LETTER Communicated by Andrew Barron

Minimization of Error Functionals over Perceptron Networks

Věra Kůrková
vera@cs.cas.cz
Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, CZ 18207

Supervised learning of perceptron networks is investigated as an optimization problem. It is shown that both the theoretical and the empirical error functionals achieve minima over sets of functions computable by networks with a given number n of perceptrons. Upper bounds on rates of convergence of these minima with n increasing are derived. The bounds depend on a certain regularity of training data expressed in terms of variational norms of functions interpolating the data (in the case of the empirical error) and the regression function (in the case of the expected error). Dependence of this type of regularity on dimensionality and on magnitudes of partial derivatives is investigated. Conditions on the data, which guarantee that a good approximation of global minima of error functionals can be achieved using networks with a limited complexity, are derived. The conditions are in terms of oscillatory behavior of the data measured by the product of a function of the number of variables d, which is decreasing exponentially fast, and the maximum of the magnitudes of the squares of the $\mathcal{L}^1$-norms of the iterated partial derivatives of the order d of the regression function or some function, which interpolates the sample of the data. The results are illustrated by examples of data with small and high regularity constructed using Boolean functions and the gaussian function.

1 Introduction

The goal of supervised learning is to adjust parameters of a neural network so that it approximates with sufficient accuracy a functional relationship between inputs and outputs known only by a sample of input-output pairs. Many learning algorithms (such as backpropagation; Werbos, 1985) iteratively decrease the average square of errors on a training set. In statistical learning theory (see, e.g., Vapnik, 1995; Cucker & Smale, 2002), such learning is modeled as a minimization of error functionals: the empirical and the expected error.

A basic model of a neural network architecture is a network with one hidden layer with Heaviside perceptrons and one linear output. Such

networks can compute all linear combinations of characteristic functions of half-spaces of $\mathbb{R}^d$, where d is the number of network inputs. The class of functions computable by such networks with an unbounded number of computational units is known to be dense in the space of continuous functions with the supremum norm, as well as in the space of square integrable functions on any compact subset of $\mathbb{R}^d$ (see, e.g., Cybenko, 1989; Hornik, Stinchcombe, & White, 1989; Ito, 1991; Mhaskar & Micchelli, 1992; Leshno, Lin, Pinkus, & Schocken, 1993), or a survey (Pinkus, 1998). In learning algorithms, either the number of computational units is chosen in advance or it is dynamically allocated during learning, but in both cases, it is constrained. Rates of approximation by perceptron networks with an increasing number of hidden units were estimated in Barron (1992, 1993), Makovoz (1996), and Kůrková, Kainen, and Kreinovich (1997). The tightness of these estimates was proven in Makovoz (1998) and the existence and discontinuity of best approximation in Kainen, Kůrková, and Vogt (1999, 2000a, 2003).

In this letter, we investigate learning of Heaviside perceptron networks as an optimization problem of minimization of the two functionals, the empirical and the expected error, over nonconvex sets of functions computable by Heaviside perceptron networks with a bounded number of units. First, we prove the existence of minima of both error functionals over sets of functions computable by these networks with any fixed number n of hidden units. We use representations of the error functionals as distance functionals and a weakened compactness property of the sets of functions computable by Heaviside perceptrons. Then we estimate the speed of decrease of the minima as n increases. We show that this speed depends on a certain regularity of training data expressed in terms of the variational norms of functions interpolating the data (in the case of the empirical error) and the regression function (in the case of the expected error).

Further, dependence of this type of regularity on the dimensionality d is studied. It is shown that the minima of error functionals over networks with n perceptrons with d inputs are bounded from above by $\frac{1}{n}$ times the maximum of the squares of the $\mathcal{L}^1_\lambda$-norms of the iterated partial derivatives of the order d of the functions, at which the global minima are achieved, multiplied by an exponentially quickly decreasing function $c(d) \lesssim \frac{\pi}{d} 2^{2-d}$ of the number of variables d.

This bound provides a quantitative description of a property of data, which guarantees that a good fit to these data can be achieved by networks with a reasonable number of hidden units. The property of the data is formulated in terms of their oscillatory behavior measured by the magnitudes of the iterated partial derivatives of the order d of the regression or some interpolating function. Because of the multiplicative factor $c(d) \lesssim \frac{\pi}{d} 2^{2-d}$, the tolerance on these magnitudes, which can be allowed so that a good fit to data is achievable using networks of a reasonable complexity, grows exponentially fast with the dimension.

The results are illustrated by examples of data chosen from the gaussian function or from symmetric Boolean functions. For such data, rates of convergence of error functionals over networks with n perceptrons are bounded by $\frac{1}{n}$ times a quadratic function of the dimension d. In particular for the data chosen from the gaussian function, our estimate of the convergence of the minima of the empirical error functional determined by these data gives some insight into the relationship of two geometrically opposite types of computational units: perceptrons (which compute plane waves) and radial basis functions (which compute radial waves).

Comparing the size of the set of characteristic functions of the Boolean cube with certain covering numbers, we prove the existence of samples of data that cannot be interpolated by functions with a small regularity expressed in terms of variational norms.

The letter is organized as follows. In section 2, minimization of error functionals is reformulated as a search for a best approximation. In section 3, existence of minima of these functionals over sets of functions computable by perceptron networks with n hidden units is proven, and upper bounds on the convergence of both error functionals are derived in terms of a variational norm tailored to Heaviside perceptrons. Section 4 investigates dependence of this norm on dimensionality and maxima of partial derivatives. Section 5 illustrates the estimates by examples of samples of data generated by Boolean real-valued functions. Section 6 is a brief discussion.

2 Learning as a Best Approximation

A standard mathematical approach to learning (see, e.g., Cucker & Smale, 2002) is in terms of minimization of two functionals, the expected and the empirical error, over various sets of functions. The expected error is determined by a nondegenerate (no nonempty open set has measure zero) probability measure ρ defined on $Z = X \times Y$, where X is a compact subset of $\mathbb{R}^d$ and Y a bounded subset of $\mathbb{R}$ ($\mathbb{R}$ denotes the set of real numbers). The measure ρ induces the marginal probability measure on X defined for every $S \subseteq X$ as $\rho_X(S) = \rho(\pi_X^{-1}(S))$, where $\pi_X : X \times Y \to X$ denotes the projection. Let $(\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2_{\rho_X}})$ denote the Lebesque space of functions satisfying $\int_X f^2 d\rho_X < \infty$. The expected error functional determined by ρ is defined for every f in $\mathcal{L}^2_{\rho_X}(X)$ as

$$\mathcal{E}_\rho(f) = \int_Z (f(x) - y)^2 \, d\rho,$$

and the empirical error functional determined by a sample of data $z = \{(u_i, v_i) \in X \times Y \mid i = 1, \ldots, m\}$ as

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(u_i) - v_i)^2.$$

Using notation from optimization theory, we denote by

$$(M, \Phi)$$

the problem of minimization of a functional $\Phi : F \to \mathbb{R}$ over M, where $M \subset F$ is called a hypothesis set.

