MATHEMATICS OF COMPUTATION
VOLUME 63, NUMBER 208
OCTOBER 1994, PAGES 607–623

NEURAL NETWORKS FOR LOCALIZED APPROXIMATION

C. K. CHUI, XIN LI, AND H. N. MHASKAR

ABSTRACT. We prove that feedforward artificial neural networks with a single hidden layer and an ideal sigmoidal response function cannot provide localized approximation in a Euclidean space of dimension higher than one. We also show that networks with two hidden layers can be designed to provide localized approximation. Since wavelet bases are most effective for local approximation, we give a discussion of the implementation of spline wavelets using multilayered networks where the response function is a sigmoidal function of order at least two.

1. INTRODUCTION

There has been much study in recent years on the question of using neural networks for approximating real-valued functions of several real variables. In particular, Cybenko [8] and Hornik, Stinchcombe, and White [11] have proved that it is possible to use a neural network with one hidden layer and a sigmoidal activation function to approximate continuous functions on any compact subset of a Euclidean space of an arbitrary dimension. In addition, Chui and Li [5], Mhaskar and Micchelli [16], Ito [13, 14], and Barron [1] also obtained such density theorems in various more or less general contexts, using different approaches, as well as studied the complexity problem in some detail. In [15], it is shown that a neural network with multiple hidden layers and a generalized sigmoidal activation function can be constructed to achieve the optimal rate of approximation for smooth nonanalytic functions and for analytic functions, a near-geometric rate independent of the dimensions of the input space. The question as to whether the same can be achieved with a single hidden layer is still open.

The objective of this paper is to investigate the possibility of constructing networks suitable for localized approximation, i.e., a network with the property that if the target function is modified only on a small subset of the Euclidean space, then only a few neurons, rather than the entire network, need to be retrained. The precise definitions will be given in the next section. We prove that

Received by the editor March 5, 1993 and, in revised form, September 7, 1993.

1991 *Mathematics Subject Classification.* Primary 41A63, 41A30, 94C99.

Key words and phrases. Neural networks, sigmoidal functions, spline functions, wavelets.

Research of the first author was supported by NSF Grant DMS 92-06928 and ARO Contract DAAH 04-93-G-0047.

Research of the second author was partially supported by the University Research Grants & Fellowships Committee at UNLV.

Research of the third author was supported by a grant from AFOSR.

if the dimension of the input space is greater than one, then such a network with one hidden layer and a Heaviside activation function cannot be constructed. In contrast, we also show that a network with two or more hidden layers can always be constructed to accomplish the task. To realize an effective local approximating network, we construct the Chui-Wang spline wavelets [7] using multilayered networks with a generalized sigmoidal activation function.

As in [15], our proofs will be constructive and the "training algorithm" will be noniterative. Hence, the usual questions about stability, settling into local minima, etc., which usually need to be discussed in connection with the more popular backpropagation networks, simply do not arise.

2. Localized approximation

In the sequel, let $s \geq 2$ be any integer and $Q := [-1, 1]^s$ be the standard cube in $\mathbf{R}^s$. Intuitively, a neural network can be said to provide localized approximation on Q, if Q can be divided into a number of subregions so that only a small number of neurons are responsible for providing approximation on each subregion. Thus, if the function to be synthesized is modified only on a small part of Q, one needs only to retrain the small number of neurons responsible for this part, rather than retraining the entire network.

To make this idea more precise, we need some terminology. For $\mathbf{x}, \mathbf{y} \in \mathbf{R}^s$, $\mathbf{x} \cdot \mathbf{y}$ denotes the inner product between $\mathbf{x}$ and $\mathbf{y}$, and $|\mathbf{x} - \mathbf{y}|$ the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. In using measure-theoretic terms such as "almost everywhere", "measurable", and so on, we refer to the s-dimensional Lebesgue measure, which will be denoted by λ. Hence, if A is a measurable subset of $\mathbf{R}^s$, and $1 \leq p \leq \infty$, the L^p norm of a measurable function $g: A \to \mathbf{R}$ is given by

$$\|g\|_{p,A} := \begin{cases} \left(\int_A |g(\mathbf{t})|^p \, d\lambda(\mathbf{t})\right)^{1/p} & \text{if } 1 \leq p < \infty, \\ \operatorname{ess\,sup}_A |g(\mathbf{t})| & \text{if } p = \infty. \end{cases} \tag{2.1}$$

The class $L^p(A)$ then consists of measurable functions g on A for which $\|g\|_{p,A} < \infty$, where two functions are identified if they are equal almost everywhere. Also, the class $C(A)$ consists of continuous functions on A which vanish at infinity. The symbol χ_A denotes the characteristic function of A, i.e., the function that takes on the value 1 on A and zero outside A.

Let $\sigma: \mathbf{R} \to \mathbf{R}$ be any function. The output of a feedforward neural network with n neurons, arranged in a single hidden layer and with response function σ, is of the form $\sum_{k=1}^{n} c_k \sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k)$. The class of all such functions will be denoted by $\Pi_{n,1,s,\sigma}$. Next, we formally define inductively the class $\Pi_{n,l,s,\sigma}$ of all possible outputs of a fully connected feedforward network with n neurons, each with response function σ, arranged in at most l hidden layers and receiving an input from $\mathbf{R}^s$. The class $\Pi_{n,1,s,\sigma}$ is already defined. Suppose that $\Pi_{n,m,s,\sigma}$ is defined for all integers $n \geq 1$ and $m < l$. A typical network with l layers is constructed as follows. Let there be p neurons in the lth layer. Each of these receives a different number of inputs. Suppose that for each k, $1 \leq k \leq p$, the output of n_k subnetworks, $\{P_{j,k}\}_{j=1}^{n_k}$, is input to the kth neuron. A typical member of $\Pi_{n,l,s,\sigma}$ is therefore of the form $\sum_{k=1}^{p} c_k \sigma(\sum_{j=1}^{n_k} w_{j,k} P_{j,k}(\mathbf{x}) + b_k)$, where, for $j = 1, \ldots, n_k$, $k = 1, \ldots, p$, the quantities c_k, $w_{j,k}$, $b_k \in \mathbf{R}$, $P_{j,k} \in \Pi_{n_{j,k},l-1,s,\sigma}$ for some integers $n_{j,k}$.

Here, we require that the total number of neurons in the circuit is at most n. We do not rule out the possibility that the output of some subnetwork may be input to more than one neuron.

In defining the notion of localized approximation, we are motivated by the approximation by piecewise constants. The simplest way to approximate an $f: Q \to \mathbf{R}$ by piecewise constants is to divide Q uniformly into smaller cubes, and define the constant value of the approximation on each cube C to be the value of f at the center of C. Hence, such an approximation can be expressed in the form $\sum a_C \chi_C$. This simple approximation scheme is obviously localized in the sense that if the target function is modified on a part of Q, only the terms in the sum corresponding to the cubes overlapping with this part need to be modified. It is now seen that the problem of localized approximation can be reduced to the problem of approximating the characteristic function of each cube by neural networks of a fixed size, independent of the degree of approximation desired. If networks with a fixed number m of neurons can be constructed to approximate the characteristic function of each cube, then this simple-minded but clearly localized approximation by piecewise constants will lead to a localized approximation by neural networks. We now observe that the class $\Pi_{n,l,s,\sigma}$ is closed under the transformation $\mathbf{x} \to w\mathbf{x}$ for any $w > 0$. Therefore, approximating the characteristic function of any cube is equivalent to approximating the characteristic function of Q. Hence, we may formulate the notion of localized approximation by a network with l hidden layers as follows.

