Bull. Austral. Math. Soc. 41A10, 41A29, 41A46
Vol. 59 (1999) [121–131]

EXTENSION OF LOCALISED APPROXIMATION BY NEURAL NETWORKS

Nahmwoo Hahm and Bum Il Hong

We prove generalised results for localised approximation by generalised translation networks. We also show the relationship between the minimum number of neurons in the generalised translation networks with one hidden layer and the desired accuracy where the target functions are in a subset $V^{1,p}([-1,1]^s)$ of the Sobolev space $W^{1,p}([-1,1]^s)$.

1. Introduction

In recent years, there has been a great deal of research in the theory of approximation of real valued functions using artificial neural networks with one hidden layers [**1, 2, 5, 6, 7, 8**]. An interesting question in the theory of neural networks is to determine the minimum number of neurons in the hidden layer of a feedforward network required to provide a preassigned accuracy in approximating each target function in a known class of functions. Mhaskar [**7**] showed that if the target function belongs to Sobolev classes, we need $O(\varepsilon^{-s})$ neurons in the hidden layer of the neural network to achieve the approximation order ε. In this case, the number of neurons in the hidden layer increases exponentially along with the dimension. This causes a serious problem in applications where it is necessary to approximate functions depending on many variables.

Chui, Li and Mhaskar [**3**] suggested a localised approximation to cure this problem. An intuitive idea of a localised approximation is the following. Assuming that the target function is defined on $[-1,1]^s$, we divide the cube into n^s small portions and construct a neural network with a small number of neurons on each small portion. To motivate a localised approximation, we consider a Lebesgue measurable function f defined on $[-1,1]^2$ which is zero almost everywhere outside a compact set $K \subset [-1,1]^2$ and we divide $[-1,1]^2$ into n^2 subsquares. If d_K is the diameter of K, then f is mainly nonzero over $(d_K/2)^2\pi n^2$ subsquares. But, d_K is not an integer and hence we consider c_1 many subsquares around the boundary of K and this leads us to infer that f is nonzero over at most $(d_K/2)^2\pi n^2 + c_1$ subsquares. Note that $(d_K/2)^2\pi n^2 + c_1 \leqslant c\big(d_K^2 + 1/n^2\big)n^2$

Received 15th July, 1998
The authors wish to acknowledge the financial support of the Korea Research Foundation made in the program year of 1997.

where $c := \max\{\pi/4, c_1\}$. Thus f is nonzero over at most $c(d_K^2 + 1/n^2)n^2$ subsquares. Then we construct a subnetwork P with $c(d_K^2 + 1/n^2)n^2 L$ neurons that has the same approximation order as the network which approximates f over $[-1,1]^2$, where L is the number of neurons that are needed to approximate f over a subsquare in which f is nonzero. This method has an advantage when the target function changes on a small part of the square since we only need to retrain neurons related to that part.

Chui, Li and Mhaskar [**4**] derived the degree of localised approximations by a neural network with the univariate squashing activation function. In this paper, we establish the degree of localised approximation by a generalised translation network.

2. Preliminaries

We adopt the following notation. In this paper, the symbol s will denote a fixed positive integer, $s \geqslant 1$. If $A \subset \mathbb{R}^s$ is a (Lebesgue) measurable set, and $f : A \to \mathbb{R}$ is a measurable function, we write

$$\|f\|_{L^p(A)} := \begin{cases} \left\{\int_A |f(\mathbf{t})|^p dt\right\}^{1/p} & \text{if } 1 \leqslant p < \infty, \\ \operatorname*{ess\,sup}_{\mathbf{t}\in A} |f(\mathbf{t})| & \text{if } p = \infty. \end{cases}$$

The space $L^p(A)$ is defined to be the class of all functions $f : A \to \mathbb{R}$ for which $\|f\|_{L^p(A)} < \infty$. As usual, we consider two functions to be equal if they are equal almost everywhere in the measure theoretic sense.

