SIAM JOURNAL ON OPTIMIZATION 14: 732--742, 2004

MINIMIZATION OF ERROR FUNCTIONALS OVER VARIABLE-BASIS FUNCTIONS [*]

PAUL C. KAINEN[†], VĚRA KŮRKOVÁ[‡], AND MARCELLO SANGUINETI[§]

Abstract. There is investigated generalized Tychonov well-posedness of the problem of minimization of error functionals over admissible sets formed by variable-basis functions, which include neural networks. For admissible sets formed by variable-basis functions of increasing complexity, rates of decrease of infima of error functionals are estimated. There are derived upper bounds on such rates that do not exibit the curse of dimensionality with respect to the number of variables of admissible functions.

Key words. generalized Tychonov well-posedness, error functionals, approximate optimization, rate of decrease of infima, complexity of admissible functions, curse of dimensionality.

AMS subject classifications. 49K40, 41A46, 41A25

1. Introduction.

Functionals defined as distances from (target) sets are called *error functionals.* Minimization of such functionals occurs in optimization tasks arising in various areas, such as system identification, machine learning, pattern recognition, etc.

In various applications, admissible solutions over which error functionals are minimized are functions depending on a large number of variables: for example, when routing strategies have to be devised for large-scale communication and transportation networks, when an optimal closed-loop control law has to be devised for a dynamical system with high-dimensional state, etc. In the last decades, complex optimization problems of this kind have been approximately solved by searching suboptimal solutions over admissible sets of functions computable by neural networks [4], [20], [21], [25], [28], [29]. Neural networks can be studied in a more general context of variable-basis functions, which also includes other nonlinear families of functions such as free-nodes splines or trigonometric polynomials with free frequencies [17]. Families of variable-basis functions are formed by linear combinations of fixed number of elements chosen from a given basis without a prespecified ordering [16], [17].

When admissible functions depend on a large number of variables, implementation of some procedures of approximate optimization may be infeasible due to the "curse of dimensionality" [3]. For example, when optimization is performed over linear combinations of fixed basis functions, the number of functions in linear combinations required to guarantee a desired optimization accuracy may grow exponentially fast

[*] Collaboration between P. C. Kainen and V. Kůrková was partially supported by NAS COBASE grant, between V. Kůrková and M. Sanguineti by the Scientific Agreement Italy-Czech Republic, Area MC 6, Project 22: "Functional Optimization and Nonlinear Approximation by Neural Networks". P. C. Kainen was partially supported by an international travel grant from Georgetown University, V. Kůrková by GA ČR grant 201/00/1482, and M. Sanguineti by the CNR-Agenzia 2000 project "New Algorithms and Methodologies for the Approximate Solution of Nonlinear Functional Optimization Problems in a Stochastic Environment."

[†] Department of Mathematics, Georgetown University, Washington, D.C. 20057-1233, USA (kainen@georgetown.edu).

[‡] 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).

with the number of variables of admissible solutions [22, pp. 232-233], [29]. However, experience has shown that some neural networks with a small number of computational units, which can be modeled as variable-basis functions with a small number of functions from the basis, often perform quite well in some optimization tasks where admissible solutions depend on a large number of variables [4], [20], [21], [25], [28], [29].

In this paper, we investigate generalized Tychonov well-posedness of the problems of minimization of error functionals over admissible sets formed by variable-basis functions and we estimate rates of decrease of infima of such problems with increasing complexity of admissible sets. As tools for such an investigation, we derive various conditions on target and admissible sets guaranteeing convergence of minimizing sequences. We show that these conditions are satisfied by target sets defined by suitable interpolation and smoothness conditions and admissible sets formed by functions computable by families of variable-basis functions that include commonly used classes of neural networks. We estimate rates of decrease of infima of error functionals over neural networks with increasing number of computational units. We derive upper bounds on such rates that do not exhibit the curse of dimensionality.

The paper is organized as follows. In Section 2, we introduce basic concepts and definitions used throughout the paper. Section 3 states conditions on sets of target functions and admissible solutions that guarantee convergence of minimizing sequences. Section 4 applies the tools developed in Section 3 to minimization of error functionals over neural networks and variable-basis functions and Section 5 gives estimates of rates of decrease of infima of such functionals with increasing number of computational units.

2. Preliminaries. In this paper, 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. For a positive integer d, a set $\Omega \subseteq \Re^d$, where $\Re$ denotes the set of real numbers, and $p \in [1, \infty)$, by $(L_p(\Omega), \|.\|_p)$ is denoted the space of measurable, real-valued functions on Ω such that $\int_\Omega |f(x)|^p \, dx < \infty$ endowed with the L_p norm. $(\mathcal{C}(\Omega), \|.\|_{\mathcal{C}})$ denotes the space of real-valued continuous functions on Ω with the supremum norm.

For a multi-index α, i.e, a d-tuple $(\alpha_1, \ldots, \alpha_d)$ of nonnegative integers, by $D^\alpha = D_1^{\alpha_1} \ldots D_d^{\alpha_d}$ is denoted a distributional derivative of order $\|\alpha\|_{l_1} = \sum_{i=1}^n \alpha_i$ [1, 1.57]. For $p \in [1, \infty)$ and an open set $\Omega \subseteq \Re^d$, the Sobolev space $(W_p^m(\Omega), \|.\|_{m,p})$ is the set of all functions $f : \Omega \to \Re$ such that $f \in (L_p(\Omega), \|.\|_p)$ and $D^\alpha f \in (L_p(\Omega), \|.\|_p)$ for $0 \leq \|\alpha\|_{l_1} \leq m$, with the norm $\|f\|_{m,p} = \left\{ \sum_{0 \leq \|\alpha\|_{l_1} \leq m} \|D^\alpha f\|_p^p \right\}^{1/p}$ [1, 3.1].

