Best approximation by linear combinations of characteristic functions of half-spaces *

Paul C. Kainen
Department of Mathematics
Georgetown University
Washington, D.C. 20057-1233, USA

Věra Kůrková
Institute of Computer Science, Academy of Sciences of the Czech Republic
182 07 Prague 8, Czech Republic

Andrew Vogt
Department of Mathematics
Georgetown University
Washington, D.C. 20057-1233, USA

*Journal of Approximation Theory, Volume 122, Number 2, June 2003, 151-159.

Proposed running head:

Half-space characteristic functions

Name and address of corresponding author:

Andrew Vogt
Department of Mathematics
Georgetown University
Washington, D.C. 20057-1233, USA

e-mail:vogt@math.georgetown.edu

Abstract

It is shown that for any positive integer n and any function in $\mathcal{L}_p([0,1]^d)$ with $p \in [1, \infty)$ there exists a best approximation by linear combinations of n characteristic functions of half-spaces. Further, sequences of such linear combinations converging in distance to the best approximation distance have subsequences converging to the best approximation, i.e., these linear combinations are an approximatively compact set.

Keywords. Best approximation, proximinal, approximatively compact, boundedly compact, Heaviside perceptron networks, plane waves.

1 Introduction

An important type of nonlinear approximation is *variable-basis approximation*, where the set of approximating functions is formed by linear combinations of n functions from a given set. This approximation scheme has been widely investigated: it includes splines with free nodes, trigonometric polynomials with free frequencies, sums of wavelets, and feedforward neural networks.

To estimate rates of variable-basis approximation, it is helpful to study properties like existence, uniqueness, and continuity of corresponding approximation operators.

Here we investigate the existence property for one-hidden-layer Heaviside perceptron networks, i.e., approximations by linear combinations of characteristic functions of closed half-spaces. Such functions are obtained by composing the Heaviside function with affine functions. We show that for all positive integers n, d in $\mathcal{L}_p([0,1]^d)$ with $p \in [1, \infty)$ there exists a best approximation mapping to the set of functions computable by Heaviside perceptron networks with n hidden and d input units. Thus for any p-integrable function on $[0,1]^d$ there is a linear combination of n characteristic functions of closed half-spaces that is nearest in the $\mathcal{L}_p$-norm. A related proposition is proved by Chui, Li, and Mhaskar in [1], where certain sequences are shown to have subsequences that converge a. e. These authors work in $\mathcal{R}^d$ rather than $[0,1]^d$ and show a. e. convergence rather than $\mathcal{L}_p$ convergence.

2 Heaviside perceptron networks

Feedforward networks compute parametrized sets of functions dependent both on the type of computational units and their interconnections. *Computational units* compute functions of two vector variables: an *input vector* and a *parameter* vector. A standard type of computational unit is the perceptron. A *perceptron* with an *activation function* $\psi : \mathcal{R} \to \mathcal{R}$ (where $\mathcal{R}$ denotes the set of real numbers) computes real-valued functions on $\mathcal{R}^d \times \mathcal{R}^{d+1}$ of the form $\psi(\mathbf{v} \cdot \mathbf{x} + b)$, where $\mathbf{x} \in \mathcal{R}^d$ is an *input* vector, $\mathbf{v} \in \mathcal{R}^d$ is an *input weight* vector, and $b \in \mathcal{R}$ is a *bias*.

The most common activation functions are sigmoidals, i.e., functions with ess-shaped graph. Both continuous and discontinuous sigmoidals are used. Here we study networks based on the archetypal discontinuous sigmoidal, namely, the *Heaviside function* ϑ defined by $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$.

