Available at WWW.MATHEMATICSWEB.ORG
POWERED BY SCIENCE @ DIRECT•

Journal of Approximation Theory 122 (2003) 151–159

JOURNAL OF Approximation Theory

http://www.elsevier.com/locate/jat

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

Paul C. Kainen,[a] Věra Kůrková,[b] and Andrew Vogt[a,*]

[a] *Department of Mathematics, Georgetown University, Box 571233, 37th and O Streets N.W., Washington, DC 20057-1233, USA*

[b] *Institute of Computer Science, Academy of Sciences of the Czech Republic, P.O. Box 5, 182 07 Prague 8, Czech Republic*

Received 23 February 2001; accepted in revised form 10 April 2003

Communicated by Will Light

Abstract

It is shown that for any positive integer n and any function f in $\mathscr{L}_p([0,1]^d)$ with $p \in [1, \infty)$ there exist n half-spaces such that f has a best approximation by a linear combination of their characteristic functions. Further, any sequence of linear combinations of n half-space characteristic functions converging in distance to the best approximation distance has a subsequence converging to a best approximation, i.e., the set of such n-fold linear combinations is 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.

*Corresponding author. Fax: +202-687-6067.
E-mail address: vogt@math.georgetown.edu (A. Vogt).

0021-9045/03/$ - see front matter © 2003 Published by Elsevier Science (USA).
doi:10.1016/S0021-9045(03)00072-8

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

We investigate the existence property for one-hidden-layer Heaviside perceptron networks. Here the approximations are by linear combinations of characteristic functions of closed half-spaces. (The characteristic function of any subset A is the function χ_A with value 1 on the subset and 0 elsewhere.) Such a characteristic function may also be described as a plane wave obtained by composing the Heaviside function with an affine function. 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 et al. [2], where certain sequences are shown to have subsequences that converge a.e. These authors work in 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 $t \geqslant 0$.

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., [3, pp. 676–681]).

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, $w_i \in \mathscr{R}$ are output weights, and $\mathbf{v}_i \in \mathscr{R}^d$ and $b_i \in \mathscr{R}$ 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, $\bigcup_{n \in \mathcal{N}_+} span_n\, H_d$ (where $\mathcal{N}_+$ denotes the set of all positive integers) is dense in $(\mathscr{C}([0,1]^d), \|.\|_{\mathscr{C}})$, the linear space of all continuous functions on $[0,1]^d$ with the supremum norm, as well as in $(\mathscr{L}_p([0,1]^d), \|.\|_p)$ with $p \in [1, \infty]$ (see, e.g., [11] or [10]). We study best approximation in $span_n\, H_d$ for a fixed n.

3. Best approximation and approximative compactness

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., $\|f - M\| = \min_{g \in M} \|f - g\|$ [14]. Clearly a proximinal subset must be closed.

A sufficient condition for proximinality of a subset M of a normed linear space $(X, \|.\|)$ is compactness. Indeed, for each $f \in X$ the functional $e_{\{f\}} : M \to \mathscr{R}$ defined by $e_{\{f\}}(m) = \|f - m\|$ is continuous [14, p. 391] and hence must achieve its minimum on the compact set M. Two generalizations of compactness also imply proximinality. A set M is called *boundedly compact* if the closure of its intersection with any bounded set is compact. A set M is called *approximatively compact* if for each $f \in X$ and any sequence $\{g_i\colon i \in \mathcal{N}_+\}$ in M such that $\lim_{i \to \infty} \|f - g_i\| = \|f - M\|$, there exists $g \in M$ such that $\{g_i\colon i \in \mathcal{N}_+\}$ converges subsequentially to g [14, p. 368]. Any closed, boundedly compact set is approximatively compact, and any approximatively compact set is proximinal [14, p. 374].

