Minimization of empirical error over perceptron networks

Věra Kůrková [1]
Institute of Computer Science, Academy of Sciences of the Czech Republic
E-mail: vera@cs.cas.cz

Abstract

Supervised learning by perceptron networks is investigated a minimization of empirical error functional. Input/output functions minimizing this functional require the same number m of hidden units as the size of the training set. Upper bounds on rates of convergence to zero of infima over networks with n hidden units (where n is smaller than m) are derived in terms of a variational norm. It is shown that fast rates are guaranteed when the sample of data defining the empirical error can be interpolated by a function, which may have a rather large Sobolev-type seminorm. Fast convergence is possible even when the seminorm depends exponentially on the input dimension.

1 Introduction

The goal of supervised learning is to adjust parameters of a neural network so that it approximates with a sufficient accuracy a functional relationship between inputs and outputs known only by a sample of input/output pairs. It is desirable that the system also generalizes well, i.e., it satisfactorily processes new data that were not used for training. Learning from data with generalization capability was studied theoretically in the framework of regularized optimization [4], [14], [10]. Theoretical results describing optimal solutions can be applied to kernel models, a special case of which are radial-basis function networks with constant width. But the most common neural networks built from perceptrons cannot be represented as kernel models.

In this paper, we investigate minimization of empirical error functionals over sets of functions computable by perceptron networks. We estimate rates of convergence of infima over networks with n hidden units to the global infimum achievable by a network with the same number of hidden units as the size of the training set.

2 Approximate minimization of empirical error

Let $\mathcal{R}$ denote the set of real numbers, Ω be a nonempty set and $z = \{(u_i, v_i) \in \Omega \times \mathcal{R}, i = 1, \dots, m\}$ be a sample of input/output pairs of data. A standard approach to learning from empirical data used, e.g., in back-propagation, is based on minimization of the *empirical error* functional defined as $\mathcal{E}_{z,V}(f) = \frac{1}{m}\sum_{i=1}^{m} V(f(u_i), v_i)$, where $V : \mathcal{R} \times \mathcal{R} \to [0, \infty)$, satisfying for all $y \in \mathcal{R}$, $V(y,y) = 0$, is called a *loss function*. The most common loss function is the *square loss* $V(f(u), v) = (f(u) - v)^2$, we denote by $\mathcal{E}_z$ the empirical error functional with this loss function, i.e., $\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^{m}(f(u_i) - v_i)^2$.

Let M be a subset of a normed linear space $(X, \|.\|)$ and $\Phi : X \to \mathcal{R}$ be a functional. Using standard notation from optimization theory, we denote by (M, Φ) the problem of minimization of Φ over M; M is called the *hypothesis set*. Elements of the set $argmin\,(M, \Phi) = \{g \in M : \Phi(g) = \inf_{g \in M} \Phi(g)\}$ are called *solutions* (or *minimum points*) of the problem (M, Φ). For $\varepsilon > 0$, elements of the set $argmin_\varepsilon(M, \Phi) = \{g \in M : \Phi(g) < \inf_{g\in M} \Phi(g) + \varepsilon\}$ are called *ε-near minimum points* of (M, Φ). A sequence $\{g_n\}$ of elements of M is called Φ-minimizing if $\lim_{n\to\infty} \Phi(g_n) = \inf_{g\in M} \Phi(g)$.

Typical hypothesis sets used in neurocomputing are sets of functions computable by *neural networks with n hidden units* of a given type *and one linear output*. Such sets are of the form $span_n\, G = \{\sum_{i=1}^{n} w_i g_i : w_i \in \mathcal{R},\ g_i \in G\}$, where G is the set of functions computable by the computational units.