Let $W^p([-1,1]^s) := W^{1,p}([-1,1]^s)$ denote the Sobolev class of degree 1. Alternatively, the class $W^p([-1,1]^s)$ consists of functions which have, at almost all points of $[-1,1]^s$, all partial derivatives up to order 1 and all of these derivatives are in $L^p([-1,1]^s)$. For $f \in W^p([-1,1]^s)$, we write

$$\|f\|_{W^p([-1,1]^s)} := \sum_{0 \leqslant \mathbf{k} \leqslant 1} \|D^{\mathbf{k}} f\|_{L^p([-1,1]^s)},$$

where, for the multi-integer $\mathbf{k} = (k_1, k_2, \cdots, k_s) \in \mathbb{Z}^s$, $0 \leqslant \mathbf{k} \leqslant 1$ means that each component of $\mathbf{k}$ is nonnegative and does not exceed 1, $|\mathbf{k}| = \sum_{i=1}^{s} |k_i|$ and

$$D^{\mathbf{k}} f = \frac{\partial^{|\mathbf{k}|} f}{\partial x_1^{k_1} \cdots \partial x_s^{k_s}}.$$

Let $V^p([-1,1]^s)$ be the set of all Lebesgue measurable functions f such that $\|f\|_{W^p([-1,1]^s)} \leqslant 1$. Then it is clear that $V^p([-1,1]^s) \subset W^p([-1,1]^s)$.

For positive integers d, s with $d \leqslant s$, the class $L_{loc}^{p,d}$ consists of all functions $g : \mathbb{R}^d \to \mathbb{R}$ such that $\int_K |g(\mathbf{x})|^p d\mathbf{x} < \infty$ for every compact subset $K \subset \mathbb{R}^d$ and $\mathcal{M}_{d\times s}$ denotes the set of all real valued $d \times s$ matrices. Let N be a positive integer, $1 \leqslant p \leqslant \infty$ and $\phi : \mathbb{R}^d \to \mathbb{R}$. A generalised translation network ψ on $\mathbb{R}^s$ with N neurons means a function of the form

$$\psi(\mathbf{x}) = \sum_{k=1}^{N} a_k \phi(A_k \mathbf{x} + \mathbf{b}_k) \tag{2.1}$$

where $A_k \in \mathcal{M}_{d\times s}$, $\mathbf{b}_k \in \mathbb{R}^d$ and $a_k \in \mathbb{R}$. The set of all such functions with a fixed N will be denoted by $\Psi_{\phi;N,s}$.

The degree of approximation of a function $f \in L^p([-1,1]^s)$ by such networks is defined by

$$E_{\phi;N,p,s}(f) := \inf\left\{\|f - P\|_{L^p([-1,1]^s)} : P \in \Psi_{\phi;N,s}\right\}.$$

If $\alpha \geqslant 1$ is not an integer, we define

$$E_{\phi;\alpha,p,s}(f) := E_{\phi;[\alpha],p,s}(f),$$

where $[\cdot]$ is the Gauss function.

We consider networks where each neuron may evaluate a different activation function. If $g_1, \dots, g_N$ are functions in $L_{loc}^{p,d}$, we define

$$E(f; \{g_1, \dots, g_N\}, p, s) := \inf\left\{\left\|f(\mathbf{x}) - \sum_{k=1}^{N} c_k g_k(A_k \mathbf{x})\right\|_{L^p([-1,1]^s)} : c_k \in \mathbb{R},\ A_k \in \mathcal{M}_{d\times s} \text{ and } \mathbf{x} \in \mathbb{R}^s\right\}$$

and

$$\varepsilon_{N,p,s}(f) := \inf\left\{E(f; \{g_1, \dots, g_N\}, p, s) : g_1, \dots, g_N \in L_{loc}^{p,d}\right\}.$$

The quantity $E(f; \{g_1, \dots, g_N\}, p, s)$ measures the degree of approximation of f when the neurons evaluate the functions $\{g_k\}$, but only on an inner product with the input variables without any thresholds b_k as in (2.1). When one considers $g_k(\mathbf{x}) := \phi(\mathbf{x} + \mathbf{b}_k)$ for k with $1 \leqslant k \leqslant N$, one may think of $\sum_{k=1}^{N} c_k g_k(A_k \mathbf{x})$ as a generalisation of the generalised translation networks. The quantity $\varepsilon_{N,p,s}(f)$ measures the degree of approximation when the neurons may evaluate any activation functions, even if the activation functions are dependent on the target function f. We define $\varepsilon_{\alpha,p,s}(f) := \varepsilon_{[\alpha],p,s}(f)$ for any real number $\alpha \geqslant 1$.

