TO APPEAR IN SIAM JOURNAL ON OPTIMIZATION, 2004

ERROR ESTIMATES FOR APPROXIMATE OPTIMIZATION BY THE EXTENDED RITZ METHOD *

VĚRA KŮRKOVÁ† AND MARCELLO SANGUINETI‡

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

1. Introduction.

In many high-dimensional optimization problems (e.g., routing in communications networks, stochastic optimal control, management of water resources, large-scale traffic networks [13, 24, 46, 81]), 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], 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 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 estimates for the multivariable case.

Since the late 1980s, neural networks became a successful alternative to linear 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 have been developed [11,

* Collaboration between V. K. and M. S. was supported by the Scientific Agreement Italy-Czech Republic, Area MC 6, Project 22. V. K. was partially supported by GA ČR Grant 201/02/0428. M. S. was partially supported by a PRIN Grant from the Italian Ministry of University and Research.

†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).

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

12, 21, 38, 45, 51, 54, 55, 56]. In a series of papers [3, 8, 9, 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], feedforward neural networks [38, 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, 18, 39, 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 problems [8], optimal control of freeway traffic [81], routing in large-scale communication networks [9, 10], optimal fault diagnosis [5], etc. 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 to the optimal solution over the whole admissible set. The upper bounds depend on the degree n 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 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 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 modelled as a 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 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 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, resp. 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 $(X, \|.\|)$ 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 $p \in [1, \infty)$, $(\mathcal{L}_p(\Omega), \|.\|_p)$ denotes the space of measurable, real-valued functions on Ω such that $\int_\Omega |f(x)|^p \, dx < \infty$ endowed with the $\mathcal{L}_p$-norm.

A ball and a sphere of radius r centered at $h \in X$ are denoted by $B_r(h, \|.\|) = \{f \in X : \|f - h\| \leq r\}$ and $S_r(h, \|.\|) = \{f \in X : \|f - h\| = r\}$, respectively. We write shortly $B_r(\|.\|) = B_r(0, \|.\|)$ and merely $B_r(h) = B_r(h, \|.\|)$, $B_r = B_r(0)$ 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 $\Phi : X \to (-\infty, +\infty]$ is called *proper* if it is not identically equal to $+\infty$. The set $dom\, \Phi = \{f \in X : \ \Phi(f) < +\infty\}$ is called the *domain of* Φ.

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

A functional Φ is *convex* on a convex set $M \subseteq X$ if for all $h, g \in M$ and all $\lambda \in [0, 1]$, $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g)$. Φ is *uniformly convex* on a convex set $M \subseteq X$ if there exists a non-negative function $\delta : \mathcal{R}_+ \to \mathcal{R}_+$, such that $\delta(0) = 0$, for all $t > 0$, $\delta(t) > 0$ and for all $h, g \in M$ and all $\lambda \in [0, 1]$, $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g) - \lambda(1 - \lambda)\delta(\|h - g\|)$. Any such function δ is called a *modulus of convexity* of Φ (see, e.g, [59])[1].

Using standard notation [34], we denote by

$$(M, \Phi)$$

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

A sequence $\{g_n\}$ of elements of M is called Φ*-minimizing over* M if $\lim_{n\to\infty} \Phi(g_n) = \inf_{g\in M} \Phi(g)$. By the definition of infimum, for any problem (M, Φ) where M is non-empty, there always exists a minimizing sequence. We denote by $argmin\, (M, \Phi) = \{g^o \in M : \Phi(g^o) = \inf_{g\in M} \Phi(g)\}$ the set of *minimum points* of the problem (M, Φ)

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

and for $\varepsilon > 0$, we denote by $argmin_\varepsilon(M,\Phi) = \{g^\varepsilon \in M : \Phi(g^\varepsilon) < \inf_{g\in M}\Phi(g) + \varepsilon\}$ the set of its *ε-near minimum points*.

The following proposition summarizes elementary properties of uniformly convex functionals.

PROPOSITION 2.1. *Let $(X, \|.\|)$ be a normed linear space, $M \subseteq X$ be convex and Φ be a uniformly convex functional on M with a modulus of convexity δ. Then the following hold:*
(i) if Ψ is convex on M, then $\Phi + \Psi$ is uniformly convex on M with a modulus of convexity δ;
(ii) if $\Phi : X \to \mathcal{R}$ then for every $f \in X$, the translated functional $\Phi(\cdot - f)$ is uniformly convex on $M - f$ with a modulus of convexity δ;
(iii) if $g^o \in argmin(M,\Phi)$, then for every $g \in M$, $\delta(\|g - g^o\|) \le \Phi(g) - \Phi(g^o)$;
(iv) if $(X, \|.\|)$ is a Hilbert space, then the functional $\|.\|^2 : X \to \mathcal{R}$ 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 $\lambda(1-\lambda)\delta(\|g-g^o\|) \le \lambda\Phi(g)+(1-\lambda)\Phi(g^o)-\Phi(\lambda g+(1-\lambda)g^o)$. As $\Phi(g^o) \le \Phi(\lambda g+(1-\lambda)g^o)$, we get $\lambda(1-\lambda)\delta(\|g-g^o\|) \le \lambda\Phi(g)+(1-\lambda)\Phi(g^o)-\Phi(g^o) = \lambda\left(\Phi(g)-\Phi(g^o)\right)$. Hence $(1-\lambda)\delta(\|g-g^o\|) \le \Phi(g)-\Phi(g^o)$. Taking the infimum over λ, we obtain $\delta(\|g-g^o\|) \le \Phi(g)-\Phi(g^o)$.
(iv) It is easy to check that for every $h, g \in X$ and $\lambda \in [0,1]$, we have $\|\lambda h+(1-\lambda)g\|^2 \le \lambda\|h\|^2 + (1-\lambda)\|g\|^2 - \lambda(1-\lambda)\|h-g\|^2$. □

The problem (M,Φ) 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,Φ) *at a minimum point* g^o is a function $\xi_{g^o} : \mathcal{R}_+ \to \mathcal{R}_+$ such that for every $t \in \mathcal{R}_+$, $\xi_{g^o}(t) = \inf_{g\in M\cap S_t(g^o)}\Phi(g) - \Phi(g^o)$. Note that the modulus of Tikhonov well-posedness is defined for any problem that has a minimum point, even when such a problem is not Tikhonov well-posed.