Standard hidden units are *perceptrons*. For $\Omega \subseteq \mathcal{R}^d$ and $\psi : \mathcal{R} \to \mathcal{R}$ we denote by $P_d(\psi, \Omega) = P_d(\psi) = \{f : \Omega \to \mathcal{R} \mid f(x) = \psi(a_i \cdot x + b_i), a_i \in \mathcal{R}^d, b_i \in \mathcal{R}\}$ the set of functions on Ω computable by perceptrons with the activation function ψ (we write $P_d(\psi)$ when Ω is clear from context). The most common activation functions are *sigmoidals*, which are monotonic increasing functions $\sigma : \mathcal{R} \to \mathcal{R}$ (i.e., for all $t_1, t_2 \in \mathcal{R}, t_1 \leq t_2$ implies $\sigma(t_1) \leq \sigma(t_2)$) satisfying $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$.

An important type of a sigmoidal is the *Heaviside function* $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. To shorten notation, we write $H_d(\Omega)$ instead of $P_d(\vartheta, \Omega)$. Note that $H_d(\Omega)$ is the *set of it characteristic functions of closed half-spaces of $\mathcal{R}^d$ intersected with Ω.*

Ito [6, p.73] proved that any function defined on a finite subset of $\mathcal{R}^d$ can be exactly represented as a function computable by a perceptron network with any sigmoidal activation function. The following theorem is a reformu-

[1] This work was partially supported by GA ČR grant 201/05/0557.

lation of Ito's result in terms of optimization theory.

Theorem 2.1 (Ito) *For all positive integers* d, m, *all samples of data* $z = \{(u_i, v_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \ldots, m\}$ *with all* u_i *distinct and all sigmoidal functions* σ, *the problem* $(span_m \mathcal{P}_d(\sigma), \mathcal{E}_z)$ *has a solution and* $\min_{f \in span_m P_d(\sigma)} \mathcal{E}_z(f) = 0$.

A drawback of this result is that the number of hidden units in the network interpolating the sample of data is equal to the size of the sample. For large samples, such networks might not be implementable. Moreover in typical applications of neural networks, a number of hidden units much smaller than the size of the training set is chosen before learning. Using such networks, only suboptimal solutions can be achieved. To compare such suboptimal solutions with the optimal one given by Theorem 2.1 we estimate rates of convergence of $\{\inf_{f \in span_n H_d(\Omega)} \mathcal{E}_z(f)\}$ to zero as n goes to m.

A useful property that allows application of several tools for investigation of approximate minimization of a functional is its continuity. The next proposition shows that continuity of the empirical error defined by a sample z follows from continuity of the evaluation functionals at the input data $u_1, \ldots, u_m$. For a normed linear space $(X, \|.\|)$ of functions on some set Ω and $x \in \Omega$, the *evaluation functional at* x, denoted by $\mathcal{F}_x$, is defined for all $f \in X$ as $\mathcal{F}_x(f) = f(x)$.

Proposition 2.2 *Let* $(X, \|.\|)$ *be a normed linear space of functions on a nonempty set* Ω, m *a positive integer,* $z = \{(u_i, v_i) \in \Omega \times \mathcal{R}, i = 1, \ldots, m\}$, $V : \Omega \times \Omega \to \mathcal{R}$ *a loss function and* $f \in X$. *If for all* $i = 1, \ldots, m$, $\mathcal{F}_{u_i}$ *is continuous at* f *and* V *is continuous at* $(f(u_i), v_i)$, *then* $\mathcal{E}_{z,V}$ *is continuous at* f.

Proof. By continuity of V we get for every $\varepsilon > 0$ some $\eta > 0$ such that $\|f(u_i) - g(u_i)\| \leq \eta$ implies $|V(f(u_i), v_i) - V(g(u_i), v_i)| \leq \varepsilon$. As all $\mathcal{F}_{u_i}$ are continuous at f, there exists $\delta > 0$ such that $\|g - f\| \leq \delta$ implies $\|f(u_i) - g(u_i)\| \leq \eta$ and hence $|\mathcal{E}_{z,V}(f) - \mathcal{E}_{z,V}(g)| = \left|\frac{1}{m}\sum_{i=1}^{m}(V(f(u_i), v_i) - V(g(u_i), v_i))\right| \leq \varepsilon$. □

It is easy to show that $\mathcal{E}_z$ is continuous on the space $\mathcal{M}(\Omega)$ of bounded measurable functions on $\Omega \subseteq \mathcal{R}^d$ with the supremum norm $\|.\|_{\sup}$ and that it is convex.