Definition 2.1. Let $l \geq 1$ be an integer. A neural network with l hidden layers and response function σ is said to provide *localized approximation* if there exists an integer $m \geq 1$ and a sequence $\{P_n\} \subseteq \Pi_{m,l,s,\sigma}$ such that

$$\lim_{n\to\infty} \|\chi_Q - P_n\|_{1,K} = 0 \tag{2.2}$$

for every compact set $K \subset \mathbf{R}^s$.

Our first goal is to show that a network with one hidden layer does not necessarily yield localized approximation.

Theorem 2.2. *Let*

$$\sigma(x) := \begin{cases} 1 & \textit{if } x \geq 0, \\ 0 & \textit{if } x < 0. \end{cases} \tag{2.3}$$

Then a neural network with one hidden layer and response function σ does not provide localized approximation.

The proof of Theorem 2.2 will consist of an elaborate compactness argument and will actually show that it is not possible to approximate any nontrivial function locally using a network with a single hidden layer and the response function (2.3). Therefore, the next question which presents itself is whether it is possible to achieve local approximation using two hidden layers. Using the ideas in [15], we prove that such an approximation is indeed possible, and we also give a rate of approximation.

In order to measure the rate of approximation, we introduce the notion of modulus of continuity. Let $I := \prod_{j=1}^{s}[a_j, b_j]$ and $f \in L^p(I)$ for some p with $1 \leq p \leq \infty$, where if $p = \infty$ then it is understood that $f \in C(I)$. For

$\mathbf{h} := (h^{(1)}, \dots, h^{(s)})$, where each $h^{(j)} \geq 0$, we let $I_{\mathbf{h}} := \prod_{j=1}^{s}[a_j + h^{(j)}, b_j - h^{(j)}]$ and define, for $\delta > 0$,

$$\omega(L^p(I), I; f, \delta) := \max_{0 \leq h^{(j)} \leq \delta, 1 \leq j \leq s} \|f(\cdot + \mathbf{h}) - f(\cdot)\|_{p, I_{\mathbf{h}}}. \tag{2.4}$$

The following theorem gives the localized analogue of Theorem 3.4 in [15].

Theorem 2.3. *Suppose that* $\sigma: \mathbf{R} \to \mathbf{R}$ *is a measurable function such that*

$$\lim_{x \to -\infty} \sigma(x) = 0, \qquad \lim_{x \to \infty} \sigma(x) = 1 \tag{2.5a}$$

and

$$|\sigma(x)| \leq M, \qquad x \in \mathbf{R}, \tag{2.5b}$$

where $M < 2s/(2s-1)$ *is a constant. Then a neural network with two hidden layers and response function* σ *provides localized approximation. Furthermore, let* $1 \leq p < \infty$ *and* $f \in L^p(Q)$. *For each integer* $n \geq 1$ *and each multi-integer* $\mathbf{m} \in \{0, 1, \dots, n-1\}^s$ *we define the cubes* $R_{\mathbf{m},n}$, $S_{\mathbf{m},n}$ *by*

$$\begin{aligned} R_{\mathbf{m},n} &:= \prod_{j=1}^{s}[-1 + m^{(j)}/n, -1 + (m^{(j)} + 1)/n], \\ S_{\mathbf{m},n} &:= \prod_{j=1}^{s}[-1 + (m^{(j)} - 1)/n, -1 + (m^{(j)} + 2)/n] \cap Q \end{aligned} \tag{2.6}$$

and set $N := N_n := (2s+1)n^s$. *Then there exists a function* $G_n(f) \in \Pi_{N,2,s,\sigma}$ *such that*

$$\|f - G_n(f)\|_{p, R_{\mathbf{m},n}} \leq 2\omega\left(L^p(S_{\mathbf{m},n}), S_{\mathbf{m},n}; f, \frac{2}{n}\right). \tag{2.7}$$

We will actually construct the operators G_n explicitly in the course of the proof. Moreover, the approximation on $R_{\mathbf{m},n}$ will be achieved by using at most $2s+1$ neurons.

3. Proofs of the theorems in §2

Until the end of the proof of Theorem 2.2, we will use the notation σ to denote only the ideal response function defined in (2.3) and Π_m to denote $\Pi_{m,1,s,\sigma}$. It is then obvious that for any $\mathbf{w} \in \mathbf{R}^s$ and $b \in \mathbf{R}$ we have

$$\sigma(\mathbf{w} \cdot \mathbf{x} + b) + \sigma((-\mathbf{w}) \cdot \mathbf{x} - b) - \sigma(\mathbf{0} \cdot \mathbf{x} + 1) = 0 \quad \text{a.e. in } \mathbf{R}^s. \tag{3.1}$$

Therefore, we first seek a canonical representation for elements on Π_m. We say that an expression of the form $\sum_{k=1}^{m} c_k \sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k)$ is in *reduced form* if

(i) the hyperspaces $\mathbf{w}_k \cdot \mathbf{x} + b_k = 0$ are all distinct,

(ii) $|\mathbf{w}_k|^2 + b_k^2 = 1$, $k = 1, \dots, m$, and

(iii) the first nonzero component of the $(s+1)$-dimensional vector $(\mathbf{w}_k, b_k)$ is positive.

Proposition 3.1. *Each* $P \in \Pi_m$ *admits almost everywhere a unique expression in reduced form.*

Proof. Let

$$P(x) = \sum_{k=1}^{m} c_k \sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k).$$

Since $\sigma(cx) = \sigma(x)$ for any $c > 0$ and $\sigma(0) = \sigma(1)$, we may assume that at most one of the $\mathbf{w}_k$'s is zero and that the corresponding b_k is equal to 1. With this convention, we may also assume that $|\mathbf{w}_k|^2 + b_k^2 = 1$ for $k = 1, \dots, m$. If any two of the hyperspaces $\mathbf{w}_k \cdot \mathbf{x} + b_k = 0$ coincide, we may rearrange the indices and assume that $\mathbf{w}_1 = -\mathbf{w}_2$ and $b_1 = -b_2$. If $\mathbf{w}_1 = \mathbf{0}$, then $\mathbf{w}_2 = \mathbf{0}$, and our convention implies that $b_1 = b_2 = 1$. Therefore, $\mathbf{w}_1 \neq \mathbf{0}$, and we let the first nonzero component of $\mathbf{w}_1$ be positive. Then, using (3.1), we get, for almost all $\mathbf{x} \in \mathbf{R}^s$,

$$\begin{aligned} &c_1\sigma(\mathbf{w}_1 \cdot \mathbf{x} + b_1) + c_2\sigma(\mathbf{w}_2 \cdot \mathbf{x} + b_2) \\ &\quad = \begin{cases} c_1 & \text{if } \mathbf{w}_1 \cdot \mathbf{x} + b_1 \geq 0, \\ c_2 & \text{if } \mathbf{w}_1 \cdot \mathbf{x} + b_1 \leq 0 \end{cases} \\ &\quad = c_2\sigma(\mathbf{0} \cdot \mathbf{x} + 1) + (c_1 - c_2)\sigma(\mathbf{w}_1 \cdot \mathbf{x} + b_1). \end{aligned} \tag{3.2}$$

Thus, any $P \in \Pi_m$ can be expressed in reduced form.

To prove the uniqueness of such an expression, it is sufficient to show that if

$$\tilde{Q}(\mathbf{x}) := \sum_{k=1}^{2m} c_k\sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k) = c \quad \text{a.e. in } \mathbf{R}^s \tag{3.3}$$

for some constant c, where none of the vectors $\mathbf{w}_k$ is zero and the expression on the left-hand side of (3.3) is in reduced form, then $c_1 = \cdots = c_{2m} = c = 0$. If this were not true, then we may assume that none of the constants c_k is zero. We shall show that one of them has to be zero, and thus arrive at a contradiction.