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

For a subset M of a normed linear space, its *affine hull* is defined as $aff\, M = \left\{\sum_{i=1}^n w_i g_i \ : \ w_i \in \mathcal{R}, \sum_{i=1}^n w_i = 1, g_i \in M, n \in \mathcal{N}_+\right\}$. An element $g \in X$ is called a *relatively 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 $aff\, M$. The set of all relative interior points of M is called the *relative interior* of M and denoted by $ri\, M$. Thus, $ri\, M = \{g \in aff\, M : \ \exists\, \epsilon > 0,\ B_\varepsilon(g) \cap aff\, M \subseteq M\}$. Note that $ri\, M$ is the interior of M as a subset of its affine hull, instead of the whole space X.

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

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

M is called *absorbing* if $dom\, p_M = X$. For every M, p_M is positively homogeneous and if M is convex, then p_M is convex, too. The following proposition states elementary properties of Minkowski functionals of convex sets containing zero, which will be used in our proofs.

PROPOSITION 2.2. *Let $(X, \|.\|)$ be a normed linear space, M be a subset of X containing 0 and $r_0 = sup\{r > 0 : B_r(\|.\|) \subseteq M\}$. 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 $0 \in int\, M$, then $dom\, p_M = X$;
(v) if $0 \in int\, M$ and if $r_0 < \infty$, then for every $f \in dom\, p_M$, $p_M(f) \leq \|f\|/r_0$;
(vi) if M is convex and $0 \in intM$, then p_M is Lipschitz on X with a 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 $\lim_{i\to\infty} \lambda_i = 1$ and for every i, $f/\lambda_i \in M$. As M is closed and $f = \lim_{i\to\infty}(f/\lambda_i)$, we have $f \in M$.

(iv) As $0 \in int\, M$, there exists $r > 0$ such that $B_r(\|.\|) \subseteq M$. So for every $f \in B_r(\|.\|)$, $p_M(f) \leq 1$. 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 $B_r(\|.\|) \subseteq M$ and every $f \in dom\, p_M$ 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 $|p_M(f) - p_M(g)| \leq \|f-g\|/r_0$. □

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

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

where for each n, X_n is an n-dimensional subspace of X. Under suitable conditions on Φ, M, and $\{X_n\}$, 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 $\lim_{n\to\infty} \Phi(g_n) = \Phi(g^o)$.

Typically, the subspaces X_n are generated by the first n elements of a subset of X with a fixed linear ordering. So this approximation scheme can be called *fixed-basis approximation* in contrast to *variable-basis approximation*, which uses nonlinear approximating sets formed by linear combinations of at most n elements of a given subset G of X. Such sets are denoted by $span_n\, G = \{\sum_{i=1}^{n} w_i g_i :\ w_i \in \mathcal{R},\ g_i \in G\}$. We call n the *degree of the variable-basis functions* in $span_n\, G$. 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, Φ) is approximated by a sequence of problems

$$\{M \cap span_n G, \Phi)\}.$$

For G formed by parameterized families of the form $G = \{g_a : a \in A\}$ with $A \subseteq \mathcal{R}^p$ this method was applied to a variety of tasks in a series of papers [3, 8, 9, 64, 65, 66, 80] and [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 $\{M \cap span_n G\}$, where G is any set.

Sets $span_n G$ 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 ε_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 $M \cap span_n G$ 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 standard learning algorithms can be applied (see, e.g., [4, 16, 18, 40, 76] and the references therein). In [5, 10, 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 ε_n-near minimum points of Φ over $M \cap span_n G$ might converge to a minimum point of Φ 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, p.V.2, Lemma 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 $\|.\|_G$, 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 $(X, \|.\|)$, G-variation is defined as the Minkowski functional of the set $cl\, conv\, (G \cup -G)$:

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

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

$$\|.\| \leq s_G \|.\|_G \,. \tag{3.1}$$

Indeed, if for $b > 0$, $f/b \in cl\, conv(G \cup -G)$, then $f/b = \lim_{\varepsilon \to 0} h_\varepsilon$, where $h_\varepsilon \in conv(G \cup -G)$ 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 $\|f\|_{1,G} = \sum_{g\in G} |f \cdot g|$ [58], [55]. 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 [12].

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

$$\|f - span_n G\| \le \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}}\,; \tag{3.2}$$

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

$$\|f - span_n\, G\| \le \frac{2^{1/\bar{p}} s_G \|f\|_G}{n^{1/\bar{q}}}\,, \tag{3.3}$$

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 [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).

4. Rates of approximate optimization over variable-basis functions. In this section, we investigate approximate solutions $\{(M \cap span_n G, \Phi)\}$ of a problem (M, Φ) 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^o be a minimum point of the problem (M, Φ) to which the extended Ritz method based on an approximation of M by sets $M \cap span_n G$ is applied. As the existence of minimum points of approximating problems $(M \cap span_n G, \Phi)$ is not guaranteed, we consider ε_n-near minimum points. To estimate the speed of convergence of these ε_n-near minimum points to the minimum point g^o of Φ over the whole M, we take advantage of Theorem 3.1. As this theorem estimates the distance of g^o from $span_n G$ but not from $M \cap span_n G$, we construct an auxiliary sequence of elements of $M \cap span_n G$ using the following technical lemma. It extends [75, Lemma 3], proven for finite-dimensional subspaces of a linear space, to subsets satisfying a kind of restricted homogeneity condition. The next 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 a Lipschitz constant of the Minkowski functional of M. When M is a ball $B_r(\|.\|)$, such a Lipschitz constant is equal to $1/r$.