Let H_d denote the set of functions on $[0,1]^d$ computable by Heaviside perceptrons, i.e.,

$$H_d = \{f : [0,1]^d \to \mathcal{R} : f(\mathbf{x}) = \vartheta(\mathbf{v} \cdot \mathbf{x} + b), \mathbf{v} \in \mathcal{R}^d, b \in \mathcal{R}\}.$$

H_d is the set of characteristic functions of closed half-spaces of $\mathcal{R}^d$ restricted to $[0,1]^d$, which is a subset of the set of plane waves (see, e.g., Courant and Hilbert [2, pp.676–681]). For $A \subseteq \mathcal{R}^d$ we denote by ξ_A the characteristic function of A, i.e., $\xi_A(x) = 1$ for $x \in A$ and $\chi_A(x) = 0$ for $x \notin A$.

The simplest type of multilayer feedforward network has one hidden layer and one linear output. Such networks with Heaviside perceptrons in the hidden layer compute functions of the form

$$\sum_{i=1}^{n} w_i \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i),$$

where n is the number of hidden units, $w_i \in \mathcal{R}$ are output weights, and $\mathbf{v}_i \in \mathcal{R}^d$ and $b_i \in \mathcal{R}$ are input weights and biases respectively.

The set of all such functions is the *set of all linear combinations of* n *elements of* H_d and is denoted by $span_n H_d$.

It is known that for all positive integers d, $\cup_{n \in \mathcal{N}_+} span_n H_d$ (where $\mathcal{N}_+$ denotes the set of all positive integers) is dense in $(\mathcal{C}([0,1]^d), \|.\|_\mathcal{C})$, the linear space of all continuous functions on $[0,1]^d$ with the supremum norm, as well as in $(\mathcal{L}_p([0,1]^d), \|.\|_p)$ with $p \in [1, \infty]$ (see, e.g., Mhaskar and Micchelli [10] or Leshno et al. [9]). We study best approximation in $span_n H_d$ for a fixed n.

3 Existence of a best approximation

Existence of a best approximation has been formalized in approximation theory by the concept of proximinal set (sometimes also called "existence" set). A subset M of a normed linear space $(X, \|.\|)$ is called *proximinal* if for every $f \in X$ the distance $\|f - M\| = \inf_{g \in M} \|f - g\|$ is achieved for some element of M, i.e., $\|f - M\| = \min_{g \in M} \|f - g\|$ (Singer [13]). Clearly a proximinal subset must be closed.

A sufficient condition for proximinality of a subset M of a normed linear space $(X, \|.\|)$ is compactness (i.e., each sequence of elements of M has a subsequence convergent to an element of M). Indeed, for each $f \in X$ the functional $e_{\{f\}} : M \to \mathcal{R}$ defined by $e_{\{f\}}(m) = \|m - f\|$ is continuous [13, p. 391] and hence must achieve its minimum on any compact set M.

Gurvits and Koiran [5] have shown that for all positive integers d the set of characteristic functions of half-spaces H_d is compact in $(\mathcal{L}_p([0,1]^d), \|.\|_p)$ with $p \in [1, \infty)$. This can be easily verified once the set H_d is reparametrized by elements of the unit sphere S^d in $\mathcal{R}^{d+1}$. Indeed, a function $\vartheta(\mathbf{v} \cdot \mathbf{x} + b)$, with the vector $(v_1, \ldots, v_d, b) \in \mathcal{R}^{d+1}$ nonzero, is equal to $\vartheta(\hat{\mathbf{v}} \cdot \mathbf{x} + \hat{b})$, where $(\hat{v}_1, \ldots, \hat{v}_d, \hat{b}) \in S^d$ is obtained from $(v_1, \ldots, v_d, b) \in \mathcal{R}^{d+1}$ by normalization. Strictly speaking, H_d is parametrized by equivalence classes in S^d since different parametrizations may represent the same member of H_d when restricted to $[0,1]^d$. Since S^d is compact, and the quotient space formed by the equivalence classes is likewise, so is H_d.

However, by extending H_d into $span_n H_d$ for any positive integer n we lose compactness since the norms are not bounded. Nevertheless compactness can be replaced by a weaker property that requires only some sequences to have convergent subsequences. A subset M of a normed linear space $(X, \|.\|)$ is called *approximatively compact* if for each $f \in X$ and any sequence $\{g_i : i \in \mathcal{N}_+\}$ in M such that $\lim_{i \to \infty} \|f - g_i\| = \|f - M\|$, there exists $g \in M$ such that $\{g_i : i \in \mathcal{N}_+\}$ converges subsequentially to g [13, p.368].

