PERGAMON

Neural Networks 13 (2000) 695–697

Neural
Networks

www.elsevier.com/locate/neunet

Letter to the editor

Best approximation by Heaviside perceptron networks

P. Kainen[a,*], V. Kůrková[b], A. Vogt[a]

[a]Department of Mathematics, Georgetown University, Washington, DC 20057-1233, USA
[b]Institute of Computer Science, Academy of Sciences of the Czech Republic, 182 07 Prague 8, Czech Republic

Received 24 May 2000; accepted 30 June 2000

Abstract

In $\mathscr{L}_p$-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. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: One-hidden-layer networks; Heaviside perceptrons; Best approximation; Metric projection; Continuous selection; Approximatively compact

1. Introduction

An important measure of the complexity of feedforward neural networks is the number of hidden units. To estimate the 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 derive implications for these networks from our recent mathematical results (Kainen, Kůrková, & Vogt, 1999b; Kainen, Kůrková, & Vogt, 2000): in $\mathscr{L}_p$-spaces with $p \in [1, \infty)$ 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 on both 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* $\psi : \mathscr{R} \rightarrow \mathscr{R}$ (where $\mathscr{R}$ denotes the set of real numbers) computes real-valued functions on $\mathscr{R}^{d+1} \times \mathscr{R}^d$ of the form $\psi(\mathbf{v}\cdot\mathbf{x} + b)$, where $\mathbf{x} \in \mathscr{R}^d$is an *input* vector, $\mathbf{v} \in \mathscr{R}^d$ is an *input weight* vector and $b \in \mathscr{R}$ 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., the set $H_d =$

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

Notice that H_d is the set of *characteristic functions of half-spaces* of $\mathscr{R}^d$ 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

$$\sum_{i=1}^{n} w_i \vartheta(\mathbf{v}_i\cdot\mathbf{x} + b_2),$$

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, 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, $\cup_{n \in \mathscr{N}_+} span_n H_d$ (where $\mathscr{N}_+$ denotes the set of all positive integers) is dense in $(\mathscr{C}([0,1]^d), \|\cdot\|_{\mathscr{C}})$, the linear space of all continuous functions on $[0,1]^d$ with the supremum norm, as well as in $(\mathscr{L}_p([0,1]^d), \|\cdot\|_p)$ with $p \in [1, \infty]$ (see, for

[*] Corresponding author.

PII: S0893-6080(00)00056-3

example, 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 $(X, \|\cdot\|)$ 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, for example, Singer, 1970). Clearly, a proximinal subset must be closed.

A sufficient condition for proximinality of a subset M of a normed linear space $(X,\|\cdot\|)$ 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 $e_{\{f\}} : M \to \mathscr{R}$ defined by $e_{\{f\}}(m) = \|m - f\|$ is continuous (see, for example, Singer, 1970, p. 391) and hence must achieve its minimum on any compact set M.

Gurvits and Koiran (1997) have shown that for all positive integers d, the set of characteristic functions of half-spaces H_d is compact in $(\mathscr{L}_p([0,1]^d), \|\cdot\|_p)$ 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 parametrization 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 $(X, \|\cdot\|)$ is called *approximatively compact* if for each $f \in X$ and any sequence $\{g_i : i \in \mathscr{N}_+\} \subseteq M$ such that $\lim_{i\to\infty}\|f - g_i\| = \|f - M\|$, there exists $g \in M$ such that $\{g_i : i \in \mathscr{N}_+\}$ converges subsequentially to g (see, for example, Singer, 1970, p. 368).

The following theorem shows that $span_n H_d$ is approximatively compact in $\mathscr{L}_p$-spaces. It extends a weaker result by Kůrková (1995), who showed that $span_n H_d$ is closed in $\mathscr{L}_p$-spaces with $p \in (1, \infty)$.

Thoerem 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 $(\mathscr{L}_p([0,1]^d), \|\cdot\|_p)$.*

The proof is based on an argument showing that any sequence of elements of $span_n H_d$ has a subsequence that converges either to an element of M or to a Dirac delta function, and the latter case cannot occur when such a sequence approximates a best approximation of some function in $\mathscr{L}_p([0,1]^d)$ (see Kainen et al., 1999b).

It is a straightforward consequence of the definitions that approximatively compact implies proximinal (see Singer, 1970, p. 382).

Corollary 3.2. *For every n, d positive integers and for every $p \in (1, \infty)$ $span_n H_d$ is a proximinal subset of $(\mathscr{L}_p([0,1]^d), \|\cdot\|_p)$.*

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

4. Uniqueness and continuity of best approximation

Let M be a subset of a normed linear space $(X, \|\cdot\|)$ and let $\mathscr{P}(M)$ denote the set of all subsets of M. The set-valued mapping $P_M : X \to \mathscr{P}(M)$ 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 $F : X \to \mathscr{P}(M)$ 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 $(X,\|\cdot\|)$ is called *strictly convex* (also called "rotund") if for all $f \neq g$ in X with $\|f\| = \|g\| = 1$, we have $\|(f + g)/2\| < 1$. This just means that the midpoint of the segment joining any two points on the unit sphere is contained in the interior of the ball. Thus, a norm is strictly convex when the unit ball is "round". It is well known that for all $p \in (1, \infty)$ $(L_p([0,1]^d), \|\cdot\|_p)$ 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 $\mathscr{L}_p([0,1]^d)$ 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 and Vogt (1999a) and Kainen et al. (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 $\mathscr{L}_p$-spaces with $p \in (1, \infty)$ satisfy the assumptions of this theorem (since the dual of $\mathscr{L}_p$ is $\mathscr{L}_q$ where $(1/p) + (1/q) = 1$ and $q \in (1, \infty)$) (see, for example, Friedman, 1982, p. 160). Hence, to show the non-existence of a continuous selection, 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 $0 > b_1 > \cdots > b_{2n} > -1$ and $\mathbf{v} = (1, 0, \ldots, 0) \in \mathscr{R}^d$. 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 $\{(t, 0, \ldots, 0) \in \mathscr{R}^d : t \in [0, 1]\}$ has $2n$ discontinuities.

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

Corollary 4.4. *In $(\mathscr{L}_p([0,1]^d), \|\cdot\|_p)$ with $p \in (1, \infty)$ for all n, d positive integers there exists a best approximation mapping from $\mathscr{L}_p([0,1]^d)$ 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 allow one to estimate worst-case errors using methods of algebraic topology (see for example, DeVore, Howard, and Micchelli, 1989). In linear approximation theory, application of such methods shows that some sets of functions defined by smoothness conditions exhibit the curse of dimensionality: the approximants converge at rate $\mathscr{O}(1/\sqrt[d]{n})$, where d is the number of variables and n the dimension of the approximating linear space (see, for example, 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, for example, 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 (1990) and later exploited by Leschno, Lin, Pinkus, and Schocken (1993) for a proof of the universal approximation property.

Acknowledgements

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. *Mathematics of Control, Signals and Systems*, *2*, 303–314.

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

Friedman, A. (1982). *Foundations of modern analysis*, New York: Dover.

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

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

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

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

Kainen, P.C., Kůrková, V., Vogt, A. (1999b). Best approximation by linear combinations of characteristic functions of half-spaces. Research report ICS-99-795.

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

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. *Mathematical Notes of the Academy of Sciences, USSR*, *8*, 776–779.