Proposition 2.3 *Let* $\Omega \subseteq \mathcal{R}^d$, $z = \{(u_i, v_i) \in \Omega \times \mathcal{R}, i = 1, \ldots, m\}$ *and* $v_{\max} = \max\{|v_1|, \ldots, |v_m|\}$. *Then* $\mathcal{E}_z : (\mathcal{M}(\Omega), \|.\|_{\sup}) \to \mathcal{R}$ *is continuous and its modulus of continuity at any* $f \in \mathcal{M}(\Omega)$ *satisfies* $\omega_f(t) \leq t(t + 2\|f\|_{\sup} + 2v_{\max})$.

In contrast to the supremum norm, in the $\mathcal{L}_2$-norm evaluation functionals need not to be continuous. Lack of continuity can be easily demonstrated for $span_2 H_d([0,1]^d)$: denoting by χ_n the characteristic function of $[0, \frac{1}{n}] \times [0,1]^{d-1}$, we get a sequence $\{n\chi_n\} \subset span_2 H_d([0,1]^d)$ with all elements having $\mathcal{L}_2$-norms equal to 1, on which the evaluation functional at zero is not bounded (and thus it is not continuous). Similarly, one can disprove continuity of evaluation functionals on $span_2 P_d(\sigma, \Omega)$ for any sigmoidal function σ.

Recently, a class of Hilbert spaces of point-wise defined functions, on which all evaluation functionals are continuous, became popular in learning theory. Such spaces are called reproducing kernel Hilbert spaces (RKHS) and they can be represented as completions of linear combinations of "translations" of kernels (symmetric positive semidefinite functions) [1], [4], [14]. Continuity of empirical error functionals on RKHSs allows one to apply theory of generalized inversion and regularization in infinite dimensional Hilbert spaces to describe solutions of the problem of minimization of the empirical error over RKHSs [10]. Moreover, continuity of empirical errors is essential for derivation of estimates of rates of approximate optimization over kernel models of bounded complexity [11], [12].

However, it is not clear, whether there exist kernels, for which sets $P_d(\sigma)$ are contained in the corresponding RKHSs. So in investigation of minimization of empirical errors over perceptron networks, we focus on the space of bounded measurable functions with the supremum norm. To derive rates of approximate optimization over $span_n H_d(\Omega)$ we take advantage of a result from nonlinear approximation theory [2] giving an upper bound on supremum norm rates of approximation by $span_n H_d(\Omega)$ in terms of a norm called *variation with respect to half-spaces*. It is a special case of *G-variation* [9] defined for any bounded nonempty subset G of a normed linear space $(X, \|.\|)$ as the Minkowski functional of the closed convex symmetric hull of G, i.e.,

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

where the closure cl is taken with respect to the topology generated by the norm $\|.\|$. Note that G-variation can be infinite (when the set on the right-hand side is empty).

Here we consider H_d-variations with respect to the topology generated by the supremum and $\mathcal{L}_2$-norm (we indicate the norm by the notation $\|.\|_{H_d,\sup}$, $\|.\|_{H_d,\mathcal{L}_2}$, resp. It is easy to check that $\|.\|_{H_d,\mathcal{L}_2} \leq \|.\|_{H_d,\sup}$ as well as $\|.\|_{P_d(\sigma),\mathcal{L}_2} \leq \|.\|_{P_d(\sigma),\sup}$ and that $\|.\|_{H_d,\mathcal{L}_2} = \|.\|_{P_d(\sigma),\mathcal{L}_2}$ for any sigmoidal σ.

Barron [2] estimated rates of approximation by $span_n H_d$ in the supremum norm (see also [3, p. 201] and [5, p. 25]).