Gurvits and Koiran [6] have shown that for all positive integers d the set of characteristic functions of half-spaces H_d is compact in $(\mathscr{L}_p([0,1]^d), \|.\|_p)$ 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 $\mathscr{R}^{d+1}$. Indeed, a function $\vartheta(\mathbf{v} \cdot \mathbf{x} + b)$, with the vector $(v_1, \ldots, v_d, b) \in \mathscr{R}^{d+1}$ 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 $(v_1, \ldots, v_d, b) \in \mathscr{R}^{d+1}$ 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 spaces formed by the equivalence classes is likewise, so is H_d.

However, $span_n\, H_d$ is not compact for any positive integer n. Nor is it boundedly compact.

The following theorem shows that $span_n\, H_d$ is approximatively compact in $\mathscr{L}_p$-spaces. It extends a result of Kůrková [9], who showed that $span_n\, H_d$ is closed in $\mathscr{L}_p$-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 $(\mathscr{L}_p([0,1]^d, \|.\|_p)$.*

To prove the theorem we need the following lemma. For a set A, let $\mathscr{P}(A)$ denote the set of all subsets of A.

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

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

Proof of Theorem 3.1. Let $f \in \mathscr{L}_p([0,1]^d)$ and let $\{\sum_{j=1}^n a_{jk} g_{jk}\colon k \in \mathscr{N}_+\}$ 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 $j = 1, \dots, n$, 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 $a_1, \dots, a_n$ such that

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

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

Decompose $\{1, \dots, n\}$ into two disjoint subsets I and J such that I consists of those j for which the sequences $\{a_{jk}\colon k \in \mathscr{N}_+\}$ have convergent subsequences, and J of those j for which the sequences $\{|a_{jk}|\colon k \in \mathscr{N}_+\}$ 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 $\{\sum_{j \in I} a_{jk} g_{jk}\colon k \in \mathscr{N}_+\}$ 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 $\{a_{jk}\colon k \in \mathscr{N}_+\}$ and $\{g_{jk}\colon k \in \mathscr{N}_+\}$ 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 $\mathscr{S}$ denote the set of all subsets of J. Decompose $\mathscr{S}$ into two disjoint subsets $\mathscr{S}_1$ and $\mathscr{S}_2$ such that $\mathscr{S}_1$ consists of those $S \in \mathscr{S}$ for which by passage to suitable subsequences $\lim_{k\to\infty} \sum_{j \in S} a_{jk} = c_S$ for some $c_S \in \mathscr{R}$, and $\mathscr{S}_2$ consists of those $S \in \mathscr{S}$ for which $\lim_{k\to\infty} |\sum_{j \in S} a_{jk}| = \infty$. Note that the empty set is in $\mathscr{S}_1$ with the convention $\sum_{j \in \phi} = 0$.

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

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

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

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

To verify (2), for each $k \in \mathscr{N}_+$ we shall decompose the integration over $[0,1]^d$ into a 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 $\{g_{jk}\colon j \in J,\ k \in \mathscr{N}_+\}$, and $\{g_j\colon j \in J\}$. The partitions are indexed by the elements of the set $\mathscr{S}$ of all subsets of J. For $k \in \mathscr{N}_+$, a partition $\{T_k(S)\colon S \in \mathscr{S}\}$ is defined by $T_k(S) = \{x \in [0,1]^d\colon (g_{jk}(x) = 1 \Leftrightarrow j \in S)\}$, and similarly a partition $\{T(S)\colon S \in \mathscr{S}\}$ is defined by $T(S) = \{x \in [0,1]^d\colon g_j(x) = 1 \Leftrightarrow j \in S\}$. Note that since for all $j = 1, \ldots, n$, $\lim_{k \to \infty} g_{jk} = g_j$ in $\mathscr{L}_p([0,1]^d)$, we have $\lim_{k \to \infty} \mu(T_k(S)) = \mu(T(S))$ for all $S \in \mathscr{S}$. Indeed, the characteristic function of $T_k(S)$, $\chi_{T_k(S)}$, equals the product $\prod_{j \in S} g_{jk} \prod_{j \notin S} (1 - g_{jk})$ and converges in $\mathscr{L}_p([0,1]^d)$ to the characteristic function of $T(S)$, $\chi_{T(S)}$, 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 $S \in \mathscr{S}$, $T_k(S)$ is just the region where for all $j \in S$ and no other $j \in J$, g_{jk} is equal to 1), we get