It is easy to see that $\tilde{Q}$ is almost everywhere equal to a piecewise constant function. To describe this function more precisely, we define the vector function $\mathbf{S} = (S^{(1)}, \dots, S^{(2m)})$ by the formula

$$S^{(j)}(\mathbf{x}) = \sigma(\mathbf{w}_j \cdot \mathbf{x} + b_j), \qquad \mathbf{x} \in \mathbf{R}^s,\ j = 1, \dots, 2m. \tag{3.4}$$

For $\mathbf{x}$, $\mathbf{y} \in \mathbf{R}^s$, we say that $\mathbf{x} \sim \mathbf{y}$ if $\mathbf{S}(\mathbf{x}) = \mathbf{S}(\mathbf{y})$. The equivalence relation $\sim$ partitions $\mathbf{R}^s$ into finitely many polytopic regions $P_1, \dots, P_M$, each having positive measure and they may intersect only at their boundaries. Each region P_k can be identified with the Boolean vector $\mathbf{S}_k = (S_k^{(1)}, \dots, S_k^{(2m)}) := \mathbf{S}(\mathbf{x}_k)$, where the choice of $\mathbf{x}_k \in P_k$ is arbitrary. The representation of the function $\tilde{Q}$ as a piecewise constant is then the following:

$$\tilde{Q}(\mathbf{x}) = \sum_{S_k^{(j)}=1} c_j \qquad \text{a.e. in } P_k,\ k = 1, \dots, M. \tag{3.5}$$

Since the hyperspaces $\mathbf{w}_k \cdot \mathbf{x} + b_k = 0$ are all nondegenerate and distinct, we see, in particular, that the hyperspace $\mathbf{w}_{2m} \cdot \mathbf{x} + b_{2m} = 0$ is a part of the boundary of two regions, say P_k and P_l such that $S_k^{(j)} = S_l^{(j)}$ for each $j = 1, \dots, 2m-1$ but $S_k^{(2m)} = 0$ and $S_l^{(2m)} = 1$. Let $\mathscr{J} := \{j\colon 1 \leq j \leq 2m-1,\ S_k^{(j)} = 1\}$. From (3.5), we also see that

$$Q(x) = \begin{cases} \displaystyle\sum_{j \in \mathscr{J}} c_j & \text{a.e. in } P_k, \\ \displaystyle\sum_{j \in \mathscr{J}} c_j + c_{2m} & \text{a.e. in } P_l. \end{cases} \tag{3.6}$$

In view of (3.3), we have $c_{2m} = 0$. This contradicts our assumption that none of the coefficients c_k is zero, thereby completing the proof of the proposition. □

When we consider sequences in Π_m, it may happen that each term of the sequence is expressed in reduced form, but the obvious limiting expression is no longer in reduced form. This leads to the following definition. Let $R_n(\mathbf{x}) = \sum_{k=1}^m c_{k,n}\sigma(\mathbf{w}_{k,n}\cdot\mathbf{x} + b_{k,n}) \in \Pi_m$ for $n = 1, 2, \ldots$. We say that $\{R_n\}$ is in *asymptotically reduced form* if

(i) each R_n is in reduced form,
(ii) $\lim_{n\to\infty}\mathbf{w}_{k,n} = \mathbf{w}_k$, $\lim_{n\to\infty} b_{k,n} = b_k$ for $k = 1, \ldots, m$, and
(iii) the hyperspaces $\mathbf{w}_k\cdot\mathbf{x} + b_k = 0$ are all distinct.

It is not clear that every sequence in Π_m, where the parameters form convergent sequences, can be rewritten in asymptotically reduced form. Nevertheless, the following result shows that this almost holds.

Proposition 3.2. *Let $P_n(\mathbf{x}) = \sum_{k=1}^m c_{k,n}\sigma(\mathbf{w}_{k,n}\cdot\mathbf{x}+b_{k,n}) \in \Pi_m$, $\lim_{n\to\infty}\mathbf{w}_{k,n} = \mathbf{w}_k$ and $\lim_{n\to\infty} b_{k,n} = b_k$ for $k = 1, \ldots, m$. Then P_n can be expressed almost everywhere as $P_n = R_n + Q_n$, where $\{R_n\} \subset \Pi_m$ is in asymptotically reduced form, $\{Q_n\} \subset \Pi_{2m}$, and*

$$\lim_{n\to\infty} \lambda(\{\mathbf{x} \in K : Q_n(\mathbf{x}) \neq 0\}) = 0 \tag{3.7}$$

for any compact set K.

Proof. By a suitable rearrangement of terms, we may assume that among the hyperspaces $H_k : \mathbf{w}_k\cdot\mathbf{x} + b_k = 0$, $1 \le k \le m$, the hyperspaces $H_1, \ldots, H_l$ are distinct. For the sake of concreteness, we assume that $\mathbf{w}_1 = 0$ and $\mathbf{w}_k \neq 0$ for $k = 2, \ldots, l$. The proof is only slightly different, and even simpler, if there is no degenerate hyperspace. Let

$$\mathscr{J}_1 := \{j : l+1 \le j \le m,\ \mathbf{w}_j = 0\}$$

and

$$\mathscr{J}_k := \{j : l+1 \le j \le m,\ \mathbf{w}_j = -\mathbf{w}_k,\ b_j = -b_k\}, \qquad k = 2, \ldots, l.$$

We may then write

$$\begin{aligned} P_n(\mathbf{x}) = {} & \left(c_{1,n} + \sum_{j\in\mathscr{J}_1} c_{j,n}\right) \\ & + \sum_{k=2}^{l}\left(c_{k,n}\sigma(\mathbf{w}_{k,n}\cdot\mathbf{x} + b_{k,n}) + \left(\sum_{j\in\mathscr{J}_k} c_{j,n}\right)\sigma(-\mathbf{w}_{k,n}\cdot\mathbf{x} - b_{k,n})\right) \\ & + \sum_{k=2}^{l}\sum_{j\in\mathscr{J}_k} c_{j,n}(\sigma(\mathbf{w}_{j,n}\cdot\mathbf{x} + b_{j,n}) - \sigma(-\mathbf{w}_{k,n}\cdot\mathbf{x} - b_{k,n})). \end{aligned} \tag{3.8}$$

We define

$$R_n(\mathbf{x}) := \left(c_{1,n} + \sum_{j\in\mathscr{I}_1} c_{j,n} + \sum_{k=2}^{l}\sum_{j\in\mathscr{I}_k} c_{j,n}\right)\sigma(\mathbf{0}\cdot\mathbf{x}+1) + \sum_{k=2}^{l}\left(c_{k,n} - \sum_{j\in\mathscr{I}_k} c_{j,n}\right)\sigma(\mathbf{w}_{k,n}\cdot\mathbf{x}+b_{k,n}) \tag{3.9}$$

and

$$Q_n(\mathbf{x}) := \sum_{k=2}^{l}\sum_{j\in\mathscr{I}_k} c_{j,n}(\sigma(\mathbf{w}_{j,n}\cdot\mathbf{x}+b_{j,n}) - \sigma(-\mathbf{w}_{k,n}\cdot\mathbf{x}-b_{k,n})). \tag{3.10}$$

Then by (3.1) we see that $P_n = R_n + Q_n$ almost everywhere. From the construction, it is clear that $\{R_n\}$ is in asymptotically reduced form. Let

$$E_{j,k} := \{\mathbf{x}\in\mathbf{R}^s\colon \sigma(\mathbf{w}_{j,n}\cdot\mathbf{x}+b_{j,n}) - \sigma(-\mathbf{w}_{k,n}\cdot\mathbf{x}-b_{k,n}) \neq 0\}.$$

If $2 \le k \le l$ and $j \in \mathscr{I}_k$, then $\lim_{n\to\infty}\mathbf{w}_{j,n} = -\mathbf{w}_k = -\lim_{n\to\infty}\mathbf{w}_{k,n}$ and $\lim_{n\to\infty} b_{j,n} = -b_k = -\lim_{n\to\infty} b_{k,n}$. Therefore, for any compact set K, it is clear that $\lim_{n\to\infty}\lambda(E_{j,k}\cap K) = 0$ for $k = 2, \dots, l$ and $j \in \mathscr{I}_k$. Hence, it follows that (3.7) holds. □

The next proposition describes a compactness property for the class Π_m.

