Minimization of Error Functionals over Variable-Basis Functions

Kainen, P.C.
2002

Institute of Computer Science
Academy of Sciences of the Czech Republic

Minimization of error functionals over variable-basis functions

Paul C. Kainen, Věra Kůrková, and Marcello Sanguineti

Technical report No. 864

October 2002

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: (+4202) 6605 1111, fax: (+4202) 86 58 57 89,

e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Minimization of error functionals over variable-basis functions[1]

Paul C. Kainen[2], Věra Kůrková[3], and Marcello Sanguineti[4]

Technical report No. 864

October 2002

Abstract:

Generalized Tychonov well-posedness is studied for minimization of error functionals defined by distance to a target set over admissible sets formed by variable-basis functions, which include neural networks. Rates of decrease of infima of such problems with increasing complexity of admissible sets are estimated. Upper bounds are derived on such rates that do not exibit the curse of dimensionality with respect to the number of variables of admissible functions.

[1] Collaboration between P.C.K. and V.K. was partially supported by NAS COBASE grant, between V. K. and M. S. by the Scientific Agreement Italy-Czech Republic, Project 22 - MC6. P.C.K. was partially supported by an international travel grant from Georgetown University, V. K. by GA ČR grant 201/00/1482, and M. S. by a CNR-Agenzia 2000 grant.

[2] Department of Mathematics, Georgetown University, Washington, D.C. 20057-1233, USA – E-mail: kainen@georgetown.edu

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

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

Keywords:

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

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

1 Introduction

Functionals expressed as distances from sets of target functions are called *error functionals*. Minimization of error functionals occurs in optimization tasks arising in a variety of areas, such as system identification, machine learning, pattern recognition, etc. ([9], [20], [21]).

In the last decades, neural networks have become a popular type of admissible sets and a widespread tool for approximate optimization ([5], [25], [27]). Such networks can be studied in the more general context of "variable-basis functions" ([17], [18]), which are linear combinations of a fixed number of functions from a basis without a prespecified ordering. The variable-basis scheme includes free-nodes splines and trigonometric polynomials with free frequencies (see the references in [18]).

For high-dimensional tasks, the implementation of approximate optimization procedures may become infeasible due to an unmanageably large number of parameters ([22, pp. 232-233], [27]). In particular, such tasks may be limited by the "curse of dimensionality" [4], i.e., an exponentially fast growth of the number of parameters of admissible functions with the number of their variables. Nevertheless, experience has shown that some neural networks of moderate complexity (which allows simulation on classical computers) perform quite well in some tasks depending on hundreds of variables ([5], [25], [27]).

In this paper, we investigate generalized Tychonov well-posedness of the problem 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 through suitable interpolation and smoothness conditions and by admissible sets consisting of functions computable by families of variable-basis functions, including commonly used classes of neural networks. Furthermore, we estimate rates of decrease of infima of error functionals over neural networks with increasing numbers of computational units and derive upper bounds on such rates. The bounds 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 generalized Tychonov well-posedness of minimization of error functionals. 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, $(X, \|.\|)$ denotes a normed linear space; we write only X when it is clear which norm is used. For a positive integer d, $\mathcal{R}$ the set of real numbers and $\Omega \subseteq \mathcal{R}^d$, by $(L_p(\Omega), \|.\|_p)$ is we denoted the space of measurable, p-th integrable real-valued functions on Ω with the L_p norm and by $(\mathcal{C}(\Omega), \|.\|_{\mathcal{C}})$ the space of real-valued continuous functions on Ω with the supremum norm.

For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$, let $D^\alpha = D_1^{\alpha_1} \ldots D_d^{\alpha_d}$ denote the distributional derivative [2, 1.57]. The Sobolev space $(W_p^m(\Omega), \|.\|_{m,p}^p)$ is the set of all functions $f : \Omega \to \mathcal{R}$ for which $D^\alpha f \in (L_p(\Omega), \|.\|_p)$ for $0 \leq |\alpha| \leq m$, with the norm $\|f\|_{m,p}^p = \sum_{0 \leq |\alpha| \leq m} \|D^\alpha f\|_p^p$. $\mathcal{B}(\{0,1\}^d)$ denotes the space of real-valued Boolean functions, i.e., functions from $\{0,1\}^d$ to $\mathcal{R}$. 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 $\|f\|_{l_2} = \sqrt{f \cdot f}$. The space $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ is isomorphic to the 2^d-dimensional Euclidean space $\mathcal{R}^{2^d}$ with the l_2-norm.