Typical hypothesis sets used in neurocomputing are sets of functions computable by neural networks with n hidden units and one linear output unit. Such sets are of the form

$$span_n G = \left\{ \sum_{i=1}^{n} w_i g_i \mid w_i \in \mathbb{R},\ g_i \in G \right\},$$

where G is the set of functions that can be computed by computational units of a given type. Note that for G linearly independent, sets $span_n G$ are not convex, and thus results from theory of convex optimization cannot be applied.

Standard computational units used in neurocomputing are perceptrons. For $X \subseteq \mathbb{R}^d$ and an activation function $\psi : \mathbb{R} \to \mathbb{R}$, they compute functions from $\mathbb{R}^d$ to $\mathbb{R}$ of the form $\psi(v \cdot x + b)$, where $v \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are parameters. A typical activation function is the Heaviside function

$$\vartheta(t) = 0 \quad \text{for} \quad t < 0 \quad \text{and} \quad \vartheta(t) = 1 \quad \text{for} \quad t \geq 0.$$

Perceptrons with the Heaviside activation function compute characteristic functions of closed half-spaces of $\mathbb{R}^d$ intersected with X. Because for all $a > 0$ and all $t \in \mathbb{R}$, $\vartheta(at) = \vartheta(t)$, one can use vectors $e \in S^{d-1}$ (where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$) as parameters of Heaviside perceptrons. We denote by $H_d(X)$ the set of functions on X computable by Heaviside perceptrons, that is,

$$H_d(X) = \{f : X \to \mathbb{R} \mid f(x) = \vartheta(e \cdot x + b), e \in S^{d-1}, b \in \mathbb{R}\}.$$

In this letter, we investigate two optimization problems,

$$(\mathcal{E}_\rho, span_n H_d(X))$$

and

$$(\mathcal{E}_z, span_n H_d(X)),$$

of minimizations of the expected and the empirical error functionals over sets of functions computable by networks with n Heaviside perceptrons.

We take advantage of representations of minimizations of these functionals in terms of best approximations of suitable functions. For the expected error, such function is the regression function f_ρ defined for $x \in X$ as

$$f_\rho(x) = \int_Y y\, d\rho(y|x),$$

where $\rho(y|x)$ is the conditional (with regard to x) probability measure on Y. It is easy to see and well known that the minimum of $\mathcal{E}_\rho$ over $\mathcal{L}^2_{\rho_X}(X)$ is achieved at f_ρ, that is,

$$\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(f_\rho).$$

Moreover, $\mathcal{E}_\rho(f) = \int_X (f(x) - f_\rho(x))^2 d\rho_X + \mathcal{E}_\rho(f_\rho) = \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \mathcal{E}_\rho(f_\rho)$ (Cucker & Smale, 2002). So $\mathcal{E}_\rho$ can be expressed as the square of the $\mathcal{L}^2_{\rho_X}$-distance from f_ρ plus a constant,

$$\mathcal{E}_\rho(f) = \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \mathcal{E}_\rho(f_\rho). \tag{2.1}$$

Thus, the search for a function minimizing $\mathcal{E}_\rho$ over any subset M of $\mathcal{L}^2_{\rho_X}(X)$ is equivalent to the search of a closest function in M to the regression function f_ρ.

Similar to the expected error, the empirical one also can be expressed in terms of a distance functional. For a sample $z = \{(u_i, v_i) \in X \times Y \mid i = 1, \ldots, m\}$, let $X_u = \{u_1, \ldots, u_m\}$ and $h_z : X_u \to Y$ be defined as

$$h_z(u_i) = v_i. \tag{2.2}$$

For any $X \subseteq \mathbb{R}^d$ containing X_u and $f : X \to \mathbb{R}$, let

$$f_u = f_{|X_u} : X_u \to \mathbb{R}$$

denote f restricted to X_u. Then $\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^m (f(u_i) - v_i)^2 = \frac{1}{m}\sum_{i=1}^m (f_u(u_i) - h_z(u_i))^2 = \frac{1}{m}\|f_u - h_z\|^2_{l^2} = \mathcal{E}_z(f_u)$. So the empirical error $\mathcal{E}_z$ can be expressed as $\frac{1}{m}$ times the square of the l^2-distance from h_z:

$$\mathcal{E}_z(f) = \frac{1}{m}\|f_u - h_z\|^2_{l^2}. \tag{2.3}$$

3 Rates of Convergence of Minima of Error Functionals

Representations (2.1) and (2.3) allow us to take advantage of tools from approximation theory to prove the existence of functions minimizing error functionals over sets of functions computable by networks with n Heaviside perceptrons and to estimate rates of convergence of these minima in terms of a certain "regularity" of data with respect to perceptrons.

Similar to the expected error $\mathcal{E}_\rho$, the empirical one $\mathcal{E}_z$ also achieves its global minimum over the whole space $\mathcal{L}^2_{\rho_X}(X)$. This minimum is equal to zero and for a sample z of a size m, $\mathcal{E}_z$ achieves this minimum at a function computable by a network with m perceptrons. This follows from a result by Ito (1992), who proved that any function defined on a finite subset of $\mathbb{R}^d$ of size m can be represented as a function computable by a network with m hidden perceptrons with any sigmoidal activation (in particular, with the Heaviside function). The following theorem is a corollary of Ito's result:

Theorem 1. *For all positive integers d, m, all $X \subseteq \mathbb{R}^d$ and all samples of data $z = \{(u_i, v_i) \in X \times \mathbb{R} \mid i = 1, \ldots, m\}$ with all u_i distinct, there exists a function $f^o \in span_m H_d(X)$ such that*

$$\mathcal{E}_z(f^o) = \min_{f \in span_m H_d(X)} \mathcal{E}_z(f) = 0.$$

However in applications, networks with a smaller number n of hidden units than the size m of the training set are used. To show that for all $n < m$, both error functionals, $\mathcal{E}_\rho$ and $\mathcal{E}_z$, achieve minima over sets $span_n H_d(X)$ of functions computable by networks with n perceptrons, we utilize a weakened compactness property requiring subsequantial convergence merely for sequences minimizing distances to some functions. Let $(F, \|.\|)$ be a normed linear space, $M \subseteq F$, and $f \in F$; then by $\|f - M\| = \inf_{g \in M} \|f - g\|$ is denoted the distance of f from M. The subset M is called approximatively compact if for any sequence $\{g_k\}$ in M such that for some $f \in F$, $\lim_{k \to \infty} \|f - g_k\| = \|f - M\|$, there exists $g \in M$ to which $\{g_k\}$ converges subsequantially (Singer, 1970). Note that approximate compactness implies that every function in F has a best approximation in M and that M is closed.