Proposition 3.3. *Let $P_n \in \Pi_m$ be such that for every compact set $K \subseteq \mathbf{R}^s$,*

$$\limsup_{n\to\infty} \|P_n\|_{1,K} \le 1. \tag{3.11}$$

Then there exists a $P \in \Pi_m$ and a subsequence Λ of integers such that

$$\lim_{n\to\infty,\, n\in\Lambda} P_n(\mathbf{x}) = P(\mathbf{x}) \quad \text{a.e. in } \mathbf{R}^s. \tag{3.12}$$

Proof. We write each P_n in reduced form, say

$$P_n(\mathbf{x}) = \sum_{k=1}^{m} c_{k,n}\sigma(\mathbf{w}_{k,n}\cdot\mathbf{x}+b_{k,n}). \tag{3.13}$$

Then there exists a subsequence Λ_1 of integers, $\mathbf{w}_k \in \mathbf{R}^s$, and $b_k \in \mathbf{R}$ such that

$$\lim_{n\to\infty,\, n\in\Lambda_1} \mathbf{w}_{k,n} = \mathbf{w}_k, \qquad \lim_{n\to\infty,\, n\in\Lambda_1} b_{k,n} = b_k, \qquad k = 1, \dots, m. \tag{3.14}$$

For each $n \in \Lambda_1$, we write $P_n = R_n + Q_n$, where

$$R_n := \sum_{k=1}^{m} a_{k,n}\sigma(\mathbf{v}_{k,n}\cdot\mathbf{x}+d_{k,n}), \tag{3.15}$$

$\{R_n\}$ is in asymptotically reduced form, and Q_n satisfies (3.7). Then we have, for a subsequence Λ_2 of Λ_1,

$$\lim_{n\to\infty,\, n\in\Lambda_2} Q_n(\mathbf{x}) = 0 \quad \text{a.e. in } \mathbf{R}^s. \tag{3.16}$$

We recall that there exist $\mathbf{v}_k \in \mathbf{R}^s$ and $d_k \in \mathbf{R}$ such that

$$\lim_{n\to\infty,\, n\in\Lambda_2} \mathbf{v}_{k,n} = \mathbf{v}_k, \qquad \lim_{n\to\infty,\, n\in\Lambda_2} d_{k,n} = d_k, \qquad k = 1, \dots, m. \tag{3.17}$$

We show that for some subsequence Λ of Λ_2, the sequences $\{a_{k,n}\}_{n\in\Lambda}$ converge. In view of (3.16), this will complete the proof. With

$$A_n := \max_{1\le k\le m} |a_{k,n}|,$$

it suffices to show that

$$\liminf_{n\to\infty,\, n\in\Lambda_2} A_n < \infty. \tag{3.18}$$

If possible, suppose that (3.18) is false. Let $\varepsilon > 0$ and K be a compact subset of $\mathbf{R}^s$, both arbitrarily given. We may assume that $\lambda(K) > \varepsilon$. In view of (3.7), there exists a set $E \subseteq K$ with $\lambda(E) \le \varepsilon/2$ such that $P_n(\mathbf{x}) = R_n(\mathbf{x})$ for $\mathbf{x} \in K\backslash E$ and all sufficiently large $n \in \Lambda_2$. Therefore, (3.11) implies that

$$\limsup_{n\to\infty,\, n\in\Lambda_2} \|R_n\|_{1,K\backslash E} \le 1. \tag{3.19}$$

Since $\lim_{n\to\infty,\, n\in\Lambda_2} A_n = \infty$, we deduce that the sequence $\{\widetilde{R}_n := A_n^{-1}R_n\}_{n\in\Lambda_2}$ converges to zero in measure on every compact subset of $\mathbf{R}^s$. In particular, there is a subsequence Λ_3 of Λ_2 such that

$$\lim_{n\to\infty,\, n\in\Lambda_3} \widetilde{R}_n(\mathbf{x}) = 0 \quad \text{a.e. in } \mathbf{R}^s. \tag{3.20}$$

In view of the definition of A_n, there is a subsequence Λ_4 of Λ_3 and numbers $a_k \in \mathbf{R}$ such that at least one of the a_k's has absolute value 1 and $\lim_{n\to\infty,\, n\in\Lambda_4}(a_{k,n}/A_n) = a_k$ for $1 \le k \le m$. We set

$$R(\mathbf{x}) := \sum_{k=1}^{m} a_k\sigma(\mathbf{v}_k\cdot\mathbf{x} + d_k).$$

Since $\{R_n\}_{n\in\Lambda_4}$ is in asymptotically reduced form, R is in reduced form. Moreover, since at least one of the a_k's is nonzero, R is not identically equal to zero. Since $\lim_{n\to\infty,\, n\in\Lambda_4} \widetilde{R}_n(\mathbf{x}) = R(\mathbf{x})$ for almost all $\mathbf{x} \in \mathbf{R}^s$, this contradicts with (3.20). This proves (3.18), and the proof is complete. □

Theorem 2.2 is quite simple to prove by using Proposition 3.3. We recall that a function ϕ: $\mathbf{R}^s \to \mathbf{R}$ is called a test function if ϕ is infinitely differentiable on $\mathbf{R}^s$ and every derivative ψ of ϕ (of arbitrary order) satisfies the condition $\sup_{\mathbf{x}\in\mathbf{R}^s} |\psi(\mathbf{x})|(1 + |\mathbf{x}|)^N < \infty$ for any integer $N \ge 0$. The class $\mathscr{D}$ of all test functions forms a locally convex space with a suitable topology [18]. A continuous linear functional on $\mathscr{D}$ is called a (tempered) distribution. We refer the reader to [18] for a detailed exposition of the properties of test functions and distributions. Here, we only recall that the Fourier transform of a test function ϕ is defined by

$$\hat{\phi}(\mathbf{x}) = (2\pi)^{-s/2}\int_{\mathbf{R}^s} \exp(-i\mathbf{x}\cdot\mathbf{t})\phi(\mathbf{t})\,d\lambda(\mathbf{t}), \qquad \mathbf{x} \in \mathbf{R}^s,$$

and that of a distribution u by

$$\hat{u}(\phi) = u(\hat{\phi}), \qquad \phi \in \mathscr{D}.$$

Moreover, if u and v are distributions such that $\hat{u} = \hat{v}$, then $u = v$.

