones–Barron estimate in a slightly reformulated way. The proof, which is a mild simplification of the argument from [12, Lemma 1, p. 934], is from [53]. By  $\text{conv}_n G$  is denoted the set of all convex combinations of at most  $n$  elements of the set  $G$ , i.e.,

$$\text{conv}_n G = \left\{ \sum_{i=1}^n w_i g_i : w_i \in [0, 1], \sum_{i=1}^n w_i = 1, g_i \in G \right\}.$$

**THEOREM A.1.** *Let  $(X, \|\cdot\|)$  be a Hilbert space,  $G$  be its bounded subset, and  $s_G = \sup_{g \in G} \|g\|$ . Then for every  $f \in \text{cl conv } G$  and for every positive integer  $n$ ,*

$$\|f - \text{conv}_n G\| \leq \sqrt{\frac{s_G^2 - \|f\|^2}{n}}.$$

*Proof.* Since the distance from  $\text{conv}_n G$  is continuous on  $(X, \|\cdot\|)$  [72, p. 391], it is sufficient to verify the statement for  $f \in \text{conv } G$ . Let  $f = \sum_{j=1}^m a_j h_j$  be a representation of  $f$  as a convex combination of elements of  $G$ . Set  $c = s_G^2 - \|f\|^2$ . We show by induction that there exists a sequence  $\{g_i\}$  of elements of  $G$  such that the barycenters  $f_n = \sum_{i=1}^n \frac{g_i}{n}$  satisfy  $e_n^2 = \|f - f_n\|^2 \leq \frac{c}{n}$ .

First we check that there exists  $g_1 \in G$  such that  $f_1 = g_1$  satisfies  $e_1^2 = \|f - f_1\|^2 \leq c$ . As  $\sum_{j=1}^m a_j \|f - h_j\|^2 = \|f\|^2 - 2 \langle f, \sum_{j=1}^m a_j h_j \rangle + \sum_{j=1}^m a_j \|h_j\|^2 \leq s_G^2 - \|f\|^2 = c$ , there must exist at least one  $j \in \{1, \dots, m\}$  for which  $\|f - h_j\|^2 \leq c$ .

Setting  $g_1 = h_j$  and assuming that we already have  $g_1, \dots, g_n$ , we derive the estimate by induction. We express  $e_{n+1}^2$  in terms of  $e_n^2$  as  $e_{n+1}^2 = \|f - f_{n+1}\|^2 = \|\frac{n}{n+1}(f - f_n) + \frac{1}{n+1}(f - g_{n+1})\|^2 = \frac{n^2}{(n+1)^2} e_n^2 + \frac{2n}{(n+1)^2} \langle f - f_n, f - g_{n+1} \rangle + \frac{1}{(n+1)^2} \|f - g_{n+1}\|^2$ .

Analogously to the first step, we consider a convex combination of the last two terms from the formula expressing  $e_{n+1}^2$  in terms of  $e_n^2$ . Thus we obtain

$$\begin{aligned}
& \sum_{j=1}^m a_j \left( \frac{2n}{(n+1)^2} \langle f - f_n, f - h_j \rangle + \frac{1}{(n+1)^2} \|f - h_j\|^2 \right) \\
&= \frac{2n}{(n+1)^2} \left\langle f - f_n, f - \sum_{j=1}^m a_j h_j \right\rangle \\
&+ \frac{1}{(n+1)^2} \left( \|f\|^2 - 2 \left\langle f, \sum_{j=1}^m a_j h_j \right\rangle + \sum_{j=1}^m a_j \|h_j\|^2 \right) \\
&= \frac{1}{(n+1)^2} \left( \sum_{j=1}^m a_j \|h_j\|^2 - \|f\|^2 \right) \leq \frac{1}{(n+1)^2} (s_G^2 - \|f\|^2) = \frac{c}{(n+1)^2}.
\end{aligned}$$

So there must exist  $j \in \{1, \dots, m\}$  such that  $\frac{2n}{(n+1)^2} \langle f - f_n, f - h_j \rangle + \frac{1}{(n+1)^2} \|f - h_j\|^2 \leq \frac{c}{(n+1)^2}$ .