Theorem 2. *Let d, m be positive integers, $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be both compact, $z = \{(u_i, v_i) \in X \times Y \mid i = 1, \ldots, m\}$ with all u_i distinct, $X_u = \{u_1, \ldots, u_m\}$, and ρ be a nondegenerate probability measure on $X \times Y$. Then for every n, there exist $f_{z,n}, f_n \in span_n H_d(X)$ such that:*

i. $\mathcal{E}_z(f_{z,n}) = \min_{f \in span_n H_d(X)} \mathcal{E}_z(f)$ *and* $h_{z,n} = f_{z,n \mid X_u}$ *satisfies* $\mathcal{E}_z(h_{z,n}) = \min_{f \in span_n H_d(X_u)} \mathcal{E}_z(f)$.

ii. $\mathcal{E}_\rho(f_n) = \min_{f \in span_n H_d(X)} \mathcal{E}_\rho(f)$.

To prove the theorem we first derive a discrete version of a result from (Kainen et al., 2003) on approximate compactness of $span_n H_d(X)$.

Theorem 3. *For all positive integers d, n, and any finite subset X of $\mathbb{R}^d$, $span_n H_d(X)$ is approximatively compact in $(l^2(X), \|.\|_{l^2})$.*

Proof. Let $X = \{x_1, \ldots, x_m\}$ and $f : X \to \mathbb{R}$ be such that $\|f - span_n H_d(X)\|_{l^2} = \lim_{k\to\infty} \|f - h_n^k\|_{l^2}$, where for each k, $h_n^k(x) = \sum_{i=1}^n w_i^k \vartheta(v_i^k \cdot x + b_i^k)$. As $\vartheta(t) = \vartheta(at)$ for all $a > 0$, without loss of generality, we can assume that $(v_i^k, b_i^k) = (v_{i1}^k, \ldots, v_{id}^k, b_i^k) \in S^d$ (here S^d denotes the unit sphere in $\mathbb{R}^{d+1}$). By compactness of S^d, there exist $v_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ such that $\{v_i^k\}$ converges subsequantially to v_i and $\{b_i^k\}$ converges subsequantially to b_i. Replacing these two sequences by suitable subsequences, we get by finiteness of X some k_0 such that for all $k \geq k_0$ and all $x_j \in X$, $\vartheta(v_i^k \cdot x_j + b_i^k) = \vartheta(v_i \cdot x_j + b_i)$. Hence $\|f - span_n H_d(X)\|_{l^2}^2 = \lim_{k\to\infty} \sum_{j=1}^m \left(\sum_{i=1}^n w_i^k \vartheta(v_i^k \cdot x_j + b_i^k) - f(x_j)\right)^2 = \lim_{k\to\infty} \sum_{j=1}^m \left(\sum_{i=1}^n w_i^k \vartheta(v_i \cdot x_j + b_j) - f(x_j)\right)^2 < \infty$.

Setting for all $j = 1, \ldots, m$, $I_j = \{i \in \{1, \ldots, n\} \mid \vartheta(v_i \cdot x_j + b_i) = 1\}$, we get $\lim_{k\to\infty} \sum_{j=1}^m \sum_{i\in I_j} \left(w_i^k - f(x_j)\right)^2 < \infty$. Thus, for all $j = 1, \ldots, m$, $\lim_{k\to\infty} \sum_{i\in I_j} \left(w_i^k - f(x_j)\right) < \infty$ and so $\lim_{k\to\infty} \sum_{i\in I_j} w_i^k = c_j < \infty$. By linearity of the mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ defined by $T(t_1, \ldots, t_n) = (\sum_{i\in I_1} t_i, \ldots, \sum_{i\in I_m} t_i)$, we prove, as in Kainen et al. (2003, lemma 3.2), that for all $i = 1, \ldots, n$ there exist $w_i \in \mathbb{R}$ such that $\lim_{k\to\infty} w_i^k = w_i$. Setting $g(x) = \sum_{i=1}^n w_i \vartheta(v_i \cdot x + b_i)$, we get $\|f - g\|_{l^2} = \|f - span_n H_d(X)\|_{l^2}$ and so $span_n H_d(X)$ is approximatively compact.

Proof of Theorem 2. (i) By theorem 3, $span_n H_d(X_u)$ is approximatively compact in $l^2(X_u)$. So $h_z : X_u \to \mathbb{R}$ defined as $h_z(u_i) = v_i$ has a best approximation $h_{z,n}(x) = \sum_{i=1}^n w_i \vartheta(e_i \cdot x + b_i)$ in $span_n H_d(X_u)$. Hence, $\|h_z - h_{z,n}\|_{l^2} = \|h_z - span_n H_d(X_u)\|_{l^2}$, and so by the representation (2.3), $\mathcal{E}_z$ achieves its minimum over $span_n H_d(X)$ at the extension $f_{z,n} \in span_n H_d(X)$ of $h_{z,n}$ defined as $f_{z,n}(x) = \sum_{i=1}^n w_i \vartheta(e_i \cdot x + b_i)$.

(ii) In Kainen et al. (2003), approximate compactness of $span_n H_d([0,1]^d)$ in $\mathcal{L}_\lambda^2([0,1]^d)$ with the Lebesque measure λ was proven. Inspection of the proof shows that $span_n H_d(X)$ is also approximately compact in $\mathcal{L}_\mu^2(X)$ for any compact subset X of $\mathbb{R}^d$ and any probabilistic measure μ on X, and in particular in $\mathcal{L}_{\rho_X}^2(X)$. Thus, the regression function f_ρ has a best approximation $f_n \in span_n H_d(X)$ and by the representation (2.1), $\mathcal{E}_\rho$ achieves its minimum over $span_n H_d(X)$ at f_n.

So for all n, both error functionals, $\mathcal{E}_\rho$ and $\mathcal{E}_z$, achieve minima over sets of functions computable by networks with n hidden Heaviside perceptrons. As the sets $span_n H_d(X)$ are not convex, such minima need not be unique.

Efficiency of utilization of networks with a smaller number of hidden units than the size of the training set depends on the speed of convergence of these minima to the global minima over $\mathcal{L}^2_{\rho_X}(X)$, which for $\mathcal{E}_\rho$ is equal to $\mathcal{E}_\rho(f_\rho)$ and for $\mathcal{E}_z$ to 0.