For $M \subseteq X$, $cl(M)$ denotes the closure of M in the norm $\|.\|$ and for $f \in X$, we let $\|f - M\| = \inf_{g \in M} \|f - g\|$. 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 when it is clear which norm is used, we write only B_r.

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 $\Phi : M \to \mathcal{R}$ 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 problem (M, Φ) is *Tychonov well-posed in the generalized sense* [8, p. 24] if each minimizing sequence converges subsequentially to an element of M. 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 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 $e_{\{h\}}$.

Recall that a nonempty subset M of a normed linear space is *compact* if every sequence has a convergent subsequence, M is *precompact* if $cl(M)$ is compact, and M 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. 383] if, for all $h \in X$, every sequence in M which minimizes the distance to h converges subsequentially to an element of M. 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 $\|h - M\| = \|h - g\|$. Proximinal implies closed since an element in the closure of a set has zero distance from such a set and so it must be equal to its best approximation.

3 Minimization of error functionals under weakened compactness

Generalized Tychonov well-posedness can be interpreted as one of the many types 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.

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 $cl(C)$ 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. As $e_C = e_{cl(C)}$, it is sufficient to prove (i) and (ii) for $cl(C)$. In both cases, we shall show that $\{g_i\}$ has a subsequence that for some $f_0 \in cl(C)$ is e_{f_0}-minimizing over M and conclude by approximative compactness of M.

(i) As $cl(C)$ is compact, it is proximinal and so for each i there exists $f_i \in cl(C)$ satisfying $e_C(g_i) = \|g_i - f_i\|$. 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 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 $i \geq i_1$, $e_{f_0}(g_i) \leq \|g_i - f_i\| + \|f_i - f_0\| < \inf_{g \in M} e_C(g) + \varepsilon \leq \inf_{g \in M} e_{f_0}(g) + \varepsilon$. Hence, $\{g_i\}$ is an e_{f_0}-minimizing sequence over M.

(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 $\lim_{i \to \infty} \|g_i - f_i\| = e_C(g_i) = \inf_{g \in M} e_C(g)$). 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 last 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 exists i_0 such that for all $i \geq i_0$, $e_C(g_i) < a + 1$ and so there exists $f_i \in C$ such that $\|g_i - f_i\| < a + 1$. By the triangle inequality, $\|g_i\| \leq \|g_i - f_i\| + \|f_i\| < a + 1 + r$. Thus for all $i \geq i_0$, $\{g_i\} \subseteq B_{r+a+1} \cap M$ and so $\{g_i\}$ has a bounded subsequence. As M is boundedly compact and closed, this subsequences converges subsequentially to some $g_0 \in M$. By continuity of e_C [26, p. 391], $\inf_{g \in M} e_C(g) = \lim_{i \to \infty} e_C(g_i) = e_C(g_0)$. □

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.

	C precompact	cl(C) boundedly compact	C bounded
M approximatively compact	Y	N	N
M approximatively compact and bounded	Y	Y	N
M boundedly compact and closed	Y	N	Y

Y = yes, N = no (by "no" we mean "there exists a counterexample"). The first entry in the first column holds by Theorem 3.1 (i), while the others in the same column hold since the conditions on M are stronger than those required there. 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 $\mathcal{R}^2$, 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. This example also contradicts the claim made in [26, p. 385, Theorem 2.3] that for two closed, boundedly compact subsets, there must exist a point in each of them such that the distance between these two points is equal to the gap between the two sets. 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. The orthogonal complement L of the unit vector (say, e_1) is approximatively compact [1, p. 23]. Then $M = L \cap B_1(\|.\|_{l_2})$ is approximatively compact and bounded. 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 computatble by variable-basis functions and neural networks.

4 Generalized Tychonov well-posedness of minimization of error functionals over variable-basis functions

Families of functions of the form $span_n\, G = \{\sum_{i=1}^n w_i g_i : \ w_i \in \mathcal{R}, \ g_i \in G\}$ 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.* Sets $span_n\, G$ model situations in which admissible functions are represented as linear combinations of any n 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 $\{\sum_{i=1}^n w_i g_i : \ w_i \in \mathcal{R}, \ g_i \in G, \sum_{i=1}^n |w_i| \leq c\}$, where $c > 0$ is the 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 $\mathcal{R}^n$ are equivalent, any norm-based constraint on the coefficients of linear combinations defines a set contained in a set of the form $conv_n\, G'$.