$$\begin{aligned}
\lim_{k \to \infty} \int_{[0,1]^d} \left| h - \sum_{j \in J} a_{jk} g_{jk} \right|^p d\mu &= \lim_{k \to \infty} \sum_{S \in \mathscr{S}} \int_{T_k(S)} \left| h - \sum_{j \in S} a_{jk} \right|^p d\mu \\
&= \lim_{k \to \infty} \left(\sum_{S \in \mathscr{S}_1} \int_{T_k(S)} \left| h - \sum_{j \in S} a_{jk} \right|^p d\mu + \sum_{S \in \mathscr{S}_2} \int_{T_k(S)} \left| h - \sum_{j \in S} a_{jk} \right|^p d\mu \right) \\
&\geqslant \lim_{k \to \infty} \sum_{S \in \mathscr{S}_1} \int_{T_k(S)} \left| h - \sum_{j \in S} a_{jk} \right|^p d\mu. \qquad (3)
\end{aligned}$$

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

$$\begin{aligned}
\lim_{k \to \infty} \sum_{S \in \mathscr{S}_1} \int_{T_k(S)} \left| h - \sum_{j \in S} a_{jk} \right|^p d\mu &= \lim_{k \to \infty} \sum_{S \in \mathscr{S}_1} \int_{T_k(S)} \left| h - \sum_{j \in S} a_j \right|^p d\mu \\
&= \sum_{S \in \mathscr{S}_1} \int_{T(S)} \left| h - \sum_{j \in S} a_j \right|^p d\mu.
\end{aligned}$$

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

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

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

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

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

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

since in (3) the first expression is equal to the last (both are equal to $||f-span_n\, H_d||_p^p$).

So we have shown that $span_n\, H_d$ is proximinal. Now we shall verify that it is even approximatively compact by showing that $\{\sum_{j\in J} a_{jk}g_{jk}\colon k \in \mathscr{N}_+\}$ converges subsequentially to $\sum_{j\in J} a_j g_j$, 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 a sum of integrals over convex regions. The left-hand side of (5) is equal to

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

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

$$\begin{aligned}
&\lim_{k\to\infty} \sum_{S\in\mathscr{S}_2} \int_{T_k(S)} \left|\sum_{j\in S} (a_{jk} - a_j g_j)\right|^p d\mu \\
&\leqslant \lim_{k\to\infty} \left(\sum_{S\in\mathscr{S}_2} \int_{T_k(S)} \left| h - \sum_{j\in S} a_{jk} \right|^p d\mu + \sum_{S\in\mathscr{S}_2} \int_{T_k(S)} \left| h - \sum_{j\in S} a_j g_j \right|^p d\mu \right) \\
&= \lim_{k\to\infty} \sum_{S\in\mathscr{S}_2} \int_{T_k(S)} \left| h - \sum_{j\in S} a_j g_j \right|^p d\mu = \sum_{S\in\mathscr{S}_2} \int_{T(S)} \left| h - \sum_{j\in S} a_j g_j \right|^p d\mu = 0
\end{aligned}$$

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

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

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

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 $c_S \chi_{T(S)} = (\sum_{j\in S} a_j g_j)\chi_{T(S)}$ in $\mathscr{L}_p([0,1]^d)$.

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 $\mathscr{L}_p([0,1]^d)$. □

Theorem 3.1 shows that a function in $\mathscr{L}_p([0,1]^d)$ 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 $\mathscr{L}_p$-distance from the function to be approximated.

