Journal of Mathematical Modelling and Algorithms (2005) **4:** 355–368
DOI: 10.1007/s10852-005-1625-z

Rates of Minimization of Error Functionals over Boolean Variable-Basis Functions★

P. C. KAINEN[1], V. KŮRKOVÁ[2,★★], and M. SANGUINETI[3,†]
[1] *Department of Mathematics, Georgetown University, Washington, DC 20057-1233, USA.*
e-mail: kainen@georgetown.edu
[2] *Institute of Computer Science, Academy of Sciences of the Czech Republic,*
Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic.
e-mail: vera@cs.cas.cz
[3] *Department of Communications, Computer, and System Sciences (DIST), University of Genoa,*
Via Opera Pia 13, 16145 Genova, Italy. e-mail: marcello@dist.unige.it

(Received: 1 February 2005)

Abstract. Approximate solution of optimization tasks that can be formalized as minimization of error functionals over admissible sets computable by variable-basis functions (i.e., linear combinations of n-tuples of functions from a given basis) is investigated. Estimates of rates of decrease of infima of such functionals over sets formed by linear combinations of increasing number n of elements of the bases are derived, for the case in which such admissible sets consist of Boolean functions. The results are applied to target sets of various types (e.g., sets containing functions representable either by linear combinations of a "small" number of generalized parities or by "small" decision trees and sets satisfying smoothness conditions defined in terms of Sobolev norms).

Mathematics Subject Classifications (2000): 49K40, 41A25, 90B99, 90C90, 05C05, 62C99.

Key words: high-dimensional optimization, minimizing sequences, Boolean decision functions, decision trees.

1. Introduction

Many tasks in operations research, control theory, statistics, management and economic sciences, etc., can be modelled as minimization of functionals satisfying certain conditions given by measured and conceptual data. Conditions defined by measured data can be formalized as minimization of distances from sets of

★ Collaboration between V.K. and P.C.K. was supported by a NSF COBASE grant and between V.K. and M.S. by a Scientific Agreement between Italy and Czech Republic, Area MC 6, Project 22 ("Functional Optimization and Nonlinear Approximation by Neural Networks").

★★ Partially supported by GA ČR Grants 201/02/0428 and 201/05/0557 and by the Institutional Research Plan AV0Z10300504.

† Partially supported by a PRIN Grant of the Italian Ministry of University and Research (Project "New Techniques for the Identification and Adaptive Control of Industrial Systems").

functions interpolating the data, whereas conditions defined by conceptual data correspond to a-priori assumptions on the smoothness properties of functions in these sets. A functional defined as a distance in a suitable metric from a given "target" set is called *error functional*. For example, in machine learning suitable target sets can be represented as decision trees [22]; in modelling tasks, target sets are classes of allowed models [28]; in pattern recognition and classification, elements of the target sets are patterns according to which classification and recognition have to be done. Recently, the need for minimizing error functionals has emerged in learning from data. In particular, the error functional equal to the distance of an element in a normed space to the ball of a certain radius in a dense subspace plays a central role in the Cucker–Smale learning theory (see, e.g., [5], [27, Chapter 3], [31]).

Approximation theory investigates rates of convergence of the simplest type of error functional, defined as a distance from a singleton. Minimization of such functionals over admissible sets formed by functions computable by neural networks have been studied using a theorem attributed to Maurey by Pisier [29] and later improved by Jones [12] and Barron [3]. Such a theorem allows one to describe sets of multivariable functions that can be approximated by nonlinear approximators belonging to a class called "*variable-basis functions*" without incurring the "*curse of dimensionality*" (i.e., an exponential growth of the number of computational units with respect to the number d of variables of admissible functions) [4].

In recent years, approximation schemes of the variable-basis type have become quite well understood theoretically (see, e.g., [8, 10, 11, 14, 17, 18, 21, 25, 26]) and widely used in applications; they include feedforward neural networks, free-nodes splines, and many other commonly used nonlinear schemes [18]. All such schemes implement input/output mappings dependent on certain parameters to be tuned. In particular, feedforward neural networks have become a widespread computational paradigm since they enjoy powerful approximating capabilities, are well-suited to distributed computing, and offer the possibility of adjusting parameters by simple and efficient nonlinear programming algorithms suitable for parallel implementation. In the last decades they have been extensively used in a variety of optimization tasks representable as approximation of nonlinear mappings between subsets of spaces of functions, possibly dependent on a very large number of variables (see, e.g., [19, 20, 28, 33] and the references therein).

In this paper, we apply the properties of variable-basis schemes to the approximate minimization of error functionals and we estimate the rates of decrease of infima over admissible sets of Boolean functions computable by certain feedforward networks. The paper is written with three objectives. First, to put into evidence some general properties of approximate optimization over variable-basis functions, which play an important role in a variety of applications (learning from data, modelling, pattern recognition and classification, etc.). Second, by exploiting such properties in the finite-dimensional context, to investigate accuracy in solving optimization tasks that can be formalized as approximate minimization of error functionals on the space of real-valued Boolean functions of d variables (i.e., functions $f : \{0,1\}^d \to \mathcal{R}$). Third, to derive upper bounds on accuracy of minimization over admissible sets of functions computable by networks with a single linear output unit and computational units corresponding to perceptrons with signum (i.e., bipolar) activation function. We give conditions that guarantee well-posedness in the generalized sense of this optimization problem. We derive upper bounds on rates of decrease of infima of error functionals over admissible sets computable by networks with increasing number of computational units. We also describe various verifiable conditions that guarantee the applicability of our results.