The following theorem shows that $span_n H_d$ is approximatively compact in $\mathcal{L}_p$-spaces. It extends a weaker result by Kůrková [8], who showed that $span_n H_d$ is closed in $\mathcal{L}_p$-spaces with $p \in (1, \infty)$.

Theorem 3.1 *For every n, d positive integers and for every $p \in [1, \infty)$ $span_n H_d$ is an approximatively compact subset of $(\mathcal{L}_p([0,1]^d, \|.\|_p)$.*

To prove the theorem we need the following lemma. For a set A $\mathcal{P}(A)$ denotes the set of all subsets of A.

Lemma 3.2 *Let m be a positive integer, $\{a_{jk} : k \in \mathcal{N}_+, j = 1, \ldots, m\}$ be m sequences of real numbers, and $\mathcal{S} \subseteq \mathcal{P}(\{1, \ldots, m\})$ be such that for each $S \in \mathcal{S}$ $\lim_{k \to \infty} \sum_{j \in S} a_{jk} = c_S$ for some $c_S \in \mathcal{R}$. Then there exist real numbers $\{a_j : j = 1, \ldots, m\}$ such that for each $S \in \mathcal{S}$ $\sum_{j \in S} a_j = c_S$.*

Proof. Let $p = card\,\mathcal{S}$ and let $\mathcal{S} = \{S_1, \ldots, S_p\}$. Define $T : \mathcal{R}^m \to \mathcal{R}^p$ by $T(x_1, \ldots, x_m) = (\sum_{j\in S_1} x_j, \ldots, \sum_{j\in S_p} x_j)$. Then T is linear, and hence its range is a subspace of $\mathcal{R}^p$ and so is a closed set. Since $(c_{S_1}, \ldots, c_{S_p}) \in cl\,T(\mathcal{R}^m) = T(\mathcal{R}^m)$, there exists $(a_1, \ldots, a_m) \in \mathcal{R}^m$ with $(c_{S_1}, \ldots, c_{S_p}) = T(a_1, \ldots, a_m)$. □

Proof of Theorem 3.1

Let $f \in \mathcal{L}_p([0,1]^d)$ and let $\{\sum_{j=1}^n a_{jk}g_{jk} : k \in \mathcal{N}_+\}$ be a sequence of elements of $span_n H_d$ such that $\lim_{k\to\infty} \|f - \sum_{j=1}^n a_{jk}g_{jk}\|_p = \|f - span_n H_d\|_p$. Since H_d is compact, by passing to suitable subsequences we can assume that for all $j = 1, \ldots, n$, there exist $g_j \in H_d$ such that $\lim_{k\to\infty} g_{jk} = g_j$ (here and in the sequel, we use the notation $\lim_{k\to\infty}$ to mean a limit of a suitable subsequence).

We shall show that there exist real numbers $a_1, \ldots, a_n$ such that

$$\|f - span_n H_d\|_p = \|f - \sum_{j=1}^n a_j g_j\|_p. \tag{1}$$

Then using (1) we shall show even that $\{\sum_{j=1}^n a_{jk}g_{jk} : k \in \mathcal{N}_+\}$ converges to $\sum_{j=1}^n a_j g_j$ in $\|.\|_p$ subsequentially.

Decompose $\{1, \ldots, n\}$ into two disjoint subsets I and J such that I consists of those j for which the sequences $\{a_{jk} : k \in \mathcal{N}_+\}$ have convergent subsequences, and J of those j for which the sequences $\{|a_{jk}| : k \in \mathcal{N}_+\}$ diverge. Again, by passing to suitable subsequences we can assume that for all $j \in I$, $\lim_{k\to\infty} a_{jk} = a_j$. Thus $\{\sum_{j\in I} a_{jk}g_{jk} : k \in \mathcal{N}_+\}$ converges subsequentially to $\sum_{j\in I} a_j g_j$.