Theorem 2.4 (Barron) *For all positive integers* d, n, *every compact* $\Omega \subset \mathcal{R}^d$ *and every* $f \in \mathcal{M}(\Omega)$,

$$\|f - span_n H_d(\Omega)\|_{\sup} \leq c\sqrt{\tfrac{d+1}{n}}\|f\|_{H_d(\Omega)\sup},$$
where c is an absolute constant.

The next theorem estimates rates of convergence of infima of a continuous functional over $span_n H_d$ to the global minimum.

Theorem 2.5 *Let $\Omega \subset \mathcal{R}^d$ be compact, $\Phi : (\mathcal{M}(\Omega), \|.\|_{\sup}) \to \mathcal{R}$ a functional such that there exists a solution f^o of the problem $(\mathcal{M}(\Omega), \Phi)$ at which Φ is continuous with the modulus of continuity ω, $\{\varepsilon_n\}$ a sequence of positive real numbers converging to zero, and $\{f_n\}$ a sequence of ε_n-minimum points of $(span_n H_d(\Omega), \Phi)$. Then for every positive integer n:*
(i) $\inf_{f \in span_n H_d(\Omega)} \Phi(f) - \Phi(f^o) \leq \omega\left(c\sqrt{\frac{d+1}{n}}\|f^o\|_{H_d(\Omega),\sup}\right)$;
(ii) if $\|f^o\|_{H_d(\Omega),\sup} < \infty$, *then* $\{f_n\}$ *is a* Φ*-minimizing sequence and* $\Phi(f_n) - \Phi(f^o) \leq \omega\left(c\sqrt{\frac{d+1}{n}}\|f^o\|_{H_d(\Omega),\sup}\right) + \varepsilon_n$.

Proof. (i) For every $\varepsilon > 0$, let $f_n^\varepsilon \in span_n H_d$ be such that $\|f^o - f_n^\varepsilon\|_{\sup} < \|f^o - span_n H_d\|_{\sup} + \varepsilon$. Then $\inf_{f \in span_n H_d} \Phi(f) - \Phi(f^o) \leq \Phi(f_n^\varepsilon) - \Phi(f^o) \leq \omega(\|f_n^\varepsilon - f^o) \leq \omega(\|f^o - span_n H_d\| + \varepsilon)$. By Theorem 2.4, infimizing over ε we get $\inf_{f \in span_n H_d} \Phi(f) - \Phi(f^o) \leq \omega\left(c\sqrt{\frac{d+1}{n}}\|f^o\|_{H_d,\sup}\right)$.

(ii) By the definition of ε_n-minimum point, $\Phi(f_n) - \Phi(f^o) \leq \inf_{f \in span_n H_d} \Phi(f) - \Phi(f^o) + \varepsilon_n$. So by (i), $\Phi(f_n) - \Phi(f^o) \leq \omega\left(c\sqrt{\frac{d+1}{n}}\|f^o\|_{H_d,\sup}\right) + \varepsilon_n$. As $\lim_{n\to\infty} \varepsilon_n = 0$ and $\|f^o\|_{H_d,\sup}$ is finite, $\{f_n\}$ is Φ-minimizing. □

Combining Theorems 2.1, 2.5 and Proposition 2.3 we get the following upper bound on rates of approximate minimization of $\mathcal{E}_z$ over $span_n H_d(\Omega)$.