LEMMA 4.1. *Let A and M be subsets of a normed linear space $(X, \|.\|)$, M be closed and convex, $0 \in M$, and $\lambda A \subseteq A$ for all $\lambda \in [0, 1)$. Then for every $g \in M$ and*

every $f \in A$ with $p_M(f) < +\infty$, there exists $h \in M \cap A$ such that
(i) $\|h-g\| \le \|f-g\| + \|g\| \left|p_M(f) - p_M(g)\right|$;
(ii) if $0 \in int\, M$, *then* $\|h-g\| \le (1+c\,\|g\|)\,\|f-g\|$, *where c is a 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 $f \in A - cl\, M$, then $f \neq 0$ and so we can set $h = \frac{p_M(g)}{p_M(f)}\, f$. Hence $p_M(h) = p_M(g) \le 1$, 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 $\|h-g\| = \left\|\frac{p_M(g)}{p_M(f)} f - g\right\| = \left\|\frac{p_M(g)}{p_M(f)}(f-g) - \left(1 - \frac{p_M(g)}{p_M(f)}\right) g\right\| \le \left|\frac{p_M(g)}{p_M(f)}\right| \|f-g\| + \left|1 - \frac{p_M(g)}{p_M(f)}\right| \|g\| < \|f-g\| + \left|\frac{p_M(f)-p_M(g)}{p_M(f)}\right| \|g\| < \|f-g\| + |p_M(f) - p_M(g)|\, \|g\|$.

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

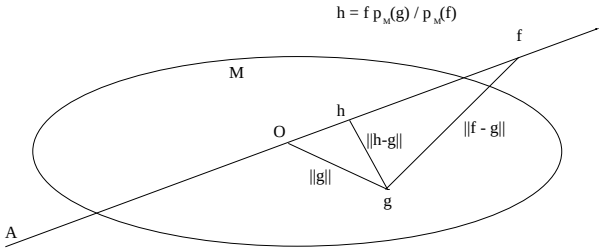

FIG. 4.1. *The construction used in the proof of Lemma 4.1.*

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 $0 \in int\, M$. 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 (see [15, pp. 68-78] and the application in Section 7.2), and when M is a ball of some radius r in the norm $\|.\|$, one can apply it to Ivanov's regularization [15, pp. 68-78]. Also the case where M is a subspace of 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 $(X, \|.\|)$ be a Hilbert space, M and G its subsets, G bounded, $s_G = \sup_{g \in G} \|g\|$, M closed, convex, and $0 \in int\, M$. Let $\Phi : X \to (-\infty, +\infty]$ be a proper functional, $g^o \in argmin\,(M, \Phi)$, Φ continuous at g^o with a modulus of continuity α, $\{\varepsilon_n\}$ be a sequence of positive reals, and $\{g_n\}$ be such that $g_n \in argmin_{\varepsilon_n}\,(M \cap span_n\, G, \Phi)$. Then p_M is Lipschitz on X, and if c is its Lipschitz constant, then the following estimates hold for every integer n:*
(i) $\inf_{g \in M \cap span_n\, G} \Phi(g) - \Phi(g^o) \le \alpha\left(\left(1 + c\|g^o\|\right)\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$;
(ii) if $\|g^o\|_G < \infty$ and $\lim_{n\to\infty} \varepsilon_n = 0$, then $\{g_n\}$ is a Φ-minimizing sequence over M and

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

(iii) if ξ is the modulus of Tikhonov well-posedness of (M, Φ) at g^o, then

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

(iv) if Φ is uniformly convex on M with a modulus of convexity δ, then

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

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

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

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

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

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

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

By Theorem 3.1 (i), we have

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

By infimizing (4.2) over ε, we obtain

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

which completes the proof of (i).

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

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

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

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

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

Theorem 4.2 shows that for $\|g^o\|_G$ finite, the approximate minimum points $\{g_n\}$ form a Φ-minimizing sequence over M and the speed of convergence of $\{\Phi(g_n)\}$ to the global minimum $\Phi(g^o)$ is bounded from above by $\alpha \left(\left(1 + c\|g^o\|\right) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n$.

When minimization is performed over the whole space, the Lipschitz constant of the Minkowski functional $p_M = p_X$ is equal to zero; thus, Theorem 4.2 gives an upper bound $\alpha \left(\sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n$, which depends on the modulus of continuity α of Φ, G-variation and the ambient space norm of g^o.

When the admissible set is a ball $B_r(\|.\|)$, the Lipschitz constant is $1/r$ and we get an upper bound $\alpha \left(\left(1 + \frac{\|g^o\|}{r}\right) \sqrt{\frac{(s_G \|g^o\|_G)^2 - \|g^o\|^2}{n}} \right) + \varepsilon_n$.

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, the estimates hold for any number d of variables of the admissible solutions. 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^o 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).

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

$\Phi(g_n) - \Phi(g^o) \leq \alpha\left(\left(1 + c\|g^o\|\right)\frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}}\right) + \varepsilon_n$;
(iii) if ξ *is the modulus of Tikhonov's well-posedness of* (M, Φ) *at* g^o*, then*
$\xi(\|g_n - g^o\|) \leq \alpha\left(\left(1 + c\|g^o\|\right)\frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}}\right) + \varepsilon_n$;
(iv) if Φ *is uniformly convex with a modulus of convexity* δ*, then*
$\delta(\|g_n - g^o\|) \leq \alpha\left(\left(1 + c\|g^o\|\right)\frac{2^{1/\bar{p}} s_G \|g^o\|_G}{n^{1/\bar{q}}}\right) + \varepsilon_n$.

