Approximation of Functions by Perceptron Networks with Bounded Number of Hidden Units

Kůrková, Věra
1994

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Approximation of functions by perceptron networks with bounded number of hidden units

Věra Kůrková

Technical report No. 600

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+422) 6605 3231 fax: (+422) 8585789
e-mail: vera@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Approximation of functions by perceptron networks with bounded number of hidden units

Věra Kůrková[1]

Technical report No. 600

Abstract

We examine the effect of constraining the number of hidden units. For one-hidden-layer networks with fairly general type of units (including perceptrons with any bounded activation function and radial-basis-function units), we show that when also the size of parameters is bounded, the best approximation property is satisfied, which means that there always exists a parameterization achieving the global minimum of any error function generated by a supremum or $\mathcal{L}_p$-norm. We also show that the only functions that can be approximated with arbitrary accuracy by increasing parameters in networks with a fixed number of Heaviside perceptrons are functions equal almost everywhere to functions that can be exactly computed by such networks. We give a necessary condition on values that such piecewise constant functions must achieve.

Keywords

Approximation of functions, One-hidden-layer neural network, Heaviside perceptron, Radial-basis-function unit, Bounded number of hidden units.

[1]This work was supported by GACR under grant 201/93/0427

1 Introduction

We consider the problem of approximating real-valued functions from a compact subset of d-dimensional space by a feedforward, single-hidden-layer neural network with a single linear output unit. Such networks, with either perceptron or radial-basis-function hidden units, are known to have the so-called "universal approximation" property as long as their activation functions satisfy rather weak conditions (Mhaskar and Micchelli, 1992, Leshno et al., 1993, and Park and Sandberg, 1991). Hence, any continuous or measurable function can be approximated with an arbitrary accuracy by such networks provided that arbitrarily large parameters and enough hidden units are available.

However in practical situations, both the size of parameters and the number of hidden units are bounded. The question of whether the universal approximation property can be achieved even with bounded parameters was answered by Stinchcombe and White (1990). They proved that there exists a constant (which depends on certain characteristics of the activation function) such that weights and biases, sufficient for universal approximation, can be bounded above inabsolute value by the constant. Recently, Hornik (1993) extended this result to an arbitrarily small bound. So, by increasing number of hidden units we can arbitrarily decrease weights and biases. However, little is known about the trade-off between the bound on parameter size and the minimum number of hidden units. In particular, the structure of spaces of functions approximable by networks with a bounded number of hidden units have not been studied.

In this paper, we first examine the effect of constraining both the size of parameters and the number of hidden units. We show that in this case, sets of functions computable by networks with fairly general type of hidden units (including perceptrons with any bounded activation function as well as radial-basis-function units) are compact. Thus, the "best approximation" property is achieved, which means that there always exists a parametrization achieving the global minimum of any error function generated by the supremum or $\mathcal{L}_p$-norm.

Further, we show that the only functions that can be approximated with arbitrary accuracy by increasing weights in networks with a *fixed* number of Heaviside perceptrons are functions differing at most on sets of measure zero from functions that can be exactly computed by such networks. To illustrate how restricted is this class we give an elementary condition on values that such a piecewise constant function must achieve on neighbouring sets from the underlying finite partition. For instance, not even such elementary functions as the characteristic functions of d-dimensional boxes (for $d \geq 2$) are included in this class. The particular result stating that characteristic functions of d-dimensional cubes for $d \geq 2$ are not in $\mathcal{L}_p$ closures of sets of functions computable on compact subsets of $\mathcal{R}^d$ by one-hidden-layer networks containing perceptrons with Heaviside activation function was proved by Chui et al. (1993). Here, we extend their results by showing that such sets not only do not contain such characteristic functions, but are even closed. We also show that the assumption that the activation function is essentially Heaviside is neccessary and give examples of functions that can be approximated with any accuracy by networks with a fixed number of hidden units having a differentiable activation function. Finally, we discuss consequences of our

results for comparisons of rates of approximation by networks with various types of hidden units.

The paper is organized as follows. In Section 2, we give necessary definitions and state our results on the best approximation property. In section 3 closures of spaces of functions computable by networks with a fixed number of hidden units are examined, while section 4 compares rates of approximation. The proofs are in Section 5.