Set $h = f - \sum_{j\in I} a_j g_j$. Since for all $j \in I$, the chosen subsequences $\{a_{jk} : k \in \mathcal{N}_+\}$ and $\{g_{jk} : k \in \mathcal{N}_+\}$ are bounded, we have $\|f - span_n H_d\|_p = \lim_{k\to\infty} \|f - \sum_{j=1}^n a_{jk}g_{jk}\|_p = \lim_{k\to\infty} \|h - \sum_{j\in J} a_{jk}g_{jk}\|_p$.

Let $\mathcal{S}$ denotes the set of all subsets of J. Decompose $\mathcal{S}$ into two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathcal{S}_1$ consists of those $S \in \mathcal{S}$ for which by passage to suitable subsequences $\lim_{k\to\infty} \sum_{j\in S} a_{jk} = c_S$ for some $c_S \in \mathcal{R}$, and $\mathcal{S}_2$ consists of those $S \in \mathcal{S}$ for which $\lim_{k\to\infty} |\sum_{j\in S} a_{jk}| = \infty$. Note that the empty set is in $\mathcal{S}_1$ with the convention $\sum_{j\in\emptyset} a_{jk} = 0$.

Using Lemma 3.2, for all $j \in \cup\mathcal{S}_1$, we get $a_j \in \mathcal{R}$ such that for all $S \in \mathcal{S}_1$, $\sum_{j\in\mathcal{S}} a_j = c_S$. For $j \in J - \cup\mathcal{S}_1$, set $a_j = 0$.

Since $\sum_{j=1}^n a_j g_j \in span_n H_d$, we have $\|f - span_n H_d\|_p \le \|f - \sum_{j=1}^n a_j g_j\|_p$ and thus to prove (1), it is sufficient to show that $\|f - span_n H_d\|_p \ge \|f - \sum_{j=1}^n a_j g_j\|_p$ or equivalently

$$\lim_{k\to\infty} \int_{[0,1]^d} \left| h - \sum_{j\in J} a_{jk}g_{jk} \right|^p d\mu \ge \int_{[0,1]^d} \left| h - \sum_{j\in J} a_j g_j \right|^p d\mu \tag{2}$$

where μ is Lebesgue measure on $[0,1]^d$.

To verify (2), for each $k \in \mathcal{N}_+$ we shall decompose the integration over $[0,1]^d$ into sum of integrals over convex regions where the functions $\sum_{j\in J} a_{jk}g_{jk}$ are constant. To describe such regions, we shall define partitions of $[0,1]^d$ determined by families of characteristic functions $\{g_{jk} : j \in J, k \in \mathcal{N}_+\}$, and $\{g_j : j \in J\}$. The partitions are indexed by the elements of the set $\mathcal{S}$ of all subsets of J. For $k \in \mathcal{N}_+$, a partition $\{T_k(S) : S \in \mathcal{S}\}$ is defined by $T_k(S) = \{x \in [0,1]^d : (g_{jk}(x) = 1 \Leftrightarrow j \in S)\}$, and similarly a partition $\{T(S) : S \in \mathcal{S}\}$ is defined by $T(S) = \{x \in [0,1]^d : g_j(x) = 1 \Leftrightarrow j \in S\}$. Notice that since for all $j = 1, \ldots, n$,

$\lim_{k\to\infty} g_{jk} = g_j$ in $\mathcal{L}_p([0,1]^d)$, we have $\lim_{k\to\infty} \mu(T_k(S)) = \mu(T(S))$ for all $S \in \mathcal{S}$. Indeed, the characteristic function of $T_k(S)$ equals the product $\prod_{j\in S} g_{jk} \prod_{j\notin S}(1-g_{jk})$ and converges in $\mathcal{L}_p([0,1]^d)$ to the characteristic function of $T(S)$, the latter equal to $\prod_{j\in S} g_j \prod_{j\notin S}(1-g_j)$.