Setting  $g_{n+1} = h_j$ , we get  $e_{n+1}^2 \leq \frac{n^2}{(n+1)^2} e_n^2 + \frac{c}{(n+1)^2}$ . It can be easily verified by induction that this recursive formula together with  $e_1^2 \leq c$  gives  $e_n^2 \leq \frac{c}{n}$ .  $\square$

In [28], the Maurey–Jones–Barron estimate was extended to  $\mathcal{L}_p$ -spaces,  $p \in (1, \infty)$ , with a more sophisticated argument, replacing inner products with peak functionals and taking advantage of Clarkson's inequalities stated in the following proposition from [42, pp. 225, 227].

**PROPOSITION A.2** (Clarkson's inequalities). *Let  $\Omega \subseteq \mathcal{R}^d$ ,  $p \in (1, \infty)$ ,  $f, g \in (\mathcal{L}_p(\Omega), \|\cdot\|_p)$ , and  $q = p/(p-1)$ . Then for  $p \in (1, 2]$*

$$(A.1) \quad \left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \leq \left( \frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p \right)^{q-1},$$

$$(A.2) \quad \left\| \frac{f+g}{2} \right\|_p^p + \left\| \frac{f-g}{2} \right\|_p^p \geq \frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p,$$

and for  $p \geq 2$

$$(A.3) \quad \left\| \frac{f+g}{2} \right\|_p^p + \left\| \frac{f-g}{2} \right\|_p^p \leq \frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p,$$

$$(A.4) \quad \left\| \frac{f+g}{2} \right\|_p^q + \left\| \frac{f-g}{2} \right\|_p^q \geq \left( \frac{1}{2} \|f\|_p^p + \frac{1}{2} \|g\|_p^p \right)^{q-1}.$$

The next estimates follow from Clarkson's inequalities.

**PROPOSITION A.3.** *Let  $\Omega \subseteq \mathcal{R}^d$ ,  $p \in (1, \infty)$ , and  $q = p/(p-1)$ ,  $\bar{p} = \min(p, q)$ .*

Then

- (i) *if  $p \in (1, 2]$ , the functional  $e_f^q$  is uniformly convex with a modulus of convexity  $\delta(t) = \frac{t^q}{2^{q-2}}$ ;*
- (ii) *if  $p \geq 2$ , the functional  $e_f^p$  is uniformly convex with a modulus of convexity  $\delta(t) = \frac{t^p}{2^{p-2}}$ ;*
- (iii) *for all  $f, g \in (\mathcal{L}_p(\Omega), \|\cdot\|_p)$ ,  $\|f+g\|_p^{\bar{p}} + \|f-g\|_p^{\bar{p}} \leq 2(\|f\|_p^{\bar{p}} + \|g\|_p^{\bar{p}})$ .*

*Proof.* (i) By [2, Lemma 2.24], for every  $1 \leq r < \infty$  and  $a, b \geq 0$ ,  $(a + b)^r \leq 2^{r-1}(a^r + b^r)$ . Thus, by (A.1) we have

$$\left\| \frac{f + g}{2} \right\|_p^q + \left\| \frac{f - g}{2} \right\|_p^q \leq 2^{q-2} \left( \left( \frac{1}{2} \|f\|_p^p \right)^{q-1} + \left( \frac{1}{2} \|g\|_p^p \right)^{q-1} \right).$$

As  $p(q - 1) = q$ , we obtain

$$\left\| \frac{f + g}{2} \right\|_p^q + \left\| \frac{f - g}{2} \right\|_p^q \leq \frac{2^{q-2}}{2^{q-1}} (\|f\|_p^q + \|g\|_p^q) = \frac{1}{2} (\|f\|_p^q + \|g\|_p^q).$$

(ii) This follows directly from (A.3).