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, Φ) by obtaining its minimum point g^o as a limit of a Φ-minimizing sequence $\{g_n\} \subseteq M$ satisfying $\lim_{n\to\infty} \Phi(g_n) = \Phi(g^o)$. Using this notion, we can rephrase our results as conditions on (M, Φ) under which the extended Ritz method has some of the properties of a direct method. By Theorems 4.2 and 4.3 for $\|g^o\|_G$ finite, any sequence $\{g_n\}$ of ε_n-minimum points of $(M \cap span_n G, \Phi)$ is Φ-minimizing and $\Phi(g^o) = \lim_{n\to\infty} \Phi(g_n) = \Phi(\lim_{n\to\infty} g_n)$. The convergence of $\{g_n\}$ to g^o is not always guaranteed (it depends on the behavior of the modulus of Tikhonov's well-posedness of (M, Φ) at g^o). 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 $M \cap span\, G$ 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 $f, g : \mathcal{N}_+ \to \mathcal{N}_+$ we write $g(n) \leq \mathcal{O}(f(n))$ when there exists $a > 0$ such that for all but finitely many $n \in \mathcal{N}_+$, $g(n) \leq a\, f(n)$.

THEOREM 5.1. *Let* $(X, \|.\|)$ *be a Hilbert space,* M *and* G *be its subsets,* G *bounded,* $s_G = \sup_{g\in G} \|g\|$*,* M *closed, convex,* $0 \in int\, M$*, and* $M \cap span\, G$ *dense in* M*. Let* $\Phi : X \to (-\infty, +\infty]$ *be a proper convex functional,* $g^o \in argmin\,(M, \Phi)$ *be such that* Φ *is bounded in its neighborhood,* $\{\varepsilon_n\}$ *be a sequence of positive reals such that* $\varepsilon_n \leq \mathcal{O}(1/\sqrt{n})$ *and* $g_n \in argmin_{\varepsilon_n}\,(M \cap span_n\, G, \Phi)$*. Then the following estimates hold:*
(i) $\inf_{g\in M\cap span_n\, G} \Phi(g) - \Phi(g^o) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$;
(ii) if $\|g^o\|_G < \infty$*, then* $\{g_n\}$ *is a* Φ*-minimizing sequence over* M *and*
$\Phi(g_n) - \Phi(g^o) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$;
(iii) if ξ *is the modulus of Tikhonov's well-posedness of* (M, Φ) *at* g^o*, then*
$\xi(\|g_n - g^o\|) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$;
(iv) if Φ *is uniformly convex with a modulus of convexity* δ*, then*
$\delta(\|g_n - g^o\|) \leq \mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2 - \|g^o\|^2}{n}}\right)$.

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

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

As M is closed, convex, $0 \in int\, M$, and $dom\, p_M = X$, we can apply Lemma 4.1

(ii) with $f = f_n^\varepsilon$, $g = g^o$, and $A = span_n\, G$ to obtain $h_n^\varepsilon \in M \cap span_n\, G$ satisfying

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

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

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

From (5.1) and (5.2) we obtain

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

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

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

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

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

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

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

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

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

reals and $\{g_n\}$ such that $g_n \in argmin_{\varepsilon_n}(M,\Phi)$, the statements of Theorem 5.1 (ii), (iii) and (iv) hold with the bounds replaced with $\mathcal{O}\left(\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}\right)+\varepsilon_n$.

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

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 [29, p. 21] (i.e., there exists a unique $g^o \in M$ such that $\|f - g^o\| = \|f - M\|$).

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 approximate optimization methods of the problem of best approximation 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 M and G be subsets of a Hilbert space $(X, \|.\|)$, G be bounded, $s_G = \sup_{g\in G}\|g\|$, M be closed, convex, $0 \in int\, M$, 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) such that the following estimates hold for every integer n:*
(i) $\inf_{g\in M\cap span_n G} e_f(g) - e_f(g^o) = \|f - M \cap span_n G\| - \|f - M\|$
$\le (1 + c\|g^o\|)\sqrt{\frac{(s_G\|g^o\|_G)^2-\|g^o\|^2}{n}}$;
(ii) if M is bounded, $\{\varepsilon_n\}$ is a sequence of positive reals, and for every n, $g_n \in argmin_{\varepsilon_n}(M \cap span_n G, e_f^2)$, then

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

Proof. As every closed convex subset of a Hilbert space is Chebyshev [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 $|e_f(h) - e_f(g)| \le \|h - g\|$. So e_f is uniformly continuous on X and its modulus of continuity is $\alpha(t) = t$. Hence, applying Theorem 4.2 (i) we obtain (i).

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

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

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 $\Omega \subseteq \mathcal{R}^d$, M and G be subsets of $(\mathcal{L}_p(\Omega), \|.\|_p)$, $p \in (1, \infty)$, G bounded, $s_G = \sup_{g \in G} \|g\|$, M closed, convex, $0 \in int\, M$, $f \in X$, $q = p/(p-1)$, $\bar{p} = \min(p, q)$, $\bar{q} = \max(p, q)$, and α_p, α_q be moduli of continuity of e_f^p, e_f^q, 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) such that the following estimates hold for every integer n:*

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

(ii) if M is bounded, $\{\varepsilon_n\}$ is a sequence of positive reals, $p \in (1, 2]$ and $g_n \in argmin_{\varepsilon_n} (M \cap span_n\, G, e_f^q)$, then

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

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

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

Proof. Since for all $p \in (1, \infty)$, $(\mathcal{L}_p(\Omega), \|.\|_p)$ is a uniformly convex space [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 $|e_f(h) - e_f(g)| \le \|h - g\|_p$. So e_f is uniformly continuous on X and its modulus of continuity is $\alpha(t) = t$. Hence, applying Theorem 4.3 (i) we obtain (i).