Using the definition of $T_k(S)$ (in particular its property guaranteeing that for all $S \in \mathcal{S}$, $T_k(S)$ is just the region where for all $j \in S$ and no other $j \in J$, g_{jk} is equal to 1), we get

$$\lim_{k\to\infty} \int_{[0,1]^d} \left| h - \sum_{j\in J} a_{jk} g_{jk} \right|^p d\mu = \lim_{k\to\infty} \sum_{S\in\mathcal{S}} \int_{T_k(S)} \left| h - \sum_{j\in S} a_{jk} \right|^p d\mu =$$

$$\lim_{k\to\infty} \left(\sum_{S\in\mathcal{S}_1} \int_{T_k(S)} \left| h - \sum_{j\in S} a_{jk} \right|^p d\mu + \sum_{S\in\mathcal{S}_2} \int_{T_k(S)} \left| h - \sum_{j\in S} a_{jk} \right|^p d\mu \right) \geq$$

$$\lim_{k\to\infty} \sum_{S\in\mathcal{S}_1} \int_{T_k(S)} \left| h - \sum_{j\in S} a_{jk} \right|^p d\mu. \tag{3}$$

Since for all $S \in \mathcal{S}$, $\lim_{k\to\infty} \mu(T_k(S)) = \mu(T(S))$ and for all $S \in \mathcal{S}_1$, $\lim_{k\to\infty} \sum_{j\in S} a_{jk} = c_S = \sum_{j\in S} a_j$, we have

$$\lim_{k\to\infty} \sum_{S\in\mathcal{S}_1} \int_{T_k(S)} \left| h - \sum_{j\in S} a_{jk} \right|^p d\mu = \lim_{k\to\infty} \sum_{S\in\mathcal{S}_1} \int_{T_k(S)} \left| h - \sum_{j\in S} a_j \right|^p d\mu =$$

$$\sum_{S\in\mathcal{S}_1} \int_{T(S)} \left| h - \sum_{j\in S} a_j \right|^p d\mu.$$

For all $S \in \mathcal{S}$, by the triangle inequality in $\mathcal{L}_p(T_k(S))$

$$\lim_{k\to\infty} \left(\int_{T_k(S)} \left| \sum_{j\in S} a_{jk} \right|^p d\mu \right)^{1/p} \leq$$

$$\lim_{k\to\infty} \left(\left(\int_{T_k(S)} \left| h - \sum_{j\in S} a_{jk} \right|^p d\mu \right)^{1/p} + \left(\int_{T_k(S)} |h|^p d\mu \right)^{1/p} \right) \leq$$

$$\lim_{k\to\infty} \left(\int_{[0,1]^d} \left| h - \sum_{j\in J} a_{jk} g_{jk} \right|^p d\mu \right)^{1/p} + \left(\int_{[0,1]^d} |h|^p d\mu \right)^{1/p} = \|f - span_n H_d\|_p + \|h\|_p.$$

Thus for all $S \in \mathcal{S}$, $\lim_{k\to\infty} \int_{T_k(S)} |\sum_{j\in S} a_{jk}|^p d\mu$ is finite. In particular this is true when $S \in \mathcal{S}_2$, for which $\lim_{k\to\infty} |\sum_{j\in S} a_{jk}|^p = \infty$, and so $\lim_{k\to\infty} \mu(T_k(S)) = 0 = \mu(T(S))$ for $S \in \mathcal{S}_2$. Thus we can replace the integration over $\cup_{S\in\mathcal{S}_1} T(S)$ by the integration over the whole of $[0,1]^d$ and so we obtain

$$\sum_{S\in\mathcal{S}_1}\int_{T(S)}\left|h-\sum_{j\in S}a_j\right|^p d\mu=\int_{[0,1]^d}\left|h-\sum_{j\in J}a_jg_j\right|^p d\mu,$$

which proves (2). Moreover, as a byproduct we even get that

$$\lim_{k\to\infty}\sum_{S\in\mathcal{S}_2}\int_{T_k(S)}\left|h-\sum_{j\in S}a_{jk}\right|^p d\mu=0, \tag{4}$$

since in (3) the left hand side is equal to the right hand side (both are equal to $\|f - span_n H_d\|_p$).