To estimate this speed, we use a result from nonlinear approximation theory, Maurey-Jones-Barron's theorem (Pisier, 1981; Jones, 1992; Barron, 1992, 1993), which implies (see Kůrková, 2003) that for every bounded subset G of a Hilbert space $(F, \|.\|)$, every $f \in F$ and every positive integer n

$$\|f - span_n G\| \leq \frac{s_G \|f\|_G}{\sqrt{n}}, \tag{3.1}$$

where $s_G = \sup_{g \in G} \|g\|$ and $\|f\|_G$ is a norm of f called G-variation. This norm is 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, that is,

$$\|f\|_G = \inf \left\{ c > 0 \mid c^{-1} f \in cl\, conv\, (G \cup -G) \right\}, \tag{3.2}$$

where the closure cl is taken with respect to the topology generated by the norm $\|.\|$ and $conv$ denotes the convex hull. Note that G-variation can be infinite (when the set on the right-hand side is empty). It was defined in Kůrková (1997) as an extension of variation with respect to half-spaces defined for $G = H_d(X)$ in Barron (1992) (for the properties of variation, see Kůrková, 2003).

We denote

$$\|.\|_{H_d(X),\mathcal{L}^2}$$

the variation with respect to $H_d(X) \subset (\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2_{\rho_X}})$ with $X \subset \mathbb{R}^d$ compact,

$$\|.\|_{H_d(X),\sup}$$

the variation with respect to $H_d(X) \subset (\mathcal{M}(X), \|.\|_{\sup})$, where $\mathcal{M}(X)$ denotes the space of all bounded measurable functions on $X \subseteq \mathbb{R}^d$ and $\|.\|_{\sup}$ denotes the supremum norm, and

$$\|.\|_{H_d(X)}$$

the variation with respect to $H_d(X) \subset l^2(X)$ with X finite (in this case, we do not need to specify the norm, as all norms on a finite-dimensional space are topologically equivalent).

The following theorem estimates speed of convergence of the minima of error functionals over sets of functions computable by networks with n Heaviside perceptrons.

Theorem 4. *Let d, m, n be positive integers, $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be compact, $z = \{(u_i, v_i) \in X \times Y \mid i = 1, \ldots, m\}$ with all u_i distinct, ρ be a nondegenerate probability measure on $X \times Y$ and f_n and $f_{z,n}$ be minimum points of the functionals $\mathcal{E}_\rho$, $\mathcal{E}_z$, resp., over $span_n H_d(X)$, and $h_{z,n}$ be a minimum point of $\mathcal{E}_z$ over $span_n H_d(X_u)$. Then:*

i. $\mathcal{E}_\rho(f_n) - \mathcal{E}_\rho(f_\rho) \leq \frac{\|f_\rho\|^2_{H_d(X),\mathcal{L}^2}}{n}$

ii. $\mathcal{E}_z(h_{z,n}) \leq \frac{\|h_z\|^2_{H_d(X_u)}}{n}$

iii. for every $h \in \mathcal{M}(X)$ interpolating the sample z, $\mathcal{E}_z(f_{z,n}) \leq \frac{\|h\|^2_{H_d(X),sup}}{n}$

Proof. i. By the representation 2.1, $\mathcal{E}_\rho(f_n) - \mathcal{E}_\rho(f_\rho) = \|f_\rho - f_n\|^2_{\mathcal{L}^2_{\rho_X}}$. As $\sup_{g \in H_d(X)} \|g\|_{\mathcal{L}^2_{\rho_X}} = 1$, by equation 3.1, $\|f_\rho - f_n\|_{\mathcal{L}^2_{\rho_X}} = \|f_\rho - span_n H_d(X)\|_{\mathcal{L}^2_{\rho_X}} \leq \frac{\|f_\rho\|_{H_d(X),\mathcal{L}^2}}{\sqrt{n}}$. Hence, $\mathcal{E}_\rho(f_n) - \mathcal{E}_\rho(f_\rho) \leq \frac{\|f_\rho\|^2_{H_d(X),\mathcal{L}^2}}{n}$.

ii. By theorem 2, $f_{z,n}|X_u = f_z$ is the best approximation in $span_n H(X_u)$ to h_z defined on X_u as $h_z(u_i) = v_i$. As $\sup_{h \in H_d(X_u)} \|h\|_{l^2} = \sqrt{m}$, by equation 3.1, $\|h_z - f_z\|_{l^2} = \|h_z - span_n H_d(X_u)\|_{l^2} \leq \sqrt{\frac{m}{n}} \|h_z\|_{H_d(X_u)}$. Hence, by equation 2.3, $\mathcal{E}_z(f_{z,n}) = \mathcal{E}_z(f_z) = \frac{1}{m}\|f_z - h_z\|^2_{l^2} \leq \frac{\|h_z\|^2_{H_d(X_u)}}{n}$.

iii. It follows from the definition of variation that for every $h \in \mathcal{M}(X)$, which interpolates the sample z, $\|h_z\|_{H_d(X_u)} \leq \|h\|_{H_d(X),\sup}$ and as $\mathcal{E}_z(f_{z,n}) = \mathcal{E}_z(h_{z,n})$, iii follows from ii.

So the minima of error functionals achievable over networks with n perceptrons decrease at least as fast as $\frac{1}{n}$ times the square of the variational norm of the regression function (or some interpolating function, resp.). When these norms are small, good approximations of the two global minima, $\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(f_\rho)$ and $\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_z(f) = 0$, can be obtained using networks with a moderate number of units.

The upper bounds from theorem 4(ii) and 4(iii) on speed of convergence of the minima of $\mathcal{E}_z$ over $span_n H_d(X)$ do not explicitly depend on the size m of the sample z. Independence from this size can be illustrated by an example of the gaussian function and a sample of data z chosen from it.

Proposition 1. *For every positive integer m and every odd positive integer d, every $X \subset \mathbb{R}^d$ compact and every sample z of the size m such that the function h_z defined as $h_z(u_i) = v_i$ is the restriction of the gaussian function $\gamma_d(x) = exp(-\|x\|^2)$ to $X_u = \{u_1, \ldots, u_m\}$,*

$$\min_{f \in span_n H_d(X)} \mathcal{E}_\rho(f) \leq \frac{4d^2}{n}.$$

Proof. Kainen, Kůrková, and Vogt (2007) proved that for d odd, $\|\gamma_d\|_{H_d(\mathbb{R}^d),\sup} \leq 2d$ (see also Cheang & Barron, 2000) for a weaker estimate depending on the size of X, which is valid also for d even). So by theorem 4 (iii), $\min_{f \in span_n H_d(X)} \mathcal{E}_\rho(f) \leq \frac{4d^2}{n}$.

Proposition 1 gives some insight into the relationship between two geometrically opposite types of computational units: gaussian radial basis functions (RBF) and Heaviside perceptrons. Minima of the empirical error functional $\mathcal{E}_z$ defined by any sample of data z chosen from the gaussian RBF unit with d inputs over networks with n Heaviside perceptrons converge to zero faster than $\frac{4d^2}{n}$. Note that the upper bound $\frac{4d^2}{n}$ grows with the dimension d only quadratically, and it does not depend on the size m of a sample.

