Approximation Bounds for Smooth Functions in $\mathbf{C}(\mathbb{R}^d)$ by Neural and Mixture Networks*

Vitaly Maiorov
Department of Mathematics
Technion, Haifa 32000
Israel

Ron Meir†
Department of Electrical Engineering
Technion, Haifa 32000
Israel

December 1996

Abstract

We consider the approximation of smooth multivariate functions in $\mathbf{C}(\mathbb{R}^d)$ by feedforward neural networks with a single hidden layer of non-linear ridge functions. Under certain assumptions on the smoothness of the functions being approximated and on the activation functions in the neural network, we present upper bounds on the degree of approximation achieved over the domain $\mathbb{R}^d$, thereby generalizing available results for compact domains. We extend the approximation results to the so-called mixture of expert architecture, which has received wide attention in recent years, showing that the same type of approximation bound may be achieved.

Key words: Neural Networks, Mixture of Experts, Approximation Bounds.

1 Introduction

A great deal of research, within both the mathematical as well as the engineering community, has been devoted in recent years to the establishment of performance bounds for approximation of functions by various types of feedforward neural networks. The many

*This research was partially supported by the Wolfson Research awards administered by the Israel Academy of Science and Humanities. Support from the Ollendorff center of the Department of Electrical Engineering at the Technion is also acknowledged.

†rmeir@dumbo.technion.ac.il

contributions made can be divided into two broad areas. First, concerning the question of denseness, Leshno *et al.* [11] have shown that any function in $\mathbf{C}(\mathbb{R}^d)$ may be approximated to arbitrary accuracy by a feedforward neural network with a single hidden layer, if and only if the activation function of the network is not a polynomial (see also [3] for a recent contribution along similar lines). This result summarizes most of the previous work on the issue of denseness (for example [4][9]). The situation concerning the rate of approximation has not yet been so conclusively answered. Several researchers have been able to demonstrate approximation error bounds for various functional classes (see for example [1][14]) approximated by neural networks, but as far as we are aware most of these results refer to approximation over compact domains.

While many open issues remain concerning approximation rates, we focus in this paper on the issue of approximation of functions defined over $\mathbb{R}^d$ by feedforward neural networks, as well as mixture networks. Two recent contributions have adressed this issue. First, in [9] several results for approximation of functions and their derivatives by feedforward neural networks have been given. This work differs from ours in that the class of functions being approximated is defined in terms of convex integral representations, following the work of Barron [1]. In this work we are concerned with the more standard Sobolev classes of functions. More recently, Delyon *et al.* [5] have considered approximation by wavelet networks, also deriving approximation results in $\mathbb{R}^d$.

In this paper we show, using tools from the theory of weighted polynomial approximation, how to extend approximation results for compactly supported functions in Sobolev space to the full Euclidean space $\mathbb{R}^d$. As a second contribution we consider a special class of networks, which we refer to as mixture networks, which impose certain stringent conditions on the activation functions. The motivation for considering this class of functions arises from the recently introduced class of networks, titled 'mixtures of experts' [10]. These system are similar to neural networks, except that they are endowed with a certain probabilistic interpretation, requiring the activations of the units in the hidden-layer to sum unity. An exact definition will be given in Section 4, while the motivation for the structure and a discussion of various applications can be found in [10].

The basic proof strategy adopted in this paper is the following. We first approximate any function f in the Sobolev space defined over the full Euclidean space $\mathbb{R}^d$ by polynomials; this is possible due to results from Freud [6] and Mhaskar [14]. Polynomials are then approximated by sigmoidal networks, using methods introduced in [11] and [14]. These results are then combined to yield the desired bounds.

The remainder of the paper is organized as follows. In Section 2 we review some existing results for weighted polynomial approximation of univariate functions, followed by an extension of these results to multivariate functions. In Section 3 we then introduce feedforward neural networks with a single hidden layer, and show how the results of Section 2 may be used to derive upper bounds for approximation by neural networks. We then introduce the mixture of expert class of networks in Section 4 and derive the appropriate approximation bounds. Finally, a short discussion is given in Section 5.

Many of the proofs of the more technical lemmas have been relegated to the appendix, so as not to interfere with the presentation of the main ideas.

Before proceeding to the main part of the paper we make some comments concerning the constants appearing in the paper. We have attempted to keep track of the various constants, arising from approximation theoretic results, by indexing them in the order in which they appear, so that one may able to trace the origin of each constant. In the final results (Theorem 3.2 and 4.1), we have lumped all the constants into a single constant. Throughout the paper we use boldface symbols (such as $\mathbf{j}$, $\mathbf{s}$ etc.) for vectors, retaining regular fonts for scalars.

2 Weighted Approximation by Polynomials

As mentioned in the Introduction, a preliminary step needed in deriving the approximation bounds by neural and mixture networks, is that of approximation by polynomials. In this section we first recapitualte some basic results from the theory of weighted polynomial approximation of univariate functions, followed by an extension of the results to multivariate functions.

Before proceeding we need to introduce some definitions and notation. Since we are working in $\mathbb{R}^d$ it is convenient to introduce a weighted norm of a function f given by:

$$\|f\|_{p,w} \stackrel{\triangle}{=} \left(\int_{\mathbb{R}^d} |w(\mathbf{x}) f(\mathbf{x})|^p d\mathbf{x} \right)^{1/p},$$

where $1 \leq p \leq \infty$, and $w(\mathbf{x})$ is a weight function to be specified below. We denote the class of functions for which $\|f\|_{p,w}$ is finite by $L_{p,w}$. The domain $\mathbb{R}^d$ is implicit throughout the paper; when referring specifically to $\mathbb{R}$ we use the notation $L_{p,w}(\mathbb{R})$.

For functions $f : \mathbb{R}^d \to \mathbb{R}$ we slightly abuse the notation by defining weight functions of the form $w(\mathbf{x}) = \prod_{i=1}^d w(x_i)$, where the argument of $w(\cdot)$ determines its domain. The class of functions we wish to approximate in this work is then defined as follows:

$$W_{p,w}^{r,d} = \left\{ f \ : \ \|f^{(\boldsymbol{\rho})}\|_{p,w} \leq M \ , |\boldsymbol{\rho}| \leq r \right\},$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_d)$ and $|\boldsymbol{\rho}| = \rho_1 + \cdots \rho_d$, and $M < \infty$. Note that a more precise notation would be $W_{p,w}^{r,d}(M)$; however, we retain the simpler notation with the constant M being implied througout. Here $f^{(\boldsymbol{\rho})}$ represents the partial derivative $\partial^{|\boldsymbol{\rho}|} f / \partial x_1^{\rho_1} \cdots \partial x_d^{\rho_d}$.

2.1 Weighted polynomial approximation of uni-variate functions

We remind the reader of some basic tools from the theory of weighted polynomial approximation of functions of a single real variable. We recapitualte only the essentials needed in the sequel; a detailed review can be found in [13].

Let a univariate weight function

$$w(x) = e^{-Q(x)},$$

be given. The function $Q(\cdot)$ is required to obey the following conditions [13]:

Assumption 2.1 *The univariate function $Q(x)$ obeys the following conditions:*
1) Q is an even function in $\mathbf{C}^2(-\infty, \infty)$.
2) Q'' is positive and non-decreasing on $(0, \infty)$.
3) There are constants c and d such that

$$1 \leq c \leq x\frac{Q''(x)}{Q'(x)} \leq d < \infty \qquad (x > 0).$$

Examples of such weight functions, which will used in this paper, are $\exp(-a|x|^\alpha)$ with $\alpha \geq 2$ and $a > 0$.

Following [13] define the one-dimensional *degree of approximation* of a function g by polynomials of degree m as follows:

$$E(g, \mathbf{P}_m, L_{p,w}) = \inf_{P \in \mathbf{P}_m} \|g - P\|_{p,w}, \tag{1}$$

where $\mathbf{P}_m$ stands for the class of degree-m algebraic polynomials. From ([13], eq. (30)) we then have a simple relationship which will be utilized in the sequel. Let g be differentiable with $wg' \in L_p(\mathbb{R})$, then

$$E(g, \mathbf{P}_m, L_{p,w}) \leq c_1\frac{q_m}{m}\, E(g', \mathbf{P}_{m-1}, L_{p,w}), \tag{2}$$

where q_m is the smallest positive real number such that

$$q_m Q'(q_m) = m. \tag{3}$$

For functions of the form $Q(x) = a|x|^\alpha$, we easily establish that

$$q_m = \left(\frac{m}{a\alpha}\right)^{1/\alpha} \qquad (Q(x) = a|x|^\alpha). \tag{4}$$

Consider now the system of polynomials, $\{P_m(t)\}$, of the one-dimensional variable t, obeying the following orthonormality conditions with respect to the weight function $w(t)$:

$$\int_{\mathbb{R}} P_n(t)P_m(t)w(t)^2 dt = \delta_{mn}. \tag{5}$$

Let $P_m(t)$ be given as

$$P_m(x) = \gamma_m t^m + \cdots + \gamma_1 t + \gamma_0 = \gamma_m \prod_{k=1}^{m}(t - t_{k,m}) \in \mathbf{P}_m \qquad (\gamma_m > 0), \tag{6}$$

where $\{t_{k,m}\}_{k=1}^{m}$ are the roots of the polynomial. Then from the work of Freud [6] (see also [13]) we have the inequalities:

$$\begin{aligned} t_{m,m} &< t_{m-1,m} < \cdots < t_{1,m} < 4q_{m-1}, \\ \frac{1}{2}q_m &\leq \frac{\gamma_{m-1}}{\gamma_m} \leq 2q_m, \end{aligned} \tag{7}$$

where q_m is given in (3).