So we have shown that $span_n H_d$ is proximinal. Now we shall verify that it is even approximatively compact by showing that $\{\sum_{j\in J}a_{jk}g_{jk} : k\in\mathcal{N}_+\}$ converges subsequentially to $\sum_{j\in J}a_jg_j$, or equivalently

$$\lim_{k\to\infty}\int_{[0,1]^d}\left|\sum_{j\in J}(a_{jk}g_{jk}-a_jg_j)\right|^p d\mu=0. \tag{5}$$

As above, we start by decomposing the integration into sum of integrals over convex regions. The left hand side of (5) is equal to

$$\lim_{k\to\infty}\sum_{S\in\mathcal{S}_1}\int_{T_k(S)}\left|\sum_{j\in S}(a_{jk}-a_jg_j)\right|^p d\mu+\sum_{S\in\mathcal{S}_2}\int_{T_k(S)}\left|\sum_{j\in S}(a_{jk}-a_jg_j)\right|^p d\mu.$$

Using the triangle inequality, (4), and $\lim_{k\to\infty}\mu(T_k(S))=0$ for all $S\in\mathcal{S}_2$, we get

$$\lim_{k\to\infty}\sum_{S\in\mathcal{S}_2}\int_{T_k(S)}\left|\sum_{j\in S}(a_{jk}-a_jg_j)\right|^p d\mu\leq$$

$$\lim_{k\to\infty}\left(\sum_{S\in\mathcal{S}_2}\int_{T_k(S)}\left|h-\sum_{j\in S}a_{jk}\right|^p d\mu+\sum_{S\in\mathcal{S}_2}\int_{T_k(S)}\left|h-\sum_{j\in S}a_jg_j\right|^p d\mu\right)=$$

$$\lim_{k\to\infty}\sum_{S\in\mathcal{S}_2}\int_{T_k(S)}\left|h-\sum_{j\in S}a_jg_j\right|^p d\mu=\sum_{S\in\mathcal{S}_2}\int_{T(S)}\left|h-\sum_{j\in S}a_jg_j\right|^p d\mu=0$$

since $\mu(T(S))=0$ for $S\in\mathcal{S}_2$.

Thus $\lim_{k\to\infty}\sum_{S\in\mathcal{S}_2}\int_{T_k(S)}|\sum_{j\in S}(a_{jk}-a_jg_j)|^p d\mu=0$, which implies that the left hand side of (5) is equal to

$$\lim_{k\to\infty}\sum_{S\in\mathcal{S}_1}\int_{T_k(S)}\left|\sum_{j\in S}(a_{jk}-a_jg_j)\right|^p d\mu=\sum_{S\in\mathcal{S}_1}\int_{T(S)}|\sum_{j\in S}a_jg_j-a_jg_j|^p d\mu=0$$

because $(\sum_{j\in S}a_{jk}g_{jk})\chi_{T_k(S)}=(\sum_{j\in S}a_{jk})\chi_{T_k(S)}$ converges to $c_S\chi_{T(S)}=(\sum_{j\in S}a_jg_j)\chi_{T(S)}$ in $\mathcal{L}_p([0,1]^d)$.

So $\lim_{k\to\infty}\sum_{j\in J} a_{jk}g_{jk} = \sum_{j\in J} a_j g_j$, the same is already known to be true when J is replaced by I, and hence also $\lim_{k\to\infty}\sum_{j=1}^n a_{jk}g_{jk} = \sum_{j=1}^n a_j g_j$ subsequentially in $\mathcal{L}_p([0,1]^d)$. □

Theorem 3.1 shows that a function in $\mathcal{L}_p([0,1]^d)$ has a best approximation among functions computable by one-hidden-layer networks with a single linear output unit and n Heaviside perceptrons in the hidden layer. In other words, in the space of parameters of networks of this type, there exists a global minimum of the error functional defined as $\mathcal{L}_p$-distance from the function to be approximated.