(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).

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

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

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

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

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

(i) if $(X, \|.\|) = (\mathcal{L}_p(\Omega), \|.\|_p)$, 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

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

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

7. Application to learning from data. Learning from a sample $\{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \ldots, m\}$ of *empirical data* can be modelled 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 Ψ is a suitable functional called *stabilizer* and γ 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 plays the role of a tradeoff between fitting to the empirical data and fitting to the properties of solutions represented by the stabilizer.

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

By the Cauchy-Schwartz inequality, for every $f \in \mathcal{H}_K(\Omega)$ and $x \in \Omega$ 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 kernel K

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

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

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

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

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

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

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

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

Thus, to compute the coefficients of the linear combination $a = (a_1, \ldots, a_m)$ 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 γ must be "large" [57]. On the other hand, a "large" γ 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 $span_m G_K$, 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 global minimum point described by the Representer Theorem. We first state basic properties of the functional $\mathcal{E}_{\gamma,K}$.

PROPOSITION 7.1. *Let Ω be a nonempty set, $K : \Omega \times \Omega \to \mathcal{R}$ be a kernel, $c_K = \sup_{x\in\Omega} \sqrt{K(x,x)}$, $\gamma > 0$, m be a positive integer, $\{(x_1, y_1), \ldots, (x_m, y_m)\} \subset (\Omega \times \mathcal{R})^m$, $\mathcal{E}_{\gamma,K}(f) = \frac{1}{m}\sum_{i=1}^m |f(x_i) - y_i|^2 + \gamma\|f\|_K^2$, and $y_{\min} = \{|y_i| : i = 1, \ldots, m\}$. 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 $a_1 = 2\big(\|f\|_K\, c_K^2 + y_{\min}\, c_K + \gamma\|f\|_K\big)$ 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^o;
(iv) for any minimum point g^o of $(M, \mathcal{E}_{\gamma,K})$ and any $f \in M$,
$\|f - g^o\|_K^2 \le \frac{|\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g^o)|}{\gamma}$.

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,

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

$$\le \left| \frac{1}{m}\sum_{i=1}^m \big(\, f(x_i) - g(x_i)\,\big)\,\big(\, f(x_i) + g(x_i) - 2y_i\,\big) \right| + \gamma \Big|\, \|f\|_K - \|g\|_K \Big| \left(\|f\|_K + \|g\|_K\right)$$

$$\le \sup_{x\in\Omega} |f(x) - g(x)| \Big| \sup_{x\in\Omega} |f(x) + g(x)| - 2y_{\min} \Big| + \gamma\, \|f - g\|_K \left(\|f\|_K + \|g\|_K\right).$$

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

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

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

(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], and 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). □

So the modulus of continuity of $\mathcal{E}_{\gamma,K}$ at any $f \in \mathcal{H}_K(\Omega)$ is bounded from above by the quadratic function $a_2t^2 + a_1t$. Note that a_2 depends on m, c_K and γ, while a_1 depends, in addition to these values, also on $\|f\|_K$ and $y_{\min}$.

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 any n (but are only useful for $n < m$, as the minimum point g^o is in $span_m G_K$).

THEOREM 7.2. *Let Ω be a nonempty set, $K : \Omega\times\Omega \to \mathcal{R}$ be a kernel, $c_K = \sup_{x\in\Omega}\sqrt{K(x,x)}$, $(\mathcal{H}_K(\Omega), \|.\|_K)$ be the RKHS defined by K, $G_K = \{K(x,.) : x \in \Omega\}$, $(x_1,\ldots,x_m) \in \Omega^m$, $(y_1,\ldots,y_m) \in \mathcal{R}^m$, $y_{\min} = \{|y_i| : i = 1,\ldots,m\}$, $\gamma > 0$, $\mathcal{E}_{\gamma,K}(f) = \frac{1}{m}\sum_{i=1}^m |f(x_i)-y_i|^2 + \gamma\,\|f\|_K$, $g^o(x) = \sum_{i=1}^m w_iK(x,x_i)$ 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 $g_n \in argmin_{\varepsilon_n}(span_n\,G_K, \mathcal{E}_{\gamma,K})$. Then for every positive integer n, the following estimates hold:*

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

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

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

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

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

where $\alpha(t) = a_2t^2 + a_1t$, $a_1 = 2\big(\|g^o\|_K\,c_K^2 + y_{\min}\,c_K + \gamma\|g^o\|_K\big)$ *and* $a_2 = c_K^2 + \gamma$.

Proof. Statements (i) and (ii) follow from Theorem 4.2 with $M = \mathcal{H}_K(\Omega)$, $c = 0$ (in this case, p_M is the constant functional equal to zero), $\Phi(f) = \mathcal{E}_{\gamma,K}(f)$, $G = G_K$, and $s_G = \sup_{x\in\Omega}\|K(x,.)\|_K = \sup_{x\in\Omega}\sqrt{\langle K(x,.), K(x,.)\rangle_K} = \sup_{x\in\Omega}\sqrt{K(x,x)} = c_K$. 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 $\alpha(t) = a_2t^2 + a_1t$.

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

(iv) follows from (7.1) and item (iii). □

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 $\{(x_i, y_i),\ i = 1, \ldots, m\}$, the kernel K, and the regularization parameter γ were derived in [57].