Finally, we need to introduce a linear operator which projects any uni-variate function $g(t)$ onto the space of polynomials. Let $p_k(t)$ be the k'th element of the orthogonal family defined in (5). In order to simplify the notation, in preparation for the the discussion of multivariate functions, let us assume that m is an odd number, in which case we follow [13] and define

$$\begin{aligned} a_k &= \int w(t)^2 g(t) p_k(t) dt \\ s_i(g,t) &= \sum_{k=1}^{i-1} a_k P_k(t), \\ \pi_m g(t) &= \frac{1}{m+1} \sum_{n=(m+3)/2}^{m+1} s_n(g,t). \end{aligned} \tag{8}$$

Note that $\pi_m g(t)$ is an algebraic polynomial of degree m. We note that by changing the order of the summations over m and k, implicit in the expression for $\pi_m g(t)$, we may express it as

$$\pi_m g(t) = \sum_{k=1}^{m} \lambda_k a_k P_k(t), \tag{9}$$

where

$$\lambda_k = \begin{cases} (m-1)/2(m+1) & \text{if } 0 \leq k \leq (m+1)/2 \\ (m-k)/(m+1) & \text{if } (m+3)/2 \leq k \leq m \end{cases} \tag{10}$$

A basic result concerning the operator π_m is the following (see eq. (28) in [13]):

$$\|g - \pi_m g\|_{p,w} \leq c_2 E(g, \mathbf{P}_m, L_{p,w}). \tag{11}$$

2.2 Approximating multivariate functions by polynomials

We now make use of the previous results in generalizing them to multivariate approximation. Let $\mathbf{m} = (m_1, \ldots, m_d)$ and consider a class of multivariate polynomials defined as follows:

$$\mathbf{P_m} = \left\{ P \ : \ P(\mathbf{x}) = \sum_{i_1=0}^{m_1} \cdots \sum_{i_d=0}^{m_d} b_{i_1,\ldots,i_d} x_1^{i_1} \cdots x_d^{i_d} \ ; \ b_{i_1,\ldots,i_d} \in \mathbb{R} \ \forall i_1, \ldots, i_d \right\}. \tag{12}$$

We bound the approximation error of any function in $W_{p,w}^{r,d}$ when approximated by polynomials from $\mathbf{P_m}$.

Lemma 2.1 *Let $f \in W^{r,d}_{p,w}$ for $1 \le p \le \infty$, where $w(\mathbf{x}) = \prod_{i=1}^{d} e^{-|x_i|^\alpha}$, $\alpha \ge 2$. Then for any $\mathbf{m} = (m_1, \ldots, m_d)$, $m_i \le m$,*

$$\inf_{P \in \mathbf{P_m}} \|f - P\|_{p,w} \le c m^{-r(1-\frac{1}{\alpha})},$$

where c is independent of f or m.

PROOF The proof generalizes the work of Freud [6] and Mhaskar [13] to $\mathbb{R}^d$. Consider first the system of orthonormal polynomials, $\{P_m(t)\}$, defined in (5). Let $\mathbf{m} = (m_1, \ldots, m_d)$, $\mathbf{x} \in \mathbb{R}^d$ and define

$$P_{\mathbf{m}}(\mathbf{x}) \stackrel{\triangle}{=} \prod_{i=1}^{d} P_{m_i}(x_i). \tag{13}$$

We then have for any $\varphi \in L_{p,w}(\mathbb{R})$

$$\begin{aligned} \varphi(t) &= \sum_{m=1}^{\infty} < \varphi, P_m > P_m(t), \\ < \varphi, P_m > &= \int_{\mathbb{R}} \varphi(t) P_m(t) w(t)^2 dt. \end{aligned}$$

In the d-dimensional case we define

$$< f, P_{m_i} >_i = < f, P_{m_i} >_i (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) = \int_{\mathbb{R}} f(\mathbf{x}) P_{m_i}(x_i) w(x_i)^2 dx_i.$$

We further define the set of *projection operators*

$$\pi_{i,m_i} f(\mathbf{x}) = \sum_{s=1}^{m_i} \lambda_s < f, P_s >_i P_s(x_i),$$

where λ_s is defined in (10). The full projector π defined through the equation

$$\pi f = \pi_{1,m_1} \pi_{2,m_2} \cdots \pi_{d,m_d} f = \sum_{s_1=1}^{m_1} \cdots \sum_{s_d=1}^{m_d} \lambda_{s_1} \cdots \lambda_{s_d} f_{s_1,\ldots,s_d} P_{s_1}(x_1) \cdots P_{s_d}(x_d),$$

where

$$f_{s_1,\ldots,s_d} = \int_{\mathbb{R}^d} \left(\prod_{i=1}^{d} w(x_i)^2 P_{s_i}(x_i) \right) f(\mathbf{x}) d\mathbf{x}.$$

We then have

$$\begin{aligned} \|f - \pi f\|_{p,w} &= \|f - \pi_{1,m_1} f + \pi_{1,m_1} f - \pi_{1,m_1} \pi_{2,m_2} f + \pi_{1,m_1} \pi_{2,m_2} f - \cdots - \pi f\|_{p,w} \\ &\le \sum_{i=1}^{d} \|\pi_0 \cdots \pi_{i-1,m_{i-1}} f - \pi_0 \cdots \pi_{i,m_i} f\|_{p,w}, \end{aligned} \tag{14}$$

where π_0 is the identity operator. Now, consider some fixed index i and let $g = \pi_0 \cdots \pi_{i-1,m_{i-1}} f$. Using (1) for the degree of approximation of the one-dimensional function $g(x)$, and viewing $\pi_{i,m_i} g = \pi_0 \cdots \pi_{i,m_i} f$ as a one-dimensional function of x_i we have,

$$\begin{aligned} \|g - \pi_{i,m_i} g\|_{p,w} &\overset{(a)}{\leq} c_2 E(g, \mathbf{P}_{m_i}, w) \\ &\overset{(b)}{\leq} c_2 c_1^{r_i} \left(\frac{q_{m_i}}{m_i}\right) \cdots \left(\frac{q_{m_i - r_i + 1}}{m_i - r_i + 1}\right) E(g^{(r_i)}, \mathbf{P}_{m_i - r_i}, L_{p,w}), \end{aligned} \tag{15}$$

where step (a) makes use of (11) while (b) is based on (2) applied r_i times.

Letting D^{r_i} represent the r_i'th partial derivative with respect to x_i, we have

$$g^{(r_i)}(\mathbf{x}) = D^{r_i} \pi_0 \cdots \pi_{i-1,m_{i-1}} f(\mathbf{x}) = \pi_0 \cdots \pi_{i-1,m_{i-1}} D^{r_i} f(\mathbf{x}),$$

and

$$\|\pi_0 \cdots \pi_{i-1,m_{i-1}} f\|_{p,w} \leq (1 + c_2) \|f\|_{p,w}.$$

The last step follows from the observation that

$$\begin{aligned} \|\pi_0 \cdots \pi_{i-1,m_{i-1}} f\|_{p,w} &\leq \|\pi_0 \cdots \pi_{i-1,m_{i-1}} f - f\|_{p,w} + \|f\|_{p,w} \\ &\leq c_2 E(f, \mathbf{P}_{m_i - 1}, L_{p,w}) + \|f\|_{p,w} \leq (1 + c_2) \|f\|_{p,w}. \end{aligned}$$

From these inequalities we conclude that

$$E(g^{(r_i)}, \mathbf{P}_{m_i - 1}, L_{p,w}) \leq \|g^{(r_i)}\|_{p,w} = \|\pi_0 \cdots \pi_{i-1,m_{i-1}} D^{r_i} f\|_{p,w} \leq (1 + c_2) \|D^{r_i} f\|_{p,w}. \tag{16}$$

By setting $r_1 = \cdots = r_d = r$ and $m_1 = \cdots = m_d = m$ we conclude from (14), (15) and (16) that

$$\|f - \pi f\|_{p,w} \leq c_2 (1 + c_2) c_1^r \sum_{i=1}^{d} \left(\frac{q_m}{m}\right) \cdots \left(\frac{q_{m-r+1}}{m - r + 1}\right) \|D^r f\|_{p,w}.$$