The upper bounds are formulated in terms of various norms (l_1, l_2, spectral norm, and a certain generalization of the concept of total variation) of elements of the target sets defining error functionals. We describe target sets for which such rates do not exhibit the curse of dimensionality and we illustrate our results by examples of target sets containing various classes of functions used in applications (e.g., functions representable by "small" decision trees, functions expressible as linear combinations of a "small" number of generalized parities, functions with bounds on their Sobolev norms).

The paper is organized as follows. In Section 2, we introduce notations and state conditions that guarantee well-posedness in the generalized sense for the problem of minimization of error functionals. Section 3 contains a short survey on approximate optimization over variable-basis functions and gives estimates of rates of decrease of infima with increasing complexity of admissible sets. In Section 4, we derive estimates of rates of decrease of such infima in the space of multivariable real-valued Boolean functions for admissible sets computable by perceptron neural networks. In Section 5, we discuss various verifiable sufficient conditions for our estimates.

2. Preliminaries

Let $(X, \|\cdot\|)$ be a normed linear space. The ball of radius r centered at $h \in X$ is denoted by $B_r(h, \|\cdot\|)$; we let $B_r(\|\cdot\|) = B_r(0, \|\cdot\|)$ and when it is clear from the context which norm is used, we write $B_r(h) = B_r(h, \|\cdot\|)$ and $B_r = B_r(0)$. A sequence is denoted by $\{x_n\} = \{x_n : n \in \mathcal{N}_+\}$, where $\mathcal{N}_+$ is the set of positive integers. We say that a sequence in a normed linear space converges *subsequentially* if it has a convergent subsequence.

For a multi-index α, i.e., a d-tuple $(\alpha_1, \dots, \alpha_d)$ of nonnegative integers, let $D^\alpha = D_1^{\alpha_1} \dots D_d^{\alpha_d}$ denote a distributional partial derivative of order $|\alpha| = \sum_{i=1}^d \alpha_i$ [1, 1.57]. For $p \in [1, \infty)$ and an open set $\Omega \subseteq \mathcal{R}^d$, the Sobolev space $(W_p^m(\Omega), \|\cdot\|_{m,p,\Omega})$ is the set of all functions $f : \Omega \to \mathcal{R}$ such that $D^\alpha f \in L_p(\Omega)$ for $|\alpha| \leqslant m$, with the norm $\|f\|_{m,p,\Omega} = \{\sum_{|\alpha|\leqslant m} \|D^\alpha f\|_p^p\}^{1/p}$ [1, 3.1].

Let $\mathcal{R}$ denote the set of real numbers. A mapping $\Phi : X \to \mathcal{R} \cup \{+\infty\}$ is called a *proper extended-real-valued functional* if Φ is not a constant equal to $+\infty$. Following [7], we denote by (M, Φ) the problem of infimizing a functional

$\Phi : M \to \mathcal{R}$ over $M \subseteq X$. M is called the set of *admissible solutions* or the *admissible set*. A sequence $\{g_i\}$ of elements of M is called *Φ-minimizing over M* if $\lim_{i\to\infty} \Phi(g_i) = \inf_{g\in M} \Phi(g)$. The set of argminima of the problem (M, Φ) is denoted by $\text{argmin}(M, \Phi) = \{h \in M : \Phi(h) = \inf_{g\in M} \Phi(g)\}$. The problem (M, Φ) is *Tikhonov well-posed in the generalized sense* [7, p. 24] if $\text{argmin}(M, \Phi)$ is not empty and each Φ-minimizing sequence over M converges subsequentially to an element of $\text{argmin}(M, \Phi)$.

For C a nonempty subset of X, the *error functional* measuring the distance from C is denoted by e_C and defined for any $h \in X$ as $e_C(h) = \|h - C\|$. We call C the *target set* or the set of *target functions*. By the triangle inequality, $e_C = e_{\text{cl}(C)}$.

For a singleton $C = \{h\} \subset X$, we write merely e_h instead of $e_{\{h\}}$. Approximation theory has studied minimization of these functionals over many types of sets of functions. Properties of minimizing sequences and their rates of convergence have been described (see, e.g., [24, 30] and the references therein).

Recall that a nonempty subset M of a normed linear space is *compact* if every sequence has a convergent subsequence, is *precompact* if $\text{cl}(M)$ is compact, and is *boundedly compact* if its intersection with any ball is precompact (equivalently, every bounded sequence in M is subsequentially convergent). Note that this definition of boundedly compact set does not require M to be closed. M is *approximatively compact* [30, pp. 368, 382] if for all $h \in X$, every sequence in M that minimizes the distance to h converges subsequentially to an element of M.

In [13, Proposition 2.1], the notion of approximatively compact set has been reformulated in terms of optimization theory as a set M such that, for every $h \in X$, the problem (M, e_h) is Tikhonov well-posed in the generalized sense. It has also been pointed out that generalized Tikhonov well-posedness can be interpreted as a type of weakened compactness of admissible sets. The following theorem from [13], which will be used to derive some of the results of this paper, shows that for error functionals generalized Tikhonov well-posedness is closely related to the concept of approximative compactness.

