Best Approximation by Linear Combinations of Characteristic Functions of Half-Spaces

Kainen, P.C.
1999

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

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

Paul Kainen   Věra Kůrková   Andrew Vogt

Technical report No. 795

Nov. 25, 1999

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodrenskou v 2, 182 07 Prague 8, Czech Republic
phone: (+4202) 6605 3231   fax: (+4202) 85 85 789
e-mail: vera@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

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

Paul Kainen[1] Věra Kůrková[2] Andrew Vogt[3]

Technical report No. 795
Nov. 25, 1999

Abstract

It is shown that in $\mathcal{L}_p$-spaces with $p \in [1, \infty)$ there exists a best approximation mapping to the set of functions computable by Heaviside perceptron networks with n hidden units; however for $p \in (1, \infty)$ such best approximation is not unique and cannot be continuous.

Keywords
One-hidden-layer networks, Heaviside perceptrons, best approximation, metric projection, continuous selection, approximatively compact.

[1]Dept. of Mathematics, Georgetown University, Washington, D.C. 20057
[2]Inst. of Computer Science, Acad. of Sci. of the Czech Republic
[3]Dept. of Mathematics, Georgetown University, Washington, D.C. 20057

1 Introduction

An important measure of the complexity of feedforward neural networks is the number of hidden units. To estimate accuracy of approximation achievable using networks with a fixed number of units, it is helpful to study properties like existence, uniqueness and continuity of approximation operators to sets of functions computable by such networks.

Here, we investigate such properties for one-hidden-layer Heaviside perceptron networks. We show that for all positive integers n, d there exists a best approximation mapping to the set of functions computable by Heaviside perceptron networks with n hidden and d input units; however, for $p \in (1, \infty)$ geometric properties (non-convexity) of sets of functions computable by such networks prevent these best approximations from being continuous.

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. Standard types of units are perceptrons. 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 $\mathcal{R}^{d+1} \times \mathcal{R}^d$ 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 discontinuous *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.

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

Notice that H_d is the set of *characteristic functions of half-spaces* of $\mathcal{R}^d$ restricted to $[0,1]^d$.

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

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

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, resp.

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., Cybenko, 1989; Hornik, Stinchcombe & White, 1989). However, for practical applications, the desired accuracy of approximation has to be achievable for n small enough to allow implementation. Thus it is useful to study approximation capabilities of the sets $span_n H_d$.

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., $\|f - M\| = \min_{g \in M} \|f - g\|$ (see e.g. Singer, 1970). 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 (see e.g. Singer, 1977, p. 391) and hence must achieve its minimum on any compact set M.

Gurvits & Koiran (1997) 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 reparameterized 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 parametrization 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, 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 $\{g_i; i \in \mathcal{N}_+\} \subseteq M$ 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 (see e.g. Singer, 1970, 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 (1995), 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 closed. Since $(c_{S_1}, \ldots, c_{S_p}) \in cl\,T(\mathcal{R}^m)$, there exists $(a_1, \ldots, a_p) \in \mathcal{R}^m$ with $(c_{S_1}, \ldots, c_{S_p}) = T(a_1, \ldots, a_p)$. □

Proof of Theorem 3.2 Let $f \in \mathcal{L}_p([0,1]^d)$ and $\{\sum_{j=1}^n a_{kj} g_{kj}; k \in \mathcal{N}_+\}$ be a sequence of elements of $span_n H_d$ such that $\lim_{k\to\infty} \|\sum_{j=1}^n a_{kj} g_{kj} - f\| = \|f - span_n H_d\|$.

Since $span_n H_d$ 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 $\lim_{k\to\infty} g_{kj} = g_j$. We shall show that there exist real numbers $a_1, \ldots, a_n$ such that $\|f - span_n H_d\| = \|f - \sum_{j=1}^n a_j g_j\|$.

Decompose $\{1, \ldots, n\}$ into two disjoint subsets F and $\bar{F}$ such that F consists of those j for which the sequence $\{a_{kj}; k \in \mathcal{N}_+\}$ has a convergent subsequence and $\bar{F}$ of those j for which the sequences $\{|a_{kj}|; k \in \mathcal{N}_+\}$ diverge. Again, by passing to suitable subsequences we can assume that $\lim_{k\to\infty} a_{kj} = a_j$ for all $j = 1, \ldots, n$.

Set $\hat{f} = f - \sum_{i\in F} a_j g_j$. Since for all $j \in F$ the chosen subsequences $\{a_{jk}; k \in \mathcal{N}_+\}$ and $\{g_{jk}; k \in \mathcal{N}_+\}$ are bounded, we have $\lim_{k\to\infty} \|\sum_{j=1}^n a_{jk} g_{kj} - f\| = \lim_{k\to\infty} \|\sum_{j\in F} a_j g_j - f + \sum_{j\in\bar{F}} a_{kj} g_{kj}\| = \lim_{k\to\infty} \|\sum_{j\in\bar{F}} a_{kj} g_{kj} - \hat{f}\|$.