Assume for the moment that $m > r(r-1)$ and denote $\beta = 1 - \frac{1}{\alpha}$. Also let $c = dMc_2(1 + c_2) c_1^r \alpha^{-r/\alpha}$. We know from (4) that $q_m = (m/\alpha)^{1/\alpha}$ from which we conclude that

$$\begin{aligned} \|f - \pi f\|_{p,w} &\leq c \frac{1}{m^\beta} \frac{1}{(m-1)^\beta} \cdots \frac{1}{(m - r + 1)^\beta} \\ &= \frac{c}{m^{r\beta}} \frac{1}{(1 - \frac{1}{m})^\beta} \cdots \frac{1}{(1 - \frac{r-1}{m})^\beta} \\ &\overset{(a)}{\leq} \frac{c}{m^{r\beta}} \frac{1}{(1 - \frac{r(r-1)}{2m})^\beta} \\ &\overset{(b)}{\leq} 2^\beta c m^{-r\beta}. \end{aligned} \tag{17}$$

In step (a) we have used the inequality ([7], p. 61)

$$\prod_{i=1}^{n}(1+a_i) \geq 1 + \sum_{i=1}^{n} a_i,$$

which is valid when all the a_i are positive or negative simultaneously and $a_i \geq -1$ for each i. Step (b) follows from the condition $m > r(r-1)$, which implies that $r(r-1)/2m < 1/2$. In order to obtain a bound valid for all m, we note that for $m \leq r(r-1)$ we always have the trivial bound $\|f - \pi f\|_{p,w} \leq M$ since $\|f\|_{p,w} \leq M$. Thus, we conclude that

$$\|f - \pi f\|_{p,w} \leq \max\left(2^{\beta} c m^{-r\beta}, M\right).$$

It is then easy to see that by defining $c' = \max\left\{c, 2^{-\beta} M (r(r-1))^{r\beta}\right\}$ we get an inequality of the desired type valid for every m. □

3 Approximation by Sigmoidal Neural Networks

We consider the approximation of functions by feedforward neural networks with a ridge function non-linearity. The approach we take is similar to that in [14], where the appropriate extensions from a compact domain to $\mathbb{R}^d$ are performed. We define the approximating structure composed of a single-hidden layer feedforward neural network with n hidden units. Formally we have

$$\mathcal{H}_n = \left\{ h \ : \ h(\mathbf{x}) = \sum_{k=1}^{n} c_k \phi(\mathbf{a}_k \cdot \mathbf{x} + b_k) \ ; \ \mathbf{a}_k \in \mathbb{R}^d, b_k, c_k \in \mathbb{R}, \ k = 1, 2, \ldots, n \right\} . \quad (18)$$

In the sequel we will need two assumptions concerning the activation function ϕ. In order to make them transparent we state them out clearly here.

Assumption 3.1 *There is a constant b such that $|\phi^{(k)}(b)| \geq c_\phi > 0$ for $k = 1, 2, \ldots$.*

This assumption is a natural one and holds, for example, for functions such as $\phi(x) = (1 + e^{-x})^{-1}, e^{-x^2}$ as well as many others (see [14] for further examples and discussion). It clearly does *not* hold for algebraic polynomials of finite degree.

Assumption 3.2 *For each finite k, there is a finite constant d_k such that*

$$\sup_{t \in \mathbb{R}} |\phi^{(k)}(t)| \leq d_k.$$

For example, it is not hard to show, by direct differentiation, for the standard sigmoidal function $\phi(t) = (1 + e^{-t})^{-1}$ that $d_k = dk!$, where d is a universal constant.

3.1 Approximating polynomials by sigmoidal networks

Let $\mathcal{F}$ and $\mathcal{G}$ represent two sets in the functional space $L_{p,w}$. We then define the distance between $\mathcal{F}$ and $\mathcal{G}$ as

$$\text{dist}\,\{\mathcal{F},\mathcal{G},L_{p,w}\} \stackrel{\triangle}{=} \sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}} \|f-g\|_{p,w}. \tag{19}$$

Before presenting the main result of this section, we define a class of multivariate algebraic polynomials with bounded coefficients.

$$\mathbf{BP}_{\mathbf{m}}(V) = \left\{ P \;:\; P(\mathbf{x}) = \sum_{0\le\mathbf{s}\le\mathbf{m}} a_{\mathbf{s}}\mathbf{x}^{\mathbf{s}} \;;\; \max_{0\le\mathbf{s}\le\mathbf{m}} |a_{\mathbf{s}}| \le V \right\}. \tag{20}$$

Here and in the sequel, the notation $0 \le \mathbf{s} \le \mathbf{m}$ should be interpreted component-wise, namely $0 \le s_i \le m_i$, $i = 1, 2, \ldots, d$. The basic result of this section is then the following.

Theorem 3.1 *Let assumptions 3.1 and 3.2 hold for the activation function ϕ. Then for every $0 < V < \infty$, $\mathbf{m} = (m_1, m_2, \ldots, m_d) \in \mathbf{Z}_+^d$, $m_i \le m$, $\epsilon > 0$ and $n > (m+1)^d$ we have*

$$\text{dist}\,\{\mathbf{BP}_{\mathbf{m}}(V), \mathcal{H}_n, L_{p,w}\} \le \epsilon.$$

Before presenting the proof of the theorem, we introduce some definitions and lemmas. The proof of the lemmas is deferred to Appendix A.

Following [14] consider the partial derivative

$$\phi^{(\mathbf{s})}(\mathbf{w}\cdot\mathbf{x}+b) \stackrel{\triangle}{=} \frac{\partial^{(|\mathbf{s}|)}}{\partial w_1^{s_1}\cdots\partial w_d^{s_d}}\left[\phi(\mathbf{w}\cdot\mathbf{x}+b)\right] = \mathbf{x}^{\mathbf{s}}\phi^{(|\mathbf{s}|)}(\mathbf{w}\cdot\mathbf{x}+b), \tag{21}$$

where $|\mathbf{s}| = s_1 + \cdots + s_d$, and $\mathbf{x}^{\mathbf{s}} = \prod_{i=1}^d x_i^{s_i}$. Thus

$$\phi^{(\mathbf{s})}(b) = \mathbf{x}^{\mathbf{s}}\phi^{(|\mathbf{s}|)}(b).$$

For any fixed b, consider a finite difference of order $\mathbf{s}$ [15]:

$$\Delta_{h,x}^{\mathbf{s}}\phi(b) = \sum_{0\le\mathbf{l}\le\mathbf{s}} (-1)^{|\mathbf{l}|}\binom{\mathbf{s}}{\mathbf{l}}\phi(h\mathbf{l}\cdot\mathbf{x}+b), \tag{22}$$

where $\binom{\mathbf{s}}{\mathbf{l}} = \prod_{i=1}^d \binom{s_i}{l_i}$. Note that $\Delta_{h,x}^{\mathbf{s}}\phi(b) \in \mathcal{H}_n$ with $n = \prod_{i=1}^d (s_i+1)$.

Lemma 3.1 *Let Assumption 3.2 be given. Then for any $N > 0$ and $h > 0$ there exists an $\epsilon_{N,h}$ such that*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \left|\phi^{(|\mathbf{s}|)}(b)\mathbf{x}^{\mathbf{s}} - h^{-|\mathbf{s}|}\Delta_{h,x}^{\mathbf{s}}\phi(b)\right| \le \begin{cases} N^{|\mathbf{s}|}\epsilon_{N,h} & , \|\mathbf{x}\|_\infty \le N, \\ |\mathbf{x}^{\mathbf{s}}|\, d_{|\mathbf{s}|} & , \|\mathbf{x}\|_\infty > N, \end{cases}$$

where $\|\mathbf{x}\|_\infty = \max_{1\le i\le d}|x_i|$ and $\epsilon_{N,h} \to 0$ for $h \to 0$.

Next, we need to upper bound the coefficients of the polynomials defined in (13).

Lemma 3.2 *Let $P_{\mathbf{m}}(\mathbf{x})$ be defined as in (13). Then if we represent*

$$P_{\mathbf{m}}(\mathbf{x}) = \sum_{0 \le \mathbf{s} \le \mathbf{m}} a_{\mathbf{s}} \mathbf{x}^{\mathbf{s}},$$

we have that

$$\max_{0 \le \mathbf{s} \le \mathbf{m}} |a_{\mathbf{s}}| \le c_3^{|\mathbf{m}|} \qquad (c_3 > 1).$$

We are now ready to prove the main result of this section.