THEOREM 2.1 ([13, Theorem 3.1]). *Let M and C be nonempty subsets of a normed linear space $(X, \|\cdot\|)$. Each of the following conditions guarantees that (M, e_C) is Tikhonov well-posed in the generalized sense*:

(i) *M is approximatively compact and C is precompact;*

(ii) *M is approximatively compact and bounded and C is boundedly compact;*

(iii) *M is boundedly compact and closed and C is bounded.*

3. Approximate Optimization over Variable-Basis Functions

An approximate solution of an optimization problem (M, Φ) by an iterative method entails the construction of a minimizing sequence converging to an element of the admissible set M. The *classical Ritz method* [9] constructs a minimizing sequence for (M, Φ) as a sequence of argminima of problems

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

where, for each n, X_n is an n-dimensional subspace of the space X and $X_n \subseteq X_{n+1}$. For conditions guaranteeing convergence of minimizing sequences in the classical Ritz method and estimates of their rates see, e.g., [6, Chapters 1 and 3], [7], and [9, Chapter 8].

We define a *generalized Ritz method* as an iterative method of approximate solution of a problem (M, Φ) by a sequence of problems

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

where $\{A_n\}$ is a nested sequence of subsets of X. In unconstrained optimization, one has $M = X$ (i.e., the set of admissible solutions is the whole space). In such a case, the Ritz method and the generalized Ritz method are iterative methods of approximate solution of problem (X, Φ) by a sequence of problems

$$\{(X_n, \Phi)\} \quad \text{and} \quad \{(A_n, \Phi)\},$$

resp.

Here, we shall consider two types of nested sequences of subsets for the generalized Ritz method. The first one is formed by linear combinations of at most n elements of a given set G,

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

while the second one is formed by convex combinations of at most n elements of G,

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

Approximation schemes of the form $\text{span}_n\, G$ and $\text{conv}_n\, G$ are called *variable-basis approximation* [17, 18]. Approximate minimization over $M \cap \text{span}_n\, G$ was introduced in [33] and called *extended Ritz method* (see also [19]).

Sets of the form $\text{span}_n\, G$ model situations in which admissible sets are formed by linear combinations of functions from a fixed basis set with unconstrained coefficients in the linear combinations. Typically, in applications such coefficients are constrained; for example, by a bound on some norm of the coefficients vector $(w_1, \ldots, w_n)$. When the norm is the l_1-norm, the corresponding functions belong to the set $\{\sum_{i=1}^{n} w_i g_i : w_i \in \mathcal{R}, g_i \in G, \sum_{i=1}^{n} |w_i| \leqslant c\}$, where $c > 0$ is the bound on the l_1-norm. It is easy to see that this set is contained in $\text{conv}_n\, G'$, where

$G' = \{rg : |r| \leqslant c,\ g \in G\}$. As any two norms on $\mathcal{R}^n$ are equivalent, any norm-based constraint on the coefficients of linear combinations defines a set contained in a set of the form $\operatorname{conv}_n G'$.

Depending on the choice of the set G, one can obtain a variety of admissible sets that include functions computable by feedforward neural networks, splines with free nodes, trigonometric polynomials with free frequencies, etc. [18].

For example, let $A \subseteq \mathcal{R}^q$, $K \subseteq \mathcal{R}^d$ and $\phi : A \times K \to \mathcal{R}$ be a function of two vector variables and let $G_\phi = \{\phi(a, \cdot) : a \in A\}$. By suitable choices of A and ϕ, one can represent as sets G_ϕ sets of functions computable by various computational units in neural networks. If $A = S^{d-1} \times \mathcal{R}$, where $S^{d-1} = \{e \in \mathcal{R}^d : \|e\| = 1\}$ is the set of unit vectors in $\mathcal{R}^d$, and $\phi((e, b), x) = \vartheta(e \cdot x + b)$, where ϑ denotes the Heaviside function, defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geqslant 0$, then G_ϕ is the set of characteristic functions of closed half-spaces of $\mathcal{R}^d$, restricted to K. If $A = [-c, c]^d \times [-c, c]$ and $\phi((v, b), x) = \psi(v \cdot x + b)$, where $\psi : \mathcal{R} \to \mathcal{R}$ is called *activation function*, then G_ϕ is the set of functions on K computable by *ψ-perceptrons* with both weights v and biases b bounded by c (a typical activation function for perceptrons is the logistic sigmoid $\psi(t) = 1/(1 + \mathrm{e}^{-t})$). If the activation function is positive and even, $A = [-c, c]^d \times [-c, c]$, and $\phi((v, b), x) = \psi(b\|x - v\|)$, where $\|\cdot\|$ is a norm on $\mathcal{R}^d$, then G_ϕ is the set of functions on K computable by *ψ-radial-basis-functions* (RBF) networks with widths b and coordinates v of centroids bounded by c (a typical activation function for RBF units is the Gaussian function $\psi(t) = \mathrm{e}^{-t^2}$).

Hence, admissible sets computable by feedforward neural networks with n computational units have the form $\operatorname{span}_n G_\phi$ or $\operatorname{conv}_n G_\phi$, depending whether the coefficients of the linear combinations are arbitrary or bounded. For many types of computational units ϕ, sets $\bigcup_{n \in \mathcal{N}_+} \operatorname{span}_n G_\phi$ are dense in the spaces of continuous or L_p functions on compacta (see [23] and the references therein).

In applications, the rate of decrease of the sequences of infima

$$\left\{ \inf_{g \in M \cap \operatorname{span}_n G} \Phi(g) \right\} \quad \text{and} \quad \left\{ \inf_{g \in M \cap \operatorname{conv}_n G} \Phi(g) \right\}$$

has to be fast enough so that functions from $\operatorname{span}_n G$ and $\operatorname{conv}_n G$, resp., are implementable. Since the union of all linear subspaces spanned by n-tuples of elements of a given set G is “much larger” than any single n-dimensional subspace, minimization of functionals over variable-basis-functions might lead to considerably faster rates than those achievable using the classical Ritz method.