Let $\mathcal{S}$ be the set of all subsets of $\bar{F}$. For each $k \in \mathcal{N}_+$ define a partition $\{T_k(S); S \in \mathcal{S}\}$ of $[0,1]^d$ by $T_k(S) = \{x \in [0,1]^d; (g_{kj}(x) = 1 \Leftrightarrow j \in S)\}$. Thus $\|\sum_{j\in\bar{F}} a_{kj} g_{kj} - \hat{f}\|^p = \sum_{S\in\mathcal{S}} \int_{T_k(S)} |\sum_{j\in S} a_{kj} - \hat{f}|^p d\mu$.

Decompose $\mathcal{S}$ into two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathcal{S}_1$ consists of those S for which passing to a suitable subsequences $\lim_{k\to\infty} \sum_{j\in S} a_{jk} = c_S$ for some $c_S \in \mathcal{R}$ and $\mathcal{S}_2$ of those S for which $\lim_{k\to\infty} |\sum_{j\in S} a_{jk}| = \infty$.

If $S \in \mathcal{S}_2$ then using triangle inequality we get $\lim_{k\to\infty} \int_{T_k(S)} |\sum_{j\in S} a_{kj} - \hat{f}|^p = \lim_{k\to\infty} |\sum_{j\in S} a_{jk}|^p \mu(T_k(S))$. Hence $\lim_{k\to\infty} \mu(T_k(S)) = 0$.

If $S \in \mathcal{S}_1$ then $\lim_{k\to\infty} \int_{T_k(S)} |\sum_{j\in S} a_{kj} - \hat{f}|^p d\mu = \lim_{k\to\infty} \int_{T_k(S)} |c_S - \hat{f}| d\mu$.

Using Lemma 3.1 we get $\{a_j, j \in \bar{F}\}$ such that for all $S \in \mathcal{S}$ $\sum_{j\in\mathcal{S}} a_j = c_S$. Hence $\lim_{k\to\infty} \|\hat{f} - \sum_{j\in S} a_{jk} g_{jk}\|^p = \|\hat{f} - \sum j \in S a_j g_j\|^p$.

Thus $\|f - span_n H_d\|^p = \lim_{k\to\infty} \|\bar{f} - \sum_{j\in\bar{F}} a_{jk} g_{jk}\|^p =$
$\sum_{S\in\mathcal{S}_1} \lim_{k\to\infty} |\sum_{j\in S} a_{kj}|^p \mu(T_k(S)) + \sum_{S\in\mathcal{S}_2} \lim_{k\to\infty} \int_{T_k(S)} |c_S - \hat{f}|^p d\mu$
$\geq \sum_{S\in\mathcal{S}_2} \lim_{k\to\infty} \int_{T_k(S)} |c_S - \hat{f}|^p d\mu = \lim_{k\to\infty} \|\sum_{j\bar{F}} a_j g_{kj} - \hat{f}\|^p$
$= \|\sum_{j\bar{F}} a_j g_j - \hat{f}\|^p = \|\sum_{j\bar{F}} a_j g_j + \sum_{j\in F} a_j g_j - f\|^p$.

Setting $h = \sum_{j\bar{F}} a_j g_j + \sum_{j\in F} a_j g_j \in span_n H_d$, we have $\|f - h\| = \|f - span_n H_d\|$. □

It is a straightforward consequence of the definitions that approximatively compact implies proximinal (see Singer, 1970).

Corollary 3.3 *For every n, d positive integers and for every $p \in [1, \infty)$ $span_n H_d$ is a proximinal subset of $(\mathcal{L}_p([0,1]^d), \|.\|_p)$.*

Thus, for any fixed number n of hidden units, 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.

4 Uniqueness and continuity of best approximation

Let M be a subset of a normed linear space $(X, \|.\|)$ and let $\mathcal{P}(M)$ denote the set of all subsets of M. The set-valued mapping $P_M : X \to \mathcal{P}(M)$ defined by $P_M(f) = \{g \in M; \|f - g\| = \|f - M\|\}$ is called the *metric projection of X onto M* and $P_M(f)$ is called the *projection of f onto M*.

Let $F : X \to \mathcal{P}(M)$ be a set-valued mapping. A *selection* from F is a mapping $\phi : X \to M$ such that for all $f \in X$, $\phi(f) \in F(f)$. A mapping $\phi : X \to M$ is called a *best approximation operator* from X to M if it is a selection from P_M.