(iii) First suppose that  $p \in (1, 2]$ . Then  $p \leq q$  and so  $\bar{p} = \min\{p, q\} = p$ . Thus, by (A.2) we have

$$(A.5) \quad \|f\|_p^p + \|g\|_p^p \leq 2 \left( \left\| \frac{f + g}{2} \right\|_p^p + \left\| \frac{f - g}{2} \right\|_p^p \right).$$

Set  $\phi = \frac{f+g}{2}$  and  $\psi = \frac{f-g}{2}$ . Then  $f = \phi + \psi$  and  $g = \phi - \psi$ . So from (A.5) we get  $\|\psi + \phi\|_p^p + \|\phi - \psi\|_p^p \leq 2(\|\phi\|_p^p + \|\psi\|_p^p) = \|\psi + \phi\|_p^{\bar{p}} + \|\psi - \phi\|_p^{\bar{p}} \leq 2(\|\psi\|_p^{\bar{p}} + \|\phi\|_p^{\bar{p}})$ , which proves (iii) for  $p \in (1, 2]$ .

Now suppose  $p \geq 2$ . Then  $p \geq q$  and so  $\bar{p} = \min\{p, q\} = q$  and  $1/(q - 1) = p - 1$ . By (A.4) we have

$$(A.6) \quad (\|f\|_p^p + \|g\|_p^p) \leq 2 \left( \left\| \frac{f + g}{2} \right\|_p^q + \left\| \frac{f - g}{2} \right\|_p^q \right)^{\frac{1}{q-1}} \\ = 2 \left( \left\| \frac{f + g}{2} \right\|_p^q + \left\| \frac{f - g}{2} \right\|_p^q \right)^{p-1}.$$

As above, set  $\phi = \frac{f+g}{2}$  and  $\psi = \frac{f-g}{2}$ . Then  $f = \phi + \psi$  and  $g = \phi - \psi$ . So, from (A.7) we get

$$(A.7) \quad \|\psi + \phi\|_p^p + \|\phi - \psi\|_p^p \leq 2 (\|\phi\|_p^q + \|\psi\|_p^q)^{p-1}.$$

Since for every  $r \in [1, \infty)$  and  $a, b \geq 0$  we have  $(a + b)^r \leq 2^{r-1}(a^r + b^r)$  [2, 2.24], setting  $a = \|\phi + \psi\|_p^q$ ,  $b = \|\phi - \psi\|_p^q$ , and  $r = p/q = p - 1$  we obtain

$$(A.8) \quad \|\psi + \phi\|_p^p + \|\phi - \psi\|_p^p = \left( \|\phi + \psi\|_p^{\frac{q}{p}} + \|\phi - \psi\|_p^{\frac{q}{p}} \right) \\ \geq \frac{1}{2^{p-2}} (\|\psi + \phi\|_p^q + \|\phi - \psi\|_p^q)^{p-1}.$$

Thus, by (A.7) and (A.8) we have  $(\|\phi + \psi\|_p^q + \|\phi - \psi\|_p^q)^{p-1} \leq 2^{p-1} (\|\phi\|_p^q + \|\psi\|_p^q)^{p-1}$ . Hence  $\|\phi + \psi\|_p^q + \|\phi - \psi\|_p^q \leq 2 (\|\phi\|_p^q + \|\psi\|_p^q) = 2 (\|\phi\|_p^{\bar{p}} + \|\psi\|_p^{\bar{p}})$ , as  $q = \bar{p}$ . This proves (iii) for  $p \geq 2$ .  $\square$

The next theorem is a slight reformulation of [28, Theorem 5]. The proof is a simplification of the argument from [28, proof of Theorem 5]. For a Banach space  $(X, \|\cdot\|)$  and  $f \in X$ , we denote by  $\Pi_f$  a *peak functional for f*, i.e., a continuous linear functional such that  $\|\Pi_f\| = 1$  and  $\Pi_f(f) = \|f\|$  [20, p. 1].

**THEOREM A.4.** Let  $\Omega \subseteq \mathcal{R}^d$  be open,  $G$  be a subset of  $(\mathcal{L}_p(\Omega), \|\cdot\|_p)$ ,  $p \in (1, \infty)$ ,  $f \in \text{cl conv } G$ , and  $r > 0$  such that  $G \subseteq B_r(f, \|\cdot\|)$ . Then for every positive integer  $n$ ,  $\|f - \text{span}_n G\|_p \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$ , where  $q = p/(p-1)$ ,  $\bar{p} = \min(p, q)$ , and  $\bar{q} = \max(p, q)$ .