Corollary 2.6 *Let $\Omega \subset \mathcal{R}^d$ be compact, $z = \{(u_i, v_i) \in \Omega \times \mathcal{R}, i = 1, \dots, m\}$, $f^o \in \mathcal{M}(\Omega)$ such that $\mathcal{E}_z(f^o) = 0$, $\{\varepsilon_n\}$ a sequence of positive reals converging to zero, $\{f_n\}$ a sequence of ε_n-minimum points of $(span_n H_d(\Omega), \mathcal{E}_z)$. Then for every n:*
(i) $\inf_{f \in span_n H_d(\Omega)} \mathcal{E}_z(f) \leq \frac{(d+1)c^2\|f^o\|^2_{H_d,\sup}}{n} + \frac{\sqrt{d+1}(2c\|f^o\|^2_{H_d,\sup} + v_{\max}\|f^o\|_{H_d,\sup})}{\sqrt{n}}$;
(ii) $\{f_n\}$ *is* $\mathcal{E}_z$*-minimizing and* $\mathcal{E}_z(f_n) \leq \frac{(d+1)c^2\|f^o\|^2_{H_d,\sup}}{n} + \frac{\sqrt{d+1}(2c\|f^o\|^2_{H_d,\sup} + v_{\max}\|f^o\|_{H_d,\sup})}{\sqrt{n}} + \varepsilon_n$, *where c is an absolute constant.*

3 Estimates of variation with respect to half-spaces

Corollary 2.6 shows that the speed of convergence of suboptimal solutions of the problem of minimization of $\mathcal{E}_z$ over the set of functions computable by networks with n Heaviside perceptrons depends on the smallest value of H_d-variation on the set of functions interpolating the data z.

To estimate H_d-variations of smooth elements of this set we take an advantage of a result from [8] bounding from above H_d-variation of a smooth function by a product of its certain Sobolev-type seminorm with
$$k_d \sim \left(\frac{e^d}{d2^{d-2}\pi^{d-1}}\right)^{1/2},$$
which as a function of d is *decreasing exponentially fast.*

For a function $f \in \mathcal{C}^d(\mathcal{R}^d)$ define
$$\|f\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}_1(\mathcal{R}^d)}.$$
For d odd and f sufficiently rapidly vanishing at infinity, an upper bound
$$\|f\|_{H_d(\mathcal{R}^d),\sup} \leq k_d \|f\|_{d,1,\infty} \qquad (1)$$
was derived in [8].

Thus by Corollary 2.6, for any sample of data z, which can be interpolated by a function f^o satisfying
$$\|f^o\|_{1,d,\infty} \leq \frac{1}{k_d} \sim \left(\frac{d2^{d-2}\pi^{d-1}}{e^d}\right)^{1/2},$$
infima of $\mathcal{E}_z$ over $span_n H_d$ converge to zero with rate
$$c^2\sqrt{\frac{d+1}{n}} + (2c + v_{\max})\frac{d+1}{n}$$
as by (1) f^o has H_d-variation at most 1.

However, there exist samples of data, which cannot be interpolated by functions with small H_d-variations. Such samples $z = \{(u_i, v_i), i = 1, \dots, m\}$ can be obtained from real-valued Boolean functions $h : \{0,1\}^d \to \mathcal{R}$ by setting $\{0,1\}^d = \{u_1, \dots, u_{2^d}\}$ and $v_i = h(u_i)$. If $f : \Omega \to \mathcal{R}$ is an extension of h, then $\|f\|_{H_d(\Omega),\sup} \geq \|h\|_{H_d(\{0,1\}^d),\sup}$.

To show that there exist functions on $\{0,1\}^d$ with $H_d(\{0,1\}^d)$-variations depending on d exponentially, we use a geometric characterization of G-variation from [13]
$$\|f\|_G \geq \frac{\|f\|^2}{\sup_{g \in G} |g \cdot f|}. \qquad (2)$$
So functions that have small inner products with all elements of G (are "almost orthogonal" to G) have large G-variations.