8. Discussion. We have derived upper bounds on rates of approximate optimization by the extended Ritz method for problems (M, Φ) having a minimum point, where Φ is continuous and M is closed, convex, 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 $1/\sqrt{n}$ 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]) have been 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-variations. 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 $\|.\|_{d/2,p}$ with "large" $((d+1)/2)$-th derivatives cannot be extended to $(d+1)$-variable functions from the unit balls in the Sobolev norms $\|.\|_{(d+1)/2,p}$. In contrast to the linear 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.

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.

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

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

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

THEOREM A.1. *Let G be a bounded subset of a Hilbert space $(X, \|.\|)$ and*

$s_G = \sup_{g\in G}\|g\|$, *then for every* $f \in cl\,conv\,G$ *and for every positive integer* n, $\|f - conv_n G\| \le \sqrt{\frac{s_G^2 - \|f\|^2}{n}}$.

Proof. Since the distance from $conv_n G$ is continuous on $(X, \|.\|)$ [72, p. 391], it is sufficient to verify the statement for $f \in conv\,G$. 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 exist a sequence $\{g_i\}$ of elements of G such that the barycenters $f_n = \sum_{i=1}^n \frac{g_i}{n}$ satisfy $e_n^2 = \|f - f_n\|^2 \le \frac{c}{n}$.

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

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 $\sum_{j=1}^m a_j\left(\frac{2n}{(n+1)^2}\langle f - f_n, f - h_j\rangle + \frac{1}{(n+1)^2}\|f - h_j\|^2\right) = \frac{2n}{(n+1)^2}\langle f - f_n, f - \sum_{j=1}^m a_j h_j\rangle + \frac{1}{(n+1)^2}\left(\|f\|^2 - 2\langle f, \sum_{j=1}^m a_j h_j\rangle + \sum_{j=1}^m a_j\|h_j\|^2\right) = \frac{1}{(n+1)^2}\left(\sum_{j=1}^m a_j g_j - \|f\|^2\right) \le \frac{1}{(n+1)^2}(s_G^2 - \|f\|^2) = \frac{c}{(n+1)^2}$. So there must exist some $j \in \{1, \dots, m\}$ such that $\frac{2n}{(n+1)^2}\langle f - f_n, f - g_{n+1}\rangle + \frac{1}{(n+1)^2}\|f - g_{n+1}\|^2 \le \frac{c}{(n+1)^2}$.

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

In [28], Maurey-Jones-Barron's 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* $\Omega \subseteq \mathcal{R}^d$, $f, g \in (\mathcal{L}_p(\Omega), \|.\|_p)$, $p \in (1, \infty)$, *and* $q = p/(p-1)$, *then for* $p \in (1, 2]$

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

$$\left\|\frac{f+g}{2}\right\|_p^p + \left\|\frac{f-g}{2}\right\|_p^p \ge \frac{1}{2}\|f\|_p^p + \frac{1}{2}\|g\|_p^p \tag{A.2}$$

and for $p \ge 2$

$$\left\|\frac{f+g}{2}\right\|_p^p + \left\|\frac{f-g}{2}\right\|_p^p \le \frac{1}{2}\|f\|_p^p + \frac{1}{2}\|g\|_p^p \tag{A.3}$$

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

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 $f, g \in (\mathcal{L}_p(\Omega), \|.\|_p)$, $\|f+g\|_p^{\bar{p}} + \|f-g\|_p^{\bar{p}} \leq 2\left(\|f\|_p^{\bar{p}} + \|g\|_p^{\bar{p}}\right)$.

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

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

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 $\|\psi+\phi\|_p^p + \|\phi-\psi\|_p^p \leq 2(\|\phi\|_p^p + \|\psi\|_p^p) = \|\psi+\phi\|_p^{\bar{p}} + \|\psi-\phi\|_p^{\bar{p}} \leq 2\left(\|\psi\|_p^{\bar{p}} + \|\phi\|_p^{\bar{p}}\right)$, 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

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

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

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

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], with $a = \|\phi+\psi\|_p^q$, $b = \|\phi-\psi\|_p^q$, and $r = p/q = p-1$ it follows

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

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

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 $(X, \|.\|)$ and $f \in X$, we denote by Π_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 $(\mathcal{L}_p(\Omega), \|.\|_p)$, $p \in (1, \infty)$, $f \in cl\, conv\, G$ and $r > 0$ be such that $G \subseteq B_r(f, \|.\|)$. Then for every positive integer n, $\|f - span_n G\|_p \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$, 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 $f \in conv\, G$. 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 exist 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 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 $G \subseteq B_r(f, \|.\|)$, 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 $g_1, \ldots, g_n$, 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 Π_n be a peak functional for $f - f_n$. Since $\sum_{j=1}^m a_j\, (f - h_j) = 0$, by linearity of Π_n we have $0 = \Pi_n \left(\sum_{j=1}^m a_j\, (f - h_j) \right) = \sum_{j=1}^m a_j\, \Pi_n(f - h_j)$. Thus, there must exist $j \in \{1, \ldots, m\}$ such that $\Pi_n(f - h_j) \leq 0$. Set $g_{n+1} = h_j$, so $\Pi_n(f - g_{n+1}) \leq 0$. Thus, by Proposition A.3 (iii) we get

$$e_{n+1}^{\bar{p}} = \|f - f_{n+1}\|_p^{\bar{p}} = \left\| \frac{n}{n+1}(f - f_n) + \frac{1}{n+1}(f - g_{n+1}) \right\|_p^{\bar{p}}$$

$$\leq 2 \left(\left\| \frac{n}{n+1}\, (f - f_n) \right\|_p^{\bar{p}} + \left\| \frac{1}{n+1}\, (f - g_{n+1}) \right\|_p^{\bar{p}} \right) \left\| \frac{n}{n+1}\, (f - f_n) - \frac{1}{n+1}\, (f - g_{n+1}) \right\|_p^{\bar{p}}. \tag{A.9}$$