*Proof.* As in the proof of Theorem A.1, it is sufficient to verify the statement for  $f \in \text{conv } G$ . Let  $f = \sum_{j=1}^m a_j h_j$  be a representation of  $f$  as a convex combination of elements of  $G$ . We show by induction that there exists a sequence  $\{g_i\}$  of elements of  $G$  such that the barycenters  $f_n = \sum_{i=1}^n \frac{g_i}{n}$  satisfy  $e_n = \|f - f_n\| \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$ .

First we check that there exists  $g_1 \in G$  such that  $f_1 = g_1$  satisfies  $e_1 = \|f - f_1\|_p \leq 2^{1/\bar{p}} r$ . This holds trivially as  $G \subseteq B_r(f, \|\cdot\|)$ , so for any  $g \in G$  we have  $\|f - g\| \leq r < 2^{1/\bar{p}} r$ . Hence we can set  $f_1 = g_1$  for any  $g_1 \in G$ .

Assume that we already have  $g_1, \dots, g_n$ . Then  $f_{n+1} = \frac{n}{n+1} f_n + \frac{1}{n+1} g_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} g_i$ . We shall express  $e_{n+1}^{\bar{p}}$  in terms of  $e_n^{\bar{p}}$ .

Let  $\Pi_n$  be a peak functional for  $f - f_n$ . Since  $\sum_{j=1}^m a_j (f - h_j) = 0$ , by linearity of  $\Pi_n$  we have  $0 = \Pi_n (\sum_{j=1}^m a_j (f - h_j)) = \sum_{j=1}^m a_j \Pi_n (f - h_j)$ . Thus, there must exist  $j \in \{1, \dots, m\}$  such that  $\Pi_n (f - h_j) \leq 0$ . Set  $g_{n+1} = h_j$ , so  $\Pi_n (f - g_{n+1}) \leq 0$ . Hence, by Proposition A.3(iii) we get

$$\begin{aligned} e_{n+1}^{\bar{p}} &= \|f - f_{n+1}\|_p^{\bar{p}} = \left\| \frac{n}{n+1} (f - f_n) + \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \\ &\leq 2 \left( \left\| \frac{n}{n+1} (f - f_n) \right\|_p^{\bar{p}} + \left\| \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \right) \\ (A.9) \quad &\times \left\| \frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}}. \end{aligned}$$

As  $\|\Pi_n\| = 1$  and  $\Pi_n (f - g_{n+1}) \leq 0$ , we have

$$\begin{aligned} \left\| \frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right\|_p &\geq \left\| \Pi_n \left( \frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right) \right\|_p \\ &\geq \left\| \Pi_n \left( \frac{n}{n+1} (f - f_n) \right) \right\|_p \\ &= \frac{n}{n+1} \|\Pi_n (f - f_n)\|_p = \frac{n}{n+1} \|f - f_n\|_p. \end{aligned}$$

Hence

$$(A.10) \quad - \left\| \frac{n}{n+1} (f - f_n) - \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \leq - \left( \frac{n}{n+1} \|f - f_n\|_p \right)^{\bar{p}}.$$

By (A.9) and (A.10),

$$\begin{aligned} e_{n+1}^{\bar{p}} &= \|f - f_{n+1}\|_p^{\bar{p}} \leq 2 \left( \left\| \frac{n}{n+1} (f - f_n) \right\|_p^{\bar{p}} + \left\| \frac{1}{n+1} (f - g_{n+1}) \right\|_p^{\bar{p}} \right) \\ &\quad - \left( \frac{n}{n+1} \|f - f_n\|_p \right)^{\bar{p}} = \frac{2}{(n+1)^{\bar{p}}} \|f - g_{n+1}\|_p^{\bar{p}} + \left( \frac{2}{n+1} \right)^{\bar{p}} \|f - f_n\|_p^{\bar{p}} \\ &= \frac{2}{(n+1)^{\bar{p}}} \|f - g_{n+1}\|_p^{\bar{p}} + \left( \frac{2}{n+1} \right)^{\bar{p}} e_n^{\bar{p}}. \end{aligned}$$