Proof of Theorem 2.2. If possible, let $m \ge 1$ be a fixed integer and $\{P_n\} \subset \Pi_m$ be a sequence which satisfies (2.2) for every compact set $K \subset \mathbf{R}^s$. Then

Proposition 3.3 implies that there exists a $P \in \Pi_m$ and a subsequence Λ of integers such that $P(\mathbf{x}) = \lim_{n\to\infty, n\in\Lambda} P_n(\mathbf{x})$ almost everywhere. It follows that $\chi_Q(\mathbf{x}) = P(\mathbf{x})$ almost everywhere. Given any finite set of hyperspaces $\mathbf{w}_k \cdot \mathbf{x} + b_k = 0$, it is easy to construct a nonnegative test function that vanishes on all of them. Therefore, a comparison of the Fourier transforms of the distributions χ_Q and P shows that $\chi_Q(\mathbf{x}) = P(\mathbf{x})$ cannot hold almost everywhere. (A more elementary proof of the fact that $\chi_Q(\mathbf{x}) \neq P(\mathbf{x})$ for any $P \in \Pi_m$ is given by Blum and Li in [2].) □

Our proof of Theorem 2.2 depends very heavily on the properties of the ideal (Heaviside) sigmoidal function. We believe, however, that the same result still holds for any other sigmoidal function which is of bounded variation. The research on this problem is postponed to a later date.

Proof of Theorem 2.3. Let $\varepsilon > 0$ and $L \geq 2$. We construct a network $N_{\varepsilon,p,L}$ with $2s+1$ neurons arranged in two hidden layers such that

$$\|\chi_Q - N_{\varepsilon,p,L}\|_{p,Q_L} \leq \varepsilon, \tag{3.21}$$

where $Q_L := [-L, L]^s$. Then the sequence $\{N_{1/n,1,n}\} \subset \Pi_{2s+1,2,s,\sigma}$ will satisfy (2.2) for every compact set $K \subset \mathbf{R}^s$. We set

$$\begin{gathered} \eta := \frac{2s-(2s-1)M}{2(2s+1)}, \quad C := s + \left(s - \frac{1}{2}\right)(M-\eta), \\ \delta := \frac{\varepsilon^p}{s2^{2s+2}(M+1)^p} \end{gathered} \tag{3.22}$$

and find positive constants A and B such that

$$\begin{aligned} &|\sigma(x) - 1| < \eta && \text{if } x \geq A, \\ &|\sigma(x) - 1| < (\varepsilon^p/2(2L)^s)^{1/p} && \text{if } x \geq B, \\ &|\sigma(x)| < \eta && \text{if } x \leq -A, \\ &|\sigma(x)| < (\varepsilon^p/2(2L)^s)^{1/p} && \text{if } x \leq -B. \end{aligned} \tag{3.23}$$

In (3.23), the constant A is fixed depending only on σ and s, and the constant B depends upon L and ε as well. Here, we assume that ε is so small that $\delta < 1$. We define

$$\begin{aligned} N_{\varepsilon,p,L}(\mathbf{x}) := \sigma\Bigg\{\frac{4B}{2s-(2s-1)M}\Bigg[&\sum_{l=1}^{s}\sigma\left(\frac{A}{\delta}(1+x^{(l)})\right) \\ &+\sum_{l=1}^{s}\sigma\left(\frac{A}{\delta}(1-x^{(l)})\right) - C\Bigg]\Bigg\}. \end{aligned} \tag{3.24}$$

If $\mathbf{x} \in [-1+\delta, 1-\delta]^s$, then for each l, $1 \leq l \leq s$, we have $(A/\delta)(1 \pm x^{(l)}) \geq A$, and we deduce from (3.23) that

$$\sum_{l=1}^{s}\sigma\left(\frac{A}{\delta}(1+x^{(l)})\right) + \sum_{l=1}^{s}\sigma\left(\frac{A}{\delta}(1-x^{(l)})\right) \geq 2s - 2s\eta = C + \frac{2s-(2s-1)M}{4},$$

and consequently that

$$|N_{\varepsilon,p,L}(\mathbf{x}) - 1| \leq \left(\frac{\varepsilon^p}{2(2L)^s}\right)^{1/p}, \qquad \mathbf{x} \in [-1+\delta, 1-\delta]^s. \tag{3.25}$$

Next, suppose that $\mathbf{x} \notin [-1-\delta, 1+\delta]^s$. Then among the $2s$ numbers $\sigma((A/\delta)(1+x^{(l)}))$ and $\sigma((A/\delta)(1-x^{(l)}))$, $1 \le l \le s$, at least one is less than η, while the rest do not exceed M. So, for $\mathbf{x} \notin [-1-\delta, 1+\delta]^s$, we have

$$\sum_{l=1}^{s} \sigma\left(\frac{A}{\delta}(1+x^{(l)})\right) + \sum_{l=1}^{s} \sigma\left(\frac{A}{\delta}(1-x^{(l)})\right) \le \eta + (2s-1)M = C - \frac{2s-(2s-1)M}{4}$$

and consequently,

$$|N_{\varepsilon,p,L}(\mathbf{x})| \le \left(\frac{\varepsilon^p}{2(2L)^s}\right)^{1/p}, \qquad \mathbf{x} \notin [-1-\delta, 1+\delta]^s. \tag{3.26}$$

If $E := [-1-\delta, 1+\delta]^s \backslash [-1+\delta, 1-\delta]^s$, then (3.25) and (3.26) imply that

$$\|\chi_Q - N_{\varepsilon,p,L}\|^p_{p,Q_L\backslash E} < \varepsilon^p/2. \tag{3.27}$$

Also, since $\lambda(E) \le 2^{2s+1}s\delta$ and $|\chi_Q(\mathbf{x}) - N_{\varepsilon,L}(\mathbf{x})| \le (M+1)$ for all $\mathbf{x} \in \mathbf{R}^s$, we have, from (3.22), that

$$\|\chi_Q - N_{\varepsilon,p,L}\|^p_{p,E} \le \varepsilon^p/2. \tag{3.28}$$

Now, the estimate (3.21) follows from (3.27) and (3.28).

Next, we write $\mathbf{Z}_n := \{0, \dots, n-1\}$. Given $f \in L^p(Q)$, we set

$$a_{\mathbf{m},n} := n^s \int_{R_{\mathbf{m},n}} f(\mathbf{t})\, d\lambda(\mathbf{t}), \qquad \mathbf{m} \in \mathbf{Z}_n^s, \tag{3.29a}$$

and

$$S_n(f, \mathbf{x}) := \sum_{\mathbf{m}\in\mathbf{Z}_n^s} a_{\mathbf{m},n}\chi_{R_{\mathbf{m},n}}(\mathbf{x}) = \sum_{\mathbf{m}\in\mathbf{Z}_n^s} a_{\mathbf{m},n}\chi_Q(2n(\mathbf{x} - c_{\mathbf{m},n})), \tag{3.29b}$$

where the vectors $c_{\mathbf{m},n}$ are defined by

$$c^{(j)}_{\mathbf{m},n} := -1 + \frac{2m^{(j)}+1}{2n}. \tag{3.30}$$

It is then easy to verify that

$$\|f - S_n(f, \cdot)\|_{p,R_{\mathbf{m},n}} \le \omega\left(L^p(S_{\mathbf{m},n}), S_{\mathbf{m},n}; f, \frac{2}{n}\right), \qquad \mathbf{m} \in \mathbf{Z}_n^s. \tag{3.31}$$

If $f = 0$ almost everywhere on Q, then we set $G_n(f, \mathbf{x}) = 0$, and (2.7) is trivial. Therefore, let us only consider the case $\|f\|_{p,Q} \ne 0$. With

$$\varepsilon := \min_{\mathbf{m}\in\mathbf{Z}_n^s} \frac{\omega(L^p(S_{\mathbf{m},n}), S_{\mathbf{m},n}; f, \frac{2}{n})}{n^s\|f\|_{p,Q}},$$

we define

$$G_n(f, \mathbf{x}) := \sum_{\mathbf{m}\in\mathbf{Z}_n^s} a_{\mathbf{m},n}N_{\varepsilon,p,4n}(2n(\mathbf{x} - c_{\mathbf{m},n})). \tag{3.32}$$

By using (3.21) and (3.31), only an easy computation is needed to verify (2.7). □

4. Implementation of spline wavelets

In this section, we consider the approximation of a spline function of higher order using a neural network with a sigmoidal function of higher order as its activation function. This will be applied to an implementation of the compactly supported spline-wavelets introduced by Chui and Wang [7].