Proof of Theorem 3.1: In order to simplify the notation we use the convention $s = |\mathbf{s}|$. Let ϵ_1' and ϵ_2' be arbitrary positive constants. We then have

$$\begin{aligned}
& \operatorname{dist}\left\{\mathbf{BP}_{\mathbf{m}}(c_3^{|\mathbf{m}|}), \mathcal{H}_n, L_{p,w}\right\}^p \\
\le\ & \left\| \sum_{0 \le \mathbf{s} \le \mathbf{m}} a_{\mathbf{s}} \mathbf{x}^{\mathbf{s}} - \sum_{0 \le \mathbf{s} \le \mathbf{m}} a_{\mathbf{s}} \frac{\Delta_{h,x}^{\mathbf{s}} \phi(b)}{h^s \phi^s(b)} \right\|_{p,w}^p \\
\le\ & (m+1)^d \max_{0 \le \mathbf{s} \le \mathbf{m}} \left\{ |a_{\mathbf{s}}| \left\| \mathbf{x}^{\mathbf{s}} - \frac{\Delta_{h,x}^{\mathbf{s}} \phi(b)}{h^s \phi^{(s)}(b)} \right\|_{p,w}^p \right\} \\
\overset{(a)}{\le}\ & (m+1)^d c_3^{|\mathbf{m}|} \max_{0 \le \mathbf{s} \le \mathbf{m}} \left\{ \left(\phi^{(s)}(b)\right)^{-p} \left[\int_{\|\mathbf{x}\|_\infty \le N} |N^s \epsilon_{N,h}|^p \, w^p(\mathbf{x}) d\mathbf{x} + \int_{\|\mathbf{x}\|_\infty > N} (|\mathbf{x}^{\mathbf{s}}| d_s)^p \, w^p(\mathbf{x}) d\mathbf{x} \right] \right\} \\
\overset{(b)}{\le}\ & c_\phi^{-p} (m+1)^d c_3^{|\mathbf{m}|} \left\{ \epsilon_1' + \max_{0 \le s \le m} d_s^p \int_{\|\mathbf{x}\|_\infty > N} \left(\prod_i |x_i|\right)^{sp} e^{-\sum_i p|x_i|^\alpha} d\mathbf{x} \right\} \\
\le\ & c_\phi^{-p} (m+1)^d c_3^{|\mathbf{m}|} \left\{ \epsilon_1' + \max_{0 \le s \le m} d_s^p \int_{\{\mathbf{x} \,:\, |x_1| > N\}} \left(\prod_i |x_i|\right)^{sp} e^{-\sum_i p|x_i|^\alpha} d\mathbf{x} \right\} \\
\le\ & c_\phi^{-p} (m+1)^d c_3^{|\mathbf{m}|} \left\{ \epsilon_1' + \max_{0 \le s \le m} d_s^p \left(\int_{|t| > N} t^{sp} e^{-pt^\alpha} dt \right) \left(\int_{\mathbb{R}} t^{sp} e^{-pt^\alpha} dt \right)^{d-1} \right\} \\
\overset{(c)}{\le}\ & c_\phi^{-p} (m+1)^d c_3^{|\mathbf{m}|} \left(\epsilon_1' + \epsilon_2'\right) \\
=\ & \epsilon_1 + \epsilon_2,
\end{aligned}$$

where $\epsilon_i = c_\phi^{-p}(m+1)^d c_3^{|\mathbf{m}|} \epsilon_i$, $i = 1, 2$. Step (a) follows from Lemmas 3.1 and 3.2, while step (b) makes use of Assumption 3.1 and the fact that $\epsilon_{N,h}$ can be made arbitrarily small by letting $h \to 0$. Step (c) uses the fact that $\int_{\mathbb{R}} t^{sp} e^{-pt^\alpha} dt$ is finite since $\alpha \ge 2$, which implies that one may choose a sufficiently large N so that the integral over the domain $|t| > N$ is arbitrarily small. Now, for each $\epsilon > 0$ choose N large enough so that $\epsilon_2 < \frac{\epsilon}{2}$, which is possible by the above argument, and $\epsilon_1 < \frac{\epsilon}{2}$, which can be achieved by taking h to be sufficiently small. □

The final result required may be obtained by combining Lemma 2.1 and Theorem 3.1.

Theorem 3.2 *Let $1 \leq p \leq \infty$ and $w(\mathbf{x}) = \prod_{i=1}^{d} e^{-|x_i|^\alpha}$, $\alpha \geq 2$. We then have*

$$\text{dist}\left\{W_{p,w}^{r,d}, \mathcal{H}_n, L_{p,w}\right\} \leq cn^{-(1-\frac{1}{\alpha})\frac{r}{d}}.$$

PROOF The theorem follows on using the triangle inequality. Clearly for any $f \in W_{p,w}^{r,d}$ and $h \in \mathcal{H}_n$

$$\|f - h\|_{p,w} \leq \|f - P_{\mathbf{m}}\|_{p,w} + \|P_{\mathbf{m}} - h\|_{p,w},$$

where $P_{\mathbf{m}} \in \mathbf{BP}_{\mathbf{m}}(c_3^{|\mathbf{m}|})$. From Lemma 2.1 the first term can be upper bounded by $cm^{(1-\frac{1}{\alpha})r}$ for some $P_{\mathbf{m}}$. From Theorem 3.1 the second term can be upper bounded by ϵ, for any $\epsilon > 0$. Taking ϵ to be of the same order as the first term, the theorem follows. □

Theorem 3.2 presents us with an upper bound on the approximation error. In order to assess the tightness of the bound we present, without proof, a recent result concerning a lower bound for the case $p = 2$. A full statement and proof can be found in [12].

Theorem 3.3 *If $d \geq 2$, $p = 2$ and $\alpha \geq 2$ we have*

$$\text{dist}\left\{W_{2,w}^{r,d}, \mathcal{H}_n, L_{2,w}\right\} \geq \frac{c'}{(n \ln n)^{\frac{r}{d-1}}} .$$

We see from Theorem 3.3 that the upper bound presented in Theorem 3.2 is not as tight as possible. At this point we leave it as an open problem whether the upper or the lower bounds can be improved.

4 Approximation by Mixture Networks

Having established approximation bounds for feedforward networks with a single hidden layer, we consider now a related class of networks, whose structure is somewhat constrained. The mixture of expert architecture has been introduced by Jordan and colleagues (see for example [10]), and has been receiving increasing attention over the past few years. As pointed out in [10], these modular networks possess several distinct advantages over the more standard feedforward neural networks. These advantages mainly have to do with more efficient training procedures and generalization capability. In this section we extend our previous work on approximation by mixture networks over compact domains [16], and derive analogous results for approximation in $\mathbb{R}^d$. As we will see the upper bounds attained are equivalent to those obtained by using standard neural networks. Coupled with the above remarks pertaining to the advantages of the mixture networks with respect to training and generalization, we believe that our results support the generally emerging view as to the superiority of modular networks over the monolithic types of structures studied in the past.

We briefly motivate the approximating structure studied in this section. Further details can be found in [10] and [16]. Consider the problem of modeling a regression function $\mathrm{E}(y|\mathbf{x})$, in a probabilistic setting. The mixture of experts model is composed of n expert networks, each of which solves a function approximation problem over a local region of the input space. A probabilistic model, that relates input vectors $\mathbf{x} \in \mathbb{R}^d$ to output vectors $y \in \mathbb{R}$, is associated with each expert. We denote the probability model of each expert as follows

$$p(y\ |\mathbf{x}; \boldsymbol{\theta}_j) \qquad j = 1, 2, \ldots, n,$$

where the $\boldsymbol{\theta}_j \in \Theta$ are parameter vectors associated with each expert. Typically, these densities are chosen from the exponential family. Thus, the overall probabilistic model assumes the form of a mixture density

$$p(y|\mathbf{x}; \Theta) = \sum_{j=1}^{n} g_j(\mathbf{x}; \boldsymbol{\theta}_g) p(y|\mathbf{x}; \boldsymbol{\theta}_j), \tag{23}$$

where

$$g_j(\mathbf{x}; \boldsymbol{\theta}_g) \geq 0 \quad \text{and} \quad \sum_{j=1}^{n} g_j(\mathbf{x}; \boldsymbol{\theta}_g) = 1. \tag{24}$$

The regression function $\mathrm{E}(y|\mathbf{x}, \Theta)$ is obtained by taking the expectation with respect to (23), giving rise to

$$\mathrm{E}(y|\mathbf{x}) = \sum_{j=1}^{n} g_j(\mathbf{x}; \boldsymbol{\theta}_j) \mu(\mathbf{x}; \boldsymbol{\theta}_j). \tag{25}$$

In order to study the approximation ability of networks of the form (25) we limit ourselves to linear functions of the form $\mu(\mathbf{x}; \boldsymbol{\theta}_j) = \boldsymbol{\theta}_j \cdot \mathbf{x} + \boldsymbol{\theta}_{j0}$. In fact, the approximation bounds will be obtained by simply retaining the constant term above, namely $\mu(\mathbf{x}; \boldsymbol{\theta}_j) = \theta_{j0}$. Note that the gating functions $g_j(\mathbf{x}; \boldsymbol{\theta}_j)$ still retain their full non-linearity. In currently ongoing work we in fact show that very little is gained by allowing $\mu(\mathbf{x}; \boldsymbol{\theta}_j)$ to be a polynomial in $\mathbf{x}$, rather than a constant as in this work.