2 The best approximation property

In this paper we consider only classes of one-hidden-layer feedforward networks with a single linear output unit. Since in any practical application, values of the inputs can vary only within certain limits, we can suppose that input vectors are within the unit cube I^d, where $I = [0,1]$ and d is the number of inputs.

We denote by $\mathcal{P}_d(\psi)$ the set of all functions computable by networks with d input units and any finite number of perceptrons with an activation function ψ in the hidden layer. So, $\mathcal{P}_d(\psi)$ is the set of all functions from $\mathcal{R}^d$ to $\mathcal{R}$ of the form

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

where k is any positive integer and $w_i, b_i \in \mathcal{R}$ and $\mathbf{v}_i \in \mathcal{R}^d$ are arbitrary. By $\mathcal{P}_d(\psi, k)$ we denote the subset of $\mathcal{P}_d(\psi)$ containing only functions computable by networks with at most k hidden units and by $\mathcal{P}_d(\psi, k, a)$ the subset of $\mathcal{P}_d(\psi, k)$ containing functions computable by networks with bounded parameters by a, i.e. satisfying $|w_i| \leq a$, $|b_i| \leq a$ and $\|\mathbf{v}_i\| \leq a$ for every $i = 1, \ldots, k$.

Analogously, for radial-basis-function networks, $\mathcal{B}_d(\psi)$ denotes the set of functions computable by networks with d inputs and any finite number of radial-basis-function hidden units with a radial function ψ and Euclidean norm $\|.\|$ on $\mathcal{R}^d$, i.e. $\mathcal{B}_d(\psi)$ is the set of all functions from $\mathcal{R}^d$ to $\mathcal{R}$ of the form

$$\sum_{i=1}^{k} w_i \psi\left(\frac{\|\mathbf{x} - \mathbf{c}_i\|}{b_i}\right)$$

where $w_i, b_i \in \mathcal{R}$ and $\mathbf{c}_i \in \mathcal{R}^d$. The subset $\mathcal{B}_d(\psi, k)$ is defined as above, while in the definition of $\mathcal{B}_d(\psi, k, a)$ we only require $|w_i| \leq a$, $|\frac{1}{b_i}| \leq a$.

Capabilities of networks to approximate functions are studied mathematically in terms of closures and dense subspaces; see, e.g. Simmons, 1963, for the basic definitions and theorems. For many functions ψ, the sets $\mathcal{P}_d(\psi)$ and $\mathcal{B}_d(\psi)$ are known to be dense in the space of continuous functions on I^d with the supremum norm $\|.\|_\infty$ topology (denoted by $\mathcal{C}(I^d)$) and in spaces of Lebesgue measurable functions on I^d with $\mathcal{L}_p$-norm $\|.\|_p$ for $p \in [1, \infty)$ (denoted by $\mathcal{L}_p(I^d)$). In neural networks terminology this capability is called the *universal approximation* property. However, one may require an arbitrarily large number of hidden units as the accuracy of the approximation increases; see recent upper estimates by Barron (1993), Darken et al. (1993) and Mhaskar and Micchelli (1993).

In practical situations, the number of hidden units is bounded by some fixed positive integer. In addition, the parameters are also bounded. Under these conditions, we show that for any function, there is a choice of network parameterization (not necessarily unique) producing an approximation with the minimum error. We call this the *best approximation* property. Our first theorem takes advantage of Ascoli's theorem (Simmons, 1993; p.126) which is a powerful classical tool for verifying the compactness of function spaces.

Theorem 2.1 *For any positive integers d, k, for any positive real a and for any bounded $\psi : \mathcal{R} \to \mathcal{R}$ both $\mathcal{P}_d(\psi, k, a)$ and $\mathcal{B}_d(\psi, k, a)$ are compact as subsets of $\mathcal{C}(I^d)$ and $\mathcal{L}_p(I^d)$ for every $p \in [1, \infty)$.*

Since any norm is continuous with respect to the topology it induces and any continuous function achieves its minimum on a compact space, we can easily derive the best approximation property for a wide class of networks.