We start by recalling the definition of the (cardinal) B-splines in the univariate case. Let $N_1 := \chi_{[0,1]}$. Then the (cardinal) B-spline of order m is defined inductively by the formula

$$N_m(x) := \int_0^1 N_{m-1}(x-t)\,dt, \qquad m = 2, 3, \dots. \tag{4.1}$$

Some of the well-known properties of B-splines are as follows (cf. [19]). The B-spline N_m is an $(m-2)$-times continuously differentiable nonnegative function which is identically equal to 0 outside of the interval $[0, m]$. On each of the subintervals $[k, k+1]$, $0 \le k \le m-1$, N_m is a polynomial of degree at most $m-1$. At any point $x \in \mathbf{R}\backslash\mathbf{Z}$, there are exactly $m-1$ integers k such that $N_m(x-k) \neq 0$. The sum of all these functions which are nonzero at x is 1. Thus, the B-splines provide a very useful tool for localized approximation.

In the multivariate setting, the simplest analogue of the B-spline is the tensor-product B-spline. For each integer $m \ge 1$ and $\mathbf{x} = (x^{(1)}, \dots, x^{(s)}) \in \mathbf{R}^s$, this is defined by the formula

$$N_m^s(\mathbf{x}) := \prod_{j=1}^{s} N_m(x^{(j)}). \tag{4.2}$$

In this section, we demonstrate that a tensor-product B-spline N_m^s can be approximated arbitrarily closely using a neural network whose activation function is a sigmoidal function of order $k \ge 2$ (cf. [16]), the size of the network depending only on k and m, and not on the degree of accuracy.

Definition 4.1. Let $\sigma\colon \mathbf{R} \to \mathbf{R}$ be a continuous function and $k \ge 0$ be an integer. We say that σ is a sigmoidal function of order k if each of the following conditions is satisfied:

$$\lim_{x\to-\infty} \frac{\sigma(x)}{x^k} = 0, \qquad \lim_{x\to\infty} \frac{\sigma(x)}{x^k} = 1, \tag{4.3a}$$

$$|\sigma(x)| \le K(1+|x|)^k, \qquad x \in \mathbf{R}, \tag{4.3b}$$

where $K \ge 1$ is a constant.

In [16], a sigmoidal function of order k is used to obtain specific networks for approximating an arbitrary continuous function. In the univariate case, the rates obtained in [16] are optimal and the approximation is local. In [15], a multilayered network was constructed to give a localized approximation with optimal rates using such functions. The following theorem is an extension of the ideas in [15].

Theorem 4.2. *Let $\varepsilon > 0$ and $m > 0$, $k \ge 2$ be integers. We let*

$$p := \lceil \log(ms - s)/\log k \rceil$$

(*i.e., the smallest integer such that* $ms - s \leq k^p$) *and define*

$$M := 2spm^s(k+1)\binom{k^p+s}{s}. \tag{4.4}$$

Then there exists a neural network $\widetilde{N}(\mathbf{x}) := \widetilde{N}(k, m, \varepsilon; \mathbf{x}) \in \Pi_{M,p,s,\sigma}$ *such that*

$$|N_m^s(\mathbf{x}) - \widetilde{N}(\mathbf{x})| \leq \varepsilon, \qquad \mathbf{x} \in [0,1]^s. \tag{4.5}$$

By Theorem 4.2, we may translate many theorems in spline approximation to the corresponding theorems in approximation using neural networks. Some applications are already given in [15]. Here, we illustrate the use of Theorem 4.2 in the implementation of the compactly supported spline wavelets. We observe that the notation which we use in the sequel is different from that used in [7]. In order to discuss these, we denote the (univariate) B-spline N_m by Φ_0 and write

$$\phi_{0;k,j} := \Phi_0(2^k x - j), \qquad k, j \in \mathbf{Z}, \tag{4.6}$$

and for each $k \in \mathbf{Z}$,

$$V_k := \operatorname{span}\{\phi_{0;k,j},\ j \in \mathbf{Z}\}, \tag{4.7}$$

where $\operatorname{span} A$ denotes the closed (in $L^2(\mathbf{R})$) linear span of A. Then $\{V_k : k \in \mathbf{Z}\}$ defines a *multiresolution analysis* of $L^2(\mathbf{R})$ in the sense that

(i) $V_k \subseteq V_{k+1}$, $k \in \mathbf{Z}$,
(ii) $\operatorname{clos}_{L^2}(\bigcup_{k\in\mathbf{Z}} V_k) = L^2(\mathbf{R})$,
(iii) $\bigcap_{k\in\mathbf{Z}} V_k = \{0\}$, and
(iv) for each $k \in \mathbf{Z}$, $\{\phi_{0;k,j} : j \in \mathbf{Z}\}$ is an unconditional basis for V_k.

We define the *wavelet space* W_k to be the orthogonal complement of V_k in V_{k+1}. It is shown in [7] that the space W_k can also be written as the linear span of the translates of a function Φ_1, similar to (4.7), as follows: Let

$$\Phi_1(x) := \sum_{j=0}^{3m-2} q_j N_m(2x - j),$$

where

$$q_j := \frac{(-1)^j}{2^{m-1}} \sum_{l=0}^{m} \binom{m}{l} N_{2m}(j - l + 1), \qquad j = 0, \ldots, 3m-2,$$

and define the wavelets *at level* k by

$$\phi_{1;k,j}(x) := \Phi_1(2^k x - j). \tag{4.8}$$

Then

$$W_k = \operatorname{span}\{\phi_{1,k,j} : j \in \mathbf{Z}\}, \qquad k \in \mathbf{Z}. \tag{4.9}$$

Among the important properties of these wavelets are the following. The wavelet Φ_1 is supported on the interval $[0, 2m-1]$. There is a refinement equation

$$\Phi_\nu(x) = \sum_{j=0}^{3m-2} q_{\nu,j}\Phi_0(2x - j), \qquad \nu = 0, 1,\ x \in \mathbf{R}, \tag{4.10}$$

where

$$q_{0,j} := 2^{m-1}\binom{m}{j}, \qquad j = 0, \dots, m,$$
$$q_{1,j} := q_j, \qquad j = 0, \dots, 3m-2$$

and $q_{\nu,j} := 0$ if the subscript j is outside the ranges prescribed above. In the reverse direction, there is a *decomposition relation*

$$\Phi_0(2x-l) = \sum_{\nu=0}^{1}\sum_{j\in\mathbf{Z}} a_{\nu;l-2j}\Phi_\nu(x-j), \tag{4.11}$$

where the constants $a_{0;j}$ and $a_{1;j}$ are determined precisely in [7].