Making the normalization in (24) explicit, we consider the approximation of functions in $W_{p,w}^{r,d}$ by mixture networks, formally defined as follows:

$$\mathcal{G}_n = \left\{ g \ : \ g(\mathbf{x}) = \frac{\sum_{k=1}^{n} c_k \phi(\mathbf{a}_k \cdot \mathbf{x} + b_k)}{\sum_{k=1}^{n} \phi(\mathbf{a}_k \cdot \mathbf{x} + b_k)} \ , \mathbf{a}_k \in \mathbb{R}^d, b_k, c_k \in \mathbb{R} \right\}. \tag{26}$$

Note that in (26) the normalization requirement (24) has been explicitly taken into account. The basic result we prove establishes upper bounds on the degree of approximation of functions in $W_{p,w}^{r,d}$ by mixture networks belonging to $\mathcal{G}_n$. Before proceeding we make a technical assumption concerning the sigmoidal function ϕ appearing in (26), namely we assume a specific form for the activation function. The motivation for this assumption is purely technical at this point. At the end of this section we discuss a generalization to other functions.

Assumption 4.1 *The activation function $\phi(\cdot)$ in (26) is given by $\phi(t) = (1 + e^{-t})^{-1}$.*

It is easy to see that this standard function obeys both assumptions 3.1 and 3.2. We then have a result analogous to Theorem 3.2.

Theorem 4.1 *For $1 \leq p \leq \infty$ and $\alpha \geq 2$, the distance between the functional spaces $W_{p,w}^{r,d}$ and the space $\mathcal{G}_n$, is upper bounded as follows:*

$$\text{dist}\left\{W_{p,w}^{r,d}, \mathcal{G}_n, L_{p,w}\right\} \leq cn^{-(1-\frac{1}{\alpha})\frac{r}{d}},$$

where the constant c is independent of n.

The proof of Theorem 4.1 requires several lemmas, which we present below. The proof of the lemmas is relegated to Appendix B, in order to retain the continuity of the presentation. First, however, we define a function $\varphi(\mathbf{x})$ which will be utilized in the proof below. Let $0 < \rho < 1$ and define

$$\varphi(\mathbf{x}) = (2\rho)^{-d} \int_{[-\rho,\rho]^d} \frac{d\mathbf{y}}{1 + \exp(-\mathbf{y} \cdot \mathbf{x})} \ . \tag{27}$$

We then have the following three lemmas.

Lemma 4.1 *For the function $\varphi(\mathbf{x})$ defined in (27), the following results hold:*
1. $\varphi(\mathbf{x}) \geq \prod_{i=1}^{d} \left(1 + e^{\rho|x_i|}\right)^{-1}$, for all $\mathbf{x} \in \mathbb{R}^d$.
2. Let $\mathbf{s} = (s_1, \ldots, s_d)$. Then $\forall f \in W_{p,w}^{r,d}$

$$\|f\varphi\|_{W_{p,w}^{r,d}} = \sum_{|\mathbf{s}| \leq r} \|D^{(\mathbf{s})}(f\varphi)\|_{p,w} \leq M_1 < \infty,$$

where $M_1 = M_1(r, \rho, d) < \infty$, i.e. $f\varphi \in W_{p,w}^{r,d}$.

The next lemma requires the introduction of an auxiliary class of functions, defined as follows:

$$\mathcal{T}_n \stackrel{\triangle}{=} \left\{ t(\mathbf{x}) \mid t(\mathbf{x}) = \sum_{0 \leq \mathbf{k} \leq \mathbf{N}} d_{\mathbf{k}} \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})},\ d_{\mathbf{k}} \in \mathbb{R},\ \mathbf{k} \in \mathbf{Z}_+^d \right\}, \tag{28}$$

where the polynomials $P_{\mathbf{k}}$ are defined in (6), and $\mathbf{N} = (N, \ldots, N)$ with $N = n^{\frac{1}{d}}$. We then have:

Lemma 4.2

$$\text{dist}\left\{W_{p,w}^{r,d}, \mathcal{T}_n, L_{p,w}\right\} \leq cn^{-(1-\frac{1}{\alpha})\frac{r}{d}} \quad .$$

Finally, we have

Lemma 4.3 *Let* $\mathbf{s} = (s_1, \ldots, s_d)$ *and* $\hat{s} = \prod_{i=1}^{d}(1+s_i)$. *Then there exist coefficients* $\{a_{\mathbf{r},\mathbf{k}}\}$ *such that for* h *sufficiently small*

$$\left\| \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} - \frac{\sum_{0\le\mathbf{r}\le\mathbf{s}} a_{\mathbf{r},\mathbf{k}}\phi(h(2\mathbf{r}-\mathbf{s})\cdot\mathbf{x})}{\hat{s}^{-1}\sum_{0\le\mathbf{r}\le\mathbf{s}}\phi(h(2\mathbf{r}-\mathbf{s})\cdot\mathbf{x})} \right\|_{\infty} \le c_{d,\phi}h,$$

for some constant $c_{d,\phi}$, *where* $\|f(\mathbf{x})\|_{\infty} = \sup_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

With lemmas 4.1, 4.2 and 4.3 at hand we are ready to prove Theorem 4.1. To do so we introduce two useful abbreviations:

$$A_h(\mathbf{x},\mathbf{k}) \stackrel{\triangle}{=} \sum_{0\le\mathbf{r}\le\mathbf{s}} a_{\mathbf{r},\mathbf{k}}\phi(h(2\mathbf{r}-\mathbf{s})\cdot\mathbf{x}) \qquad ; \qquad B_h(\mathbf{x}) \stackrel{\triangle}{=} \hat{s}^{-1}\sum_{0\le\mathbf{r}\le\mathbf{s}} \phi(h(2\mathbf{r}-\mathbf{s})\cdot\mathbf{x}). \quad (29)$$

Proof of Theorem 4.1: Consider the function $g_{N,s,h}(\cdot) \in \mathcal{G}_n$, defined as follows:

$$g_{N,s,h}(\mathbf{x}) = \frac{\sum_{\mathbf{k}\le\mathbf{N}} d_{\mathbf{k}} \sum_{0\le\mathbf{r}\le\mathbf{s}} a_{\mathbf{r},\mathbf{k}}\phi(h(2\mathbf{r}-\mathbf{s})\cdot\mathbf{x})}{\hat{s}^{-1}\sum_{0\le\mathbf{r}\le\mathbf{s}}\phi(h(2\mathbf{r}-\mathbf{s})\cdot\mathbf{x})} = \frac{\sum_{\mathbf{k}\le\mathbf{N}} d_{\mathbf{k}} A_h(\mathbf{x},\mathbf{k})}{B_h(\mathbf{x})},$$

where $\mathbf{N} = (N, N, \ldots, N) = (n^{1/d}, \ldots, n^{1/d})$, $\mathbf{k}, \mathbf{r}, \mathbf{s} \in \mathbf{Z}_+^d$ and $a_{\mathbf{r},\mathbf{k}}, d_{\mathbf{k}} \in \mathbb{R}$. Recalling the definition of the function $\varphi(\cdot)$ in (27), and using the triangle inequality we have

$$\begin{aligned} \|f(\mathbf{x}) - g_{N,s,h}(\mathbf{x})\|_{p,w} &\le \left\| f(\mathbf{x}) - \sum_{\mathbf{k}\le\mathbf{N}} d_{\mathbf{k}} \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} \right\|_{p,w} \\ &+ \left\| \sum_{\mathbf{k}\le\mathbf{N}} d_{\mathbf{k}} \left(\frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} - \frac{\sum_{\mathbf{k}\le\mathbf{N}} d_{\mathbf{k}} A_h(\mathbf{x},\mathbf{k})}{B_h(\mathbf{x})} \right) \right\|_{p,w}. \end{aligned} \quad (30)$$

Since (30) holds for any $P_{\mathbf{k}}$, we may take the minimum over $P_{\mathbf{k}} \in \mathbf{P_N}$ and make use of Lemmas 4.2 and 4.3 to obtain

$$\inf_{a_{\mathbf{r},\mathbf{k}}\in\mathbb{R}} \|f(\mathbf{x}) - g_{N,s,h}(\mathbf{x})\|_{p,w} \le cn^{-(1-\frac{1}{\alpha})\frac{r}{d}} + \sum_{\mathbf{k}\le\mathbf{N}} |d_{\mathbf{k}}| c_{d,\phi} h.$$

Setting $h \le n^{-(1-\frac{1}{\alpha})\frac{r}{d}} / c_{d,\phi} \sum_{\mathbf{k}\le\mathbf{N}} |d_{\mathbf{k}}|$ the theorem follows. □

In accordance with assumption 4.1 the proof so far applies specifically to the standard sigmoidal function $\phi(t) = (1+e^{-t})^{-1}$. It is not hard to see from the proof in this section and in Appendix B that the result can in fact be generalized to functions obeying the following conditions.