We shall derive estimates of the rate of approximate infimization by variable-basis functions using a result from nonlinear approximation theory, called Maurey–Jones–Barron theorem (see [3, 12, 29]). Let G be a bounded subset of a Hilbert space and $s_G = \sup_{g \in G} \|g\|$. Such a theorem states that for any $f \in \operatorname{cl} \operatorname{conv} G$ and any positive integer n, one has $\|f - \operatorname{conv}_n G\| \leqslant \sqrt{\frac{s_G^2 - \|f\|^2}{n}}$. We refer to this theorem as MJB theorem and to its estimate as MJB bound.

Here we use reformulation of MJB theorem in terms of a norm, called G-variation and denoted by $\|\cdot\|_G$, which has been defined in [15] for a subset G of a normed linear space $(X, \|\cdot\|)$, as the Minkowski functional of the set $\mathrm{cl}\,\mathrm{conv}(G \cup -G)$, where $\mathrm{conv}\, G = \{\sum_{i=1}^n w_i g_i : w_i \in [0,1], \sum_{i=1}^n w_i = 1, g_i \in G, n \in \mathcal{N}_+\}$ denotes the convex hull of G. Thus,

$$\|f\|_G = \inf\bigl\{c \in \mathcal{R}_+ : c^{-1} f \in \mathrm{cl}\,\mathrm{conv}(G \cup -G)\bigr\}.$$

G-variation is a norm on the subspace $\{f \in X : \|f\|_G < \infty\} \subseteq X$; for its properties see [15, 17] and [18]. Roughly speaking, G-variation of f represents how much the set G should be "dilated" so that f is contained in the closure of the symmetric convex hull of G.

When G is an orthonormal basis of a separable Hilbert space (i.e., a Hilbert space with a countable dense subset), then G-variation can be expressed using *l_1-norm with respect to* G defined, for $f \in X$, as $\|f\|_{1,G} = \sum_{g\in G} |f \cdot g|$. It has been shown in [21] and [17] that for any orthonormal basis G of a separable Hilbert space, G-variation is equal to l_1-norm with respect to G. Thus the notion of G-variation is a generalization of the notion of l_1-norm. It is also generalization of the concept of total variation studied in integration theory, since for functions of one variable variation with respect to perceptrons coincides, up to a constant, with the notion of total variation [2].

MJB bound reformulated in terms of G-variation states that for any bounded subset G of a Hilbert space $(X, \|\cdot\|)$, any $f \in X$ and any positive integer n, one has

$$\|f - \mathrm{span}_n\, G\| \leqslant \|f - \mathrm{conv}_n\, G(r)\| \leqslant \frac{r s_G}{\sqrt{n}}, \tag{1}$$

where $r = \|f\|_G$ and $G(r) = \{wg : g \in G, w \in \mathcal{R}, |w| \leqslant r\}$.

The following theorem gives upper bounds on the speed of decrease of infima of an error functional e_C over $\mathrm{span}_n\, G$ with n increasing, in terms of the infimum of G-variations of the functions in the target set C.

THEOREM 3.2. *Let C, G and M be nonempty subsets of a Hilbert space $(X, \|\cdot\|)$ such that both $r = \inf_{f\in C} \|f\|_G$ and $s_G = \sup_{g\in G} \|g\|$ are finite. Then the following hold*:

(i) *For every positive integer n,*

$$\inf_{g\in \mathrm{span}_n G} e_C(g) \leqslant \inf_{g\in \mathrm{conv}_n G(r)} e_C(g) \leqslant \frac{r s_G}{\sqrt{n}}.$$

(ii) *If for some positive integer n_0, $\mathrm{conv}_{n_0} G(r) \subseteq M$, then for every $n \geqslant n_0$,*

$$\inf_{g\in M\cap \mathrm{span}_n G} e_C(g) \leqslant \inf_{g\in M\cap \mathrm{conv}_n G(r)} e_C(g) \leqslant \frac{r s_G}{\sqrt{n}}.$$

(iii) *If* $(X, \|\cdot\|)$ *is separable and* G *is its orthonormal basis, then for every positive integer* n*,*

$$\inf_{g\in \operatorname{span}_n G} e_C(g) \leqslant \inf_{g\in \operatorname{conv}_n G(r)} e_C(g) \leqslant \frac{rs_G}{2\sqrt{n}}.$$

Proof. (i) For each $t > r$, choose $f_t \in C$ such that $r \leqslant \|f_t\|_G < t$. By MJB bound (1) for every n, $\|f_t - \operatorname{conv}_n G(t)\| \leqslant ts_G/\sqrt{n}$ and so there exists a sequence $\{g_{t,i}\} \subset \operatorname{conv}_n G(t)$ such that $\|f_t - \operatorname{span}_n G\| = \lim_{i\to\infty} \|f_t - g_{t,i}\| \leqslant ts_G/\sqrt{n}$. As $f_t \in C$, we have $e_C(g_{t,i}) \leqslant e_{f_t}(g_{t,i}) = \|f_t - g_{t,i}\|$ and hence $\inf_{g\in \operatorname{conv}_n G(t)} e_C(g) \leqslant ts_G/\sqrt{n}$. Since $\operatorname{conv}_n G(r) = \cap\{\operatorname{conv}_n G(t) : t > r\}$, we have

$$\inf_{g\in \operatorname{span}_n G} e_C(g) \leqslant \inf_{g\in \operatorname{conv}_n G(r)} e_C(g) \leqslant \frac{rs_G}{\sqrt{n}}.$$

Part (ii) follows directly from (i) as for all $n \geqslant n_0$, $M \cap \operatorname{conv}_n G(r) = \operatorname{conv}_n G(r)$.