For a Lebesgue measurable function f on $\mathbb{R}^s$, we denote by $\Delta(f)$ the diameter of the essential support of f. For $f \in V^p([-1,1]^s)$, and positive numbers r and t, we write

$$\Lambda(f, r, t) := \left((\Delta(f))^s + r^{-s}\right)t.$$

3. Main Results

In this section, we state two theorems that are main results of this paper.

Let $1 \leqslant p \leqslant \infty$, and let d, s, n be integers with $1 \leqslant d \leqslant s$ and $n \geqslant 1$. For $\phi \in L_{loc}^{p,d}$, we define

$$\widetilde{E}_{\phi;n,p,s}\big(V^p([-1,1]^s)\big) := \inf\Big\{t > 0 : \sup_{f \in V^p([-1,1]^s)} E_{\phi;\Lambda(f,n,t),p,s}(f) \leqslant 1/n\Big\}.$$

The quantity $\widetilde{E}_{\phi;n,p,s}\big(V^p([-1,1]^s)\big)$ gives the minimum number of neurons in a network with one hidden layer, each evaluating the activation function ϕ, that are necessary to guarantee localised approximation of every function in $V^p([-1,1]^s)$ up to order $1/n$.

Theorem 3.1. *Let d, s be integers such that $1 \leqslant d \leqslant s$, and let $1 \leqslant p \leqslant \infty$. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be infinitely many times continuously differentiable in some open sphere in $\mathbb{R}^d$. Assume that for some $\mathbf{b}$ in this open sphere,*

$$D^{\mathbf{k}}\phi(\mathbf{b}) \neq 0 \tag{3.1}$$

for any $\mathbf{k} \in \mathbb{Z}^d$ and $\mathbf{k} \geqslant 0$. Then, for $\delta > 0$, we have

$$\widetilde{E}_{\phi;n,p,s}\big(V^p([-1,1]^s)\big) \leqslant cn^{s+\delta}, \qquad n \geqslant 1,$$

where c is a positive number depending on δ, p and s but is independent of n.

This theorem shows that localisation provides a near optimal order of approximation.

Remark. Some examples where (3.1) is satisfied are the following, where for $\mathbf{x} \in \mathbb{R}^d$, we write $|\mathbf{x}| = \left(\sum_{j=1}^{d} x_j^2\right)^{1/2}$:

The squashing function	$\phi(x) = \big(1 + e^{-x}\big)^{-1}, \quad d = 1,$
Generalised multiquadrics	$\phi(\mathbf{x}) = \big(1 + \lvert\mathbf{x}\rvert^2\big)^{\alpha}, \alpha \neq \mathbb{Z}, \quad d \geqslant 1,$
The Gaussian function	$\phi(\mathbf{x}) = \exp\big(-\lvert\mathbf{x}\rvert^2\big), \quad d \geqslant 1,$
Thin plate splines	$\phi(\mathbf{x}) = \begin{cases} \lvert\mathbf{x}\rvert^{2q-d}\log\lvert\mathbf{x}\rvert & \text{if } d \text{ is even} \\ \lvert\mathbf{x}\rvert^{2q-d} & \text{if } d \text{ is odd} \end{cases}$ for $d \geqslant 1$ and $q \in \mathbb{Z}$ with $q > d/2$.

For $1 \leqslant p \leqslant \infty$, we define

$$\widetilde{\varepsilon}_{n,p,s}\big(V^p([-1,1]^s)\big) := \inf\Big\{t > 0 : \sup_{f \in V^p([-1,1]^s)} \varepsilon_{\Lambda(f,n,t),p,s}(f) \leqslant 1/n\Big\}.$$

This quantity shows the minimum number of neurons in a network with one hidden layer, each evaluating $\{g_k\}$, that are necessary to approximate every function in $V^p([-1,1]^s)$ locally up to order $1/n$.