By $\mathcal{B}(\{0,1\}^d)$ is denoted the space of real-valued Boolean functions, i.e., functions from $\{0,1\}^d$ to $\Re$. 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 l_2-norm $\|f\|_{l_2} = \sqrt{f \cdot f}$. The space $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ is isomorphic to the 2^d-dimensional Euclidean space $\Re^{2^d}$ with the l_2-norm.

For $M \subseteq (X, \|.\|)$, $cl(M)$ denotes the closure of M in the topology induced by the norm $\|.\|$. For $f \in X$, we write $\|f - M\| = \inf_{g \in M} \|f - g\|$. A ball of radius r centered at $h \in (X, \|.\|)$ is denoted by $B_r(h, \|.\|) = \{f \in X : \|f - h\| \leq r\}$. We write $B_r(\|.\|)$ for $B_r(0, \|.\|)$ and merely B_r when it is clear which norm is used.

For brevity, sequences are denoted by $\{h_i\}$ instead of $\{h_i : i \in \mathcal{N}_+\}$, where $\mathcal{N}_+$ is the set of positive integers. When there is no ambiguity, the same notation is used for a sequence and its subsequences. A sequence converges *subsequentially* if it has a convergent subsequence.

Following [8], we denote by (M, Φ) the problem of infimizing a functional $\Phi : M \to \Re$ over a subset M of 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 *Tychonov well-posed in the generalized sense* [8, 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 $e_C(h) = \|h - C\|$, where $\|h - C\| = \inf_{f\in C} = \|h - f\|$. We call C the *target set* or the set of *target functions*. By the triangle inequality, $e_C = e_{cl(C)}$. For a singleton $C = \{h\} \subset X$, we write e_h instead of $e_{\{h\}}$.

For error functionals, the definition of generalized Tychonov well-posedness can be simplified as stated in the following proposition.

Proposition 2.1. *Let M, C be nonempty subsets of a normed linear space $(X, \|.\|)$. Then (M, e_C) is Tychonov well-posed in the generalized sense if and only if every sequence in M which minimizes e_C converges subsequentially to an element of M.*

Proof. Let $\{g_i\}$ be a subsequence of an e_C-minimizing sequence converging to $g^o \in M$. By continuity of e_C [26, p. 391], $\inf_{g\in M} e_C(g) = \lim_{i\to\infty} e_C(g_i) = e_C(\lim_{i\to\infty} g_i) = e_C(g_0)$. Thus, $g^o \in \text{argmin}(M, e_C)$ and so (M, e_C) is Tychonov well-posed in the generalized sense. The "only if " statement follows directly from the definition of generalized Tychonov well-posedness. □

Recall that a nonempty subset M of a normed linear space is *compact* if every sequence has a convergent subsequence, is *precompact* if $cl(M)$ 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* [26, p. 382] if, for all $h \in X$, every sequence in M that minimizes the distance to h converges subsequentially to an element of M.

By Proposition 2.1, the notion of approximatively compact set can be reformulated in terms of optimization theory as a set M such that, for every $h \in X$, the problem (M, e_h) is Tychonov well-posed in the generalized sense. A subset M of a normed linear space X is *proximinal* (or an *existence set*) if for any $h \in X$ there exists $g \in M$ such that $\|h - M\| = \|h - g\|$. In decreasing degree of strength, a subset of a normed linear space may be compact, boundedly compact, approximatively compact, and proximinal [26, pp. 368, 382-383]. Each implies the next with the exception that bounded compactness only implies approximative compactness for closed sets; proximinal implies closed [26, p. 382].

3. Minimization of error functionals under weakened compactness. Generalized Tychonov well-posedness can be interpreted as a type of weakened compactness of admissible sets. The following theorem shows that for error functionals it is closely related to the concept of approximative compactness studied in approximation theory [26, p. 382].

Theorem 3.1. *Let M, C be nonempty subsets of a normed linear space $(X, \|.\|)$. Each of the following conditions guarantees that (M, e_C) is Tychonov 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.

Proof. Let $\{g_i\}$ be an e_C-minimizing sequence over M. By Proposition 2.1 it is sufficient to show that $\{g_i\}$ converges subsequentially to $g^o \in M$.

(i) Since $e_C = e_{cl(C)}$, it is sufficient to consider $cl(C)$. As $cl(C)$ is compact, it is proximinal and so there exists a sequence $\{f_i\} \subseteq cl(C)$ such that for every i, $e_C(g_i) = \|f_i - g_i\|$. Again by compactness, the sequence $\{f_i\}$ converges subsequentially to $f_0 \in cl(C)$. Replacing $\{f_i\}$ and $\{g_i\}$ with the corresponding subsequences, for every $\varepsilon > 0$ we get $i_0 \in \mathcal{N}_+$ such that for all $i \geq i_0$, $\|f_i - f_0\| < \varepsilon/2$. As $\{g_i\}$ is e_C-minimizing over M, there exists $i_1 \geq i_0$ such that for all $i \geq i_1$, $e_C(g_i) \leq \inf_{g \in M} e_C(g) + \varepsilon/2$. So, for all $i \geq i_1$, $e_{f_0}(g_i) \leq \|g_i - f_i\| + \|f_i - f_0\| = e_C(g_i) + \|f_i - f_0\| < \inf_{g \in M} e_C(g) + \varepsilon \leq \inf_{g \in M} e_{f_0}(g) + \varepsilon$. Hence, $\{g_i\}$ is an e_{f_0}-minimizing sequence over M. By approximative compactness of M, there exists $g^o \in M$ such that $\{g_i\}$ converges subsequentially to g^o.