1. $\phi(t) = (1+e^{-h(t)})^{-1}$, $t \in \mathbb{R}$.
2. $\phi(t)$ satisfies the conditions in assumption 3.1.
3. $h(t)$ is a non-decreasing function on $\mathbb{R}$.
4. $0 \le h(t) \le 1 + |t|^{\alpha}$ for every $t \in \mathbb{R}$.

We remark that while the proof given holds with little modification under the conditions listed above, the present method of proof does not permit generalization to a wider class of activation functions as in Section 3, for the case of the standard neural network.

5 Discussion

In this paper we have been concerned with deriving upper bounds on the approximation error attained by feedforward neural networks and mixture of expert architectures in approximating smooth functions, belonging to some well-defined Sobolev class. While much previous work in this field has been concerned with approximation over compact domains. The main contribution of this work is in the establishment of bounds which hold over the full Euclidean space $\mathbb{R}^d$. Although some results along these lines exist (see for example [5] and [9]), they either do not apply to the class of functions considered in this paper, or consider different types of approximation functions. Moreover, the method of proof suggested in this paper is entirely different.

It may be argued that the issue of generalizing approximation bounds from bounded domains to $\mathbb{R}^d$ is mainly of technical interest. We note, however, that recent results in the literature of model selection [2] make extensive use of these results in deriving guaranteed performance bounds for various learning algorithms. Since many problems of practical interest are characterized by a-priori unbounded domains it is crucial to obtain upper bounds on the approximation error in this more general setting, in order that the performance of the algorithms be properly assessed.

Acknowledgments We are grateful to H.N. Mhaskar for useful discussions and for bringing reference [13] to our attention.

References

[1] A.R. Barron, "Universal Approximation Bound for Superpositions of A Sigmoidal Function," *IEEE Trans. Inf. Theory*, vol. IT-39, pp. 930-945, 1993.

[2] A.R. Barron, L. Birge and P. Massart, "Risk Bounds for Model Selection via Penalization", Yale University preprint, June 1995.

[3] T. Chen, H. Chen and R. Liu, "Approximation Capability in $\mathbf{C}(\bar{\mathbb{R}}^n)$ by Multilayer Feedforward Networks and Related Problems", *IEEE Trans. Neural Networks*, vol. 6, no. 1:25-30, 1995.

[4] G. Cybenko, "Approximations by Superpositions of a Sigmoidal Function", *Math. Contr. Signals, Systems*, vol. 2:303-314, 1989.

[5] B. Delyon, A. Judistky and A. Benveniste, "Accuracy Analysis for Wavelet Approximations", *IEEE Trans. Neural Networks*, vol. 6, no. 2, 1995.

[6] G. Freud, "On Polynomial Approximations with Respect to General Weights", in *Lecture Notes*, vol. 399 (H.G. Garnir *et al.* Eds.), pp. 149-179, Springer Verlag, Berlin, 1974.

[7] G. Hardy, J.E. Littlewood and G. Polya, *Inequalities*, Cambridge Mathematical Library, Second Edition, 1952.

[8] K. Hornik, M. Stinchcombe and H. White, "Universal Approximation of an Unknown Function and its Derivatives Using Multilayer Feedforward Networks", *Neural Networks*, vol. 3:551-560, 1990.

[9] K. Hornik, M. Stinchcombe, H. White and P. Auer, "Degree of Approximation Results for Feedforward Networks Approximating Unknown Mappings and their Derivatives", *Neural Computation*, vol. 6: 1262-1275, 1994.

[10] Jordan, M.I. and Jacobs, R.A. "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation*, vol. 6:181-214, 1994.

[11] Leshno, M., Lin, V., Pinkus, A. and Schocken, S. "Multilayer Feedforward Networks with a Polynomial Activation Function Can Approximate any Function", *Neural Networks*, vol. 6, 861-867, 1993.

[12] V. Maiorov, "On Best Approximation by Ridge Functions", submitted to J. APprox. Theory, 1996.

[13] H.N. Mhaskar, "Weighted Polynomial Approximation", J. Approx. Theory, 46:100-110, 1986.

[14] H.N. Mhaskar, "Neural Networks for Optimal Approximation of Smooth and Analytic Functions", *Neural Computation*, 164-177, 1995.

[15] A.F. Timan, *Theory of Approximation of Functions of a Real Variable*, Macmillan, New York, 1963.

[16] A. Zeevi, R. Meir and V. Maiorov, "Error Bounds for Functional Approximation and Estimation Using Mixtures of Experts", Technical Report #CC-132, Technion, Israel Institute of Technology.

Appendix A

We prove the lemmas stated in Section 3.

Proof of Lemma 3.1: We need to upper bound the difference between the derivative $\phi^{(\mathbf{s})}$ and the finite difference $h^{-|\mathbf{s}|}\Delta^{\mathbf{s}}_{h,x}\phi(b)$. We first represent the finite difference (22)

as an integral operator. Let $\mathbf{s} = (s_1, s_2, \ldots, s_d)$. Then for any function $g(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^d$, let

$$I_h^q g(\mathbf{y}) \stackrel{\triangle}{=} \int_0^h \cdots \int_0^h \phi^{(|\mathbf{s}|)} \left[\left((\tau_1 + \cdots + \tau_{s_1}) y_1 + \cdots + (\tau_{|\mathbf{s}|-s_d+1} + \cdots \tau_{|\mathbf{s}|}) y_d \right) + b \right] d\tau_1 \cdots d\tau_{|\mathbf{s}|}. \tag{31}$$

One can then show (see Lemma A.1 below) that

$$\Delta_{h,x}^{\mathbf{s}} \phi(b) = \mathbf{x}^{\mathbf{s}} I_h^{|\mathbf{s}|} \phi(\mathbf{x}). \tag{32}$$

We then have

$$\begin{aligned} A_x &\stackrel{\triangle}{=} \phi^{(\mathbf{s})}(b) - h^{-|\mathbf{s}|} \Delta_{h,x}^{\mathbf{s}} \phi(b) \\ &= \mathbf{x}^{\mathbf{s}} \left(\phi^{(|\mathbf{s}|)}(b) - h^{-|\mathbf{s}|} I_h^{\mathbf{s}} \phi^{(|\mathbf{s}|)}(\mathbf{x}) \right), \end{aligned} \tag{33}$$

where we have used (21) and (32). Let N be a fixed positive number. Using (31) and the mean value theorem we conclude that there is a $\xi \in [0, |\mathbf{t} \cdot \mathbf{x}|]$, where $\mathbf{t} = (s_1 h, \ldots, s_d h)$, such that

$$I_h^{|\mathbf{s}|} \phi(\mathbf{x}) = h^{|\mathbf{s}|} \phi^{(|\mathbf{s}|)}(b + \xi). \tag{34}$$

If $\|\mathbf{x}\|_\infty \leq N$ we then have from (33) and (34)

$$|A_x| \leq N^{|\mathbf{s}|} |\phi^{(|\mathbf{s}|)}(b) - \phi^{(|\mathbf{s}|)}(b + \xi)| \leq N^{|\mathbf{s}|} \epsilon_{N,h}, \tag{35}$$

where $\xi \in [0, |\mathbf{t} \cdot \mathbf{x}|] \in [0, h|\mathbf{s}|N]$. Note that $\epsilon_{N,h}$ can be made arbitrarily small by letting h approach zero. In order to upper bound A_x in the case $\|\mathbf{x}\|_\infty > N$, we make use of Assumption 3.2 yielding

$$|A_x| \leq 2|\mathbf{x}^{\mathbf{s}}| \sup_{t \in \mathbb{R}} |\phi^{|\mathbf{s}|}(t)| \leq 2 d_{|\mathbf{s}|} |\mathbf{x}^{\mathbf{s}}|. \tag{36}$$