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 $A \subseteq \mathcal{R}^q$, $\phi : A \times [0,1]^d \to \mathcal{R}$ 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 computational units. 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 $t \geq 0$, then we shall denote such G_ϕ by H_d, as it the set of characteristic functions of closed half-spaces of $\mathcal{R}^d$, restricted to $[0,1]^d$.

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, denoted by $P_d(\psi, c)$, is the set of functions on $[0, 1]^d$ computable by *ψ-perceptrons* with both weights v and biases b bounded by c. $P_d(\psi)$ denotes the corresponding set with no bounds on the parameters values. The most common activation functions are *sigmoidals*, i.e., bounded measurable functions $\sigma : \mathcal{R} \to \mathcal{R}$ such that $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to+\infty} \sigma(t) = 1$ (e.g., the logistic sigmoid $\sigma(t) = 1/(1 + \exp(-t))$ or the hyperbolic tangent). 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 $\|.\|$ is a norm on $\mathcal{R}^d$, then G_ϕ, denoted by $F_d(\psi, c)$, is the set of functions on $[0, 1]^d$ 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 $\psi(t) = e^{-t^2}$). $F_d(\psi)$ denotes the corresponding set with no bounds on the parameters values.

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 apply:*

(i) C is bounded and $M = span_n\, G_\phi$ with G_ϕ finite-dimensional for any positive integer n;

(ii) $(X, \|.\|) = (\mathcal{C}([0, 1]^d), \|.\|_\mathcal{C})$, C is bounded and either $M = conv_n P_d(\psi, c)$ or $M = conv_n F_d(\psi, c)$ with $c > 0$ and ψ continuous, for any positive integers d and n ;

(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 and $M = conv_n\, H_d$ for any positive integers d and n.

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 continuous, compactness in $(\mathcal{C}([0, 1]^d), \|.\|_\mathcal{C})$ was proved in [13].

(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. It was shown in [12, Theorem 3.1] that $M = span_n\, H_d$ is approximatively compact in $(L_p([0,1]^d), \|.\|_p)$, $p \in [1,\infty)$. 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 $(L_p([0,1]^d), \|.\|_p)$, $p \in [1,\infty)$, was proved in [10]. 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 Gaussian) 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 [23].

Theorem 3.1 can be combined with various conditions guaranteeing precompactness of C, such as an interpolation and a smoothness conditions, which can model neural networks learning from data described by input/output pairs of intervals (in L_p-spaces, such intervals should have sufficiently large measure) 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, $b > 0$, and suppose that C is the family of all continuous real-valued functions on $[0,1]^d$ which satisfy the following two conditions:*

1) (smoothness) On $(0,1)^d$ all first order partial derivatives are continuous and bounded by b in absolute value.

2) (interpolation) There are given closed intervals $X_j \subset [0,1]^d$ and $Y_j \subset \mathcal{R}$ with some Y_j bounded and for all $j = 1,\ldots,k$, $f(X_j) \subseteq Y_j$.

Then, $(conv_n P_d(\psi, c), e_C)$ and $(conv_n F_d(\psi, c), e_C)$ are Tychonov well-posed in the generalized sense in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ for every $c > 0$ and ψ continuous and $(span_n\, H_d, e_C)$ is Tychonov well-posed in the generalized sense in $(L_p([0,1]^d), \|.\|_p), p \in [1,\infty)$.

Proof. We shall prove that C is precompact in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$, which implies precompactness in $(L_p([0,1]^d), \|.\|_p)$, $p \in [1,\infty)$. To this end, it is sufficient to check that C satisfies the conditions of the Ascoli-Arzelá theorem [2, Theorem 1.30], which states that $C \subseteq (\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ is precompact if it is equibounded and equicontinuous on $(0,1)^d$. Equicontinuity follows from the Mean Value Theorem and Cauchy-Schwarz inequality, which imply that for all $f \in C$ and $x, y \in (0,1)^d$, $|f(x) - f(y)| = |\nabla f(z) \cdot (x-y)| \leq \|\nabla f(z)\| \, \|x-y\| \leq b\sqrt{d}\, \|x-y\|$. Let Y_j be bounded, i.e., there exists $a > 0$ such that $Y_j \subseteq [-a, a]$. Choose some $x_j \in X_j$. Applying the inequality just derived, 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 C is equibounded. Thus C is precompact in $(\mathcal{C}([0,1]^d), \|.\|_{\mathcal{C}})$ and the statements follow from Proposition 4.1 (ii) and (iii). □

For $C \subset (L_p([0,1]^d), \|.\|_p)$, precompactness results in $(L_p([0,1]^d), \|.\|_p)$ can be derived using L_p versions of Ascoli-Arzelá theorem (see, e.g., [2, Th. 2.21]). Note that the conditions of smoothness and interpolation required by Propositions 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 could 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 n for which admissible functions have a moderate complexity. We shall derive estimates of such rates using a result from approximation theory by Maurey [24], Jones [11], and Barron [3]. 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 [14] 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 [18] and [19]. In [17] and [19] 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 denote $G(t) = \{wg : g \in G, w \in \mathcal{R}, |w| \leq t\}$.