When M is proximinal, then $P_M(f)$ is non-empty for all $f \in X$ and so there exists a best approximation mapping from X to M. The best approximation need not be unique. When it is unique, M is called a *Chebyshev set* (or "unicity" set). Thus M is Chebyshev if for all $f \in X$ the projection $P_M(f)$ is a singleton.

Let us recall that a normed linear space $(X, \|.\|)$ is called *strictly convex* if for all $f \neq g$ in X with $\|f\| = \|g\| = 1$ we have $\|\frac{f+g}{2}\| < 1$. It is well known that for all $p \in (1, \infty)$ $(\mathcal{L}_p([0,1]^d), \|.\|_p)$ is strictly convex.

In the previous section, we have noted that for all positive integers n, d and $p \in [1, \infty)$ there exists a best approximation mapping from $\mathcal{L}_p([0,1]^d)$ to $span_n H_d$. The following theorem implies for p in the open interval $(1, \infty)$ that if among such best approximations there is a continuous one, then best approximation is unique.

Theorem 4.1 *In a strictly convex normed linear space, any subset with a continuous selection from its metric projection is Chebyshev.*

For the proof and extensions to non-strictly convex spaces see Kainen, Kůrková & Vogt (1999), (2000).

To apply Theorem 4.1 to $span_n H_d$ we shall use the following geometric characterization of Chebyshev sets with continuous best approximation by Vlasov (1970).

Theorem 4.2 *In a Banach space with strictly convex dual, every Chebyshev subset with continuous metric projection is convex.*

It is well known that $\mathcal{L}_p$-spaces with $p \in (1, \infty)$ satisfy the assumptions of this theorem (since the dual of $\mathcal{L}_p$ is $\mathcal{L}_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $q \in (1, \infty)$). So it is sufficient to verify that $span_n H_d$ is not convex.

Proposition 4.3 *For all n, d positive integers, $span_n H_d$ is not convex.*

To verify nonconvexity of $span_n H_d$ consider $2n$ parallel half-spaces with the characteristic functions $g_i(\mathbf{x}) = \vartheta(\mathbf{v} \cdot \mathbf{x} + b_i)$, where $0 > b_1 > \ldots > b_{2n} > -1$ and $\mathbf{v} = (1, 0, \ldots, 0) \in \mathcal{R}^d$. Then $\frac{1}{2} \sum_{i=1}^{2n} g_i$ is a convex combination of two elements of

$span_n H_d$, $\sum_{i=1}^{n} g_i$ and $\sum_{i=n+1}^{2n} g_i$, but it is not in $span_n H_d$, since its restriction to the one-dimensional set $\{(t, 0, \ldots, 0) \in \mathcal{R}^d; t \in [0,1]\}$ has $2n$ discontinuities.

Summarizing results of this section and of the previous one, we get the following corollary.

Corollary 4.4 *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.*

5 Discussion

We have shown that convenient properties of projection operators such as uniqueness and continuity are not satisfied by Heaviside perceptron networks with a fixed number of hidden units. These properties would allow one to estimate worst-case errors using methods of algebraic topology (see e.g. DeVore, Howard & Micchelli, 1989). In linear approximation theory, application of such methods shows that for some sets of functions defined by smoothness conditions exhibit the curse of dimensionality: the approximants converge at rate $\mathcal{O}(\frac{1}{\sqrt[d]{n}})$, where d is the number of variables and n is the dimension of the approximating linear space (see, e.g. Pinkus, 1986). Our results show that these arguments are not applicable to approximation by Heaviside perceptron networks.

Note that the results from Section 3 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 & Poggio (1990) and later exploited by Leschno et al. (1993) for a proof of the universal approximation property.

Acknowledgement

V. Kůrková was partially supported by GA ČR grant 201/99/0092 and collaboration of V. Kůrková and A. Vogt was supported by an NRC COBASE grant.

References

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

DeVore, R., Howard, R., & Micchelli, C. (1989). Optimal nonlinear approximation. *Manuscipta Mathematica* 63, 469–478.

Girosi, F. & Poggio, T. (1990).

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

Hornik, K., Stichcome, M., & White, M. (1989). Multlilayer feedforward networks are universal approximators. *Neural Networks* 2, 359–366.

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

Kainen, P.C., Kůrková, V., & Vogt, A. (2000). Geometry and topology of continuous best and near best approximations. *J. Approx. Theory*, to appear.

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

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

Pinkus, A. (1986). *n-Width in Approximation Theory*. Berlin: Springer.

Singer, I. (1970). *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*. Berlin: Springer.

Vlasov, L. P. (1970). Almost convex and Chebyshev sets. *Math. Notes Acad. Sci. USSR* 8, 776–779.