As  $e_n = \|f - f_n\| \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$ , we get

$$\begin{aligned} e_{n+1}^{\bar{p}} &\leq \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}} + \left(\frac{n}{n+1}\right)^{\bar{p}} \left(\frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}\right)^{\bar{p}} = \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}} \left(1 + \frac{n^{\bar{p}}}{n^{\bar{p}/\bar{q}}}\right) \\ &= \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}} \left(1 + n^{\bar{p}-\bar{p}/\bar{q}}\right). \end{aligned}$$

It can be easily verified that  $\bar{p} - \frac{\bar{p}}{\bar{q}} = 1$  in both cases,  $\bar{p} = p$  (and so  $\bar{q} = q = \frac{p}{p-1}$ ), and  $\bar{p} = q$  (and so  $\bar{q} = p$ ). Thus  $e_{n+1}^{\bar{p}} \leq \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}}} (n+1)$ . As  $\bar{p} - 1 = \frac{\bar{p}}{\bar{q}}$  for both  $\bar{p} = p$  (hence  $\bar{q} = q$ ) and  $\bar{p} = q$  (hence  $\bar{q} = p$ ), we get  $e_{n+1}^{\bar{p}} \leq \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}-1}} = \frac{2 r^{\bar{p}}}{(n+1)^{\bar{p}/\bar{q}}}$ , i.e.,  $e_{n+1} \leq \frac{2^{1/\bar{p}} r}{(n+1)^{1/\bar{q}}}$ .  $\square$

Theorem 3.1 is a corollary of Theorems A.1 and A.4 in terms of  $G$ -variation. For  $t > 0$ , we define  $G(t) = \{wg : g \in G, w \in \mathcal{R}, |w| \leq t\}$ .

*Proof of Theorem 3.1.* (i) As  $\text{span}_n G \supseteq \text{conv}_n G$ , by Theorem A.1 applied to  $G(\|f\|_G)$  we have  $\|f - \text{span}_n G\| \leq \|f - \text{conv}_n G(\|f\|_G)\| \leq \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}}$ .

(ii) As  $\text{span}_n G \supseteq \text{conv}_n G$ , by applying Theorem A.4 to  $G(\|f\|_G)$  we get  $\|f - \text{span}_n G\| \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$  for every  $r$  such that  $G(\|f\|_G) \subseteq B_r(f, \|\cdot\|)$ . Set  $r = 2 s_G \|f\|_G$ . By (3.1), for every  $h \in G(\|f\|_G)$  we have  $\|h - f\| \leq \|h\| + \|f\| \leq s_G \|h\|_G + s_G \|f\|_G \leq 2 s_G \|f\|_G$ . So  $G(\|f\|_G) \subseteq B_{2r\|f\|_G}(f, \|\cdot\|)$ , and hence  $\|f - \text{span}_n G\| \leq \frac{2^{1/\bar{p}} s_G \|f\|_G}{n^{1/\bar{q}}}$ .

**Acknowledgments.** The authors thank R. Zoppoli (University of Genoa) for stimulating their interest in the theoretical investigation of approximate optimization by the extended Ritz method. They are also grateful to P. C. Kainen and A. Vogt (both of Georgetown University) for fruitful comments and discussions.

#### REFERENCES

- [1] E. AARTS AND J. KORST, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*, John Wiley and Sons, New York, 1989.
- [2] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [3] A. ALESSANDRI, M. BAGLIETTO, T. PARISINI, AND R. ZOPPOLI, *A neural state estimator with bounded errors for nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 2028–2042.
- [4] A. ALESSANDRI, M. SANGUINETI, AND M. MAGGIORE, *Optimization-based learning with bounded error for feedforward neural networks*, IEEE Trans. Neural Networks, 13 (2002), pp. 261–273.
- [5] A. ALESSANDRI AND M. SANGUINETI, *Optimization of approximating networks for optimal fault diagnosis*, Optim. Methods Softw., 20 (2005), pp. 235–260.
- [6] W. ALT, *On the approximation of infinite optimization problems with an application to optimal control problems*, Appl. Math. Optim., 12 (1984), pp. 15–27.
- [7] N. ARONSZAJN, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.
- [8] M. BAGLIETTO, T. PARISINI, AND R. ZOPPOLI, *Numerical solutions to the Witsenhausen counterexample by approximating networks*, IEEE Trans. Automat. Control, 46 (2001), pp. 1471–1477.
- [9] M. BAGLIETTO, T. PARISINI, AND R. ZOPPOLI, *Distributed-information neural control: The case of dynamic routing in traffic networks*, IEEE Trans. Neural Networks, 12 (2001), pp. 485–502.
- [10] M. BAGLIETTO, M. SANGUINETI, AND R. ZOPPOLI, *Facing the curse of dimensionality by the extended Ritz method in stochastic functional optimization: Dynamic routing in traffic networks*, in High Performance Algorithms and Software for Nonlinear Optimization, G. Di Pillo and A. Murli, eds., Kluwer Academic Publishers, Norwell, MA, 2003, pp. 23–56.