Corollary 2.2 *For every positive integers d, k, for every positive a, for every bounded continuous $\psi : \mathcal{R} \to \mathcal{R}$, for every $p \in [1, \infty]$ and for every $f \in \mathcal{C}(I^d)$ (or $\mathcal{L}_p(I^d)$) there exists $f_1 \in \mathcal{P}_d(\psi, k, a)$ and $f_2 \in \mathcal{B}_d(\psi, k, a)$ such that $\|f - f_1\|_p = \min\{\|f - g\|_p; g \in \mathcal{P}_d(\psi, k, a)\}$ and $\|f - f_2\|_p = \min\{\|f - g\|_p; g \in \mathcal{B}_d(\psi, k, a)\}$.*

3 Closures of spaces of functions computable by perceptron type networks with fixed number of hidden units

Even when the parameters are not bounded, the following theorem shows that the set of functions computable by networks with k hidden units is closed provided that the units are simple threshold perceptrons having the Heaviside activation function ϑ defined by $\vartheta(x) = 0$ for $x < 0$ and $\vartheta(x) = 1$ for $x \geq 0$. For $p \in [1, \infty)$ we denote by $cl_{\mathcal{L}_p}(X)$ the closure of a subset X in $\mathcal{L}_p(I^d)$ with $\mathcal{L}_p$ topology.

Theorem 3.1 *For every positive integers d, k, for every $p \in [1, \infty)$ and for every $f \in cl_{\mathcal{L}_p}(\mathcal{P}_d(\vartheta, k))$ there exists $g \in \mathcal{P}_d(\vartheta, k)$ such that $\|f - g\|_p = 0$.*

So, $cl_{\mathcal{L}_p}(\mathcal{P}_d(\vartheta, k)$ contains only functions equal almost everywhere to functions from $\mathcal{P}_d(\vartheta, k)$. Since in $\mathcal{L}_p(I^d)$ functions differing on sets of measure zero are identified, $\mathcal{P}_d(\vartheta, k))$ is closed as a subspace of $\mathcal{L}_p(I^d)$.

Thus, the functions that can be approximated arbitrarily well by networks with a fixed number of Heaviside perceptrons are exactly the functions equal almost everywhere to functions actually computed by such networks. These functions form a very narrow class - containing only functions that are piecewise constant on partitions of I^d generated by cozero hyperplanes of affine functions $\mathbf{v}_i \cdot \mathbf{x} + b_i$ given by weights $\mathbf{v}_i$ and biases b_i of hidden units. Even among such piecewise constant functions, only a very limited class belongs to $\mathcal{P}_d(\vartheta, k)$.

Blum and Li (1991) gave a necessary condition for a piecewise constant function to be in $\mathcal{P}_d(\vartheta, k)$, namely that it must not contain a "saddle point". Our next theorem extends their characterization. Consider a family of hyperplanes $\{H_i; i = 1, \ldots, k\}$ and by H_{i0} and H_{i1} denote the two halfspaces separated by H_i. We call a function $f : I^d \to \mathcal{R}$ *piecewise constant with respect to* $\{H_i; i = 1, \ldots, k\}$ if for every $\nu : \{1, \ldots, k\} \to \{0, 1\}$ f is constant on $\bigcap\{H_{i\nu(i)}; i = 1, \ldots, k\}$. Suppose that for every maximal subfamily $\{H_i; i \in K\}$ of $\{H_i; i = 1, \ldots, k\}$ with a non-empty intersection there exists a constant c_K such that for every pair of antipodal sectors into which $\{H_i; i \in K\}$ cuts an open neighbourhood U of $\bigcap\{H_i; i = 1, \ldots, k\}$ nonintersecting H_j for $j \notin K$, the sum of values of f on these two sectors is equal to c_K. It is straightforward to check that this property is invariant with respect to interchange between H_i0 and H_i1. We call a piecewise constant function satisfying this property *balanced*.

Theorem 3.2 *For every positive integer d, k and for every $f \in \mathcal{P}_d(\vartheta, k)$ there exists a a family of hyperplanes $\{H_i; i = 1, \ldots, k\}$ such that f is piecewise constant and balanced with respect to $\{H_i; i = 1, \ldots, k\}$.*

Using Theorem 3.2 it is easy to verify that for $d \geq 2$ $\mathcal{P}_d(\vartheta, k)$ contains no characteristic function of a d-dimensional box: among the pairs of antipodal sectors into which the hyperplanes forming faces of a box cut I^d, only one pair has sum of values equal to 1, while for all other the sum of values equals 0.