The Lemma follows by combining (35) and (36). □

Proof of Lemma 3.2: The proof makes use of the inequalities (7) of the polynomials $P_m(t)$, defined in (5). The polynomial defined in (6) in terms of its roots can also be expressed in the form

$$P_m(t) = b_m t^m + \cdots + b_1 t + b_0 = b_m \prod_{j=1}^{m} (t - t_{j,m}).$$

In order to bound the coefficients $\{b_j\}$ we expand the polynomial $P_m(t)$ in terms of its roots, and equate coefficients. We then have

$$b_m = \gamma_m \qquad ; \qquad b_j = \gamma_m \sum_{\substack{k_1, \cdots, k_{m-j} \\ k_i \neq k_j}}^{m} (-1)^{m-j} t_{k_1,m} \cdots t_{k_{m-j},m},$$

from which we infer, using (7), that

$$\begin{aligned} |b_j| &\leq \binom{m}{j} \max_{k_1,\ldots,k_{m-j}} |\gamma_m t_{k_1,m} \cdots t_{k_{m-j},m}| \\ &\leq 2^m \frac{2^m}{\sqrt{m!}} \left(4\sqrt{m-1}\right)^{m-j} \\ &\leq 16^m e^{m/2} \frac{\left(\sqrt{m-1}\right)^{m-j}}{\sqrt{m!}} \\ &\leq c_3^m, \end{aligned}$$

where we made use of Stirling's approximation for $m!$. The theorem then follows (with $c_3 = 16\sqrt{e}$) on using (13). □

Lemma A.1 *The finite difference (22) is related to the integral operator (31) through*

$$\Delta_{h,x}^{\mathbf{s}} \phi(b) = \mathbf{x}^{\mathbf{s}} I_h^{|\mathbf{s}|} \phi(\mathbf{x}).$$

PROOF We prove the statement by induction. Let $\mathbf{s} = (1, 0, \ldots, 0)$. Then

$$I_h^{(\mathbf{s})} \phi(\mathbf{x}) = \int_0^h \phi'(\tau_1 x_1 + b) d\tau_1 = x_1^{-1} \int_b^{b+hx_1} \phi'(y) dy = x_1^{-1} \left(\phi(b + hx_1) - \phi(b)\right) = x_1^{-1} \Delta_{h,x}^{(\mathbf{s})} \phi(b).$$

Assume now the statement is correct for $\mathbf{s} = (s_1, s_2, \ldots, s_d)$; we show it also holds for $\mathbf{s}' = (s_1 + 1, s_2, \ldots, s_d)$. Let $\mathbf{e}_i$ be the standard unit vectors, namely $e_{i,j} = \delta_{ij}$. Then

$$\begin{aligned} I_h^{(\mathbf{s}')} \phi(\mathbf{x}) &= \int_0^h \cdots \int_0^h \phi^{|\mathbf{s}'|} \left((\tau_1 + \cdots + \tau_{s_1} + \tau_{s_1+1}) x_1 + \cdots + b\right) d\tau_1 \cdots d\tau_{|\mathbf{s}|+1} \\ &= \int_0^h \cdots \int_0^h \left\{ \frac{\partial}{\partial \tau_{s_1+1}} \phi^{|\mathbf{s}|} \left((\tau_1 + \cdots + \tau_{s_1} + \tau_{s_1+1}) x_1 + \cdots + b\right) \right\} d\tau_1 \cdots d\tau_{|\mathbf{s}|+1} \\ &= x_1^{-1} \int_0^h \cdots \int_0^h \Big[\phi^{|\mathbf{s}|} \left((\tau_1 + \cdots + \tau_{s_1} + h) x_1 + \cdots + b\right) \\ &\quad - \phi^{|\mathbf{s}|} \left((\tau_1 + \cdots + \tau_{s_1}) x_1 + \cdots + b\right) \Big] d\tau_1 \cdots d\tau_{|\mathbf{s}|} \\ &= x_1^{-1} \left[I_h^{(\mathbf{s})} \phi(\mathbf{x} + hx_1 \mathbf{e}_1) - I_h^{(\mathbf{s})} \phi(\mathbf{x}) \right] \\ &= x_1^{-1} \Delta_{h,x}^{(\mathbf{s}')} \phi(b) \ . \end{aligned} \tag{37}$$

The induction step proceeds similarly for the other components. □

Appendix B

We prove the lemmas stated in Section 4.

Proof of Lemma 4.1: For all $\mathbf{x} \in \mathbb{R}^d$ and $y \in [-\rho, \rho]^d$ we have $|\mathbf{y} \cdot \mathbf{x}| \leq \rho(|x_1| + \cdots + |x_d|)$ and hence

$$\varphi(\mathbf{x}) \geq \frac{1}{1 + \exp\left(\rho(|x_1| + \cdots + |x_d|)\right)} \geq \frac{1}{(1 + e^{\rho|x_1|}) \cdots (1 + e^{\rho|x_d|})},$$

which establishes the first part of the lemma. The second part of the lemma is proved as follows.

$$\begin{aligned}
\|f\varphi\|_{W^{r,d}_{p,w}} &= \sum_{|\alpha|\le r}\left(\int_{\mathbb{R}^d}|D^\alpha(f\varphi)(\mathbf{x})|^p w(\mathbf{x})^p d\mathbf{x}\right)^{1/p} \\
&\overset{(a)}{=} \sum_{|\alpha|\le r}\left(\int_{\mathbb{R}^d}\left|\sum_{\alpha'+\alpha''=\alpha} A_{\alpha',\alpha''}[D^{\alpha'}fD^{\alpha''}\varphi](\mathbf{x})\right|^p w(\mathbf{x})^p d\mathbf{x}\right)^{1/p} \\
&\overset{(b)}{\le} \sum_{|\alpha|\le r}\sum_{\alpha'+\alpha''=\alpha}|A_{\alpha',\alpha''}|\left(\int_{\mathbb{R}^d}|[D^{\alpha'}fD^{\alpha''}\varphi](\mathbf{x})|^p w(\mathbf{x})^p d\mathbf{x}\right)^{1/p} \\
&\overset{(c)}{\le} \sum_{|\alpha|\le r}\sum_{\alpha'+\alpha''=\alpha}|A_{\alpha',\alpha''}|\ \|D^{\alpha'}f(\mathbf{x})\|_{p,w}\|D^{\alpha''}\varphi(\mathbf{x})\|_\infty \\
&\overset{(d)}{\le} c_{r,\phi}\sum_{|\alpha|\le r}\|D^\alpha f\|_{p,w} \\
&\overset{(e)}{=} c_{r,\phi}\|f\|_{W^{r,d}_{p,w}} \le M_1, \qquad (38)
\end{aligned}$$

where (a) follows from the chain rule of differentiation and the coefficients $A_{\alpha',\alpha''}$ depend only on α' and α''. Steps (b) and (c) follow from Minkowski's and Hölder's inequalities (with $p=1$ and $q=\infty$), respectively. Step (d) follows from the boundedness of the derivatives of $\phi(\mathbf{w}\cdot\mathbf{x})$, that is

$$\|D^{\alpha''}\varphi\|_\infty = \left\|(2\rho)^{-d}\int_{[\rho,\rho]^d} D_x^{\alpha''}\phi(\mathbf{w}^T\mathbf{x})d\mathbf{w}\right\|_\infty \le c_\phi$$

and rewriting the summation over the derivatives of f. Finally, step (e) is established by the assumption of $f\in W^{r,d}_{p,w}$ so that $\|f\|_{W^{r,d}_{p,w}}\le M$. □

Proof of Lemma 4.2: From the definition of $\|\cdot\|_{p,w}$ we have

$$D = \left\|f(\mathbf{x}) - \sum_{\mathbf{k}\le\mathbf{N}} d_{\mathbf{k}}\frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})}\right\|_{p,w} = \left\|f(\mathbf{x})\varphi(\mathbf{x}) - \sum_{\mathbf{k}\le\mathbf{N}} d_{\mathbf{k}}P_{\mathbf{k}}(\mathbf{x})\right\|_{p,\frac{w}{\varphi}}. \qquad (39)$$

Now, consider the function $\frac{w}{\varphi}$, which based on Lemma 4.1 can be bounded as follows

$$\frac{w(\mathbf{x})}{\varphi(\mathbf{x})} \le \prod_{i=1}^d \left\{e^{-|x_i|^\alpha}\left(1+e^{\rho|x_i|}\right)\right\}.$$

Since $|t|\le 1+|t|^\alpha$ for $t\in\mathbb{R}$ and $\alpha\ge 2$, it follows that

$$1+e^{|t|}\le 2e^{|t|}\le 2e^{1+|t|^\alpha} = (2e)e^{|t|^\alpha}.$$

We therefore conclude that

$$\frac{w(\mathbf{x})}{\varphi(\mathbf{x})} \le \prod_{i=1}^d\left((2e)e^{-|x_i|^\alpha}e^{\rho^\alpha|x_i|^\alpha}\right) = (2e)^d\prod_{i=1}^d e^{-a_\rho|x_i|^\alpha} \qquad (a_\rho = 1-\rho^\alpha).$$