Part (iii) is proven analogously to part (i) with MJB bound replaced by the bound $rs_G/(2\sqrt{n})$, which holds when G is an orthonormal basis of a separable space (see [21, Theorem 2.7] and [17, Theorem 3]). □

Note that when C and $\operatorname{span}_n G$ or $\operatorname{conv}_n G(r)$ satisfy the assumptions of Theorem 2.1 (see [13] for examples of such cases), the problems $(\operatorname{span}_n G, e_C)$ and $(\operatorname{conv}_n G(r), e_C)$ are Tikhonov well-posed in the generalized sense, so the infima considered in Theorem 3.2 are achieved.

4. Rates of Approximate Optimization by Real-Valued Boolean Variable-Basis Functions

In this section, we apply the results from the previous section to approximate minimization in the space $\mathcal{B}(\{0,1\}^d)$ of real-valued Boolean functions. This space is endowed with the standard inner product defined for $f, g \in \mathcal{B}(\{0,1\}^d)$, as $f \cdot g = \sum_{x\in\{0,1\}^d} f(x)g(x)$, which induces the norm $\|f\| = \|f\|_{l_2} = \sqrt{f\cdot f}$. The space $(\mathcal{B}(\{0,1\}^d, \|\cdot\|)$ is isomorphic to the 2^d-dimensional Euclidean space $\mathcal{R}^{2^d}$ with the l_2-norm.

The following corollary gives conditions on the subsets C, M, and G of $\mathcal{B}(\{0,1\}^d)$ guaranteeing that the problems $(M\cap \operatorname{conv}_n G, e_c)$ and $(M\cap \operatorname{span}_n G, e_c)$ are Tikhonov well-posed in the generalized sense.

COROLLARY 4.3. *Let* d *be a positive integer and* C, M, G *be subsets of* $\mathcal{B}(\{0,1\}^d)$ *such that* C *is bounded,* M *compact, and* G *finite. Then for every positive integer* n*, the problems* $(M \cap \operatorname{conv}_n G, e_C)$ *and* $(M \cap \operatorname{span}_n G, e_C)$ *are Tikhonov well-posed in the generalized sense.*

Proof. Since G is finite, $\operatorname{conv}_n G$ is compact and $\operatorname{span}_n G$ is boundedly compact and closed. As M is a compact subset of a finite-dimensional space, $M \cap \operatorname{conv}_n G$

and $M \cap \mathrm{span}_n\, G$ are compact and closed boundedly compact, resp. So by Theorem 2.1(iii) both problem $(M \cap \mathrm{conv}_n\, G, e_C)$ and $(M \cap \mathrm{span}_n\, G, e_C)$ are Tikhonov well-posed in the generalized sense. □

An important class of Boolean variable-basis functions are functions computable by perceptron feedforward neural networks. We consider perceptrons with the signum (bipolar) activation function, defined as $\mathrm{sgn}(t) = -1$ for $t < 0$ and $\mathrm{sgn}(t) = 1$ for $t \geqslant 0$, instead of more common Heaviside function that assigns zero to negative numbers. Let $\bar{H}_d$ denotes the set of functions on $\{0,1\}^d$ computable by signum perceptrons, i.e., $\bar{H}_d = \{f : \{0,1\}^d \to \mathcal{R} : f(x) = \mathrm{sgn}(v \cdot x + b), v \in \mathcal{R}^d, b \in \mathcal{R}\}$.

Taking advantage of the equivalence between G-variation and l_1-norm with respect to an orthonormal countable basis G in any separable Hilbert space [18], we shall estimate variation with respect to signum perceptrons using variations with respect to two orthonormal bases of $\mathcal{B}(\{0,1\}^d)$. The first one is the *Euclidean orthonormal basis*, defined as $E_d = \{e_u : u \in \{0,1\}^d\}$, where $e_u(u) = 1$ and for every $x \in \{0,1\}^d$ with $x \neq u$, $e_u(x) = 0$. The second one is the *Fourier orthonormal basis* (see, e.g., [32]) defined as $F_d = \left\{f_u : u \in \{0,1\}^d, f_u(x) = \frac{1}{\sqrt{2^d}}(-1)^{u\cdot x}\right\}$. Every $f \in \mathcal{B}(\{0,1\}^d)$ can be represented as $f(x) = \frac{1}{\sqrt{2^d}} \sum_{u \in \{0,1\}^d} \hat{f}(u)(-1)^{u\cdot x}$, where $\hat{f}(u) = \frac{1}{\sqrt{2^d}} \sum_{x \in \{0,1\}^d} f(x)(-1)^{u\cdot x}$. The l_1-norm with respect to the Fourier basis, $\|f\|_{1,F_d} = \|\hat{f}\|_{l_1} = \sum_{u \in \{0,1\}^d} |\hat{f}(u)|$, called the *spectral norm*, is equal to F_d-variation (see [17] and [21]). For a subset $I \subset \{0,1\}^d$, I-parity is defined by $p_I(u) = 1$ if $\sum_{i \in I} u_i$ is odd, and $p_I(u) = 0$ otherwise. If we interpret the output 1 as -1 and 0 as 1, then the elements of the Fourier basis F_d correspond to the generalized parity functions.