To show that the assumption of Theorem 3.1 (that the activation function is Heaviside), is necessary consider a function of the form

$$\lim_{n\to\infty} w_n \left(\psi(\mathbf{v}_n \cdot \mathbf{x} + b_n) - \psi(\mathbf{v} \cdot \mathbf{x} + b)\right), \tag{0.1}$$

where $\lim_{n\to\infty} \mathbf{v}_n = \mathbf{v}$, $\lim_{n\to\infty} b_n = b$ and $\{w_n; n \in \mathcal{N}\}$ is diverging.

In the case when $\psi = \vartheta$ is the Heaviside function we get in (1) a sequence of functions the Lebesgue measures of the supports of which are converging to zero. Such a sequence either converges to a distribution or to a finite function that is almost everywhere zero.

However, when supports of functions $\psi(\mathbf{v}_n \cdot \mathbf{x} + b_n) - \psi(\mathbf{v} \cdot \mathbf{x} + b)$ do not converge to zero, we may get non-trivial functions in $cl_{\mathcal{L}_p}(\mathcal{P}_d(\vartheta, 2))$ that are not in $\mathcal{P}_d(\vartheta, 2)$. If ψ is differentiable, we thus get for instance functions of the form

$$\lim_{n\to\infty} \left(n\psi\left((v + \frac{1}{n})x + b\right) - n\psi(vx + b)\right) = \frac{\partial\psi(vx + b)}{\partial v} = x\psi'(vx + b)$$

and

$$\lim_{n\to\infty} \left(n\psi(vx + b + \frac{1}{n}) - n\psi(vx + b)\right) = \frac{\partial\psi(vx + b)}{\partial b} = \psi'(vx + b).$$

Girosi and Poggio (1990) noted that for the logistic sigmoid $\lambda = \frac{1}{1+\exp(-x)}$ the function $\frac{1}{2(1+\cosh(x))}$ can be approximated with any accuracy by a network with only two hidden units. Indeed, it is easy to verify that

$$\frac{1}{2(1 + \cosh(x))} = \lim_{n\to\infty} \left(n\lambda(x) - n\lambda(x + \frac{1}{n})\right).$$

4 Comparison of rates of approximation

Characterization of closures of spaces of functions computable by networks with a bounded number of hidden units might be useful for comparing the rates of approximation of functions from the same class by networks with different types of units. Let $\mathcal{F}$ and $\mathcal{G}$ be sets of functions computable by one-hidden-layer networks with two different types of units. By $\mathcal{F}(n)$ and $\mathcal{G}(n)$ denote subsets of $\mathcal{F}$ and $\mathcal{G}$, resp., containing functions computable by networks with n hidden units. Let $\mathcal{S} \subseteq cl_{\mathcal{L}_p}(\mathcal{F})$ and $\mathcal{S} \subseteq cl_{\mathcal{L}_p}(\mathcal{G})$, i.e. functions from $\mathcal{S}$ can be approximated by networks of both types. We say that the *rate of approximation of functions from $\mathcal{S}$ by $\mathcal{F}$ is related to the rate of approximation by functions from $\mathcal{G}$* if there exists a mapping $r : \mathcal{N} \to \mathcal{N}$ ($\mathcal{N}$ denotes the set of natural numbers) such that for every $f \in \mathcal{S}$ and for every ε if there exists $f_n \in \mathcal{F}(n)$ such that $\|f - f_n\|_p < \varepsilon$ then there exists $g_{r(n)} \in \mathcal{G}(r(n))$ such that $\|f - g_{r(n)}\|_p < \varepsilon$. The function r can grow arbitrarily fast; we do not restrict to linear or polynomial functions.

Notice that if $\mathcal{S}$ contains a function from $\mathcal{F}$, i.e. a function f computable by a network of the first type, then $f \in cl_{\mathcal{L}_p}\mathcal{G}(n)$. So, understanding which functions computable by one type of networks are in closures of sets of functions computable by networks of another type with fixed number of hidden units enables comparison of rates of approximation.

We have shown in Theorem 3.1 that $\mathcal{P}_d(\vartheta, k)$ is closed in $\mathcal{L}_p(I^d)$ and so, no sufficiently large class of functions can be approximated by one-hidden-layer networks with another type of units than Heaviside perceptrons with a rate of approximation related to the rate of approximation by perceptron networks.

5 Mathematical proofs

All limits involving functions are in $\mathcal{L}_p(I^d)$.