Since $0 < \rho < 1$ clearly $0 < a_\rho < 1$. Consider then the function

$$Q(t) = a_\rho |t|^\alpha \qquad (\alpha \geq 2,\ a_\rho > 0).$$

It is easy to verify that this function satisfies the conditions in Assumption 2.1. Since from Lemma 4.1 $f\varphi \in W_{p,w}^{r,d}$ we conclude from (39), Lemma 4.1 and Lemma 2.1 that for any $f \in W_{p,w}^{r,d}$

$$\inf_{P_{\mathbf{k}} \in \mathbf{P_N}} \left\| f(\mathbf{x}) - \sum_{\mathbf{k} \leq \mathbf{N}} d_{\mathbf{k}} \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} \right\|_{p,w} \leq cN^{-(1-\frac{1}{\alpha})r} \leq cn^{-(1-\frac{1}{\alpha})\frac{r}{d}}.$$

Taking the supremum over $f \in W_{p,w}^{r,d}$ yields the desired result. □

Proof of Lemma 4.3:

$$\left\| \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x})} - \frac{A_h(\mathbf{k},\mathbf{x})}{B_h(\mathbf{x})}) \right\|_{p,w} \leq \underbrace{\left\| \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x}) B_h(\mathbf{x})} (\varphi(\mathbf{x}) - B_h(\mathbf{x})) \right\|_{p,w}}_{J_1} + \underbrace{\left\| \frac{1}{B_h(\mathbf{x})} (P_{\mathbf{k}}(\mathbf{x}) - A_h(\mathbf{k},\mathbf{x})) \right\|_{p,w}}_{J_2}.$$

We will now upper bound J_1 and J_2 separately. Clearly

$$J_1 \leq \|\varphi(\mathbf{x}) - B_h(\mathbf{x})\|_\infty \left\| \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x}) B_h(\mathbf{x})} \right\|_{p,w}. \tag{40}$$

In order to proceed we derive an upper bound on $\|\varphi(\mathbf{x}) - B_h(\mathbf{x})\|_\infty$. Let $\rho = \hat{s}^{\frac{1}{d}} h$ and recall that $\phi(t) = (1 + e^{-t})^{-1}$. We have

$$\begin{aligned}
& \|\varphi(\mathbf{x}) - B_h(\mathbf{x})\|_\infty \\
\overset{(a)}{=} & \left\| (2\rho)^{-d} \int_{[-\rho,\rho]^d} \phi(\mathbf{w} \cdot \mathbf{x}) d\mathbf{w} - (\hat{s})^{-1} \sum_{0 \leq \mathbf{r} \leq \mathbf{s}} \phi[h(2\mathbf{r} - \mathbf{s}) \cdot \mathbf{x}] \right\|_\infty \\
\overset{(b)}{=} & \left\| \hat{s}^{-1} (2h)^{-d} \sum_{0 \leq \mathbf{r} \leq \mathbf{s}} \int_{[0,2h]^d} \phi([(\mathbf{w} + h(2\mathbf{r} - \mathbf{s})] \cdot \mathbf{x}) d\mathbf{w} - (\hat{s})^{-1} \sum_{0 \leq \mathbf{r} \leq \mathbf{s}} \phi[h(2\mathbf{r} - \mathbf{s}) \cdot \mathbf{x}] \right\|_\infty \\
\overset{(c)}{=} & \hat{s}^{-1} (2h)^{-d} \left\| \sum_{0 \leq \mathbf{r} \leq \mathbf{s}} \int_{[0,2h]^d} \{\phi([(\mathbf{w} + h(2\mathbf{r} - \mathbf{s})] \cdot \mathbf{x}) - \phi[h(2\mathbf{r} - \mathbf{s}) \cdot \mathbf{x}]\} d\mathbf{w} \right\|_\infty \\
\overset{(d)}{\leq} & c_{d,\phi} h
\end{aligned} \tag{41}$$

In step (a) we simply plug in the definition of φ so that it appears explicitly in the expression. Step (b) consists of partitioning the integration region $[-\hat{s}^{1/d}h, \hat{s}^{1/d}h]^d$ into cells of size $[0, 2h]^d$. The number of these cells is equal to the cardinality of $\mathbf{s}$ (i.e., the number of terms in the summation). In step (c) we represent the second term as an integral over $\mathbf{w}$ in the region $[0, 2h]^d$, and utilize the linearity of the integration

operator. Step (d) follows from the mean value theorem, applied to the integrand (i.e., the difference of sigmoid functions). Formally we have

$$\|\phi([(\mathbf{w}+h(2\mathbf{r}-\mathbf{s})]^T\mathbf{x}) - \phi[h(2\mathbf{r}-\mathbf{s})^T\mathbf{x}]\| \leq c\|\nabla\phi\|_\infty\|\mathbf{w}\|_1 \leq c_{d,\phi}h,$$

where the second inequality follows from the definition of $\mathbf{w} \in [0,2h]^d$, thus $\|\mathbf{w}\|_1 \leq 2dh$. We have thus shown that

$$\|\varphi(\mathbf{x}) - B_h(\mathbf{x})\|_\infty \leq c_{d,\phi}h. \tag{42}$$

Moreover, from Lemma 4.1 using $\rho \leq 1$ we have

$$\varphi(\mathbf{x}) \geq \prod_{i=1}^d \frac{1}{1+e^{\rho|x_i|}} \geq \prod_{i=1}^d \frac{1}{1+e^{|x_i|}}.$$

In order to lower bound $B_h(\mathbf{x})$, we note that for h sufficiently small we have $hs_i \leq 1$, $i = 1,2,\ldots,d$, from which

$$\begin{aligned}
B_h(\mathbf{x}) &= \hat{s}^{-1} \sum_{0\leq\mathbf{r}\leq\mathbf{s}} \phi(h(2\mathbf{r}-\mathbf{s})\cdot\mathbf{x}) \\
&\geq \hat{s}^{-1} \sum_{0\leq\mathbf{r}\leq\mathbf{s}} \frac{1}{1+\exp\left(h\sum_i |s_i x_i|\right)} \\
&= \frac{1}{1+\exp\left(h\sum_i |s_i x_i|\right)} \\
&\geq \prod_{i=1}^d \frac{1}{1+\exp\left(h|s_i x_i|\right)} \\
&\geq \prod_{i=1}^d \frac{1}{1+e^{|x_i|}}.
\end{aligned} \tag{43}$$

Hence we conclude that

$$\left\| \frac{P_\mathbf{k}(\mathbf{x})}{\varphi(\mathbf{x})B_h(\mathbf{x})} \right\|_{p,w} \leq \left\{ \int_{\mathbb{R}^d} \left[|P_\mathbf{k}(\mathbf{x})| \prod_{i=1}^d e^{-|x_i|^\alpha} \left(1+e^{|x_i|}\right)^2 \right]^p d\mathbf{x} \right\}^{1/p}.$$

The polynomial $P_\mathbf{k}(\mathbf{x})$ can be expressed as

$$P_\mathbf{k}(\mathbf{x}) = \sum_{0\leq\mathbf{j}\leq\mathbf{k}} b_\mathbf{j}\mathbf{x}^\mathbf{j},$$

and thus we conclude from Minkowski's inequality ([7], p. 148) that

$$\begin{aligned}
&\left\{ \int_{\mathbb{R}^d} \left[|P_\mathbf{k}(\mathbf{x})| \prod_{i=1}^d e^{-|x_i|^\alpha} \left(1+e^{|x_i|}\right)^2 \right]^p d\mathbf{x} \right\}^{1/p} \\
\leq &\sum_{0\leq\mathbf{j}\leq\mathbf{k}} |b_\mathbf{j}| \left\{ \int_{\mathbb{R}^d} \left| \mathbf{x}^\mathbf{j} \prod_{i=1}^d e^{-|x_i|^\alpha} \left(1+e^{|x_i|}\right)^2 \right|^p d\mathbf{x} \right\}^{1/p} \leq \tilde{c},
\end{aligned}$$

where $\tilde{c}$ is a constant dependent only on $\mathbf{k}$, p, α and d. We thus have

$$\left\| \frac{P_{\mathbf{k}}(\mathbf{x})}{\varphi(\mathbf{x}) B_h(\mathbf{x})} \right\|_{p,w} \leq \tilde{c}. \tag{44}$$

From (40), (42) and (44) we thus establish that

$$J_1 \leq \tilde{c} c_{d,\phi} h. \tag{45}$$

In order to upper bound J_2, we first observe that $B_h(\mathbf{x})$ may be lower bounded as in (43). The proof then proceeds by upper bounding $\|P_{\mathbf{k}}(\mathbf{x}) - A_h(\mathbf{x}, \mathbf{k})\|_{p,w}$. This can be done along the lines of the proof of Theorem 3.1, yielding the result that for any $\epsilon > 0$

$$J_2 = \|P_{\mathbf{k}}(\mathbf{x}) - A_h(\mathbf{x}, \mathbf{k})\|_{p, \frac{w}{B_h}} \leq \epsilon.$$

Keeping in mind the arbitariness of ϵ, and combining the result with the upper bound on J_1 given in (45), the theorem follows. $\square$