Next proposition investigates Tikhonov well-posedness and estimates rates of approximate solution of (M, e_C) by a generalized Ritz method with $M = \mathcal{B}(\{0,1\}^d)$ and A_n equal to linear or convex combinations of certain Boolean functions. Moreover, the proposition gives conditions on target sets, which guarantee rates of minimization of error functionals of the order of $\mathcal{O}(\frac{1}{\sqrt{n}})$ for any number of variables d. By G^0 is denoted the set of normalized elements of G with respect to the norm $\|\cdot\|$ (note that $E_d^0 = E_d$ and $F_d^0 = F_d$). We call $\|f\|_{G^0}$ *normalized* G-*variation of* f. We use G^0-variation in our estimates as, for every $f \in X$, we have $\|f\| \leqslant \|f\|_{G^0}$ (i.e., the unit ball in G^0-variation is contained in the unit ball in $\|\cdot\|$) and $\|f\|_{G^0} \leqslant \|f\|_G \sup_{g \in G} \|g\| = \|f\|_G s_G$ [15].

PROPOSITION 4.4. *Let d be a positive integer, $r > 0$, and C be a bounded subset of $\mathcal{B}(\{0,1\}^d)$. Then the following hold*:

(i) *If $C \cap B_r(\|\cdot\|_{\bar{H}_d^0}) \neq \emptyset$, then for every positive integer n, the problems* $(\mathrm{span}_n\, \bar{H}_d, e_C)$ *and* $(\mathrm{conv}_n\, \bar{H}_d(r), e_C)$ *are Tikhonov well-posed in the generalized sense and*

$$\min_{g\in \operatorname{span}_n \bar{H}_d} e_C(g) \leqslant \min_{g\in \operatorname{conv}_n \bar{H}_d(r)} e_C(g) \leqslant \frac{r}{\sqrt{n}}.$$

(ii) *If* $C \cap B_r(\|\cdot\|_{1,F_d}) \neq \emptyset$, *then for every positive integer* n, *the problems* $(\operatorname{span}_{dn+1} \bar{H}_d, e_C)$ *and* $(\operatorname{conv}_{dn+1} \bar{H}_d(r), e_C)$ *are Tikhonov well-posed in the generalized sense and*

$$\min_{g\in \operatorname{span}_{dn+1} \bar{H}_d} e_C(g) \leqslant \min_{g\in \operatorname{conv}_{dn+1} \bar{H}_d(r)} e_C(g) \leqslant \frac{r}{2\sqrt{n}}.$$

(iii) *If* $C \cap B_r(\|\cdot\|_{1,E_d}) \neq \emptyset$, *then for every positive integer* n, *the problems* $(\operatorname{span}_{n+1} \bar{H}_d, e_C)$ *and* $(\operatorname{conv}_{n+1} \bar{H}_d(r), e_C)$ *are Tikhonov well-posed in the generalized sense and*

$$\min_{g\in \operatorname{span}_n \bar{H}_d} e_C(g) \leqslant \min_{g\in \operatorname{conv}_n \bar{H}_d(r)} e_C(g) \leqslant \frac{r}{2\sqrt{n-1}}.$$

Proof. (i) The statement follows from Corollary 4.3 and Theorem 3.2(i).

(ii) It is easy to verify that every function from the Fourier basis F_d can be expressed as a linear combination of at most $d+1$ signum perceptrons [21]. Indeed, for every $u, x \in \{0,1\}^d$ one has $(-1)^{u\cdot x} = \frac{1+(-1)^d}{2} + \sum_{j=1}^{d}(-1)^j \operatorname{sgn}(u\cdot x - j + \frac{1}{2})$. Moreover, any linear combination of n elements of F_d belongs to $\operatorname{span}_{dn+1} \bar{H}_d$, since all of the n occurrences of the constant function can be expressed by a single perceptron. As $\|\tilde{f}\|_1 = \|f\|_{1,F_d} = \|f\|_{F_d}$, the statement follows from Corollary 4.3 and Theorem 3.2(iii).

(iii) It is easy to check that for any $u \in \{0,1\}^d$, $e_u(x)$ is expressible as $\frac{\operatorname{sgn}(v\cdot x+b)+1}{2}$ for appropriate v and b [21]. Analogously as in the proof of (ii), adding several occurrences of the constant function together, one obtains a representation of every linear combination of n functions of the Euclidean basis as an element of $\operatorname{span}_{n+1} \bar{H}_d$. As $\|f\|_{1,E_d} = \|f\|_{E_d}$, the statement follows from Corollary 4.3 and Theorem 3.2(iii). □

By Proposition 4.4, “fast” rates of minimization are guaranteed when target sets contain a function with either “small” variation with respect to signum perceptrons or “small” spectral norm or “small” norm with respect to the Euclidean basis. Depending on which of these norms is smaller or for which an estimate is available, one of the conditions (i), (ii), and (iii) of Proposition 4.4 can be applied.

5. Discussion

Deriving upper bounds on rates of approximation from Theorem 3.2 and Proposition 4.4 requires to estimate G-variation.

Upper bounds obtained via Proposition 4.4(ii) and (iii) require to estimate variation with respect to the orthonormal sets F_d and E_d, respectively. This may exhibit limitations with respect to upper bounds derived via Proposition 4.4(i) combined with estimates of variation with respect to the set $\bar{H}_d$: examples of functions for which $\bar{H}_d^0$-variation grows linearly with d while both F_d-variation and E_d-variation grow exponentially are given in [21]. In [3, pp. 941–942], upper bounds on $\bar{H}_d$-variation were derived in via estimates of a spectral norm (see also [22]).