THEOREM 3.2. *For $1 \leqslant p \leqslant \infty$ and $1 \leqslant d < s$, we have*

$$\widetilde{\varepsilon}_{n,p,s}\big(V^p([-1,1]^s)\big) \geqslant cn^s \log n, \qquad n \geqslant 1,$$

where c is a positive number depending on p and s, but independent of n.

This theorem shows that the number of neurons in a network with one hidden layer is at least $cn^s \log n$ and implies that the optimal order n^s cannot be reached by localisation.

4. PROOFS

To prove Theorem 3.1, we use the B-spline functions. The B-spline of order 0 is defined to be the characteristic function of the interval $[0,1]$ and is denoted by N_0. For any integer $m \geqslant 1$, N_m is defined recursively by

$$N_m(x) := \int_0^1 N_{m-1}(x-t)\,dt, \qquad x \in \mathbb{R}.$$

Then N_m has a compact support $[0,m]$ and N_m is $(m-1)$ times continuously differentiable. In addition, $\|N_m\|_{L^1([0,m])} = \|N_m\|_{L^\infty([0,m])} = 1$ and $\|N_m\|_{L^p([0,m])} \leqslant 1$ for $1 < p < \infty$.

The tensor product B-splines are defined by

$$N_m^s(\mathbf{x}) := \prod_{k=1}^{s} N_m(x_k), \qquad \mathbf{x} = (x_1, \dots, x_s) \in \mathbb{R}^s.$$

We prepare with two lemmas to prove Theorem 3.1.

LEMMA 4.1. *Let ϕ satisfy the conditions of Theorem 3.1, and let m be an integer larger than $(s^2+s+\delta)/\delta$. For any integer $n \geqslant 1$ and multi-integer $\mathbf{k}$, there exists a $\varphi \in \Pi_{\phi;[n^\delta],s}$ such that*

$$\|N_m^s - \varphi\|_{L^\infty([-1,1]^s)} \leqslant cn^{-s-1},$$

where c is a positive constant depending on δ, m and s.

PROOF: The result is immediate from [**7**, Theorem 2.1] when we replace $m-1$ by r. ∎

According to [**9**], there exists a bounded linear operator $T : W^p([-1,1]^s) \to W^p([-2,2]^s)$ such that $T(f) = f$ almost everywhere on $[-1,1]^s$. Since T is a bounded linear operator, there exists a constant c such that

$$\|T(f)\|_{W^p([-2,2]^s)} \leqslant c\|f\|_{W^p([-1,1]^s)}$$

for every $f \in W^p([-1,1]^s)$. Let ψ be an infinitely many times continuously differentiable function that takes the value 1 on $[-1,1]^s$ and 0 outside of $[-3/2,3/2]^s$. Then the function $T(f)\psi$ equals f on $[-1,1]^s$ almost everywhere and is identically 0 outside $[-3/2,3/2]^s$ and $\|T(f)\psi\|_{W^p([-2,2]^s)} \leqslant c\|f\|_{W^p([-1,1]^s)}$. We denote the extension $T(f)\psi$ of f by the symbol $\widetilde{f}$.

Now we have the following lemma.

LEMMA 4.2. *Let $I_{\mathbf{k},n}$ be the cube with centre $(2\mathbf{k}+1)/(2n)$ and side $1/n$ for any integer $n \geqslant 1$ and multi-integer $\mathbf{k} \in \mathbb{Z}^s$, where $(2\mathbf{k}+1)/(2n) = \big((2k_1+1)/2n, \cdots, (2k_s+1)/2n\big)$. Define*

$$S_{n,m}(f)(\mathbf{x}) = n^s \sum_{-m-1 \leqslant \mathbf{k} \leqslant n-1} \left(\int_{I_{\mathbf{k},n}} \widetilde{f}(\mathbf{t})\,d\mathbf{t}\right) N_m^s(n\mathbf{x}-\mathbf{k}), \tag{4.1}$$

where $-m-1 \leqslant \mathbf{k} \leqslant n-1$ means that $-m-1 \leqslant k_i \leqslant n-1$ for $1 \leqslant i \leqslant s$. Then, for any $f \in V^p([-1,1]^s)$ and $n \geqslant 1$,

$$\|f - S_{n,m}(f)\|_{L^p([-1,1]^s)} \leqslant \frac{c}{n}$$

where c is a positive constant depending only on s.