For a Hilbert space $(X, \|.\|)$ we define on its unit ball S_1 a pseudometrics $\rho_X(f,g) = \arccos|f \cdot g|$, which measures the distance as the minimum of the two angles between f and g and between f and $-g$ (it is a pseudometrics as the distance of antipodal vectors is zero). For $\alpha > 0$, let $\mathcal{N}_\alpha(S_1)$ denote the α-covering number of S_1 with respect to ρ_X, i.e., the size of the smallest α-net in S_1. The next proposition shows that when for some α close to $\pi/2$, the cardinality of G is smaller than $\mathcal{N}_\alpha(S_1)$, then in S_1 there exists a function with a "large" G-variation. It also guarantees existence a function with G-variation at least $\frac{1}{\varepsilon}$ for any subset G of the unit sphere S^{m-1} in $\mathcal{R}^m$ of smaller cardinality than the ε-*quasiorthogonal dimension* $dim_\varepsilon m$ of $\mathcal{R}^m$. For $\varepsilon > 0$, $dim_\varepsilon m$ was defined in [7] as the maximal number of vectors which are pairwise ε-*quasiorthogonal*, i.e., $|u \cdot v| \leq \varepsilon \|u\| \|v\|$.

Proposition 3.1 *(i) If G is a subset of the unit sphere S_1 in a Hilbert space X and $\alpha \in [0, \pi/2]$ is such that $card\, G < \mathcal{N}_\alpha(S_1)$ with respect to the pseudometrics ρ_X, then there exists $f \in S_1$ with $\|f\|_G \geq 1/\cos\alpha$.*
(ii) If $G \subset S^{m-1} \subset \mathcal{R}^m$ such that $card\, G < dim_\varepsilon m$, then there exists $f \in S^{m-1}$ with $\|f\|_G \geq \frac{1}{\varepsilon}$.

Proof. (i) If $card\, G < \mathcal{N}_\alpha(S_1)$, then there exists $f \in S_1$ such that for all $g \in G$, $\rho_X(f,g) \geq \alpha$ and hence $|f \cdot g| \leq \cos\alpha$. Then by (2) $\|f\|_G \geq \frac{1}{\sup_{g\in G}|f\cdot g|} \geq 1/\cos\alpha$.
(ii) follows from (i) as $dim_\varepsilon m \leq \mathcal{N}_{\arccos(\varepsilon)}(S^{m-1})$. □

Theorem 3.2 *For every positive integer d there exists a sample $z = \{(u_i, v_i) : i = 1, \dots, 2^d\} \subset \{0,1\}^d \times \mathcal{R}$ such that for every $\Omega \supseteq \{0,1\}^d$ and every $f : \mathcal{R}^d \to \mathcal{R}$ such that $\mathcal{E}_z(f) = 0$, $\|f\|_{H_d(\Omega),\sup} \geq \frac{2^{(d-1)/2}}{d\sqrt{ln2}}$.*

Proof. It was shown in [7] that $dim_\varepsilon m \geq e^{m\varepsilon^2/2}$. On the other hand, $card\, H_d(\{0,1\}^d) = 2^{d^2 - d\log_2 d + O(d)}$ [15]. Denoting $H_d^o(\{0,1\}^d)$ the set of normalized elements of $H_d(\{0,1\}^d)$ with respect to l_2-norm on $\mathcal{R}^{2^d}$, we get $\|.\|_{H_d(\{0,1\}^d),\sup} \geq \|.\|_{H_d(\{0,1\}^d),l_2} \geq \|.\|_{H_d^o(\{0,1\}^d),l_2}$. As $card\, H_d^o(\{0,1\}^d) < 2^{d^2}$, for $\varepsilon = \frac{\sqrt{d ln2}}{2^{(d-1)/2}}$, $card\, H_d^o(\{0,1\}^d) < e^{(2^d\varepsilon^2)/2}$ and hence by Proposition 3.1 (ii) there exists a function $h \in S^{2^d-1}$ with $\|h\|_{H_d(\{0,1\}^d),\sup} \geq \|h\|_{H_d^o(\{0,1\}),l_2} \geq \frac{2^{(d-1)/2}}{d\sqrt{ln2}}$. Let $(u_1, \dots, u_{2^d}) = \{0,1\}^d$, $v_i = f(u_i)$ and $z = \{(u_i, v_i) : i = 1, \dots, m\}$. Then for every $f : \Omega \to \mathcal{R}$, for which $\mathcal{E}_z(f) = 0$, $\|f\|_{H_d(\Omega),\sup} \geq \frac{2^{(d-1)/2}}{d\sqrt{ln2}}$. □

Note that by (1) for every d odd and $f^o \in \mathcal{C}^d(\mathcal{R}^d)$ sufficiently rapidly vanishing at infinity interpolating the sample described in Theorem 3.2, $\|f^o\|_{1,d,\infty} \geq \left(\frac{2^{2d-3}\pi^{d-1}}{de^d ln2}\right)^{1/2}$.