- [11] A. R. BARRON, *Neural net approximation*, in Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems, New Haven, CT, 1992, pp. 69–72.
- [12] A. R. BARRON, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. Theory, 39 (1993), pp. 930–945.
- [13] R. W. BEARD AND T. W. MCLAIN, *Successive Galerkin approximation algorithms for nonlinear optimal and robust control*, Internat. J. Control, 71 (1998), pp. 717–743.
- [14] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
- [15] M. BERTERO, *Linear inverse and ill-posed problems*, Adv. Electron. Electron. Phys., 75 (1989), pp. 1–120.
- [16] D. P. BERTSEKAS, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.
- [17] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.
- [18] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.
- [19] W. E. BOSARGE, JR., O. G. JOHNSON, R. S. MCKNIGHT, AND W. P. TIMLAKE, *The Ritz-Galerkin procedure for nonlinear control problems*, SIAM J. Numer. Anal., 10 (1973), pp. 94–111.
- [20] D. BRAESS, *Nonlinear Approximation Theory*, Springer, Berlin, 1986.
- [21] L. BREIMAN, *Hinging hyperplanes for regression, classification, and function approximation*, IEEE Trans. Inform. Theory, 39 (1993), pp. 999–1013.
- [22] H. BREZIS, *Analyse Fonctionnelle—Théorie et Applications*, Masson, Paris, 1983.
- [23] F. C. CHEN AND H. KHALIL, *Adaptive control of a class of nonlinear discrete-time systems using multilayer neural networks*, IEEE Trans. Automat. Control, 40 (1995), pp. 791–801.
- [24] V. C. P. CHEN, D. RUPPERT, AND C. A. SHOEMAKER, *Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming*, Oper. Res., 47 (1999), pp. 38–53.
- [25] F. CUCKER AND S. SMALE, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2001), pp. 1–49.
- [26] J. W. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [27] J. W. DANIEL, *The Ritz-Galerkin method for abstract optimal control problems*, SIAM J. Control, 11 (1973), pp. 53–63.
- [28] C. DARKEN, M. DONAHUE, L. GURVITS, AND E. SONTAG, *Rate of approximation results motivated by robust neural network learning*, in Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory, The Association for Computing Machinery, New York, 1993, pp. 303–309.
- [29] F. DEUTSCH, *Best Approximation in Inner Product Spaces*, Springer, New York, 2001.
- [30] R. A. DEVORE, R. HOWARD, AND C. MICHELLI, *Optimal nonlinear approximation*, Manuscripta Math., 63 (1989), pp. 469–478.
- [31] R. A. DEVORE AND G. G. LORENTZ, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer, Berlin, 1993.
- [32] R. A. DEVORE AND V. N. TEMLYAKOV, *Nonlinear approximation by trigonometric sums*, J. Fourier Anal. Appl., 2 (1995), pp. 29–48.
- [33] A. L. DONTCHEV, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci. 52, Springer, Berlin, Heidelberg, 1983.
- [34] A. L. DONTCHEV AND T. ZOLEZZI, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer, Berlin, Heidelberg, 1993.
- [35] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, Oxford, American Elsevier, New York, 1976.
- [36] U. FELGENHAUER, *On Ritz type discretizations for optimal control problems*, in Proceedings of the 18th IFIP-ICZ Conference, Res. Notes in Math. 386, Chapman and Hall, London, 1999, pp. 91–99.
- [37] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
- [38] F. GIROSI, M. JONES, AND T. POGGIO, *Regularization theory and neural networks architectures*, Neural Computation, 7 (1995), pp. 219–269.
- [39] D. E. GOLDBERG, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA, 1989.
- [40] L. GRIPPO, *Convergent on-line algorithms for supervised learning in neural networks*, IEEE Trans. Neural Networks, 11 (2000), pp. 1284–1299.
- [41] W. W. HAGER, *The Ritz-Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.