In [21], it was shown that linear combination of a "small" number of generalized parities have "small" variation. More precisely, let C be a subset of $\mathcal{B}(\{0,1\}^d)$ containing a function f with at most m Fourier coefficients nonzero and with $\|f\| \leqslant c$. Proceeding as in [21, p. 655], we obtain $f = \sum_{i=1}^m w_i g_i$, where $g_i \in F_d$. Hence, $\|f\|_{F_d} = \|\tilde{f}\|_1 = \|f\|_{1,F_d} = \sum_{i=1}^m |w_i|$. By the Cauchy–Schwarz inequality one has $\sum_{i=1}^m |w_i| \leqslant \|w\|\|u\|$, where $w = (w_1, \dots, w_m)$ and $u = (u_1, \dots, u_m)$, with $u_i = \operatorname{sgn}(w_i)$. As $\|w\| = \|f\| \leqslant c$ and $\|u\| \leqslant \sqrt{m}$, we have $\|f\|_{1,F_d} \leqslant c\sqrt{m}$. Thus C contains a function f with $\|f\|_{1,F_d} \leqslant c\sqrt{m}$. So Proposition 4.4(ii) implies that, when e_C is minimized over the set of d-variable Boolean functions computable by networks with $dn + 1$ signum perceptrons, where $n \geqslant \frac{c^2 m}{4\varepsilon^2}$, then its minimum is bounded from above by ε. As the number $\frac{dc^2 m}{4\varepsilon^2} + 1$ of perceptrons needed for an accuracy ε grows with d linearly, the curse of dimensionality is avoided.

Another application of Proposition 4.4 is to decision trees, which play an important role in machine learning (see, e.g., [22] and the references therein). Recall that a *decision tree* is a binary tree with labeled nodes and edges. The *size* of a decision tree is the number of its leaves. A function $f : \{0,1\} \to \mathcal{R}$ is representable by a decision tree if there exists a tree with internal nodes labeled by variables $x_1, \dots, x_d$, all pairs of edges outgoing from a node labeled by 0s and 1s, and all leaves labeled by real numbers, such that f can be computed by this tree as follows. The computation starts at the root and after reaching an internal node labeled by x_i, continues along the edge whose label coincides with the actual value of the variable x_i; finally a leaf is reached and its label is equal to $f(x_1, \dots, x_d)$. Let C be a subset of $\mathcal{B}(\{0,1\}^d)$ containing a function f such that, for all $x \in \{0,1\}^d$, $f(x) \neq 0$, f is representable by a decision tree of size s, and $\frac{\max_{x\in\{0,1\}^d} |f(x)|}{\min_{x\in\{0,1\}^d} |f(x)|}\|f\| \leqslant b$. According to [21, Theorem 3.4], $\|f\|_{1,F_d} = \|\tilde{f}\|_1 \leqslant sb$. An error functional defined by such target sets achieves for the minimum a value bounded from above by ε when minimization is performed over admissible sets of d-variable Boolean functions computable by neural networks with $dn + 1$ signum perceptrons, where $n \geqslant (\frac{sb}{2\varepsilon})^2$. Thus Proposition 4.4(ii) implies that for target sets containing a function with sb bounded by a constant independent of d or growing "slowly" with d, the curse of dimensionality in minimization over Boolean perceptron networks is avoided.

Upper bounds on variation in the Boolean case can be derived from upper bounds on variations of suitable extensions of Boolean functions to a domain Ω containing $[0,1]^d$, since $\bar{H}_d$ can be obtained by restricting to the Boolean cube $\{0,1\}^d$ the set H_d of functions computable by signum perceptrons defined on Ω. For target sets C containing a sufficiently smooth function, this can be combined with the possibility of embedding balls in Sobolev norms into balls of proper radii in H_d-variation. More precisely, let Ω be an uniformly C^s-regular domain in $[0, 1]^d$ (for the definition of C^s-regular domain, see [1, p. 67]),

$$a = \inf\{\|h\|_{2,s,\Omega} : h \in C_{|\Omega}\},$$

and $b = (\int_{\mathcal{R}^d} (1 + \|\omega\|^{2(s-1)})^{-1} \mathrm{d}\omega)^{1/2}$ (for any $x \in \mathcal{R}^d$ and $d \geqslant 1$, $\|x\|$ denotes its Euclidean norm). Hence a is a lower bound on the Sobolev norm of functions in $C_{|\Omega}$. Let

$$\mathcal{E} : (W_2^s(\Omega), \|\cdot\|_{2,s,\Omega}) \to (W_2^s(\mathcal{R}^d), \|\cdot\|_{2,s,\mathcal{R}^d})$$

be an extension operator such that for all $f \in (W_2^s(\Omega), \|\cdot\|_{2,s,\Omega})$ one has $(\mathcal{E}f)_{|\Omega} = f$ a.e. in Ω and $\|\mathcal{E}f\|_{2,s,\mathcal{R}^d} \leqslant c\|f\|_{2,s,\Omega}$, where $c > 0$ is a constant depending on s and Ω; see [1, pp. 83–84]. For every $\varepsilon > 0$, suppose that C contains a function in the ball of radius $a + \varepsilon/c$ in $W_2^s(\Omega)$. Thus, $\mathcal{E}f$ is in the ball of radius $ac + \varepsilon$ in $W_2^s(\mathcal{R}^d)$. Since ε can be arbitrarily small, $\mathcal{E}f$ is in the ball of radius ac in $W_2^s(\mathcal{R}^d)$. Arguing as in [3, pp. 935, 941], we obtain that

$$B_{ac}(\|\cdot\|_{2,s,\mathcal{R}^d})_{|\Omega} \subseteq B_{2abc}(\|\cdot\|_{H_d})_{|\Omega},$$

where $b = (\int_{\mathcal{R}^d} (1 + \|\omega\|^{2(s-1)})^{-1} \mathrm{d}\omega)^{1/2}$; b is finite as $2(s - 1) > d$. Combining this with Proposition 4.4(i), one obtains upper bounds on $(\mathrm{span}_n \bar{H}_d, e_C)$ and $(\mathrm{conv}_n \bar{H}_d(r), e_C)$ formulated in terms of the smallest Sobolev norm of elements of the target set C.