The following theorem gives a reformulation of Maurey-Jones-Barron's theorem and its extension to L_p-spaces from [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 ist 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 (i) see [14], for the proof of (ii) see [15], and for the proof of (iii) see [19, Theorem 2.7] and [17, Theorem 3].

As a corollary of Theorem 5.1, we get the following upper bounds on rates of decrease of infima of e_C 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 separable 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. (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 $\|f_t - conv_n\, G(t)\| \leq t\, s_G/\sqrt{n}$ and so there exists a sequence $\{g_{t,i}\} \subset conv_n\, G(t)$ such that $\|f_t - conv_n G(r)\| = \lim_{i\to\infty} \|f_t - g_{t,i}\| \leq t\, s_G/\sqrt{n}$. As $f_t \in C$, we have $e_C(g_{t,i}) \leq e_{f_t}(g_{t,i}) = \|f_t - g_{t,i}\|$ and hence $\inf_{g\in conv_n\, G(t)} e_C(g) \leq t\, s_G/\sqrt{n}$. Since $conv_n G(r) = \cap\{conv_n G(t) : t > r\}$, we have $\inf_{g\in span_n G} e_C(g) \leq \inf_{g\in conv_n\, G(r)} e_C(g) \leq r\, s_G/\sqrt{n}$.

(ii) and (iii) are proved analogously to part (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 do not exhibit the curse of dimensionality for functions in balls of fixed radius in G-variation. However, such balls may depend on the number of variables of functions in G [19].

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.

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

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

(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 $B_r(\|.\|_{2,s,\mathcal{R}^d})_{|\Omega}$ and $B_r(\|.\|_{H_d})_{|\Omega}$ denote the restrictions to Ω of the balls of radius r in the Sobolev norm $\|.\|_{2,s,\mathcal{R}^d}$ and in H_d-variation, resp. Using the technique exploited in [3, pp. 935, 941], one obtains $B_a(\|.\|_{2,s,\mathcal{R}^d})_{|\Omega} \subseteq B_{2\,a\,b}(\|.\|_{H_d})_{|\Omega}$, where $b = \left(\int_{\mathcal{R}^d} (1 + \|\omega\|^{2(s-1)})^{-1}\, d\omega\right)^{1/2}$ is finite as $2(s-1) > d$.

By [6, Chap. 9] there exists an extension operator $\mathcal{P} : (W_2^s(\Omega), \|.\|_{2,s,\Omega}) \to (W_2^s(\mathcal{R}^d), \|.\|_{2,s,\mathcal{R}^d})$ such that for all $f \in (W_2^s(\Omega), \|.\|_{2,s,\Omega})$, $(\mathcal{P}\, f)_{|\Omega} = f$ and $\|\mathcal{P}\, f\|_{2,s,\mathcal{R}^d} \leq c\, \|f\|_{2,s,\Omega}$, where c is a constant depending only on Ω. Hence there exists $f \in C$ such that $\mathcal{P}f \in B_{a\,c}(\|.\|_{2,s,\mathcal{R}^d})$. As $B_a(\|.\|_{2,s,\mathcal{R}^d})_{|\Omega} \subseteq B_{2\,a\,b}(\|.\|_{H_d})_{|\Omega}$, we have $f \in B_{2\,a\,b\,c}(\|.\|_{H_d})_{|\Omega}$. Since $s_{H_d} \leq 1$, the statement follows from by Corollary 5.2 (i) with $r = 2abc$ and $s_G = 1$.

(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 [16], 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 functions.