- [42] E. HEWIT AND K. STROMBERG, *Abstract Analysis*, Springer, Berlin, 1965.
- [43] R. B. HOLMES, *Approximating best approximations*, Nieuw Archief voor Wiskunde, XIV(3), (1966), pp. 106–113.
- [44] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Lecture Notes in Math. 257, Springer, Berlin, 1972.
- [45] L. K. JONES, *A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training*, Ann. Statist., 20 (1992), pp. 608–613.
- [46] S. A. JOHNSON, J. R. STEDINGER, C. SHOEMAKER, Y. LI, AND J. TEJADA-GUIBERT, *Numerical solution of continuous-state dynamic programs using linear and spline interpolation*, Oper. Res., 41 (1993), pp. 484–500.
- [47] A. JUDITSKY, H. HJALMARSSON, A. BENVENISTE, B. DELYON, L. LJUNG, J. SJÖBERG, AND Q. ZHANG, *Nonlinear black-box models in system identification: Mathematical foundations*, Automatica, 31 (1995), pp. 1725–1750.
- [48] P. C. KAINEN, V. KŮRKOVÁ, AND M. SANGUINETI, *Minimization of error functionals over variable-basis functions*, SIAM J. Optim., 14 (2003), pp. 732–742.
- [49] P. C. KAINEN, V. KŮRKOVÁ, AND A. VOGT, *An integral formula for Heaviside neural networks*, Neural Network World, 10 (2000), pp. 313–319.
- [50] P. C. KAINEN, V. KŮRKOVÁ, AND A. VOGT, *Upper Bounds on Variation with Respect to Half-Spaces*, Research Report ICS-2004-900, Institute of Computer Science, Prague, 2004.
- [51] V. KŮRKOVÁ, *Dimension-independent rates of approximation by neural networks*, in Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, K. Warwick and M. Kárný, eds., Birkhäuser, Boston, 1997, pp. 261–270.
- [52] V. KŮRKOVÁ, *Neural networks as universal approximators*, in The Handbook of Brain Theory and Neural Networks, M. Arbib, ed., MIT Press, Cambridge, MA, 2002, pp. 1180–1183.
- [53] V. KŮRKOVÁ, *High-dimensional approximation and optimization by neural networks*, in Advances in Learning Theory: Methods, Models, and Applications, J. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, eds., IOS Press, Amsterdam, 2003, pp. 69–88.
- [54] V. KŮRKOVÁ, P. C. KAINEN, AND V. KREINOVICH, *Estimates of the number of hidden units and variation with respect to half-spaces*, Neural Networks, 10 (1997), pp. 1061–1068.
- [55] V. KŮRKOVÁ AND M. SANGUINETI, *Bounds on rates of variable-basis and neural-network approximation*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2659–2665.
- [56] V. KŮRKOVÁ AND M. SANGUINETI, *Comparison of worst case errors in linear and neural network approximation*, IEEE Trans. Inform. Theory, 48 (2002), pp. 264–275.
- [57] V. KŮRKOVÁ AND M. SANGUINETI, *Learning with generalization capability by kernel methods of bounded complexity*, J. Complexity, to appear.
- [58] V. KŮRKOVÁ, P. SAVICKÝ, AND K. HLAVÁČKOVÁ, *Representations and rates of approximation of real-valued Boolean functions by neural networks*, Neural Networks, 11 (1998), pp. 651–659.
- [59] E. S. LEVITIN AND B. T. POLYAK, *Convergence of minimizing sequences in conditional extremum problems*, Dokl. Akad. Nauk SSSR, 168 (1966), pp. 764–767.
- [60] A. LINNEMANN, *Convergent Ritz approximations of the set of stabilizing controllers*, Systems Control Lett., 36 (1998), pp. 151–156.
- [61] K. S. NARENDRA AND S. MUKHOPADHYAY, *Adaptive control using neural networks and approximate models*, IEEE Trans. Neural Networks, 8 (1997), pp. 475–485.
- [62] K. S. NARENDRA AND K. PARTHASARATHI, *Identification and control of dynamical systems using neural networks*, IEEE Trans. Neural Networks, 4 (1990), pp. 4–26.
- [63] J. M. ORTEGA, *Numerical Analysis: A Second Course*, Classics Appl. Math. 3, SIAM, Philadelphia, 1990.
- [64] T. PARISINI, M. SANGUINETI, AND R. ZOPPOLI, *Nonlinear stabilization by receding-horizon neural regulators*, Internat. J. Control, 70 (1998), pp. 341–362.
- [65] T. PARISINI AND R. ZOPPOLI, *Neural networks for feedback feedforward nonlinear control systems*, IEEE Trans. Neural Networks, 5 (1994), pp. 436–449.
- [66] T. PARISINI AND R. ZOPPOLI, *Neural approximations for multistage optimal control of nonlinear stochastic systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 889–895.
- [67] A. PINKUS, *n-Widths in Approximation Theory*, Springer, Berlin, Heidelberg, 1985.
- [68] G. PISIER, *Remarques sur un résultat non publié de B. Maurey*, Séminaire d'Analyse Fonctionnelle 1980-81, Exposé no. V, École Polytechnique, Centre de Mathématiques, Palaiseau, France, pp. V.1–V.12.
- [69] T. POGGIO AND S. SMALE, *The mathematics of learning: Dealing with data*, Notices Amer. Math. Soc., 50 (2003), pp. 536–544.
- [70] B. T. POLAK, *Existence theorems and convergence of minimizing sequences in extremum prob-*