On the other hand, there exist samples z, the sizes of which influence the magnitudes of the variations of the functions h_z defined as $h_z(u_i) = v_i$. For example, for any positive integer k, consider $X = [0, 2k]$, $Y = [-1, 1]$ and the sample $z = \{(2i, 1), (2i + 1, -1) \mid i = 0, \ldots, k - 1\}$ of the size $m = 2k$. Then one can easily verify that $\|h_z\|_{H_d(X)} = 2k$ (for functions of one variable, variation with respect to half-spaces is up to a constant equal to their total variation; see Barron, 1992; Kůrková et al., 1997). This example indicates that the more the data "oscillate," the larger the variation of functions, which interpolate them. In the next section, we investigate such oscillatory behavior in terms of $\mathcal{L}^1$-norms of iterated partial derivatives of the order d of interpolating functions.

4 Regularity of Samples of High-Dimensional Data

By theorem 4, fast convergence of error functionals is guaranteed for such samples z, which can be interpolated by functions with small variational norms. Reciprocals of squares of these norms can play roles of measures of certain regularity of data with respect to perceptron networks. The smaller the norms, the more regular are the data.

More insight into such regularity of data can be obtained from a relationship between the variational norm, magnitudes of partial derivatives, and the number of variables. For a function f that can be represented as an integral of the form of a neural network with a "continuum" of Heaviside perceptrons,

$$f(x) = \int_{S^{d-1} \times \mathbb{R}} w_f(e, b) \vartheta(e \cdot x + b) d\nu db, \tag{4.1}$$

variation with respect to half-spaces is bounded from above by the $\mathcal{L}^1_\lambda$-norm of the weighting function w_f (Kůrková et al., 1997).

For all compactly supported functions $f \in \mathcal{C}^d(\mathbb{R}^d)$ with d odd, the integral representation 4.1 was derived in Kůrková et al. (1997) with $w_f = a_d \int_{H_{e,b}} (D_e^{(d)} f)(y) dy$, where $a_d = (-1)^{k-1}(1/2)(2\pi)^{1-d}$, $D_e^{(d)} f$ denotes the directional derivative of f in the direction e iterated d times, de is the $(d-1)$-dimensional volume element on S^{d-1}, and dy is likewise on a hyperplane $H_{e,b} = \{x \in \mathbb{R}^d \mid x \cdot e + b = 0\}$. The representation 4.1 was extended to rapidly vanishing functions in Kainen, Kůrková, and Vogt (2006). Thus, the $\mathcal{L}^1_\lambda$-norm of w_f is bounded from above by the product of a function $k(d) \sim \left(\frac{4\pi}{d}\right)^{1/2} \left(\frac{e}{2\pi}\right)^{d/2}$, which is decreasing exponentially fast with the number of variables d, and a Sobolev seminorm defined as

$$\|f\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}^1_\lambda(\mathbb{R}^d)},$$

where λ denotes the Lebesque measure, $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index with nonnegative integer components, $D^\alpha = (\partial/\partial x_1)^{\alpha_1} \ldots (\partial/\partial x_d)^{\alpha_d}$ and $|\alpha| = \alpha_1 + \cdots + \alpha_d$ (see Kainen et al., 2007). So we have

$$\|f\|_{H_d(\mathbb{R}^d),\sup} \leq k(d) \|f\|_{d,1,\infty}. \tag{4.2}$$

Note that for large d, the seminorm $\|f\|_{1,d,\infty}$ is much smaller than the Sobolev norm,

$$\|f\|_{d,1} = \sum_{|\alpha| \leq d} \|D^\alpha f\|_{\mathcal{L}^1_\lambda(\mathbb{R}^d)},$$

as instead of the summation of 2^d iterated partial derivatives of f over all α with $|\alpha| \leq d$, merely their maximum over α with $|\alpha| = d$ is taken. A function f is in the ball of radius r in the seminorm $\|.\|_{d,1,\infty}$ if all the $\mathcal{L}^1_\lambda$-norms of its iterated partial derivatives $D^\alpha f$ with $|\alpha| = d$ are at most r, while it is in the ball of radius r in the Sobolev norm $\|.\|_{d,1}$ if the sum of 2^d terms (the $\mathcal{L}^1_\lambda$-norms of all its iterated partial derivatives with $|\alpha| \leq d$) is bounded by r. For example, in the case when all these $\mathcal{L}^1_\lambda$-norms are the same, they cannot be larger than $\frac{r}{2^d}$. So for large d, such function is almost constant, as its first derivatives are very small. On the other hand, balls in the seminorm $\|.\|_{1,d,\infty}$ contain many more functions, as the bound r applies merely to the $\mathcal{L}^1_\lambda$-norms of the derivatives D^α with $|\alpha| = d$.

Using relationship 4.2 between the variation with respect to half-spaces and the Sobolev seminorm, we get the following upper bound on the speed of decrease of minima of the empirical error functional over sets of functions computable by networks with n Heaviside perceptrons. It is formulated for functions sufficiently rapidly vanishing at infinity, to which the integral representation 4.1 was extended in Kainen, Kůrková, and Vogt (2000b) and Kainen et al. (2006). A function $f \in \mathcal{C}^d(\mathbb{R}^d)$ is said to be of a

weakly controlled decay if for every α, $0 \le |\alpha| < d$, $\lim_{\|x\|\to\infty}(D^\alpha f)(x) = 0$ and there exists $\varepsilon > 0$ satisfying for each multi-index α with $|\alpha| = d$, $\lim_{\|x\|\to\infty}(D^\alpha f)(x)\|x\|^{d+1+\varepsilon} = 0$. Note that the class of functions with weakly controlled decay contains all d-times continuously differentiable functions with compact support as well as all functions from the Schwartz class $\mathcal{S}(\mathbb{R}^d)$ (all $\mathcal{C}^\infty(\mathbb{R}^d)$ functions, which with all their iterated partial derivatives are rapidly decreasing (Strichartz, 2003; Adams & Fournier, 2003). In particular, it contains the gaussian function $\gamma_d(x) = \exp(-\|x\|^2)$.

Theorem 5. *For all positive integers n, m, all odd positive integers d, every compact subset X of $\mathbb{R}^d$, every sample $z = \{(u_i, v_i) \in X \times \mathbb{R} \mid i = 1, \ldots, m\}$ with all u_i distinct, and every $h : \mathbb{R}^d \to \mathbb{R}$ of a weakly controlled decay interpolating the sample z,*

$$\min_{f \in span_n H_d(\mathbb{R})} \mathcal{E}_z(f) \le \frac{c(d)\|h\|^2_{d,1,\infty}}{n},$$

where $c(d) \sim \frac{4\pi}{d}\left(\frac{e}{2\pi}\right)^d < \frac{4\pi}{d2^d}$.