(ii) As $cl(C)$ is boundedly compact and closed, it is proximinal and so there exists a sequence $\{f_i\} \subseteq cl(C)$ such that for every i, $e_C(g_i) = \|f_i - g_i\|$. By the triangle inequality, $\|f_i\| \leq \|f_i - g_i\| + \|g_i\|$. Both sequences, $\{\|g_i\|\}$ and $\{\|f_i - g_i\|\}$, are bounded: the first one by boundedness of M and the second one as $\{\|f_i - g_i\|\}$ is convergent (since $\lim_{i\to\infty} \|g_i - f_i\| = \lim_{i\to\infty} e_C(g_i) = \inf_{g \in M} e_C(g)$). By closedness and bounded compactness of $cl(C)$, there exists $f_0 \in cl(C)$ to which $\{f_i\}$ converges subsequentially and so we can proceed as in the second part of the proof of (i).

(iii) As C is bounded, there exists $r > 0$ such that $C \subseteq B_r$. Let $a = \inf\{\|f - g\| : f \in C, g \in M\}$. Then there exist $i_0 \in \mathcal{N}_+$ and $b > 0$ such that for all $i \geq i_0$, $e_C(g_i) < a + b$ and so there exist $i_1 \geq i_0$, $f_i \in C$, and $b' \geq b$ such that for all $i \geq i_1$, $\|g_i - f_i\| < a + b'$. By the triangle inequality, $\|g_i\| \leq \|g_i - f_i\| + \|f_i\| < a + b' + r$. Thus for all $i \geq i_1$, $\{g_i\} \subseteq B_{a+b'+r} \cap M$ and so $\{g_i\}$ has a bounded subsequence. As M is boundedly compact and closed, this subsequence converges subsequentially to $g^o \in M$. □

The following table summarizes the conditions on M and C assumed in Theorem 3.1, which guarantee that (M, e_C) is Tychonov well-posed in the generalized sense.

Table 3.1
Conditions on M and C guaranteeing Tychonov well-posedness in the generalized sense of (M, e_C). Y = yes, N = no (by "no" we mean "there exists a counterexample").

	C precompact	C boundedly compact	C bounded
M approximatively compact	Y	N	N
M boundedly compact and closed	Y	N	Y
M approximatively compact and bounded	Y	Y	N

The first entry in the first column holds by Theorem 3.1 (i), while the other two entries in the same column hold since there the conditions on M are stronger than those required in the first entry. In the second column, Theorem 3.1 (ii) justifies the "yes" entry, while "yes" in the third column holds by Theorem 3.1 (iii). Both "no" entries in the second column are shown by the following counterexample. In the Euclidean space $\Re^2$, let C be the the x-axis and M the graph of the exponential function. Then M and C are boundedly compact and closed and hence approximatively compact. But no e_C-minimizing sequence in M has a convergent subsequence.

The "no" entries in the third column are demonstrated by the following example. Let $(l_2, \|.\|_{l_2})$ be the Hilbert space of square-summable sequences and $\{e_i\}$ be its orthonormal basis. Let L denote the orthogonal complement of the unit vector (say, e_1) and let $M = L \cap B_1(\|.\|_{l_2})$. As every closed convex subset of a uniformly convex Banach space is approximatively compact [5, p. 25], M is a bounded approximatively compact set. Let $C = w\, e_1 + M$, where w is any nonzero real number. Then C is closed and bounded. The sequence $\{e_2, e_3, ...\}$ in M satisfies for all $j \geq 2$, $\|e_j - C\| = |w|$ and so it is e_C-minimizing over M but it has no convergent subsequence.

Theorem 3.1 will be used in the next section to investigate generalized Tychonov well-posedness of (M, e_C), for admissible sets M computable by variable-basis functions and, as a particular case, by neural networks.

4. Convergence of minimizing sequences formed by variable-basis functions. Sets of functions of the form $span_n\, G = \{\sum_{i=1}^n w_i g_i :\ w_i \in \Re,\ g_i \in G\}$ and $conv_n\, G = \{\sum_{i=1}^n w_i g_i :\ w_i \in [0,1],\ \sum_{i=1}^n w_i = 1,\ g_i \in G\}$ are called *variable-basis functions* [16], [17]. Sets $span_n\, G$ model situations in which admissible functions are represented as linear combinations of any n-tuple of functions from G, with unconstrained coefficients in the linear combinations. In many applications such coefficients are constrained by a bound on a norm of the coefficients vector $(w_1, \ldots, w_n)$. When such a norm is the l_1-norm, the corresponding functions belong to the set $\{\sum_{i=1}^n w_i g_i\ :\ w_i \in \Re,\ g_i \in G, \sum_{i=1}^n |w_i| \leq c\}$, where $c > 0$ is a given bound on the l_1-norm. It is easy to see that this set is contained in $conv_n G'$, where $G' = \{rg : |r| \leq c\,,\, g \in G\}$. As any two norms on $\Re^n$ are equivalent, every norm-based constraint on the coefficients of linear combinations defines a set contained in a set of the form $conv_n\, G'$.

Depending on the choice of the set G, one can obtain a variety of admissible sets that include functions computable by neural networks, splines with free nodes, trigonometric polynomials with free frequencies, etc. For simplicity, we shall consider functions defined on $[0,1]^d$. Let $A \subseteq \Re^q$, $\phi : A \times [0,1]^d \to \Re$ be a function of two vector variables, and $G_\phi = \{\phi(a, \cdot) : a \in A\}$.