- lems with restrictions*, Dokl. Akad. Nauk SSSR, 166 (1966), pp. 72–75.
- [71] D. E. RUMELHART, G. E. HINTON, AND R. J. WILLIAMS, *Learning internal representation by error propagation*, in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, I: Foundations*, D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, eds., MIT, Cambridge, MA, 1986, pp. 318–362.
- [72] I. SINGER, *Best Approximation in Normed Linear Spaces by Elements of Subspaces*, Springer, Berlin, 1970.
- [73] H. R. SIRISENA AND F. S. CHOU, *Convergence of the control parametrization Ritz method for nonlinear optimal control problems*, J. Optim. Theory Appl., 29 (1979), pp. 369–382.
- [74] J. SJÖBERG, Q. ZHANG, L. LJUNG, A. BENVENISTE, P.-Y. GLORENNEC, B. DELYON, H. HJALMARSSON, AND A. JUDITSKY, *Nonlinear black-box modeling in system identification: A unified overview*, Automatica, 31 (1995), pp. 1691–1724.
- [75] V. B. TJUHTIN, *An error estimate for approximate solutions in one-sided variational problems*, Vestnik Leningrad Univ. Math., 14 (1982), pp. 247–254.
- [76] P. TSENG, *Incremental gradient(-projection) method with momentum term and adaptive step-size rule*, SIAM J. Optim., 8 (1998), pp. 506–531.
- [77] J. F. TRAUB AND A. G. WERSCHULZ, *Complexity and Information*, Cambridge University Press, Cambridge, UK, 1998.
- [78] A. A. VLADIMIROV, YU. E. NESTEROV, AND YU. N. CHEKANOV, *On uniformly convex functionals*, Vestnik Moskov. Univ. Ser. 15 Vychisl. Mat. Kibernet., 3 (1978), pp. 12–23 (in Russian); *Moscow University Computational Mathematics and Cybernetics* (1979), pp. 10–21 (in English).
- [79] G. YIN, *Rates of convergence for a class of global stochastic optimization algorithms*, SIAM J. Optim., 10 (1999), pp. 99–120.
- [80] 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., Birkhäuser, Boston, 1992, pp. 193–210.
- [81] R. ZOPPOLI, M. SANGUINETI, AND T. PARISINI, *Approximating networks and extended Ritz method for the solution of functional optimization problems*, J. Optim. Theory Appl., 112 (2002), pp. 403–440.