Proof. It follows from Kainen et al. (2007), (theorems 3.3 and 4.2 and corollary 3.4) that for all d odd and all h of a weakly controlled decay

$$\|h\|_{H_d(\mathbb{R}^d),\sup} \le k(d)\|h\|_{d,1,\infty},$$

where $k(d) \sim \left(\frac{4\pi}{d}\right)^{1/2}\left(\frac{e}{2\pi}\right)^{d/2}$. So by theorem 4(iii), the statement follows.

Thus, for any sample of data z, which can be interpolated by a function $h \in \mathcal{C}^d(\mathbb{R}^d)$ vanishing sufficiently quickly at infinity with maxima of the $\mathcal{L}^1_\lambda$-norms of its partial derivatives of the order $|\alpha| = d$, which do not exceed an exponentially increasing upper bound,

$$\|h\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}^1_\lambda(\mathbb{R}^d)} \le \frac{1}{k(d)} \sim \left(\frac{d}{4\pi}\right)^{1/2}\left(\frac{2\pi}{e}\right)^{d/2} < \left(\frac{d}{\pi}2^{d-2}\right)^{1/2},$$

the minima of the empirical error $\mathcal{E}_z$ over networks with n Heaviside perceptrons decrease to zero rather quickly—at least as fast as $\frac{1}{n}$. So, for example, when for $d > 4\pi$, all $\mathcal{L}^1_\lambda$-norms of the partial derivatives of the order d are smaller than $2^{d/2}$, convergence faster than $\frac{1}{n}$ is guaranteed.

Theorem 5 gives some quantitative insight into the role of smoothness in preventing the "curse of dimensionality" as suggested by Barron (1993) and Vapnik (1995). Indeed, existence of partial derivatives up to the order equal to the dimension d is among assumptions in theorem 5 (a function of a quickly controlled decay of d variables is in $\mathcal{C}^d(\mathbb{R}^d)$), but the magnitudes of these derivatives measured by their $\mathcal{L}^1_\lambda$-norms can be rather large. They

can even increase exponentially fast with the dimension d. The more the data "oscillate," the larger are the magnitudes of derivatives of functions that interpolate them. But even for rather "oscillatory" data, for which all interpolating functions have these magnitudes as large as $\left(\frac{d}{\pi}2^{d-2}\right)^{1/2}$, the empirical error functional converges with the rate faster than $\frac{1}{n}$.

So rather than smoothness expressed merely in terms of the existence of higher-order derivatives, a kind of regularity expressed in terms of magnitudes of these derivatives plays an essential role in the growth of network complexity with the number of variables. A good match to training data (a sufficiently small minima of error functionals) is theoretically achievable by reasonably small networks even when these magnitudes grow exponentially fast with the number of variables.

5 Regularity of Samples Defined by Boolean Functions

Samples with interesting properties can be obtained from Boolean functions. Choose a linear ordering

$$\{u_1, \ldots, u_{2^d}\}$$

of the set of vectors from $\{0,1\}^d$. For a real-valued Boolean function $h : \{0,1\}^d \to \mathbb{R}$, define a sample of pairs of data,

$$z_h = \{(u_i, v_i) \mid i = 1, \ldots, 2^d\}, \quad \text{where} \quad v_i = h(u_i). \tag{5.1}$$

A function $h : \{0,1\}^d \to \mathbb{R}$ is called symmetric when for all $x, y \in \{0,1\}^d$, $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ implies $h(x) = h(y)$ (Savický, 1994). So the values of a symmetric function are invariant under permutations of entries of vectors in $\{0,1\}^d$. The next upper bound on the speed of convergence of minima of the empirical error functional defined by a sample generated by a symmetric function is based on a modification of an estimate of variation with respect to signum perceptrons of a symmetric function from Kůrková, Savický, and Hlaváčková (1998).

Proposition 2. *For every odd positive integer d and every symmetric function $h : \{0,1\}^d \to \{-1,1\}$,*

$$\min_{f \in span_n H_d(\{0,1\}^d)} \mathcal{E}_{z_h}(f) \leq \frac{(2d+1)^2}{n}.$$

Proof. As h is symmetric, there exists $\phi : \{0, \ldots, d\} \to \{-1, 1\}$ such that h can be represented as $h(x) = \phi(\sum_{i=1}^d x_i) = \phi(x \cdot 1)$, where 1 denotes the vector $(1, \ldots, 1) \in \{0,1\}^d$. It is easy to check that $h(x) = \phi(0) + \sum_{j=1}^d (\phi(j) -$

$\phi(j-1))\vartheta(x\cdot 1-j)$. Thus, $\|h\|_{H_d(\{0,1\}^d)} \le |\phi(0)| + \sum_{j=1}^{d} |\phi(j)-\phi(j-1)| \le 2d+1$ and by theorem 4(ii), the statement follows.

So samples z_h obtained from symmetric functions $h : \{0,1\}^d \to \mathbb{R}$ are quite regular with respect to perceptron networks. Minima of the empirical error functionals $\mathcal{E}_{z_h}$ over $span_n H_d(\{0,1\}^d)$ are smaller than or equal to $\frac{(2d+1)^2}{n}$. Note that this upper bound grows with d only quadratically.

On the other hand, there exist samples of data that have high irregularity with respect to perceptron networks. We prove their existence using covering numbers in the angular pseudometrics δ_F defined on the unit ball S_1 of any Hilbert space $(F, \|.\|)$ as

$$\delta_F(f,g) = \arccos|f\cdot g|.$$

This pseudometrics defines 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).

Let $S_r = \{f \in F \mid \|f\| = r\}$ denote the sphere of the radius r in $(F, \|.\|)$. For $\alpha > 0$, a subset A of S_1 is an α-net with respect to δ_F if for every $f \in S_1$ there exists $g \in A$ such that $\delta_F(f,g) < \alpha$. For F finite dimensional, let $\mathcal{N}_\alpha(S_1)$ denote the α-covering number of S_1 with respect to δ_F, that is, the size of the smallest α-net in S_1.

For $g \in F$, let $g^o = \frac{g}{\|g\|}$, and for $G \subseteq F$, let $G^o = \{g^o \mid g \in G\}$. 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 the sphere S_r of the radius equal to the supremum of the norms of elements of G, there exists a function with a "large" G-variation.