Combining Theorem 3.1 with Theorem 2.2 of [8] (see also [7]), we note that while such best approximation operators exist from $\mathscr{L}_p([0,1]^d)$ to $span_n\, H_d$, they cannot be continuous for $p \in (1, \infty)$.

4. Discussion

In Proposition 3.3 of [2] the authors show that any sequence $\{P_k\}$ in $span_n\, H_d$ (domain taken to be R^d here), with the property that $\limsup_k ||P_k||_{\mathscr{L}_1(K)} \leqslant 1$ for every compact set K in R^d, has a subsequence converging a.e. in R^d to a member of $span_n\, H_d$. Although the proof techniques in [2] do have some overlap with those used here, the results there are different. A.e. convergence need not imply $\mathscr{L}_p$ convergence for $p \in [1, \infty)$: the sequence $P_k = (k)^{\frac{1}{p}} \chi_A$, where $A = [0, \frac{1}{k}] \times [0,1]^{d-1}$, converges a.e. in $\mathscr{L}_p(R^d)$ but has no convergent subsequence in the $\mathscr{L}_p$-norm.

Since the sequence $\{P_k\}$ above is bounded and lies in $span_1\, H_d$ (with respect to $\mathscr{L}_p([0,1]^d)$), it also illustrates that $span_1\, H_d$, and hence $span_n\, H_d$, are 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 [14, pp. 368–369].

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 $P_d(\psi) = \{f : [0,1]^d \to \mathscr{R} : f(\mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x} + b), \mathbf{v} \in \mathscr{R}^d, b \in \mathscr{R}\}$) are not closed and hence cannot be proximinal. This was first observed by Girosi and Poggio [5] and later exploited by Leshno et al. [10] 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 [4,12,13]) that give lower bounds on such errors and depend on continuity of best approximation operators are not applicable by the remarks at the end of Section 3.

Cheang and Barron [1] show that linear combinations of characteristic functions of closed half-spaces with relatively few terms can yield good approximations of such functions as the characteristic function χ_B of a ball. However, χ_B is not approximated by the linear combination itself but rather by the characteristic function of the set where the linear combination exceeds a certain threshold. This amounts to replacing a linear output in the corresponding neural network by a threshold unit.

Acknowledgments

V. Kůrková was partially supported by GA ČR Grants 201/99/0092 and 201/02/0428. Collaboration of P. Kainen, V. Kůrková, and A. Vogt was supported by NRC COBASE grants and by a Faculty Development grant from Georgetown University.

References

[1] G.H.L. Cheang, A.R. Barron, A better approximation for balls, J. Approx. Theory 104 (2000) 183–203.
[2] C.K. Chui, X. Li, H.N. Mhaskar, Neural networks for localized approximation, Math. Comput. 63 (1994) 607–623.
[3] R. Courant, D. Hilbert, Methods of Mathematical Physics, Vol. II, Wiley, New York, 1962.
[4] R. DeVore, R. Howard, C. Micchelli, Optimal nonlinear approximation, Manuscripta Math. 63 (1989) 469–478.
[5] F. Girosi, T. Poggio, Networks and the best approximation property, Bio. Cybernet. 63 (1990) 169–176.
[6] L. Gurvits, P. Koiran, Approximation and learning of convex superpositions, J. Comput. System Sci. 55 (1997) 161–170.
[7] P.C. Kainen, V. Kůrková, A. Vogt, Geometry and topology of continuous best and near best approximations, J. Approx. Theory 105 (2000) 252–262.

[8] P.C. Kainen, V. Kůrková, A. Vogt, Continuity of approximation by neural networks in L^p spaces, Ann. Oper. Res. 101 (2001) 143–147.

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

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

[11] H.N. Mhaskar, C. Micchelli, Approximation by superposition of sigmoidal and radial basis functions, Adv. Appl. Math. 13 (1992) 350–373.

[12] A. Pinkus, n-Width in Approximation Theory, Springer, Berlin, 1989.

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

[14] I. Singer, Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces, Springer, Berlin, 1970.