By suitable choices of A and ϕ, one can represent by G_ϕ sets of functions computable by various types of so-called *neural networks* with computational unit ϕ. If $A = S^{d-1} \times \Re$, where $S^{d-1} = \{e \in \Re^d : \|e\| = 1\}$ is the set of unit vectors in $\Re^d$, 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 $t \geq 0$, then we shall denote such a set G_ϕ by H_d, as it is the set of characteristic functions of closed half-spaces of $\Re^d$, restricted to $[0,1]^d$. Functions in H_d are called *Heaviside perceptrons* ; functions in $span_n\, H_d$ and $conv_n\, H_d$ are called *Heaviside perceptron networks.*

If $A = [-c,c]^d \times [-c,c]$ and $\phi((v,b),x) = \psi(v \cdot x + b)$, where $\psi : \Re \to \Re$ is called *activation function*, b is called *bias* and the components of v are called *weights*, then G_ϕ, denoted by $P_d(\psi, c)$, is the set of functions on $[0,1]^d$ computable by ψ-*perceptrons* with both biases and weights bounded by c. $P_d(\psi)$ denotes the corresponding set with no bounds on the parameters values. Functions in $span_n\, P_d(\psi, c)$, $conv_n\, P_d(\psi, c)$, $span_n\, P_d(\psi)$, and $conv_n\, P_d(\psi)$ are called ψ-*perceptron networks.* The most common activation functions in perceptrons are *sigmoidals*, i.e., bounded measurable functions $\sigma : \Re \to \Re$ such that $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to+\infty} \sigma(t) = 1$ (e.g., the logistic sigmoid $\sigma(t) = 1/(1 + \exp(-t))$ and the hyperbolic tangent). If the activation function ψ is positive and even, $A = [-c,c]^d \times [-c,c]$, and $\phi((v,b),x) = \psi(b\|x - v\|)$, where $\|.\|$ is a norm on $\Re^d$, b is called *width* and v is called *centroid*, then G_ϕ, denoted by $F_d(\psi, c)$, is the set of functions on $[0,1]^d$ computable by *ψ-radial-basis-functions* with both widths and centroids bounded by c (a typical activation function for RBF units is the Gaussian function $\psi(t) = e^{-t^2}$). $F_d(\psi)$ denotes the corresponding set with no bounds on the parameters values. Functions in $span_n\, F_d(\psi, c)$, $conv_n\, F_d(\psi, c)$, $span_n\, F_d(\psi)$, and $conv_n\, F_d(\psi)$ are called *ψ-radial-basis-functions (RBF) networks.* The number n of hidden units in ψ-perceptron networks and ψ-RBF networks can be considered as a measure of the network complexity, as the number of network parameters depends on n linearly.

The following proposition applies Theorem 3.1 to admissible sets computable by neural networks.

Proposition 4.1. *Let $(X, \|.\|)$ be a normed linear space and C, M its subsets. The problem (M, e_C) is Tychonov well-posed in the generalized sense if any of the following conditions holds:*
(i) C is bounded and $M = conv_n\, G_\phi$ or $M = span_n\, G_\phi$ where n is a positive integer and G_ϕ is finite-dimensional;
(ii) $(X, \|.\|) = (\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$, C is bounded and $M = conv_n P_d(\psi, c)$ or $M = conv_n F_d(\psi, c)$ where $c > 0$, ψ is bounded and continuous, and d, n are positive integers;
(iii) $(X, \|.\|) = (L_p([0,1]^d), \|.\|_p)$, $p \in [1, \infty)$, C is precompact and $M = span_n\, H_d$, or else C is bounded, $M = conv_n\, H_d$ and d, n are positive integers.

Proof. (i) If G_ϕ is finite-dimensional (e.g., if the set A of parameters of ϕ is finite), then it is straightforward that $span_n\, G_\phi$ is boundedly compact and closed. So we conclude by Theorem 3.1 (iii).

(ii) By Theorem 3.1 (iii), it is sufficient to check that in all these cases M is boundedly compact and closed. Since the convex hull of a compact set G is compact and $conv_n\, G$ is closed

in $conv\, G$, compactness of $M = conv_n\, G$ follows from compactness of G. For $G = P_d(\psi, c)$ and $G = F_d(\psi, c)$ with $c > 0$ and ψ bounded and continuous, compactness of $conv_n\, G$ in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ has been proved in [12].

(iii) If C is precompact and $M = span_n\, H_d$, then by Theorem 3.1 (i) it is sufficient to check that M is approximatively compact. Approximative compactness of $M = span_n\, H_d$ in $(L_p([0,1]^d), \|.\|_p)$, $p \in [1, \infty)$, was shown in [11]. If C is bounded and $M = conv_n\, H_d$, then by Theorem 3.1 (iii) it is sufficient to prove that M is boundedly compact and closed. Compactness of $G = H_d$ in $(L_2([0,1]^d), \|.\|_2)$ was proved in [9] and inspection of the argument shows that it also holds for L_p-spaces with $p \in [1, \infty)$. Since the convex hull of a compact set G is compact and $conv_n\, G$ is closed in $conv\, G$, compactness of $conv_n\, H_d$ follows from compactness of H_d. □

Note that for neural networks with differentiable hidden unit functions (e.g., perceptrons with logistic sigmoid or RBF with the Gaussian activation function) the sets $span_n G_\phi$ are not approximatively compact in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ or in $(L_p([0,1]^d), \|.\|_p)$, because they are not even closed (it was shown in [23] for perceptron networks and the arguments used there can be extended to Gaussian RBF networks).

Theorem 3.1 can be combined with various conditions guaranteeing precompactness of the target set C, such as interpolation and smoothness conditions, which model neural network learning from data described by input/output pairs and constraints given by physical considerations or feasibility of implementation. The following proposition establishes precompactness of such target sets.