In the multivariate setting, the simplest way to generalize the Chui-Wang wavelets is again by using tensor products. Thus, we write $\mathbf{Z}_2^s := \{0, 1\}^s$ and define the wavelets (resp. scaling functions when $\mathbf{p} = \mathbf{0}$) by

$$\Phi_{\mathbf{p}}^s(\mathbf{x}) = \prod_{j=0}^{s}\Phi_{p^{(j)}}(x^{(j)}), \qquad \mathbf{p} \in \mathbf{Z}_2^s, \tag{4.12}$$

where $\mathbf{p} = (p^{(1)}, \dots, p^{(s)})$ and

$$\phi^s_{\mathbf{p};k,\mathbf{j}}(\mathbf{x}) = \Phi_{\mathbf{p}}^s(2^k\mathbf{x}-\mathbf{j}), \qquad k \in \mathbf{Z}, \mathbf{j} \in \mathbf{Z}^s, \mathbf{p} \in \mathbf{Z}_2^s. \tag{4.13}$$

The refinement and decomposition equations (4.10) and (4.11) take on the form

$$\phi^s_{\mathbf{p};k,\mathbf{j}}(\mathbf{x}) = \sum_{0\le\mathbf{l}\le 3m-2} q_{\mathbf{p},\mathbf{l}}\phi^s_{\mathbf{0};k+1,2\mathbf{j}+\mathbf{l}}(\mathbf{x}), \qquad \mathbf{p} \in \mathbf{Z}_2^s, \tag{4.14}$$

and

$$\phi^s_{\mathbf{0};k+1,\mathbf{l}}(\mathbf{x}) = \sum_{\mathbf{p}\in\mathbf{Z}_2^s}\sum_{\mathbf{j}\in\mathbf{Z}^s} a_{\mathbf{p};\mathbf{l}-2\mathbf{j}}\phi^s_{\mathbf{p};k,\mathbf{j}}(\mathbf{x}), \tag{4.15}$$

where

$$q_{\mathbf{p},\mathbf{l}} := \prod_{j=1}^{s} q_{p^{(j)},l^{(j)}}, \qquad a_{\mathbf{p},\mathbf{l}} := \prod_{j=1}^{s} a_{p^{(j)},l^{(j)}}, \tag{4.16}$$

and the symbol $0 \le \mathbf{l} \le 3m-2$ means that all components of $\mathbf{l}$ are between 0 and $3m-2$.

In order to implement the wavelets $\Phi_{\mathbf{p}}^s$ using a neural network, we observe that at any level $k \ge 0$ only the scaling functions $\{\phi^s_{\mathbf{0};k+1,\mathbf{j}}\}_{-m-1\le\mathbf{j}\le 2^{k+1}}$ are nonzero on $[0, 1]^s$. We approximate these, using Theorem 4.2, by $\tilde{\phi}_{\mathbf{0},\varepsilon;k+1,\mathbf{j}}$ and then define $\phi^s_{\mathbf{p};k,\mathbf{j}}$ for $\mathbf{p} \in \mathbf{Z}_2^s\backslash\{\mathbf{0}\}$ using (4.14) to yield the networks $\tilde{\phi}_{\mathbf{p},\varepsilon;k,\mathbf{j}}$. The functions $\phi^s_{\mathbf{0};k,\mathbf{j}}$ are then defined in two different ways. Theorem 4.2 ensures that the difference between these two implementations can be made arbitrarily small **using only a fixed number of neurons**, *independent of the accuracy desired*. Therefore, given a wavelet expansion of a function, we may use the networks so constructed to directly synthesize the function within essentially the same margin of accuracy as the expansion itself. The total number of neurons required in the process is proportional to the necessary number of wavelets. Similarly, we observe that only finitely many terms in (4.15) are nonzero if $\mathbf{x} \in [0, 1]^s$. Therefore, the networks $\tilde{\phi}_{\mathbf{p}, \varepsilon; k, \mathbf{j}}$ can be used to construct $\phi^s_{\mathbf{0}; k+1, \mathbf{j}}$ to any degree of accuracy. Again, the total number of neurons required in this decomposition is proportional to the number of wavelets which enter into the expansion.

We now turn to the proof of Theorem 4.2.

Proof of Theorem 4.2. This proof proceeds in several steps.

Step I. Given an integer $p \geq 1$, we construct a network $P(A, p, \varepsilon; x) \in \Pi_{p, 1, 1, \sigma}$ with the property that

$$|x_+^{k^p} - P(A, p, \varepsilon; x)| \leq \varepsilon, \qquad |x| \leq A. \tag{4.17}$$

This is done in [15]; we merely sketch the proof. Let $\delta := (\varepsilon/2^{k+1}K)^{1/k}$. Find $B > 0$ such that

$$|\sigma(x)| \leq \varepsilon |x|^k, \quad x < -B, \qquad |\sigma(x) - x^k| \leq \varepsilon |x|^k, \quad x > B.$$

Also, set

$$P(1, 1, \varepsilon; x) := (\delta/B)^k \sigma(Bx/\delta). \tag{4.18}$$

It is easy to verify (cf. [15]) that (4.17) is satisfied with $A = 1$ and $p = 1$. If we set

$$P(A, 1, \varepsilon; x) := A^k P(1, 1, A^{-k}\varepsilon; x/A), \tag{4.19}$$

then it is clear that (4.17) is satisfied with $p = 1$. Since σ is uniformly continuous on any compact interval, we may find $\eta > 0$ such that

$$|P(1, 1, \varepsilon/2; x) - P(1, 1, \varepsilon/2; y)| \leq \varepsilon/2, \qquad |x - y| \leq \eta, \ |x|, |y| \leq 2.$$

We define $P(1, l, \varepsilon; x)$ inductively as follows:

$$P(1, l, \varepsilon; x) := P(1, 1, \varepsilon/2; P(1, l-1, \eta; x)), \qquad l \geq 2. \tag{4.20}$$

It is shown in [15] that $P(1, p, \varepsilon; x)$ satisfies (4.17) with $A = 1$. We now define

$$P(A, p, \varepsilon; x) := A^{k^p} P(1, p, A^{-k^p}\varepsilon; x/A). \tag{4.21}$$

Step II. We define

$$\widetilde{P}(A, p, \varepsilon; x) := P(A, p, \varepsilon/2; x) + (-1)^{k^p} P(A, p, \varepsilon/2; -x). \tag{4.22}$$

The network so defined has $2p$ neurons arranged in p layers and satisfies

$$|x^{k^p} - \widetilde{P}(A, p, \varepsilon; x)| \leq \varepsilon, \qquad |x| \leq A. \tag{4.23}$$

Step III. We construct a network $Q(A, \varepsilon; x) \in \Pi_{k+1, 1, 1, \sigma}$ such that

$$|x_+ - Q(A, \varepsilon; x)| \leq \varepsilon, \qquad |x| \leq A. \tag{4.24}$$

We obtain numbers α_μ, $0 \leq \mu \leq k$, which solve the system of equations

$$\sum_{\mu=0}^{k} \alpha_\mu \mu^{k-\nu} = \frac{1}{k}\delta_{\nu, 1}, \qquad \nu = 0, 1, \ldots, k.$$

Then it is readily verified that

$$\sum_{\mu=0}^{k} \alpha_\mu (x + \mu)^k = x, \qquad x \in \mathbf{R}. \tag{4.25}$$

Let

$$N \geq \max\left\{\frac{2(2k)^k \sum |\alpha_\mu|}{\varepsilon}, k\right\}.$$

In view of (4.25), if $x > 0$ or $x \leq -k/N$, then

$$\sum_{\mu=0}^{k} \alpha_\mu N^{k-1}\left(x + \frac{\mu}{N}\right)_+^k = \frac{1}{N}\sum_{\mu=0}^{k} \alpha_\mu (Nx+\mu)_+^k = Nx_+/N = x_+.$$

If $-k/N \leq x \leq 0$, then the above choice of N implies that

$$\left|\sum_{\mu=0}^{k} \alpha_\mu N^{-k-1}(x+\mu/N)_+^k\right| \leq \frac{(2k)^k}{N}\sum |\alpha_\mu| \leq \varepsilon/2.$$

Thus, we have

$$\left|\sum_{\mu=0}^{k} \alpha_\mu N^{k-1}\left(x+\frac{\mu}{N}\right)_+^k - x_+\right| \leq \varepsilon/2, \quad x \in \mathbf{R}. \tag{4.26}$$

We now let $\eta := \varepsilon(2N^{k-1}\sum|\alpha_\mu|)^{-1}$ and define

$$Q(A, \varepsilon; x) := \sum_{\mu=0}^{k} N^{k-1}\alpha_\mu P(A+1, 1, \eta; x+\mu/N). \tag{4.27}$$

From (4.26) and (4.17), it is easy to see that (4.24) is satisfied.