PROOF: See Mhaskar [4]. ∎

Lemma 4.1 and Lemma 4.2 can be used to prove Theorem 3.1.

PROOF OF THEOREM 3.1: We choose an integer m larger than $(s^2+s+\delta)/\delta$ and M such that the summand in (4.1) does not exceed $M\Lambda(f,n,n^s)$. We construct a neural network

$$P(\mathbf{x}) = n^s \sum_{-m-1 \leqslant \mathbf{k} \leqslant n-1} \left(\int_{I_{\mathbf{k},n}} \widetilde{f}(\mathbf{t})\,d\mathbf{t}\right) \varphi(n\mathbf{x}-\mathbf{k}).$$

Then P is a network having at most $M\Lambda(f,n,n^{s+\delta})$ neurons by Lemma 4.1. In addition, we have, by Lemma 4.1 and Lemma 4.2,

$$\begin{aligned}\|f-P\|_{L^p([-1,1]^s)} &\leqslant \|f - S_{n,m}(f)\|_{L^p([-1,1]^s)} + \|S_{n,m}(f) - P\|_{L^p([-1,1]^s)} \\ &\leqslant \frac{c}{n} + n^s \left\{ \sum_{-m-1 \leqslant \mathbf{k} \leqslant n-1} \left(\int_{I_{\mathbf{k},n}} \widetilde{f}(\mathbf{t})\,d\mathbf{t}\right)^p \right\}^{1/p} \|N_m^s - \varphi\|_{L^\infty([-1,1]^s)} \\ &\leqslant \frac{c}{n} + \frac{c}{n} \\ &:= \frac{c}{n}.\end{aligned}$$

This completes the proof. ∎

Since Hölder's inequality says that $\|f\|_{L^1([-1,1]^s)} \leqslant \|f\|_{L^p([-1,1]^s)}$ holds for a Lebesgue integrable function f and $1 \leqslant p$, it is enough to show that Theorem 3.2 holds for $p = 1$. For this purpose we prove the following two lemmas.

LEMMA 4.3. *Let $m, s \geqslant 1$ be integers and $0 < a < 2^{-m}s^{-1/2}$. Define*

$$B(a) := \{\mathbf{x} \in \mathbb{R}^s : |x_i| < a\}.$$

If F is a function from $\mathbb{R}^s$ into $[0, \infty)$, and is zero outside $B(a)$, then

$$\varepsilon_{m,1,s}(F) = \|F\|_{L^1([-1,1]^s)}.$$

PROOF: Let $A_1, \ldots, A_m$ be $d \times s$ matrices and let $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be unit vectors in $\mathbb{R}^s$ such that $A_i\mathbf{u}_i = 0$, $1 \leqslant i \leqslant m$. Let

$$\chi_a(\mathbf{x}) := \begin{cases} 1 & \text{if } \mathbf{x} \in B(a), \\ 0 & \text{otherwise.} \end{cases}$$

We define

$$f_1(\mathbf{x}) = \chi_a(\mathbf{x}) - \chi_a\big(\mathbf{x} + \sqrt{s}a\mathbf{u}_1\big)$$

and inductively, for $2 \leqslant n \leqslant m$,

$$f_n(\mathbf{x}) = f_{n-1}(\mathbf{x}) - f_{n-1}\big(\mathbf{x} + 2^{n-1}\sqrt{s}a\mathbf{u}_n\big).$$

Then, directly from the construction of f_n for $1 \leqslant n \leqslant m$ we have

$$|f_n(\mathbf{x})| \leqslant 1, \quad \text{for } \mathbf{x} \in \mathbb{R}^s \quad \text{and} \quad f_n(\mathbf{x}) = 0, \quad \text{for} \quad \mathbf{x} \notin B\big(2^n\sqrt{s}a\big).$$