Proposition 4.2. *Let d, n, k be positive integers and C be a set of continuous*

functions defined on $[0,1]^d$, satisfying the following two conditions:
1) (smoothness) there exists $b > 0$ such that on $(0,1)^d$ all first order partial derivatives of all elements of C are continuous and bounded by b in absolute value;
2) (interpolation) there exist closed intervals $X_j \subset [0,1]^d$ and $Y_j \subset \Re$, $j = 1, \dots, k$ such that some Y_j is bounded and for all $j = 1, \dots, k$, $f(X_j) \subseteq Y_j$.
Then for every $c > 0$ and ψ bounded and continuous, $(conv_n P_d(\psi, c), e_C)$ and $(conv_n F_d(\psi, c), e_C)$ are Tychonov well-posed in the generalized sense in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ and $(conv_n\, H_d, e_C)$ and $(span_n\, H_d, e_C)$ are Tychonov well-posed in the generalized sense in $(L_p([0,1]^d), \|.\|_p), p \in [1, \infty)$.

Proof. Since precompactness in the space $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ implies precompactness in $(L_p([0,1]^d), \|.\|_p)$, $p \in [1, \infty)$, it is sufficient to check that C satisfies the assumptions of the Ascoli-Arzelá Theorem [1, Theorem 1.30], i.e., elements of C are equibounded and equicontinuous on $(0,1)^d$. Equicontinuity follows from the Mean Value Theorem [6, p. 79] and Cauchy-Schwarz inequality, which imply that for all $f \in C$, all $x \in (0,1)^d$, and all h such that for every $t \in [0,1]$, $x + th \in (0,1)^d$, there exists $\tau \in (0,1)$ such that $|f(x+th) - f(x)| = |\nabla f(x+\tau h) \cdot h| \leq \|\nabla f(x+\tau h)\|\, \|h\| \leq b\sqrt{d}\, \|h\|$. Taking j such that Y_j is bounded and $a > 0$ such that $Y_j \subseteq [-a, a]$, and choosing $x_j \in X_j$ we apply the inequality just derived. Thus for every $f \in C$ and every $x \in (0,1)^d$ we have $|f(x) - f(x_j)| \leq b\sqrt{d}\, \|x - x_j\| \leq b\, d$. Hence, $f(x) \in [-a - b\, d, a + b\, d]$ and so functions in C are equibounded on $(0,1)^d$. Thus C is precompact in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ and the statements follow from Proposition 4.1 (ii) and (iii). □

Note that precompactness in $(L_p([0,1]^d), \|.\|_p)$ can also be derived using L_p versions of Ascoli-Arzelá theorem (see, e.g., [1, Th. 2.21]). Note that the conditions of smoothness and interpolation required by Proposition 4.2 may be incompatible, i.e., C could be empty. In this case, one must either increase the size of the intervals Y_j or increase the bound on the derivatives. Alternatively, some interval constraints should be discarded.

5. Rates of decrease of infima with increasing complexity of admissible sets of variable-basis functions. In applications, the rate of decrease of infima of an error functional over $conv_n\, G$ and $span_n\, G$ should be fast enough to achieve a desirable accuracy for small values of n, such that admissible functions have a moderate complexity. We shall derive estimates of such rates using a result from approximation theory by Maurey [24], Jones [10], and Barron [2]. Here we shall use its reformulation in terms of a norm tailored to a given basis G. Such a norm, called *G-variation* and denoted by $\|.\|_G$, was introduced in [13] for a subset G of a normed linear space $(X, \|.\|)$ as the Minkowski functional of the set $cl\, conv\, (G \cup -G)$. Thus, $\|f\|_G = \inf\left\{c > 0 : \ c^{-1} f \in cl\, conv\, (G \cup -G)\right\}$. G-variation is a norm on the subspace $\{f \in X : \|f\|_G < \infty\} \subseteq X$; for its properties see [15], [17] and [18]. In [16] and [18] it has been shown that when G is an orthonormal basis of a separable Hilbert space, G-variation is equal to the *l_1-norm with respect to G*, defined for $f \in X$ as $\|f\|_{1,G} = \sum_{g \in G} |f \cdot g|$. For $t > 0$, we define $G(t) = \{wg : g \in G, w \in \Re, |w| \leq t\}$.

The following theorem reformulates in terms of G-variation Maurey-Jones-Barron's theorem [24], [10], [2] and its extension to L_p-spaces [7].

Theorem 5.1. *Let $(X, \|.\|)$ be a normed linear space, G its bounded subset and $s_G = \sup_{g \in G} \|g\|$. For every $f \in X$ and every positive integer n, the following hold:*
(i) if $(X, \|.\|)$ is a Hilbert space, then

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

(ii) if $(X, \|.\|) = (L_p([0,1]^d), \|.\|_p)$*,* $p \in (1, \infty)$*, then*

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

where $q = p/(p-1)$*,* $\bar{p} = \min(p,q)$*, and* $\bar{q} = \max(p,q)$*;*
(iii) if $(X, \|.\|)$ *is a separable Hilbert space and* G *its orthonormal basis, then*

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

For the proof of Theorem 5.1 (i) and (ii) see [13] and [14], resp.; for the proof of Theorem 5.1 (iii) see [18, Theorem 2.7] and [16, Theorem 3].

As a corollary of Theorem 5.1, we obtain the following upper bounds on rates of decrease of infima of error functionals over $span_n\, G$, with n increasing.