4 Discussion

We have shown that fast convergence of infima of the empirical error functional $\mathcal{E}_z$ over networks with n Heaviside perceptrons to zero can be achieved for samples that can be interpolated by functions f^o with the Sobolev seminorm $\|f^o\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f^o\|_{\mathcal{L}_1(\mathcal{R}^d)}$ depending exponentially on the input dimension. Note that the seminorm $\|f^o\|_{1,d,\infty}$ is much smaller than the Sobolev norm $\|f^o\|_{d,1} = \sum_{|\alpha|\leq d} \|D^\alpha f^o\|_{\mathcal{L}_1(\mathcal{R}^d)}$ as instead of summation of iterated partial derivatives of f over all α with $|\alpha| \leq d$ only their maximum over α with $|\alpha| = d$ is taken.

We have also shown that there exist samples of data constructed using special Boolean functions, for which the Sobolev seminorms of interpolating functions are even larger than the exponential size allowed for fast convergence described in Corollary 2.6.

The proof of Proposition 3.1 is existential, but in [13] a lower bound $\mathcal{O}(2^{d/6})$ on $H_d(\{0,1\}^d)$-variation was derived for a concrete function, namely the "inner product modulo 2".

References

[1] Aronszajn, N. (1950). Theory of reproducing kernels. Transactions of AMS, 68: 337-404.

[2] Barron, A. R. (1992). Neural net approximation. In: Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems (pp. 69–72).

[3] Cheang, G. H. L., Barron, A. R. (2000). A better approximation for balls. Journal of Approximation Theory 104: 183–203.

[4] Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. Bulletin of AMS 39: 1-49.

[5] Donahue, M. J., Gurvits, L., Darken, C., Sontag, E. (1997). Rates of convex approximation in non-Hilbert spaces. Constructive Approximation 13: 187–220.

[6] Ito, Y. (1992). Finite mapping by neural networks and truth functions. Math. Scientist 17: 69–77.

[7] Kainen, P. C., Kůrková, V. (1993). Quasiorthogonal dimension of Euclidean spaces. Applied Math. Letters 6: 7–10, 1993.

[8] Kainen, P. C., Kůrková, V., Vogt, A. (2004). A Sobolev-type upper bound for rates of approximation by linear combinations of plane waves. Research Report ICS-2003-900, Institute of Computer Science, Prague.

[9] Kůrková, V. (2003). High-dimensional approximation and optimization by neural networks. Chapter 4 In: Advances in Learning Theory: Methods, Models and Applications. (Eds. J. Suykens et al.) (pp. 69-88). IOS Press, Amsterdam.

[10] Kůrková, V. (2004). Learning from data as an inverse problem. In: Proceedings of COMPSTAT 2004 (Ed. J. Antoch) (pp. 1377–1384). Physica-Verlag, Heidelberg.

[11] Kůrková, V., Sanguineti, M. (2004). Error estimates for approximate optimization by the extended Ritz method. SIAM Journal on Optimization (to appear).

[12] Kůrková, V., Sanguineti, M. (2004). Learning with generalization capability by kernel methods of bounded complexity. Journal of Complexity (to appear).

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

[14] Poggio T., Smale, S. (2003). The mathematics of learning: dealing with data. Notices of AMS 50: 536-544.

[15] Shläfli, L. (1950). Gesamelte mathematische abhandlungen. Band 1. Basel, Verlag Birkhäuzer.