Step IV. We construct a network

$$R(A, m, \varepsilon; \mathbf{x}) = R_{s,k}(A, m, \varepsilon; \mathbf{x}) \in \Pi_{M,p+1,s,\sigma}$$

such that

$$|\mathbf{x}_+^{m-1} - R(A, m, \varepsilon; \mathbf{x})| \leq \varepsilon, \qquad \max_{1\leq j\leq s}|x^{(j)}| \leq A, \tag{4.28}$$

where, with $p = \lceil \log(ms-s)/\log k\rceil$, we have used the notation

$$M := 2ps(k+1)\binom{k^p+s}{s}.$$

Following [6], we choose numbers a_i and b_i and vectors $\mathbf{w}_i \in \mathbf{R}^s$ such that

$$\mathbf{x}^{m-1} = \sum_{i=1}^{N} a_i(\mathbf{w}_i \cdot \mathbf{x} + b_i)^{k^p}, \tag{4.29}$$

where $N := \binom{k^p+s}{s}$. With

$$L := \max_{1\leq i\leq N}\left(\sum_{j=1}^{s}|w_i^{(j)}| + |b_i|\right), \qquad \eta := \left(k^p(L+1)^{k^p}\sum|a_i|\right)^{-1},$$

we define

$$R(1, m, \varepsilon; \mathbf{x}) := \sum_{i=1}^{N} a_i\widetilde{P}\left(L+1, p, \eta; \sum_{j=1}^{s} w_i^{(j)} Q(1, \eta; x^{(j)}) + b_i\right). \tag{4.30}$$

If $\max_{1\le j\le s}|x^{(j)}|\le 1$, then (4.24) implies that

$$\sum_j \left| w_i^{(j)} Q(1,\eta;x^{(j)}) + b_i - \left(\sum_j w_i^{(j)} x_+^{(j)} + b_i\right)\right| \\ \le \sum_j |w_i^{(j)}||Q(1,\eta;x^{(j)}) - x_+^{(j)}| \le L\eta .$$

We have $|\sum w_i^{(j)} x_+^{(j)} + b_i| \le L$, and hence, for all sufficiently small $\eta > 0$,

$$\left|\sum_j w_i^{(j)} Q(1,\eta;x^{(j)}) + b_i\right| \le L+1 .$$

Therefore,

$$\left|\left(\sum_j w_i^{(j)} Q(1,\eta;x^{(j)}) + b_i\right)^{k^p} - \left(\sum_j w_i^{(j)} x_+^{(j\,.)} + b_i\right)^{k^p}\right| \\ \le k^p(L+1)^{k^p-1}L\eta . \tag{4.31}$$

In view of (4.23), we obtain, for $1\le i\le N$,

$$\left|\tilde{P}\left(L+1,p,\eta;\sum_j w_i^{(j)} Q(1,\eta;x^{(j)}) + b_i\right) \right. \\ \left. - \left(\sum_j w_i^{(j)} Q(1,\eta;x^{(j)}) + b_i\right)^{k^p}\right| \le \eta . \tag{4.32}$$

The estimate (4.28) follows from (4.29), (4.30), (4.31), and (4.32) in the case when $A = 1$. For the general A, we define

$$R(A,m,\varepsilon;\mathbf{x}) := A^{ms-s}R(1,m,A^{-ms+s}\varepsilon;\mathbf{x}/A) . \tag{4.33}$$

Step V. The quantity $N_m^s(\mathbf{x})$ can be written as a linear combination of m^s quantities of the form $(\mathbf{x}-\mathbf{j})_+^{m-1}$. Hence, Theorem 4.2 follows from Step IV above. □

Bibliography

1. A. R. Barron, *Universal approximation bounds for superposition of a sigmoidal function*, preprint, November 1990.
2. E. K. Blum and L. K. Li, *Approximation theory and neural networks*, Neural Networks **4** (1991), 511–515.
3. S. M. Caroll and S. M. Dickinson, *Construction of neural nets using the Radon transform*, preprint, 1990.
4. T. P. Chen, H. Chen, and R. W. Liu, *A constructive proof of approximation by superposition of sigmoidal functions for neutral networks*, preprint, 1990.
5. C. K. Chui and X. Li, *Approximation by ridge functions and neural networks with one hidden layer*, J. Approx. Theory **70** (1992), 131–141.
6. ———, *Realization of neural networks with one hidden layer*, Multivariate approximations: From CAGD to Wavelets (K. Jetter and F. Utreras, eds.), World Scientific Publ., Singapore, 1993, pp. 77–89.

7. C. K. Chui and J. Z. Wang, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc. **330** (1992), 903–916.

8. G. Cybenko, *Approximation by superposition of sigmoidal functions*, Math. Control Signals Systems **2** (4) (1989), 303–314.

9. W. Dahmen and C. A. Micchelli, *Some remarks on ridge functions*, Approx. Theory Appl. **3** (1987), 139–143.

10. K. I. Funahashi, *On the approximate realization of continuous mappings by neural networks*, Neural Networks **2** (1989), 183–192.

11. K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural Networks **2** (1989), 359–366.

12. B. Irie and S. Miyake, *Capabilities of three layered perceptrons*, IEEE Internat. Conf. on Neural Networks **1** (1988), 641–648.

13. Y. Ito, *Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory*, Neural Networks **4** (1991), 385–394.

14. ——, *Approximation of functions on a compact set by finite sums of a sigmoid function without scaling*, Neural Networks **4** (1991), 817–826.

15. H. N. Mhaskar, *Approximation properties of a multilayered feedforward artificial neural network*, Adv. in Comput. Math. **1** (1993), 61–80.

16. H. N. Mhaskar and C. A. Micchelli, *Approximation by superposition of a signmoidal function*, Adv. in Appl. Math. **13** (1992), 350–373.

17. T. Poggio and F. Girosi, *Regularization algorithms for learning that are equivalent to multilayer networks*, Science **247** (1990), 978–982.

18. W. Rudin, *Functional analysis*, McGraw-Hill, New York, 1973.

19. I. J. Schoenberg, *Cardinal spline interpolation*, CBMS-NSF Conf. Series in Appl. Math. #12, SIAM, Philadelphia, PA, 1973.

20. M. Stinchcombe and H. White, *Universal approximation using feedforward network with non-sigmoid hidden layer activation functions*, Proc. Internat. Joint Conference on Neural Networks (1989), 613–618, San Diego, SOS printing.

21. ——, *Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights*, IEEE Internat. Conf. on Neural Networks **3** (1990), III-7–III-16.

DEPARTMENT OF MATHEMATICS, CENTER FOR APPROXIMATION THEORY, TEXAS A&M UNIVERSITY, COLLEGE STATION, TEXAS 77843
E-mail address: `cat@math.tamu.edu`

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF NEVADA, LAS VEGAS, NEVADA 89154
E-mail address: `xinlixin@nevada.edu`

DEPARTMENT OF MATHEMATICS, CALIFORNIA STATE UNIVERSITY, LOS ANGELES, CALIFORNIA 90032
E-mail address: `hmhaskar@calstatela.edu`