In addition, by mathematical induction,

$$\int_{[-1,1]^s} f_n(\mathbf{x})F(\mathbf{x})\,d\mathbf{x} = \|F\|_{L^1([-1,1]^s)} \tag{4.2}$$

and

$$\int_{[-1,1]^s} f_n(\mathbf{x})g(A_j\mathbf{x})\,d\mathbf{x} = 0, \qquad 1 \leqslant j \leqslant n, \;\; g \in L^{1,d}_{loc}. \tag{4.3}$$

For $n = 1$, we have

$$\begin{aligned}\int_{[-1,1]^s} f_1(\mathbf{x})F(\mathbf{x})\,d\mathbf{x} &= \int_{[-1,1]^s} \chi_a(\mathbf{x})F(\mathbf{x})\,d\mathbf{x} - \int_{[-1,1]^s} \chi_a\big(\mathbf{x} + \sqrt{s}a\mathbf{u}_1\big)F(\mathbf{x})\,d\mathbf{x} \\ &= \int_{[-1,1]^s} \chi_a(\mathbf{x})F(\mathbf{x})\,d\mathbf{x} \\ &= \|F\|_{L^1([-1,1]^s)}\end{aligned}$$

and

$$\begin{aligned}\int_{[-1,1]^s} f_1(\mathbf{x})g(A_1\mathbf{x})\,d\mathbf{x} &= \int_{B(a)} g(A_1\mathbf{x})\,d\mathbf{x} - \int_{B(a)} g\big(A_1\big(\mathbf{x}-\sqrt{s}a\mathbf{u}_1\big)\big)\,d\mathbf{x}\\ &= \int_{B(a)} g(A_1\mathbf{x})\,d\mathbf{x} - \int_{B(a)} g(A_1\mathbf{x})\,d\mathbf{x}\\ &= 0.\end{aligned}$$

If we assume that (4.2) and (4.3) hold for $n = k$ with $k < m$, we have, for $n = k+1$,

$$\begin{aligned}\int_{[-1,1]^s} f_{k+1}(\mathbf{x})F(\mathbf{x})\,d\mathbf{x} &= \int_{[-1,1]^s} f_k(\mathbf{x})F(\mathbf{x})\,d\mathbf{x} - \int_{[-1,1]^s} f_k\big(\mathbf{x}+2^k\sqrt{s}a\mathbf{u}_{k+1}\big)F(\mathbf{x})\,d\mathbf{x}\\ &= \int_{[-1,1]^s} f_{k+1}(\mathbf{x})F(\mathbf{x})\,d\mathbf{x}\\ &= \|F\|_{L^1([-1,1]^s)},\end{aligned}$$

since $\mathbf{x}+2^k\sqrt{s}a\mathbf{u}_{k+1} \notin B(a)$ for $x \in B(a)$, and

$$\begin{aligned}&\int_{[-1,1]^s} f_{k+1}(\mathbf{x})g(A_j\mathbf{x})\,d\mathbf{x}\\ &\quad= \int_{[-1,1]^s} f_k(\mathbf{x})g(A_j\mathbf{x})\,d\mathbf{x} - \int_{[-1,1]^s} f_k\big(\mathbf{x}+2^k\sqrt{s}a\mathbf{u}_{\mathbf{k+1}}\big)g(A_j\mathbf{x})\,d\mathbf{x}\\ &\quad= \int_{[-1,1]^s} f_k(\mathbf{x})g(A_j\mathbf{x})\,d\mathbf{x} - \int_{[-1,1]^s} f_k(\mathbf{x})g\big(A_j\big(\mathbf{x}-2^k\sqrt{s}a\mathbf{u}_{\mathbf{k+1}}\big)\big)\,d\mathbf{x}\\ &\quad= 0,\end{aligned}$$

since $g\big(A_j\big(\mathbf{x}-2^k\sqrt{s}a\mathbf{u}_{\mathbf{k+1}}\big)\big) \in L^{1,d}_{loc}$ for $1 \leqslant j \leqslant k$ and $A_{k+1}\mathbf{u}_{k+1} = 0$ for $j = k+1$.