Combining Theorem 3.1 with results from [7], we get the following corollary.

Corollary 3.3 *In $(\mathcal{L}_p([0,1]^d), \|.\|_p)$ with $p \in (1,\infty)$ for all n, d positive integers there exists a best approximation mapping from $\mathcal{L}_p([0,1]^d)$ to $span_n H_d$, but no such mapping is continuous.*

4 Discussion

In Proposition 3.3 of [1] the authors show that any sequence $\{P_k\}$ in $span_n H_d$, with the property that $\limsup_k \|P_k\|_{\mathcal{L}_1(K)} \leq 1$ for every compact set K in $\mathcal{R}^d$, has a subsequence converging a. e. in $\mathcal{R}^d$ to a member of $span_n H_d$. Although the proof techniques in [1] do have some overlap with those used here, the results there are different. A. e. convergence need not imply $\mathcal{L}_p$ convergence for $p \in [1,\infty)$: the sequence $P_k = (k)^{\frac{1}{p}}\chi_{[0,\frac{1}{k}]}$ converges a. e. in $\mathcal{L}_p(\mathcal{R}^1)$ but has no convergent subsequence in the $\mathcal{L}_p$-norm. Since this sequence is bounded and has no convergent subsequences, it also illustrates that $span_n H_d$ is not boundedly compact. Another example of an approximatively compact set that is not boundedly compact is any closed infinite-dimensional subspace of a uniformly convex Banach space.

Theorem 3.1 cannot be extended to perceptron networks with differentiable activation functions, e.g., the logistic sigmoid or hyperbolic tangent. For such functions, sets $span_n P_d(\psi)$ (where $P_d(\psi) = \{f : [0,1]^d \to \mathcal{R} : f(\mathbf{x}) = \psi(\mathbf{v}\cdot\mathbf{x} + b), \mathbf{v} \in \mathcal{R}^d, b \in \mathcal{R}\}$) are not closed and hence cannot be proximinal. This was first observed by Girosi and Poggio [4] and later exploited by Leschno et al. [9] for a proof of the universal approximation property.

Theorem 3.1 does not offer any information on the error of the best approximation. Estimates in the literature (DeVore, Howard, and Micchelli [3], Pinkus [11], Pinkus [12]) that give lower bounds on such errors and depend on continuity of best approximation operators are not applicable because of Corollary 3.3.

Acknowledgment

V. Kůrková was partially supported by GA ČR grants 201/99/0092 and 201/02/0428. Collaboration of V. Kůrková and P. C. Kainen was supported in part by an NRC COBASE grant.

References

[1] C. K. Chui, X. Li, and H. N. Mhaskar, Neural networks for localized approximation. *Math. of Computation* **63** (1994), 607–623.

[2] R. Courant and D. Hilbert, “Methods of Mathematical Physics”, vol. II, Wiley, New York, 1962.

[3] R. DeVore, R. Howard, and C. Micchelli, Optimal nonlinear approximation. *Manuscripta Math.* **63** (1989), 469–478.

[4] F. Girosi and T. Poggio, Networks and the best approximation property, *Biological Cybernetics* **63** (1990), 169–176.

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

[6] P. C. Kainen, V. Kůrková, and A. Vogt, Approximation by neural networks is not continuous. *Neurocomputing* **29** (1999), 47–65.

[7] P. C. Kainen, V. Kůrková, and A. Vogt, Geometry and topology of continuous best and near best approximations. *J. Approx. Theory* **105** (2000), 252–262.

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

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

[10] H. N. Mhaskar and C. Micchelli, Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Math.* **13** (1992), 350–373.

[11] A. Pinkus, “n-Width in Approximation Theory”, Springer-Verlag, Berlin, 1989.

[12] A. Pinkus, Approximation theory of the MLP model in neural networks. *Acta Numerica* **8** (1999), 143–195.

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