Proof of Theorem 2.1

By Ascoli's theorem (see e.g. Simmons, 1963, p.126) a set of mappings between compact metric spaces is compact in the topology of uniform convergence if and only if it is equicontinuous. Since both $\mathcal{P}_d(\psi, k, a)$ and $\mathcal{B}_d(\psi, k, a)$ contain sums of k functions bounded by $a\|\psi\|_\infty$, they map compact I^d into a compact subset of $\mathcal{R}$. Equicontinuity follows from the joint bound a on parameters. So, $\mathcal{P}_d(\psi, k, a)$ and $\mathcal{B}_d(\psi, k, a)$ are compact subsets of $\mathcal{C}(I^d)$. The topology of uniform convergence is finer (contains more open sets) than any of the $\mathcal{L}_p$ topologies and so $\mathcal{P}_d(\psi, k, a)$ and $\mathcal{B}_d(\psi, k, a)$ are compact in $\mathcal{L}_p(I^d)$, too.

Proof of Theorem 3.1

Let $f \in cl_{\mathcal{L}_p}(\mathcal{P}_d(\vartheta, k))$. Then there exists a sequence $\{f_n; n \in \mathcal{N}\} \subset \mathcal{P}_d(\vartheta, k)$ converging to f in $\mathcal{L}_p$-norm on I^d. Since $\vartheta(t) = \vartheta(at)$ for every $a > 0$, we may assume that for every $n \in \mathcal{N}$

$$f_n(\mathbf{x}) = \sum_{i=1}^{k_n} w_{ni}\vartheta(\mathbf{v}_{ni} \cdot \mathbf{x} + b_{ni}) + c$$

with $(\mathbf{v}_{ni}, b_{ni}) = (v_{ni1}, \ldots, v_{nid}, b_{ni})$ being a unit vector in $\mathcal{R}^{d+1}$ for every $i = 1, \ldots, k$ and $v_{nij} > 0$ for the first non-zero v_{nij} $(j = 1, \ldots, d)$. Compactness of the unit ball in $\mathcal{R}^{d+1}$ guarantees that we can choose from every sequence a converging one. Finiteness of the set $\{k_n; n \in \mathcal{N}\}$ (all $k + n \geq k$) implies that there exists $l \leq k$ and an infinite $\mathcal{M} \subseteq \mathcal{N}$ and $\mathbf{v}_i \in \mathcal{R}^d$, $b_i \in \mathcal{R}$ such that for every $i = 1, \ldots, l$ $\lim\limits_{\substack{n\to\infty \\ n\in\mathcal{M}}} b_{ni} = b_i$, $\lim\limits_{\substack{n\to\infty \\ n\in\mathcal{M}}} \mathbf{v}_{ni} = \mathbf{v}_i$, and the sequence $\{w_{ni}; n \in \mathcal{M}\}$ is either diverging or there exists $w_i \in \mathcal{R}$ such that $\lim\limits_{\substack{n\to\infty \\ n\in\mathcal{M}}} w_{ni} = w_i$.

First, reorder $\{1, \ldots, l\}$ in such a way that for $i = 1, \ldots, m$ the sequences $\{w_{ni}; n \in \mathcal{M}\}$ are diverging and for $i = m+1, \ldots, l$ the sequences $\{w_{ni}; n \in \mathcal{M}\}$ are converging.

Put $g(\mathbf{x}) = \sum_{i=m+1}^{k} w_i \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i) + c$. Since for $i = m+1, \ldots, l$ the sequences $\{w_{ni}; n \in \mathcal{N}\}$ are converging

$$g(\mathbf{x}) = \lim_{\substack{n\to\infty \\ n\in\mathcal{M}}} \sum_{i=m+1}^{l} w_{ni} \vartheta(\mathbf{v}_{ni} \cdot \mathbf{x}_i + b_{ni}) + c.$$

Hence

$$f(\mathbf{x}) - g(\mathbf{x}) = \lim_{\substack{n\to\infty \\ n\in\mathcal{M}'}} \sum_{i=1}^{m} w_{ni} \vartheta(\mathbf{v}_{ni} \cdot \mathbf{x} + b_{ni}).$$

Let p be the number of distinct pairs among $\{(\mathbf{v}_1, b_1), \ldots, (\mathbf{v}_m, b_m)\}$. Reorder $\{1, \ldots, m\}$ in such a way that for $i = 1, \ldots, p$ the pairs $\{(\mathbf{v}_1, b_1), \ldots, (\mathbf{v}_p, b_p)\}$ are distinct and for every $i = 1, \ldots, p$, put $K_i = \{j \in \{1, \ldots, m\}; (\mathbf{v}_j, b_j) = (\mathbf{v}_i, b_i)\}$. For every $n \in \mathcal{M}$ put $\hat{w}_{ni} = \sum_{j \in K_i} w_{nj}$.