In the following we apply Corollary 5.2 to admissible sets of Boolean functions in $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$. The next proposition gives conditions on target sets 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., $\bar{H}_d = \{f : \{0,1\}^d \to \mathcal{R} : f(x) = \mathrm{sgn}(v \cdot x + b), v \in \mathcal{R}^d, b \in \mathcal{R}\}$. We estimate variation with respect to signum perceptrons using variation with respect to the the *Fourier orthonormal basis* defined as $F_d = \left\{f_u \,:\, u \in \{0,1\}^d,\, f_u(x) = \frac{1}{\sqrt{2^d}}(-1)^{u \cdot x}\right\}$. 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, $\|f\|_{1,F_d} = \|\hat{f}\|_{l_1} = \sum_{u \in \{0,1\}^d} |\hat{f}(u)|$, is called the *spectral norm.*

Proposition 5.4 *Let d be a positive integer, $r > 0$, and let C be a bounded subset of $(\mathcal{B}(\{0,1\}^d), \|.\|_{l_2})$ such that $C \cap B_r(\|.\|_{F_d}) \neq \emptyset$. 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* $\min_{g \in span_{dn+1}\, \bar{H}_d} e_C(g) \leq \min_{g \in conv_{dn+1}\, \bar{H}_d(r)} e_C(g) \leq \frac{r}{2\sqrt{n}}$.

Proof. 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 [19]. 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 ([17], [19]), we have $\|\tilde{f}\|_1 = \|f\|_{1,F_d} = \|f\|_{F_d}$ and the statement follows from Proposition 4.1 (i) and Corollary 5.2 (iii). □

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. The next two propositions describe target sets for which minimization of error functionals over admissible sets computable by Boolean signum perceptrons 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 $\|f\| \leq c$. 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 $g_i \in F_d$. Then $\|f\|_{F_d} = \|\tilde{f}\|_1 = \|f\|_{1,F_d} = \sum_{i=1}^{m} |w_i|$. 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 $\|w\|_2 = \|f\| \leq c$ and $\|u\|_2 \leq \sqrt{m}$, we have $\|f\|_{1,F_d} \leq c\sqrt{m}$. Thus C contains a function f with $\|f\|_{1,F_d} \leq c\sqrt{m}$ 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 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.

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 [20].

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 $x_1, \ldots, x_d$, 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 $f(x_1, \ldots, x_d)$.

Proposition 5.6 *Let d, s be positive integers, $b \geq 0$, and let 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 $\frac{\max_{x \in \{0,1\}^d} |f(x)|}{\min_{x \in \{0,1\}^d} |f(x)|} \|f\| \leq b$. 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 [19, Theorem 3.4] (which extends [20, Lemma 5.1]) we have $\frac{\|\tilde{f}\|_1}{\|f\|_2} \leq s \frac{\max_{x \in \{0,1\}^d} |f(x)|}{\min_{x \in \{0,1\}^d} |f(x)|}$. Hence, we get $\|f\|_{1,F_d} = \|\tilde{f}\|_1 \leq sb$ and the so 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 its 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.

Bibliography

[1] N. I. Achieser, *Theory of Approximation*, Dover, New York, 1992 (orig. Frederick Ungar Publishing Co., New York, 1956).

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

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

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

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

[6] H. Brezis, *Analyse Fonctionnelle - Théorie et Applications*, Masson, Paris, 1983.

[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. The Association for Computing Machinery, New York, N.Y., pp. 303-309, 1993.

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

[9] F. Girosi, *Regularization Theory, Radial Basis Functions and Networks*, in From Statistics to Neural Networks. Theory and Pattern Recognition Applications, J.H. Friedman, V. Cherkassky, H. Wechsler Eds, Subseries F, Computer and System Sciences, pp. 166-187. Springer-Verlag, Berlin, 1993.

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

[11] 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, vol. 20, pp. 608–613, 1992.

[12] P. C. Kainen, V. Kůrková, and A. Vogt, *Best approximation by Heaviside perceptron networks*, Neural Networks, vol. 13, pp. 645–647, 2000.

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

[14] 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., pp. 261–270, Birkhäuser, 1997.

[15] V. Kůrková, *High-dimensional approximation by neural networks*, in Learning Theory and Practice, J. Suykens, Ed., to appear.

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

[17] V. Kůrková and M. Sanguineti, M., *Bounds on Rates of Variable–Basis and Neural–Network Approximation*, IEEE Trans. on Information Theory, vol. 47, pp. 2659-2665, 2001.

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

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

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

[21] K.S. Narendra, J. Balakrishnan, and K. M. Ciliz, *Adaptation and learning using multiple models, switching, and tuning*, IEEE Control Systems Magazine, vol. 15, pp. 37-51, 1995.

[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 non-polynomial activation can approximate any function*, Neural Networks, vol. 6, pp. 861–867, 1993.

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

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

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

[27] 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, vol. 112, pp. 403-440, 2002.