Proposition 3. *Let G be a bounded subset of a finite-dimensional Hilbert space $(F, \|.\|)$ with $\sup_{g\in G}\|g\| = r$ and $\alpha \in [0, \pi/2]$ be such that card $G < \mathcal{N}_\alpha(S_1)$ with respect to the pseudometrics δ_F. Then there exists $f \in S_r$ such that*

$$\|f\|_G \ge \frac{1}{cos\alpha}.$$

Proof. As card G^o = card $G < \mathcal{N}_\alpha(S_1)$, there exist some $h \in S_1$ such that for all $g^o \in G^o$, $\delta_F(h, g^o) \ge \alpha$ and $|h\cdot g^o| \le \cos\alpha$.

It follows from a geometric characterization of G-variation derived in Kůrková et al. (1998, theorem 2.2) that for all $f \in F$, $\|f\|_G \ge \frac{\|f\|^2}{\sup_{g\in G}|g\cdot f|} = \frac{\|f\|}{\sup_{g\in G}(|g^o\cdot f^o|\,\|g\|)}$. In particular, for $f = rh$, $\|f\|_G \ge \frac{r}{\sup_{g\in G}(|g^o\cdot h|\,\|g\|)} \ge \frac{r}{r\cos\alpha} = \frac{1}{\cos\alpha}$.

So functions that have small inner products with all elements of G (i.e., are "almost orthogonal" to G) have large G-variations. Thus, for every set G of a smaller cardinality than the α-covering number of the unit sphere in the angular pseudometrics δ_F, there exists a function with the magnitudes of its G-variation at least $\frac{1}{\cos\alpha}$.

Applying proposition 3 to the space $(\mathbb{R}^{2^d}, \|.\|_{l^2})$, which is equivalent to the space of all real valued functions on $\{0,1\}^d$, we prove the existence of Boolean real-valued functions with large variations with respect to half-spaces. First we state a corollary of proposition 3 for subsets G of the unit sphere S^{m-1} in $(\mathbb{R}^m, \|.\|_{l^2})$ that have cardinality smaller than the maximal number of pairwise nearly orthogonal vectors in $\mathbb{R}^m$. For $\varepsilon > 0$, two vectors $u, v \in S^{m-1}$ are ε-quasiorthogonal if

$$|u \cdot v| \leq \varepsilon \|u\|_{l^2} \|v\|_{l^2}.$$

The largest number of ε-quasiorthogonal vectors in S^{m-1} was called in Kainen and Kůrková (1993) the ε- quasiorthogonal dimension of $\mathbb{R}^m$ and denoted

$$\dim_\varepsilon m.$$

Corollary 1. *Let m be a positive integer, $\varepsilon > 0$ and G be a bounded subset of $\mathbb{R}^m$ such that* $\operatorname{card} G < dim_\varepsilon m$ *and* $\sup_{g\in G} \|g\| = r$. *Then there exists $f \in S_r$ with*

$$\|f\|_G \geq \frac{1}{\varepsilon}.$$

Proof. It follows from the definition of δ_F that $\dim_\varepsilon m \leq \mathcal{N}_{\arccos\varepsilon}(S^{m-1})$. Hence, $\operatorname{card} G < \mathcal{N}_{\arccos\varepsilon}(S_r)$ and so by proposition 3, there exists $f \in S^{m-1}$ such that $\|f\|_G \geq \frac{1}{\varepsilon}$.

The next theorem utilizes a lower bound on the quasiorthogonal dimension from Kainen and Kůrková (1993), which guarantees for a fixed ε an exponential growth of $\dim_\varepsilon m$ with m increasing. We prove the existence of a sample z_h obtained from a function $h : \{0,1\}^d \to \mathbb{R}$ with $H_d(\{0,1\}^d)$-variation depending on the number of variables exponentially by comparing the "large" size of $\dim_\varepsilon 2^m$ with the "small" size of the set of the characteristic functions of half-spaces intersected with the Boolean cube $\{0,1\}^d$. It follows from a classical result by Shläfli (1950) that the cardinality of the set $H_d(\{0,1\}^d)$ is smaller than 2^{d^2} (so it is rather small in contrast to the size 2^{2^d} of the set of all subspaces of the Boolean cube $\{0,1\}^d$).

Theorem 6. *For every positive integer d, there exists a function $h : \{0,1\}^d \to \mathbb{R}$ such that for every $X \subseteq \mathbb{R}^d$ with $\{0,1\}^d \subseteq X$ and every $f : X \to \mathbb{R}$ interpolating the sample $z_h = \{(u_i, v_i) \mid i = 1, \ldots, 2^d\} \subset \{0,1\}^d \times \mathbb{R}$, where $v_i = h(u_i)$,*

$$\|f\|_{H_d(X),sup} \geq \|h\|_{H_d(\{0,1\}^d)} \gtrsim \frac{2^{(d-1)/2}}{d\sqrt{ln\, 2}}.$$

Proof. Kainen and Kůrková (1993) showed that for all positive integers m and all $\varepsilon > 0$,

$$\dim_\varepsilon m \geq \frac{2^{m-1}}{B(m, \lambda_{m,\varepsilon})},$$

where $\lambda_{m,\varepsilon} = \lceil \frac{m(1-\varepsilon)}{2} - 1 \rceil$ and $B(m, \lambda_{m,\varepsilon}) = \sum_{i=0}^{\lambda_{m,\varepsilon}} \binom{m}{i}$. By Fine (1999) and by Stirling's formula, $\frac{2^{m-1}}{B(m,\lambda_{m,\varepsilon})} \approx e^{m\varepsilon^2/2}$. So

$$\dim_\varepsilon m \gtrsim e^{m\varepsilon^2/2},$$

and in particular $\dim_\varepsilon 2^d \gtrsim e^{2^{d-1}\varepsilon^2}$.

On the other, hand by Shläfli (1950),

$$\text{card } H_d(\{0,1\}^d) \leq 2^{d^2 - d\log_2 d + \mathcal{O}(d)} < 2^{d^2}.$$

So by corollary 1, for every $\varepsilon > 0$ for which $2^{d^2} \leq e^{2^{d-1}\varepsilon^2} \lesssim \dim_\varepsilon 2^d$, there exists a function $h \in S^{2^d - 1}$ with $\|h\|_{H_d(\{0,1\}^d)} \geq \frac{2^{(d-1)/2}}{d\sqrt{\ln 2}}$.

Thus, for every $f : X \to \mathbb{R}$ interpolating the sample z_h, $\|f\|_{H_d(X),\sup} \geq \|h\|_{H_d(\{0,1\}^d)} \geq \frac{2^{(d-1)/2}}{d\sqrt{\ln 2}}$.

Note that for d odd, every $f \in \mathcal{C}^d(\mathbb{R}^d)$ of a weakly controlled decay, which interpolates the sample z_h from theorem 6, must have large Sobolev seminorm, because by equation 4.2, $\|f\|_{1,d,\infty} \geq \frac{2^{(d-1)/2}}{k(d)d\sqrt{\ln 2}} \gtrsim \left(\frac{2^{2d-3}\pi^{d-1}}{\ln 2 d e^d}\right)^{1/2}$.