Corollary 5.2. *Let* $(X, \|.\|)$ *be a normed linear space and* G*,* C *its subsets such that* $r = \inf_{f \in C} \|f\|_G$ *and* $s_G = \sup_{g \in G} \|g\|$ *are finite. For every* $f \in X$ *and every positive integer* n*, the following hold:*
(i) if $(X, \|.\|)$ *is a Hilbert space, then*

$$\inf_{g \in span_n G} e_C(g) \leq \inf_{g \in conv_n\, G(r)} e_C(g) \leq \frac{r}{\sqrt{n}} s_G;$$

(ii) if $(X, \|.\|) = (L_p([0,1]^d), \|.\|_p)$*,* $p \in (1, \infty)$*, then*

$$\inf_{g \in span_n G} e_C(g) \leq \inf_{g \in conv_n\, G(r)} e_C(g) \leq \frac{r\, 2^{1/\bar{p}+1}}{n^{1/\bar{q}}} s_G.$$

(iii) if $(X, \|.\|)$ *is a separable Hilbert space and* G *is its orthonormal basis, then*

$$\inf_{g \in span_n G} e_C(g) \leq \inf_{g \in conv_n\, G(r)} e_C(g) \leq \frac{r}{2\sqrt{n}} s_G.$$

Proof. `Shortened proof to be checked` (i) For each $t > r$, choose $f_t \in C$ such that $r \leq \|f_t\|_G < t$. By Theorem 5.1 (i), for every n we have $\|f_t - conv_n\, G(t)\| \leq t\, s_G/\sqrt{n}$. Thus, $\inf_{g \in conv_n G(t)} e_C(g) = \inf_{g \in conv_n G(t)} \inf_{f \in C} \|g - f\| \leq \inf_{g \in conv_n\, G(t)} \|g - f_t\| = \|f_t - conv_n\, G(t)\| \leq t\, s_G/\sqrt{n}$. Since $conv_n G(r) = \bigcap \{conv_n G(t) : t > r\}$, we obtain $\inf_{g \,\in\, span_n G} e_C(g) \leq \inf_{g \in conv_n\, G(r)} e_C(g) \leq r\, s_G/\sqrt{n}$.

(ii) and (iii) are proved analogously to (i) using Theorem 5.1 (ii) and (iii), resp. □

When applied to spaces of functions of d variables, the bounds from Theorem 5.1 and Corollary 5.2 show that for functions in balls of fixed radii in G-variation the curse of dimensionality does not occur. However, the shape of such balls may depend on the number of variables [14], [17], [18].

The following proposition applies Corollary 5.2 to admissible functions computable by Heaviside perceptron networks and target sets containing a sufficiently smooth function. The proof exploits the possibility of embedding balls in certain Sobolev norms into balls of proper radii in H_d-variation. For a set S of functions $f : \mathcal{D} \to \Re$, $\mathcal{D} \subseteq \Re^d$, and a set $\Omega \subseteq \Re^d$, $S_{|\Omega}$ denotes the set of functions whose elements are restrictions to Ω of functions in S. For the sake of clarity, in the following we shall write the norm in the Sobolev spaces $W_2^s(\Re^d)$ and $W_2^s(\Omega)$ specifying also the domain of their functions, i.e., we shall write $\|.\|_{2,s,\Re^d}$ and $\|.\|_{2,s,\Omega}$.

Proposition 5.3. *Let d, s be positive integers, $s \geq \lfloor d/2 \rfloor + 2$, $\Omega \subset [0,1]^d$ be an open ball in $l_2(\Re^d)$, $C \subset (L_2([0,1]^d), \|.\|_2)$, $a = \inf\{a' > 0 : C_{|\Omega} \cap B_{a'}(\|.\|_{2,s,\Omega}) \neq \emptyset\}$, and $b = \left(\int_{\Re^d}(1+\|\omega\|^{2(s-1)})^{-1}\, d\omega\right)^{1/2}$. Then there exists $c > 0$ depending on s and Ω such that in $(L_2(\Omega), \|.\|_2)$, for $r = 2\,a\,b\,c$ and every positive integer n,*

(i) $\inf_{g \in span_n H_d} e_C(g) \leq \inf_{g \in conv_n\, H_d(r)} e_C(g) \leq \frac{r}{\sqrt{n}}$;

(ii) if C is precompact, then $(span_n\, H_d, e_C)$ is Tychonov well-posed in the generalized sense and $\min_{g \in span_n\, H_d} e_C(g) \leq \min_{g \in conv_n\, H_d(r)} e_C(g) \leq \frac{r}{\sqrt{n}}$.

Proof. (i) Let $B_r(\|.\|_{2,s,\Re^d})_{|\Omega}$ and $B_r(\|.\|_{H_d})_{|\Omega}$ denote the balls of radii r in the Sobolev norm $\|.\|_{2,s,\Re^d}$ and in H_d-variation, whose functions are the restrictions to Ω of functions in $B_r(\|.\|_{2,s,\Re^d})$ and $B_r(\|.\|_{H_d})$, resp.

Using the technique exploited in [2, pp. 935, 941], one obtains for every $\rho > 0$, $B_\rho(\|.\|_{2,s,\Re^d})_{|\Omega} \subseteq B_{2\,\rho\,b}(\|.\|_{H_d})_{|\Omega}$, where $b = \left(\int_{\Re^d}(1+\|\omega\|^{2(s-1)})^{-1}\, d\omega\right)^{1/2}$ is finite as $2(s-1) > d$.

By [1, 4.24-4.29] there exists an extension operator $\mathcal{P} : (W_2^s(\Omega), \|.\|_{2,s,\Omega}) \to (W_2^s(\Re^d), \|.\|_{2,s,\Re^d})$ such that for all $f \in (W_2^s(\Omega), \|.\|_{2,s,\Omega})$, $(\mathcal{P}\,f)_{|\Omega} = f$ a.e. in Ω and $\|\mathcal{P}\,f\|_{2,s,\Re^d} \leq c\,\|f\|_{2,s,\Omega}$, where c is a constant depending on s and Ω. Since by hypothesis $C_{|\Omega} \bigcap B_a(\|.\|_{2,s,\Omega}) \neq \emptyset$, there exists $f \in C_{|\Omega}$ such that $\mathcal{P}f \in B_{a\,c}(\|.\|_{2,s,\Re^d})$. As for every $\rho > 0$, $B_\rho(\|.\|_{2,s,\Re^d})_{|\Omega} \subseteq B_{2\,\rho\,b}(\|.\|_{H_d})_{|\Omega}$, taking $\rho = ac$ we have $f \in B_{2\,a\,b\,c}(\|.\|_{H_d})_{|\Omega}$. Since in $(L_2(\Omega), \|.\|_2)$ we have $s_{H_d} \leq 1$, the statement follows from by Corollary 5.2 (i) with $r = 2abc$.