All the examples of minimization of error functionals discussed above share a common feature, which has a deep meaning: a fixed accuracy ε of approximate minimization can be guaranteed for any value d of the dimension (number of variables) by requiring that the target set C contains at least one “sufficiently smooth” function (it may happen that the larger d, the more restrictive such a requirement becomes). In other words, the “curse of dimensionality” in minimization of error functionals over variable-basis functions can be mitigated by the “blessing of smoothness”.

References

1. Adams, R. A.: *Sobolev Spaces*, Academic Press, New York, 1975.
2. Barron, A. R.: Neural net approximation, in K. Narendra (ed.), *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, Yale University Press, 1992, pp. 69–72.
3. Barron, A. R.: Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. on Inform. Theory* **39** (1993), 930–945.
4. Bellman, R.: *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
5. Cucker, F. and Smale, S.: On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* **39** (2001), 1–49.
6. Daniel, J. W.: *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
7. Dontchev, A. L. and Zolezzi, T.: *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, Berlin, 1993.

8. Donahue, M. J., Gurvits, L., Darken, C. and Sontag, E.: Rates of convex approximation in non-Hilbert spaces, *Constr. Approx.* **13** (1997), 187–220.
9. Gelfand, I. M. and Fomin, S. V.: *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
10. Girosi, F. and Anzellotti, G.: Rates of convsergence for radial basis functions and neural networks, in R. J. Mammone (ed.), *Artificial Neural Networks for Speech and Vision*, Chapman & Hall, London, 1993, pp. 97–114.
11. Gurvits, L. and Koiran, P.: Approximation and learning of convex superpositions, *J. Comput. System Sci.* **55** (1997), 161–170.
12. Jones, L. K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Ann. Statist.* **20** (1992), 608–613.
13. Kainen, P. C., Kůrková, V. and Sanguineti, M.: Minimization of error functionals over variable-basis functions, *SIAM J. Optim.* **14** (2003), 732–742.
14. Kainen, P. C., Kůrková, V. and Vogt, A.: Continuity of approximation by neural networks in $\mathcal{L}_p$-spaces, *Ann. Oper. Res.* **101** (2001), 143–147.
15. Kůrková, V.: Dimension-independent rates of approximation by neural networks, in K. Warwick and M. Kárný (eds), *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality*, Birkhäuser, Boston, MA, 1997, pp. 261–270.
16. 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** (1997), 1061–1068.
17. Kůrková, V. and Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation, *IEEE Trans. on Inform. Theory* **47** (2001), 2659–2665.
18. Kůrková, V. and Sanguineti, M.: Comparison of worst case errors in linear and neural network approximation, *IEEE Trans. on Inform. Theory* **48** (2002), 264–275.
19. Kůrková, V. and Sanguineti, M.: Error estimates for approximate optimization by the extended Ritz method, *SIAM J. Optim.* **15** (2005), 461–487.
20. Kůrková, V. and Sanguineti, M.: Learning with generalization capability by kernel methods of bounded complexity, *J. Complexity*, in press.
21. Kůrková, V., Savický, P. and Hlaváčková, K.: Representations and rates of approximation of real-valued Boolean functions by neural networks, *Neural Networks* **11** (1998), 651–659.
22. Kushilevicz, E. and Mansour, Y.: Learning decision trees using the Fourier spectrum, *SIAM J. Comput.* **22** (1993), 1331–1348.
23. Leshno, M., Pinkus, A. and Schocken, S.: Multilayer feedforward networks with a non-polynomial activation function can approximate any function, *Neural Networks* **6** (1993), 861–867.
24. Lorentz, G. G., v. Golitschek, M. and Makovoz, Y.: *Constructive Approximation. Advanced Problems*, Springer-Verlag, 1996.
25. Makovoz, Y.: Uniform approximation by neural networks, *J. Approx. Theory* **95** (1998), 215–228.
26. Mhaskar, H. N. and Micchelli, C. A.: Dimension-independent bounds on the degree of approximation by neural networks, *IBM J. Res. Devel.* **38** (1994), 277–283.
27. Micchelli, C. A., Xu, Y. and Ye, P.: Cucker Smale learning theory in Besov spaces, in J. Suykens, G. Horváth, S. Basu, C. Micchelli and J. Vanderwalle (eds), *Advances in Learning Theory: Methods, Models, and Applications*, Nato Science Series, IOS Press, Amsterdam, 2003.
28. Narendra, K. S., Balakrishnan, J. and Ciliz, K. M.: Adaptation and learning using multiple models, switching, and tuning, *IEEE Control Systems Magazine* **15** (1995), 37–51.
29. Pisier, G.: Remarques sur un resultat non publié de B. Maurey, in *Seminaire d'Analyse Fonctionelle*, vol. I, no. 12, École Polytechnique, Centre de Mathématiques, Palaiseau, France, 1980–81.

30. Singer, I.: *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer-Verlag, Berlin, 1970.
31. Smale, S. and Zhou, D.-X.: Estimating the approximation error in learning theory, *Analysis and Applications* **1** (2003), 1–25.
32. Weaver, H. J.: *Applications of Discrete and Continuous Fourier Analysis*, Wiley, New York, 1983.
33. 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** (2002), 403–440.