It is easy to verify, that there exists an infinite $\mathcal{M}' \subseteq \mathcal{M}$ such that for every $i \in \{1, \ldots, p\}$ the sequence $\{\hat{w}_{ni}; n \in \mathcal{M}'\}$ is either converging or diverging. Reorder $\{1, \ldots, p\}$ in such a way that for $i = 1, \ldots, q$ there exist $\hat{w}_i \in \mathcal{R}$ such that $\lim\limits_{\substack{n\to\infty \\ n\in\mathcal{M}'}} \hat{w}_{ni} = \hat{w}_i$ and for $i = q+1, \ldots, p$ the sequences $\{\hat{w}_{ni}; n \in \mathcal{M}'\}$ are diverging.

Put $h(\mathbf{x}) = \sum_{i=1}^{q} \hat{w}_i \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i)$ and for every $i = 1, \ldots, p$ and for every $n \in \mathcal{N}$ put

$$\phi_{ni}(\mathbf{x}) = \sum_{j \in K_i} w_{nj} \left(\vartheta(\mathbf{v}_{nj} \cdot \mathbf{x} + b_{nj}) - \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i) \right).$$

So, we have

$$f(\mathbf{x}) - g(\mathbf{x}) - h(\mathbf{x}) = \lim_{\substack{n\to\infty \\ n\in\mathcal{M}'}} \left(\sum_{i=q+1}^{p} \hat{w}_{ni} \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i) + \sum_{i=1}^{p} \phi_{ni}(\mathbf{x}) \right).$$

Putting for every $n \in \mathcal{M}'$ $c_n = \max\{|\hat{w}_{ni}|; i = q+1, \ldots, p\}$ and for every $i = q+1, \ldots, p$ $a_{ni} = \frac{\hat{w}_{ni}}{c_n}$, we have

$$\lim_{\substack{n\to\infty \\ n\in\mathcal{M}'}} \left(\sum_{i=q+1}^{p} a_{ni} \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i) + \sum_{i=1}^{p} \frac{\phi_{ni}(\mathbf{x})}{c_{ni}} \right) = \lim_{\substack{n\to\infty \\ n\in\mathcal{M}'}} \frac{f(\mathbf{x}) - g(\mathbf{x}) - h(\mathbf{x})}{c_n}.$$

Since $f-g-h$ is bounded in $\mathcal{L}_p$-norm on I^d and since $\lim_{\substack{n\to\infty \\ n\in\mathcal{M}'}} c_n = \infty$, we have

$$\lim_{\substack{n\to\infty \\ n\in\mathcal{M}'}} \left(\sum_{i=q+1}^{p} a_{ni}\vartheta(\mathbf{v}_i\cdot\mathbf{x}+b_i) + \sum_{i=1}^{p} \frac{\phi_{ni}(\mathbf{x})}{c_{ni}} \right) = 0.$$

It is easy to verify that there exists an infinite $\mathcal{M}'' \subseteq \mathcal{M}$ such that for every $i = q+1,\ldots,p$ there exist $a_i \in \mathcal{R}$ with $\lim_{\substack{n\to\infty \\ n\in\mathcal{M}''}} a_{ni} = a_i$ and for at least one i either $a_i = 1$ or $a_i = -1$.