By the definition of $\varepsilon_{m,1,s}(F)$, it is clear that $\varepsilon_{m,1,s}(F) \leqslant \|F\|_{L^1([-1,1]^s)}$. On the other hand, if we take $g_1, \dots, g_m \in L^{1,d}_{loc}$ and $f_1, \dots, f_m$ as above, then we have

$$\begin{aligned}\Big\|F(\mathbf{x}) - \sum_{k=1}^m c_k g_k(A_k\mathbf{x})\Big\|_{L^1([-1,1]^s)} &\geqslant \int_{[-1,1]^s} \Big|F(\mathbf{x})f_m(\mathbf{x}) - \sum_{k=1}^m c_k g_k(A_k\mathbf{x})f_m(\mathbf{x})\Big|\,d\mathbf{x}\\ &= \int_{[-1,1]^s} \big|F(\mathbf{x})f_m(\mathbf{x})\big|\,d\mathbf{x}\\ &= \|F\|_{L^1([-1,1]^s)}.\end{aligned}$$

This completes the proof. ∎

Lemma 4.4. *Let $0 < a < 1$ and define, for $\mathbf{x} \in \mathbb{R}^s$,*

$$F(\mathbf{x}) = \Big(1 - \frac{\|\mathbf{x}\|^2}{sa^2}\Big)\chi_a(\mathbf{x})$$

where $\|\mathbf{x}\|^2 = \sum_{i=1}^{s} x_i^2$. *Then there exist constants* α *and* β *depending only on* s *such that*

$$\|F\|_{L^1([-1,1]^s)} = \alpha a^s \quad \text{and} \quad \|F\|_{W^1([-1,1]^s)} \leqslant \beta a^{s-1}.$$

PROOF: We observe that F is supported on $B(a)$ and is continuously differentiable. Therefore

$$\begin{aligned}
\|F\|_{L^1([-1,1]^s)} &= \int_{[-1,1]^s} \left(1 - \frac{\|\mathbf{x}\|^2}{sa^2}\right) \chi_a(\mathbf{x})\, d\mathbf{x} \\
&= \int_{B(a)} \left(1 - \frac{\|\mathbf{x}\|^2}{sa^2}\right) d\mathbf{x} \\
&= \int_{B(a)} d\mathbf{x} - \frac{1}{sa^2} \sum_{i=1}^{s} \int_{B(a)} x_i^2 \, dx_1 \ldots dx_s \\
&= 2^s a^s - \frac{1}{sa^2} \sum_{i=1}^{s} \frac{2a^3}{3} 2^{s-1} a^{s-1} \\
&:= \alpha a^s.
\end{aligned}$$

Since $\int_{[-1,1]^s} \left|\frac{\partial F}{\partial x_i}\right| d\mathbf{x} = \int_{B(a)} \left|(2x_i/sa^2)\right| d\mathbf{x} = (2^s a^{s-1})/s$, we have

$$\begin{aligned}
\|F\|_{W^1([-1,1]^s)} &= \|F\|_{L^1([-1,1]^s)} + \sum_{i=1}^{s} \int_{[-1,1]^s} \left|\frac{\partial F}{\partial x_i}\right| d\mathbf{x} \\
&= \alpha a^s + \frac{2^s a^{s-1}}{s} s \\
&\leqslant \alpha a^{s-1} + 2^s a^{s-1} \\
&:= \beta a^{s-1}.
\end{aligned}$$

This completes the proof. ∎

Now we prove Theorem 3.2 using Lemma 4.3 and Lemma 4.4.

PROOF OF THEOREM 3.2: Let $N := \widetilde{\varepsilon}_{n,1,s}(V_{1,s})$, and $a := (2\beta)/(\alpha n)$, where α, β are the constants in Lemma 4.4. Let m be the largest integer less than $\left((\sqrt{s}a)^s + n^{-s}\right)N$. This implies that

$$m \leqslant \left(\left(\sqrt{s}a\right)^s + \frac{1}{n^s}\right) N \leqslant \left(\frac{(2\sqrt{s}\beta)^s + \alpha^s}{\alpha^s n^s}\right) N. \tag{4.4}$$

Assume that $0 < a < 2^{-m}s^{-1/2}$ as in Lemma 4.3. We define

$$F(\mathbf{x}) = \left(1 - \frac{\sum_{i=1}^{s} x_i^2}{sa^2}\right)\chi_a(\mathbf{x}).$$