(ii) follows from (i) and Proposition 4.1 (iii). □

Proposition 5.3 extends the existential statement from Proposition 4.1 (iii) to a quantitative result: it gives an upper bound on $\min_{g \in span_n\, H_d} e_C(g)$ formulated in terms of the smallest Sobolev norm of elements of the target set C. As for any continuous non-decreasing sigmoidal function σ $P_d(\sigma)$-variation is equal to H_d-variation [15], the same estimate as in Proposition 5.3 (i) holds for $(span_n\, P_d(\sigma), e_C)$ and $(conv_n\, P_d(\sigma)(r), e_C)$ for any such sigmoidal function.

In the following we apply Corollary 5.2 to admissible sets of Boolean functions in $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$. We shall give conditions on target sets C, which guarantee rates of minimization of e_C of order $\mathcal{O}(1/\sqrt{n})$ for any number of variables d, for admissible sets of functions in $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ computable by perceptron neural networks with the signum activation function, defined as sgn(t) $= -1$ for $t < 0$ and sgn(t) $= 1$ for $t \geq 0$. $\bar{H}_d$ denotes the set of functions on $\{0,1\}^d$ computable by signum perceptrons, i.e., $\bar{H}_d = \{f : \{0,1\}^d \to \Re : f(x) = \mathrm{sgn}(v \cdot x + b),\ v \in \Re^d,\ b \in \Re\}$. We estimate variation with respect to signum perceptrons using variation with respect to the *Fourier orthonormal basis* defined as $F_d = \left\{f_u : u \in \{0,1\}^d,\ f_u(x) = \frac{1}{\sqrt{2^d}}(-1)^{u \cdot x}\right\}$ [27]. Every real-valued Boolean function 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 the Fourier coefficients $\hat{f}(u)$ are given by $\hat{f}(u) = \frac{1}{\sqrt{2^d}} \sum_{x \in \{0,1\}^d} f(x)(-1)^{u \cdot x}$. 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. The l_1-norm with respect to the Fourier basis, defined as $\|f\|_{1,F_d} = \|\hat{f}\|_{l_1} = \sum_{u \in \{0,1\}^d} |\hat{f}(u)|$, is called the *spectral norm*.

Next proposition gives an upper bound on the rate of decrease of infima of error functionals over perceptron neural networks, in terms of the smallest spectral norm of elements of the target set C.

Proposition 5.4. *Let d be a positive integer, $r > 0$, and C be a bounded subset of $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$, $a = \inf\{a' > 0 : C \cap B_{a'}(\|.\|_{1,F_d}) \neq \emptyset\}$. For every positive integer*

n, the problems $(span_{dn+1}\,\bar{H}_d, e_C)$ and $(conv_{dn+1}\,\bar{H}_d(r), e_C)$ are Tychonov well-posed in the generalized sense and $\min_{g \in span_{dn+1}\,\bar{H}_d} e_C(g) \leq \min_{g \in conv_{dn+1}\,\bar{H}_d(a)} e_C(g) \leq \frac{a}{2\sqrt{n}}$.

Proof. It is easy to verify that every function of the Fourier basis F_d can be expressed as a linear combination of at most $d+1$ signum perceptrons [18]. Indeed, for every $u, x \in \{0,1\}^d$ one has $(-1)^{u \cdot x} = \frac{1+(-1)^d}{2} + \sum_{j=1}^{d}(-1)^j \mathrm{sgn}(u \cdot x - j + \frac{1}{2})$. Moreover, any linear combination of n elements of F_d belongs to $span_{dn+1}\bar{H}_d$, since all of the n occurrences of the constant function can be expressed by a single perceptron. As for any orthonormal basis of a separable Hilbert space G-variation is equal to l_1-norm with respect to G [16], [18], we have $\|f\|_{F_d} = \|f\|_{1,F_d}$ and the statement follows from Proposition 4.1 (i) and Corollary 5.2 (ii). □

According to Proposition 5.4, rates of minimization of order $\mathcal{O}(1/\sqrt{n})$ independent on the number d of variables, are guaranteed when target sets contain a function with "small" spectral norm. Next two propositions describe target sets for which minimization of error functionals over admissible sets computable by Boolean signum perceptron networks does not exhibit the curse of dimensionality. The first result considers target sets whose elements can be expressed as linear combinations of a "small" number of generalized parities.

Proposition 5.5. *Let d, n, and m be positive integers, $m \leq 2^d$, $c > 0$, and C be a subset of $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ such that C contains a function f with at most m Fourier coefficients nonzero and with $\|f\|_{l_2} \leq c$. The problems $(span_{dn+1}\,\bar{H}_d, e_C)$ and $(conv_{dn+1}\,\bar{H}_d(\sqrt{m}), e_C)$ are Tychonov well-posed in the generalized sense and* $\min_{g \in span_{dn+1}\,\bar{H}_d} e_C(g) \leq \min_{g \in conv_{dn+1}\,\bar{H}_d(\sqrt{m})} e_C(g) \leq \frac{c}{2}\sqrt{\frac{m}{n}}$.