As $\|\Pi_n\| = 1$ and $\Pi_n(f - g_{n+1}) \leq 0$, we have $\| \frac{n}{n+1}\, (f - f_n) - \frac{1}{n+1}\, (f - g_{n+1}) \|_p \geq \left\| \Pi_n \left(\frac{n}{n+1}\, (f - f_n) - \frac{1}{n+1}\, (f - g_{n+1}) \right) \right\|_p \geq \left\| \Pi_n \left(\frac{n}{n+1}\, (f - f_n) \right) \right\|_p = \frac{n}{n+1} \left\| \Pi_n (f - f_n) \right\|_p = \frac{n}{n+1} \|f - f_n\|_p$. Hence

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

By (A.9) and (A.10), $e_{n+1}^{\bar{p}} = \|f - f_{n+1}\|_p^{\bar{p}} \leq 2 \left(\| \frac{n}{n+1}\, (f - f_n) \|_p^{\bar{p}} + \| \frac{1}{n+1}\, (f - g_{n+1}) \|_p^{\bar{p}} \right) - \left(\frac{n}{n+1} \|f - f_n\|_p \right)^{\bar{p}} = \frac{2}{(n+1)^{\bar{p}}} \|f - g_{n+1}\|_p^{\bar{p}} + \left(\frac{2}{n+1} \right)^{\bar{p}} \|f - f_n\|_p^{\bar{p}} = \frac{2}{(n+1)^{\bar{p}}} \|f - g_{n+1}\|_p^{\bar{p}} + \left(\frac{2}{n+1} \right)^{\bar{p}} e_n^{\bar{p}}$. As $e_n = \|f - f_n\| \leq \frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}}$, we get $e_{n+1}^{\bar{p}} \leq \frac{2\, r^{\bar{p}}}{(n+1)^{\bar{p}}} + \left(\frac{n}{n+1} \right)^{\bar{p}} \left(\frac{2^{1/\bar{p}} r}{n^{1/\bar{q}}} \right)^{\bar{p}} = \frac{2\, r^{\bar{p}}}{(n+1)^{\bar{p}}} \left(1 + \frac{n^{\bar{p}}}{n^{\bar{p}/\bar{q}}} \right) = \frac{2\, r^{\bar{p}}}{(n+1)^{\bar{p}}} \left(1 + n^{\bar{p} - \bar{p}/\bar{q}} \right)$. 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., $e_{n+1} \leq \frac{2^{1/\bar{p}} r}{(n+1)^{1/\bar{q}}}$. □

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 $span_n\, G \supseteq conv_n\, G$, by Theorem A.1 applied to $G(\|f\|_G)$ we have $\|f - span_n\, G\| \leq \|f - conv_n\, G(\|f\|_G)\| \leq \sqrt{\frac{(s_G\, \|f\|_G)^2 - \|f\|^2}{n}}$.
(ii) As $span_n\, G \supseteq conv_n\, G$, by applying Theorem A.4 to $G(\|f\|_G)$ we get $\|f - span_n\, G\| \leq \frac{2^{1/\bar{p}}\, r}{n^{1/\bar{q}}}$ for every r such that $G(\|f\|_G) \subseteq B_r(f, \|.\|)$. 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 $G(\|f\|_G) \subseteq B_{2r\|f\|_G}(f, \|.\|)$, hence $\|f - span_n\, G\| \leq \frac{2^{1/\bar{p}}\, s_G\, \|f\|_G}{n^{1/\bar{q}}}$.

REFERENCES

[1] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*, John Wiley & Sons, 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] Alessandri, A., Sanguineti, M., and Maggiore, M., *Optimization-based learning with bounded error for feedforward neural networks*, IEEE Trans. Neural Networks, 13, pp. 261-273, 2002.
[5] Alessandri, A. and Sanguineti, M., *Optimization of approximating networks for optimal fault diagnosis*, Optimization Methods and Software, to appear (2004).
[6] W. Alt, *On the approximation of infinite optimization problems with an application to optimal control problems*, Appl. Math. Optim., 12, pp. 15-27, 1984.
[7] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society, 68, pp. 337-404, 1950.
[8] M. Baglietto, T. Parisini, and R. Zoppoli, *Numerical solutions to the Witsenhausen counterexample by approximating networks*, IEEE Trans. Automat. Control, 46, pp. 1471–1477, 2001.
[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, pp. 485–502, 2001.
[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, pp. 22-55, 2003.
[11] A.R. Barron, *Neural net approximation*, in Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems, pp. 69–72, 1992.
[12] A.R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. Theory, 39, pp. 930-945, 1993.
[13] R. W. Beard and T. W. McLain, *Successive Galerkin approximation algorithms for nonlinear optimal and robust control*, Int. J. Contr., 71, pp. 717–743, 1998.
[14] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, 1957.
[15] M. Bertero, *Linear inverse and ill-posed problems*, Advances in Electronics and Electron Physics, 75, pp. 1-120, 1989.
[16] D. P. Bertsekas, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7, 913-926, 1997.
[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, pp. 94-111, 1973.
[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, pp. 999-1013, 1993.
[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, pp. 791-801, 1995.

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

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

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

[27] Daniel, J. W., *The Ritz-Galerkin method for abstract optimal control problems*, SIAM J. Control, 11, pp. 53-63, 1973.

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

[29] Deutsch, F., *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.

[30] DeVore, R., Howard, R., and Micchelli, C., *Optimal nonlinear approximation*, Manuscripta Mathematica, 63, pp. 469–478, 1989.