Since for $i = q+1,\ldots,p$ the sequences $\{a_{ni}; n \in \mathcal{M}''\}$ are converging

$$\sum_{i=q+1}^{p} a_i\vartheta(\mathbf{v}_i\cdot\mathbf{x}+b_i) = \lim_{\substack{n\to\infty \\ n\in\mathcal{M}''}} \sum_{i=q+1}^{p} a_{ni}\vartheta(\mathbf{v}_i\cdot\mathbf{x}+b_i).$$

So, we have

$$\sum_{i=q+1}^{p} a_i\vartheta(\mathbf{v}_i\cdot\mathbf{x}+b_i) = -\lim_{\substack{n\to\infty \\ n\in\mathcal{M}''}} \sum_{i=1}^{p} \frac{\phi_{ni}(\mathbf{x})}{c_{ni}}.$$

Since ϑ is the Heaviside function and for every $i = 1,\ldots,p$ and for every $j \in K_i$ $\lim_{\substack{n\to\infty \\ n\in\mathcal{M}''}} \mathbf{v}_{nj} = \mathbf{v}_i$ and $\lim_{\substack{n\to\infty \\ n\in\mathcal{M}''}} b_{nj} = b_i$ the Lebesgue measures of the supports of the functions ϕ_{ni} are converging to zero. So, we have

$$\| \sum_{i=q+1}^{p} a_i\vartheta(\mathbf{v}_i\cdot\mathbf{x}+b_i)\|_p = 0.$$

This can only happen when either $p = q$ or all $a_i = 0$; otherwise, we would get a contradiction with the uniqueness of expression in reduced form (i.e., with all pairs $(\mathbf{v}_i, b_i)$ inducing different cozero hyperplanes) proved by Chui et al., 1993, Proposition 3.1.). Since there exists i with $a_i \neq 0$, we have $p = q$ and so

$$f-g-h = \lim_{\substack{n\to\infty \\ n\in\mathcal{M}''}} \sum_{i=1}^{p} \phi_{ni}.$$

Since the Lebesgue measures of the supports of the functions on the right side converge to zero, we have $\|f-g-h\|_p = 0$. Hence f is equal almost everywhere to a function from $\mathcal{P}_d(\vartheta,k)$, namely to $g+h$.

Proof of Theorem 3.2

Let $f \in \mathcal{P}_d(\vartheta,k)$. Then there exists $l \leq k$ and affine transformations $\alpha_i : \mathcal{R}^d \to \mathcal{R}$, $i = 1,\ldots,l$ such that $f = \sum_{i=1}^{l} w_i\vartheta\circ\alpha_i + c$. Let H_i denotes the cozero hyperplane of α_i. For a maximal subfamily $\{H_i; i \in K\}$ with $\bigcap\{H_i; i \in K\}$ non-empty take a neighbourhood U such that $U \cap H_i = \emptyset$ for $i \notin K$. For every $i \in \{1,\ldots,l\}$ put $a_i = \vartheta\circ\alpha_i/U$. Put $c_k = \sum_{i\in K} w_i + 2\sum_{i\notin K} a_i$.

For every mapping $\nu : K \to \{0,1\}$ define $\bar{\nu} : K \to \{0,1\}$ by $\bar{\nu}(i) = 1-\nu(i)$. Each ν characterizes one of the sectors into which U is cutted by $\{H_i; i \in k\}$, and $\bar{\nu}$ characterizes the antipodal sector. Since the value of f on the sector characterized by ν is $\sum_{i\in K} w_i\nu_i + \sum_{i\notin K} a_i$ and the value of f on the antipodal sector is $\sum_{i\in K} w_i\bar{\nu}_i + \sum_{i\notin K} a_i$, their sum is equal to $\sum_{i\in K} w_i + 2\sum_{i\notin K} a_i = c_K$.

References

Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory* **39,3**,

Blum, E. K., & Li, L. K. (1991). Approximation theory and feedforward networks. *Neural Networks*, **4**, 511-515.

Chui, C. K., Li, X., & Mhaskar, H. N. (1993). Neural networks for localized approximation. Preprint.

Darken, C., Donahue, M., Gurvits, L., & Sontag, E. (1993). Rate of approximation results motivated by robust neural network learning. In *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory* (pp.303-309). New York: Association for Computing Machinery.

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

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks*, **6**, 1069-1072.

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

Mhaskar, H. N., & Micchelli, C. A. (1992). Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, **13**, 350-373.

Mhaskar, H. N., & Micchelli, C. A. (1993). Dimension-independent bounds on the degree of approximation by neural networks. Research Report IBM RC 18981.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, **3**, 246-257.

Simmons, F. G. (1963). *Introduction to Topology and Modern Analysis*. New York: McGraw-Hill.

Stinchcombe, M., & White, H. (1990). Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights. In *Proceedings of IJCNN* (pp. III.7-16). New York: IEEE Press.