The proof of theorem 6 is existential, but in Kůrková et al. (1998), 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."

6 Discussion

We have proved the existence of the minima of the expected and the empirical error functionals over networks with n Heaviside perceptrons and derived upper bounds on rates of convergence of these minima to the global ones over $\mathcal{L}^2_{\rho_X}(X)$. Our bounds are of the form $\frac{1}{n}$ times the squares of the

maxima of the $\mathcal{L}^1_\lambda$-norms of the iterated partial derivatives of the order d of functions, at which the global minima are achieved (the regression function in the case of the expected error and any function interpolating the sample of data in the case of the empirical one), multiplied by an exponentially quickly decreasing function $c(d) \lesssim \frac{\pi}{d} 2^{2-d}$ of the number of variables d.

These bounds give quantitative insight into the role of dimensionality in learning of neural networks. They imply conditions on data that guarantee that a good approximation of global minima of error functionals can be achieved using networks with limited complexity. Our estimates suggest that the Sobolev seminorm $\|f\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}^1_\lambda(\mathbb{R}^d)}$ of the regression or some interpolating function f can play a role in measuring oscillatory behavior of data. With an increasing dimension d, even data for which this norm is increasing exponentially fast but does not exceed $\frac{\pi}{d} 2^{2-d}$ can be processed efficiently by networks with a reasonable number of perceptrons.

The best-approximation property of the sets $span_n H_d$ is rather exceptional among sets of linear combinations of n perceptrons. For many activation functions ψ, the closures of the sets $span_n G_\psi$ (where G_ψ denotes the set of functions computable by perceptrons with the activation function ψ) contain polynomials of degrees increasing with n. Leshno et al. (1993) used these polynomials to prove the universal approximation property of perceptron networks with quite general nonpolynomial activations.

Nevertheless, for any bounded set G, a weaker version of theorem 4 holds with the values of the functionals $\mathcal{E}_\rho$ and $\mathcal{E}_z$ at their minimum points replaced with the infima of these functionals over the sets $span_n G$.

The reason for restricting our focus in this letter to the case of the Heaviside activation function is in the application of the integral representation $f(x) = \int_{S^{d-1}\times\mathbb{R}} w_f(e,b)\vartheta(e \cdot x + b)dvdb$. We do not know whether an analogous formula also holds for sigmoidal plane waves.

A preliminary version of some of the results from this letter were published in conference proceedings (Kůrková, 2005).

Acknowledgments

This work was partially supported by GA ČR grant 201/05/0557 and the Institutional Research Plan AV0Z10300504. I thank P. C. Kainen and A. Vogt for fruitful discussions.

References

Adams, R. A., & Fournier, J. J. F. (2003). *Sobolev spaces*. Orlando, FL: Academic Press.

Barron, A. R. (1992). Neural net approximation. In K. Narenda (Ed.), *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems* (pp. 69–72). New Haven, CT: Yale University Press.

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

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

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of AMS, 39*, 1–49.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control and Signals Systems, 2*, 303–314.

Fine, T. L. (1999). *Feedforward neural networks methodology*. New York: Springer-Verlag.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multi-layer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Ito, Y. (1991). Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. *Neural Networks, 4*, 385–394.

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

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

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

Kainen, P. C., Kůrková, V., & Vogt, A. (1999). Approximation by neural networks is not continuous. *Neurocomputing, 29*, 47–56.

Kainen, P. C., Kůrková, V., & Vogt, A. (2000a). Geometry and topology of continuous best and near best approximations. *Journal of Approximation Theory, 105*, 252–262.

Kainen, P. C., Kůrková, V., & Vogt, A. (2000b). An integral formula for Heaviside neural networks. *Neural Network World, 10*, 313–320.

Kainen, P. C., Kůrková, V., & Vogt, A. (2003). Best approximation by linear combinations of characteristic functions of half-spaces. *Journal of Approximation Theory, 122*, 151–159.

Kainen, P. C., Kůrková, V., & Vogt, A. (2006). Integral combinations of Heavisides. Submitted, Research report ICS–966. Available online at http://www.cs.cas.cz/research/publications.shtml.

Kainen, P. C., Kůrková, V., & Vogt, A. (2007). A Sobolev-type upper bound for rates of approximation by linear combinations of plane waves. *Journal of Approximation Theory, 147*, 1–10.

Kůrková, V. (1997). Dimension-independent rates of approximation by neural networks. In K. Warwick & M. Kárný (Eds.), *Computer-intensive methods in control and signal processing: The curse of dimensionality* (pp. 261–270). Boston: Birkhauser.

Kůrková, V. (2003). High-dimensional approximation and optimization by neural networks. In J. Suykens, G. Horvath, S. Basu, C. Micchelli, & J. Vandewalle (Eds.), *Advances in learning theory: Methods, models and applications* (pp. 69–88). Amsterdam: IOS Press.

Kůrková, V. (2005). Minimization of empirical error functional over perceptron networks. In B. Ribeiro, R. F. Albrecht, A. Dobnikar, D. W. Pearson, & N. C. Steele (Eds.), *Adaptive and natural computing algorithms* (pp. 46–49). Berlin: Springer-Verlag.

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

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.

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a non-polynomial activation can approximate any function. *Neural Networks, 6*, 861–867.

Makovoz, Y. (1996). Random approximants and neural networks. *Journal of Approximation Theory, 85*, 98–109.

Makovoz, Y. (1998). Uniform approximation by neural networks. *Journal of Approximation Theory, 95*, 215–228.

Mhaskar, H. N., & Micchelli, C. A. (1992). Approximation by superposition of a sigmoidal function. *Advances in Applied Mathematics, 13*, 350–373.

Pinkus, A. (1998). Approximation theory of the MPL model in neural networks. *Acta Numerica, 8*, 277–283.

Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. In *Séminaire d'Analyse Fonctionnelle 1980–81*. Palaiseau, France: Ecole Polytechnique, Centre de Mathématiques.

Savický, P. (1994). On the bent Boolean functions that are symmetric. *European Journal of Combinatorics, 15*, 145–168.

Shläfli, L. (1950). *Gesamelte mathematische abhandlungen*. Basel: Birkhäuzer.

Singer, I. (1970). *Best approximation in normed linear spaces by elements of linear subspaces*. Berlin: Springer-Verlag.

Strichartz, R. S. (2003). *A guide to distribution theory and Fourier transforms*. Singapore: World Scientific.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Werbos, P. J. (1985). Backpropagation: Basics and new developments. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 134–139). Cambridge, MA: MIT Press.

Received May 5, 2006; accepted January 19, 2007.