[31] DeVore, R. A. and Lorentz, G. G., *Constructive approximation*, Grundlehren der Mathematischen Wissenschaften, 303, Springer-Verlag, Berlin, 1993.

[32] DeVore, R. A. and Temlyakov, V. N., *Nonlinear approximation by trigonometric sums*, The J. of Fourier Analysis and Applications, 2, pp. 29–48, 1995.

[33] Dontchev, A. L., *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Information Sciences, 52, Springer-Verlag, Berlin Heidelberg, 1983.

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

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

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

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

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

[39] Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[40] Grippo, L., *Convergent on-line algorithms for supervised learning in neural networks*, IEEE Trans. Neural Networks, 11, pp. 1284-1299, 2000.

[41] Hager, W. W., *The Ritz-Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12, pp. 854-867, 1975.

[42] Hewit, E., Stromberg, K., *Abstract Analysis*, Springer-Verlag, Berlin, 1965.

[43] Holmes, R. B, *Approximating best approximations*, Nieuw Archief voor Wiskunde, XIV(3), pp. 106-113, 1966.

[44] Holmes, R. B, *A Course on Optimization and Best Approximation*, Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1972.

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

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

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

[48] Kainen, P. C., Kůrková, V., and Sanguineti, M., *Minimization of error functionals over variable-basis functions*, SIAM J. Optim., 14, pp. 732-742, 2003.

[49] Kainen, P.C., Kůrková, V., and Vogt, A., *An integral formula for Heaviside neural networks*, Neural Network World, 10, pp. 313–319, 2000.

[50] Kainen, P. C., Kůrková, V., and Vogt, A., *Upper bounds on variation with respect to half-spaces*, Research Report ICS-2003-900, Institute of Computer Science, Prague, 2003.

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

[52] Kůrková, V., *Neural networks as universal approximators*, in The Handbook of Brain Theory and Neural Networks, M. Arbib, ed., Cambridge, MIT Press, 2002, pp. 1180–1183.

[53] Kůrková, V., *High-dimensional approximation and optimization by neural networks*, in Advances in Learning Theory: Methods, Models, and Applications, J. Suykens et al., eds., Chapter 4 (pp. 69–88), NATO Science Series III: Computer & Systems Sciences, vol. 190. IOS Press, Amsterdam, 2003.

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

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

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

[57] Kůrková, V. and Sanguineti, M., *Learning with generalization capability by kernel methods of bounded complexity*, Research Report ICS-2003-901, Institute of Computer Science, Prague, 2003

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

[59] Levitin, E. S. and Polyak, B. T., *Convergence of minimizing sequences in conditional extremum problems*, Dokl. Akad. Nauk SSSR, 168, n. 5, pp. 764-767, 1966.

[60] Linnemann, A., *Convergent Ritz approximations of the set of stabilizing controllers*, System & Control Letters, 36, pp. 151-156, 1998.

[61] Narendra, K.S. and Mukhopadhyay, S., *Adaptive control using neural networks and approximate models*, IEEE Trans. Neural Networks, 8, pp. 475-485, 1997.

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

[63] Ortega, J. M., *Numerical Analysis: A Second Course*, SIAM, Philadelphia, 1990.

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

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

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

[67] Pinkus, A., *n-Widths in Approximation Theory*, Springer-Verlag, Berlin Heidelberg, 1985.

[68] Pisier, G., *Remarques sur un résultat non publié de B. Maurey*, Séminaire d'Analyse Fonctionnelle 1980-81, Exposé no. V, pp. V.1-V.12, École Polytechnique, Centre de Mathématiques, Palaiseau, France.

[69] Poggio, T., and Smale, S., *The mathematics of learning: dealing with data*, Notices of the AMS, 50, n. 5, pp. 536-544, 2003.

[70] Polak, B. T., *Existence theorems and convergence of minimizing sequences in extremum problems with restrictions*, Dokl. Akad. Nauk SSSR, 166, n. 2, pp. 72-75, 1966.

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

[72] Singer, I., *Best approximation in normed linear spaces by elements of subspaces*, Springer-Verlag, Berlin, 1970.

[73] Sirisena, H. R. and Chou, F. S., *Convergence of the control parametrization Ritz method for nonlinear optimal control problems*, J. Optim. Theory Appl., 29, pp. 369–382, 1979.

[74] Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Glorennec, P.-Y., Delyon, B., Hjalmarsson, H., and Juditsky, A., *Nonlinear black-box modeling in system identification: a unified overview*, Automatica, 31, pp. 1691–1724, 1995.

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

[76] Tseng, P., *Incremental gradient(-projection) method with momentum term and adaptive stepsize rule*, SIAM J. Optim., 8, pp. 506-531, 1998.

[77] Traub, J. F. and Werschulz, A. G., *Complexity and Information*, Cambridge University Press, Cambridge, 1998.

[78] Vladimirov, A. A., Nesterov, Yu. E., and Chekanov, Yu. N., *On uniformly convex func-*

tionals, Vestnik Moskovskogo Universiteta. Seriya 15 - Vychislitel'naya Matematika i Kibernetika, 3, pp. 12-23, 1979. (English translation: *Moscow University Computational Mathematics and Cybernetics*, pp. 10-21, 1979).

[79] Yin, G., *Rates of convergence for a class of global stochastic optimization algorithms*, SIAM J. Optim., 10, pp. 99-120, 1999.

[80] Zoppoli, R. and Parisini, T., *Learning techniques and neural networks for the solution of N-stage nonlinear nonquadratic optimal control problems*, in Systems, Models and Feedback: Theory and Applications, A. Isidori and T. J. Tarn, eds., pp. 193-210. Birkhäuser, Boston, 1992.

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