Note that the diameter of the essential support of $F/\|F\|_{W^1([-1,1]^s)}$ is $\sqrt{s}a$ and $F/\|F\|_{W^1([-1,1]^s)}$ belongs to $V^p([-1,1]^s)$. In addition, (4.4) implies that m is the largest integer less than $\Lambda\bigl(F/\|F\|_{W^1([-1,1]^s)}, n, N\bigr)$. According to the definition of $\widetilde{\varepsilon}_{n,1,s}\bigl(V^p([-1,1]^s)\bigr)$, we have

$$\varepsilon_{\Lambda\left(F/\|F\|_{W^1([-1,1]^s)},n,N\right),1,s}\left(\frac{F}{\|F\|_{W^1([-1,1]^s)}}\right) = \varepsilon_{m,1,s}\left(\frac{F}{\|F\|_{W^1([-1,1]^s)}}\right) \leqslant \frac{1}{n}.$$

By Lemma 4.4,

$$\varepsilon_{m,1,s}(F) \leqslant \frac{\|F\|_{W^1([-1,1]^s)}}{n} \leqslant \frac{\beta a^{s-1}}{n}. \tag{4.5}$$

On the other hand, we have, by Lemma 4.3 and Lemma 4.4,

$$\varepsilon_{m,1,s}(F) = \|F\|_{L^p([-1,1]^s)} = \alpha a^s. \tag{4.6}$$

Therefore, (4.5) and (4.6) show that

$$\alpha a^s \leqslant \frac{\beta a^{s-1}}{n}, \qquad \text{that is,} \quad \frac{2\beta}{n} = \alpha a \leqslant \frac{\beta}{n}.$$

This is a contradiction to the assumption and hence

$$a \geqslant 2^{-m}s^{-1/2}. \tag{4.7}$$

Let $c_1 := (2\beta\sqrt{s})/\alpha$. Then (4.7) implies that

$$\log n \leqslant m\log 2 + \log c_1 \leqslant c_2 m$$

for some constant $c_2 \in \mathbb{R}$. Therefore, $\log n/c_2 \leqslant m$ and

$$\begin{aligned} N &\geqslant mn^s\left(\frac{\alpha^s}{(2\sqrt{s}\beta)^s + \alpha^s}\right) \\ &\geqslant \frac{1}{c_2}\log n\, n^s\left(\frac{\alpha^s}{(2\sqrt{s}\beta)^s + \alpha^s}\right) \\ &:= cn^s\log n \end{aligned}$$

where $c = \alpha^s \Big/ \Bigl(c_2\bigl((2\sqrt{s}\beta)^s + \alpha^s\bigr)\Bigr)$. This completes the proof. ∎

References

[1] A.R. Barron, 'Universal approximation bounds for superposition of a Sigmoidal function', *IEEE Trans. Inform. Theory* **39** (1993), 930–945.

[2] C.K. Chui and X. Li, 'Approximation by Ridge functions and neural networks with one hidden layer', *J. Approx. Theory* **70** (1992), 131–141.

[3] C.K. Chui, X. Li and H. N. Mhaskar, 'Localized approximation by neural networks', *Math. Comp.* **63** (1994), 607–623.

[4] C.K. Chui, X. Li and H.N. Mhaskar, 'Limitations of the approximation capabilities of neural networks with one hidden layer', *Adv. Comput. Math.* **5** (1996), 233–243.

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

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

[7] H.N. Mhaskar, 'Neural networks for optimal approximation of smooth and analytic functions', *Neural Computation* **8** (1996), 164–177.

[8] H.N. Mhaskar and N. Hahm, 'Neural networks for functional approximation and system identification', *Neural Computation* **9** (1997), 143–159.

[9] E.M. Stein, *Singular integrals and differentiability properties of functions* (Princeton Univ. Press, Princeton, 1970).

Institute of Natural Sciences
Kyung Hee University, Yongin
Kyunggi 449-701
Korea
e-mail: nhahm@lion.inchon.ac.kr

Department of Mathematics
Kyung Hee Univeristy, Yongin
Kyunggi 449-701
Korea
e-mail: bihong@nms.kyunghee.ac.kr