Proof. Let $f \in C$ be such that $f = \sum_{i=1}^{m} w_i g_i$, where $g_i \in F_d$ are the Fourier coefficients. Then $\|f\|_{F_d} = \|f\|_{1,F_d} = \|\hat{f}\|_{l_1} = \sum_{i=1}^{m}|w_i|$. By the Cauchy-Schwarz inequality $\sum_{i=1}^{m}|w_i| \leq \|w\|_2\|u\|_2$, where $w = (w_1, \ldots, w_m)$ and $u = (u_1, \ldots, u_m)$, with $u_i = \mathrm{sgn}(w_i)$. As $\|w\|_2 = \|f\|_{l_2} \leq c$ and $\|u\|_2 \leq \sqrt{m}$, we have $\|f\|_{1,F_d} \leq c\sqrt{m}$. Thus C contains a function f with $\|f\|_{1,F_d} \leq c\sqrt{m}$, so $a = \inf\{a' > 0 : C \cap B_{a'}(\|.\|_{1,F_d}) \neq \emptyset\} \leq c\sqrt{m}$ and the statement follows by Proposition 5.4. □

For C satisfying the assumptions of Proposition 5.5, if e_C is minimized over the set of d-variable Boolean functions computable by networks with $dn+1$ signum perceptrons, where $n \geq \frac{c^2 m}{4\varepsilon^2}$, then the minimum is bounded from above by ε. As the number $d\frac{c^2 m}{4\varepsilon^2} + 1$ of perceptrons needed for an accuracy ε grows with d linearly, the curse of dimensionality is avoided.

An interesting class of target sets, for which minimization of error functionals can be efficiently performed over sets of functions computable by a "moderate" number of Boolean signum perceptrons, are functions representable by "small" decision trees. Such trees play an important role in machine learning [19].

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 $f : \{0,1\}^d \to \Re$ is representable by a decision tree if there exists such a tree with internal nodes labeled by variables $x_1, \ldots, x_d$, all pairs of edges outgoing from a node are labeled by 0s and 1s, and all leaves are labeled by real numbers, so that f can be computed 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 $f(x_1, \ldots, x_d)$.

Proposition 5.6. *Let d, s be positive integers, $b \geq 0$, and C be a subset of*

$(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ *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* $\frac{\max_{x\in\{0,1\}^d}|f(x)|}{\min_{x\in\{0,1\}^d}|f(x)|}\|f\|_{l_2} \leq b$. *Then the problems* $(span_{dn+1}\, \bar{H}_d, e_C)$ *and* $(conv_{dn+1}\, \bar{H}_d(sb), e_C)$ *are Tychonov well-posed in the generalized sense and* $\min_{g\in span_{dn+1}\, \bar{H}_d} e_C(g) \leq \min_{g\in conv_{dn+1}\, \bar{H}_d(sb)} e_C(g) \leq \frac{sb}{2\sqrt{n}}$.

Proof. By [18, Theorem 3.4] (which extends [19, Lemma 5.1]), the hypotheses imply that $\frac{\|\hat{f}\|_{l_1}}{\|f\|_{l_2}} \leq s\frac{\max_{x\in\{0,1\}^d}|f(x)|}{\min_{x\in\{0,1\}^d}|f(x)|}$, so we get $\|f\|_{1,F_d} = \|\hat{f}\|_{l_1} \leq sb$. Thus C contains a function f with $\|f\|_{1,F_d} \leq sb$, so $a = \inf\{a' > 0 : C \cap B_{a'}(\|.\|_{1,F_d}) \neq \emptyset\} \leq sb$ and the statement follows from Proposition 5.4. □

For C satisfying the assumptions of Proposition 5.6, if e_C is minimized over the set of d-variable Boolean functions computable by networks with $dn + 1$ signum perceptrons, where $n \geq \left(\frac{sb}{2\varepsilon}\right)^2$, then the minimum is bounded from above by ε. As the number $d\left(\frac{sb}{2\varepsilon}\right)^2 + 1$ of perceptrons needed for an accuracy ε grows with d linearly, the curse of dimensionality is avoided.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. on Information Theory, 39 (1993), pp. 930–945.

[3] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

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

[5] D. Braess, *Nonlinear Approximation Theory*, Springer-Verlag, Berlin Heidelberg, 1986.

[6] R. Courant, *Differential and Integral Calculus*, Vol. II, Wiley, 1988.

[7] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, *Rates of approximation results motivated by robust neural network learning*, Proc. 6th Annual ACM Conference on Computational Learning Theory (1993), The Association for Computing Machinery, New York, pp. 303-309.

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

[9] L. Gurvits and P. Koiran, *Approximation and learning of convex superpositions*, J. of Computer and System Sciences, 55 (1997), pp. 161–170.

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

[11] P. C. Kainen, V. Kůrková, and A. Vogt, *Best approximation by linear combinations of characteristic functions of half-spaces*, J. of Approximation Theory, (2003), to appear.

[12] V. Kůrková, *Approximation of functions by perceptron networks with bounded number of hidden units*, Neural Networks, 8 (1995), pp. 745–750.

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

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

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

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

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

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

[19] E. Kushilevicz and Y. Mansour, *Learning decision trees using the Fourier spectrum*, SIAM J. Comput., 22 (1993), pp. 1331-1348.

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

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

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

[23] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation can approximate any function*, Neural Networks, 6 (1993), pp. 861–867.

[24] G. Pisier, *Remarques sur un resultat non publié de B. Maurey*, in Séminaire d'Analyse Fonctionelle 1980-81, vol. I, no. 12, École Polytechnique, Centre de Mathématiques, Palaiseau.

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

[26] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer-Verlag, Berlin, 1970.

[27] H. J. Weaver, *Applications of Discrete and Continuous Fourier Analysis*, Wiley, New York, 1983.

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

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