On the near optimality of the stochastic approximation of smooth functions by neural networks [*]

V.E. Maiorov [a,**] and R. Meir [b,***]

[a] *Department of Mathematics, Technion, Haifa 32000, Israel*
[b] *Department of Electrical Engineering, Technion, Haifa 32000, Israel*

Received November 1997; revised November 1999
Communicated by T. Poggio

We consider the problem of approximating the Sobolev class of functions by neural networks with a single hidden layer, establishing both upper and lower bounds. The upper bound uses a probabilistic approach, based on the Radon and wavelet transforms, and yields similar rates to those derived recently under more restrictive conditions on the activation function. Moreover, the construction using the Radon and wavelet transforms seems very natural to the problem. Additionally, geometrical arguments are used to establish lower bounds for two types of commonly used activation functions. The results demonstrate the tightness of the bounds, up to a factor logarithmic in the number of nodes of the neural network.

Keywords: stochastic approximation, neural networks, upper bounds, lower bounds

1. Introduction

Investigations of function approximation of multi-variate real-valued functions by neural networks have, in recent years, assumed an important place in the general theory of non-linear function approximation (see, for example, [2,5,15], to name but a few). General theorems pertaining to the density of typical neural networks in various functional spaces can be found in [15,20,21].

Going beyond density issues, the various methods for obtaining upper bounds on approximation rates can be broken up into two broad classes. The first class, composed of stochastic methods, is based on the demonstration that some well-defined stochastic procedure yields an approximating structure, which *on the average* possesses some approximation rate. Standard arguments then lead to the conclusion that there

[*] This research was partially supported by a grant from the Israel Science Foundation. Support from the Ollendorff Center of the Department of Electrical Engineering at the Technion is also acknowledged.
[**] The author was partially supported by the Center for Absorption in Science, Ministry of Immigrant Absorption, State of Israel.
[***] Part of this work was done while the author was visiting the Isaac Newton Institute, Cambridge, England.

exists a certain function in the approximating class with the desired approximation rate. A second, more classical, approach is based on the construction of a specific algorithm for which the desired approximation rates can be established. We refer to this class of methods as deterministic.

Within the stochastic approach, Barron [2] considered the approximation of functions by neural networks, and obtained upper bounds on approximation rates for classes defined by the Fourier transform. More recently, Delyon et al. [9], developing the method of Barron [2], obtained an upper bound for approximation of functions from the Sobolev class $W_p^{r,d}$ by the linear combination of wavelet functions.

We observe that the stochastic methods alluded to above usually make use of some integral representation of the function being approximated, using a kernel depending on the structure of the particular approximation method used, followed by an application of the Monte Carlo method. Barron [2] makes use of the Fourier representation of functions, while Delyon et al. [9] utilize the multi-dimensional wavelet representation of functions.

In this paper we consider the neural network manifold

$$H_n(\varphi) = \left\{ h(x) = \sum_{k=1}^{n} c_k \varphi(a_k \cdot x + b_k)\colon\ a_k \in \mathbb{R}^d,\ c_k, b_k \in \mathbb{R},\ \forall k \right\}, \qquad (1)$$

where φ is some sigmoidal function on $\mathbb{R}$, and $a_k \cdot x$ is the inner product of two vectors a_k and x in $\mathbb{R}^d$. Note that $H_n(\varphi)$ has the following invariance property with respect to affine transformations, namely if $h(x) \in H_n(\varphi)$ then $h(Ax + t) \in H_n(\varphi)$ for any non-degenerate matrix A and vector $t \in \mathbb{R}^d$. Expressing the function to be approximated by an integral representation, combining both the Radon [13] and wavelet [8] transforms, we obtain an upper bound for the approximation error of functions from the Sobolev class by use of the manifold $H_n(\varphi)$, having order $n^{-r/d+\varepsilon}$, where ε is an arbitrary positive number. The techniques used for obtaining our estimates of the upper bound make use of the methods of Delyon et al. [9], as well as basic properties of the Radon and wavelet transforms.

Deterministic methods of approximation of functions by neural networks are considered in the works of Mhaskar and Micchelli [22], Chui et al. [4], and Mhaskar [20]. The latter work considered the approximation of the Sobolev class $W_p^{r,d}$ by the manifold $H_n(\varphi)$ for appropriate φ, and has given the best possible upper bound of order $n^{-r/d}$, but using functions φ from a rather restrictive class. DeVore et al. [7] obtained an upper bound on the rate of approximation of the Sobolev class $W_p^{r,d}$ for the special case of two-dimensional functions, $d = 2$, for a wide class of sigmoidal functions. Recently, Petrushev [24] has extended these results to the class $W_2^{r,d}$, $d > 2$.

The problem of deriving lower bounds on the approximation rates by neural networks has received considerably less emphasis. An important step, however, was taken by DeVore et al. [6], who showed that with the additional requirement of continuity of the approximation operator, a lower bound of order $n^{-r/d}$ can be obtained for approximating the Sobolev class $W_p^{r,d}$ by non-linear n-dimensional manifolds. More recently, Maiorov [16] and Maiorov and Pinkus [18] have obtained an asymptotically *tight* lower bound on the rate of approximation which is of order $n^{-r/(d-1)}$.

In this work we obtain a lower bound of order $(n \ln n)^{-r/d}$, without the additional continuity restriction. Here we consider sigmoidal functions of two types: piecewise polynomial functions and the standard function of the form $\varphi(x) = (1 + e^{-x})^{-1}$. These results make use of known results from the analysis of multi-variate algebraic polynomials ([30,31,14], see review in [28]).

Before presenting the main results of this work, we should stress that the general conclusions presented here have broad applicability beyond the field of function approximation. It has become clear in recent years that efficient and robust strategies exist, whereby multi-variate functions may be estimated using samples of the function values at a finite set of points (see, for example, [29]). It turns out that the error incurred by such schemes is composed of two parts, the first being the approximation error discussed above, and the second being a stochastic error resulting from the finiteness of the sample used to estimate an appropriate approximating function. Thus, any attempt to deliver useful performance bounds for function estimation requires as a first step the establishment of tight bounds on the approximation error, of the form discussed in this work and the other papers alluded to above.

The remainder of the paper is organized as follows. We begin in section 2 by expressing a general function as an integral representation, using properties of the Radon and wavelet transforms. This integral is then approximated in section 3 by a finite sum and upper bounds are derived on the approximation. Section 4 is devoted to the computation of a lower bound for the case of piecewise polynomial activation functions. Finally, we extend the lower bound results in section 5 to the widely studied case of the standard sigmoidal activation function.

2. An integral representation for neural networks

In this section we pursue the line of thought discussed in section 1, pertaining to a stochastic approach to approximation. These methods are based on two basic ingredients. First, the function of interest is expressed as a convex integral of the form

$$f(x) = \int \Phi(x, z) w(z) \, \mathrm{d}z,$$

where Φ and w depend on the function f, and the non-negative function $w(z)$ can be thought of as a probability density function for the "random variable" z, namely $\int w(z) \, \mathrm{d}z = 1$. The second step then corresponds to approximating the integral by a finite sum, as in the theory of Monte Carlo integration (see, for example, [2]). Standard results from the latter theory then yield upper bounds on the rates of convergence of the approximation to the exact value of the function. Following the work of Barron [2], various authors have recently utilized this basic idea in constructing neural network approximations. In this section we follow the line of work of Delyon et al. [9], who applied it to wavelet networks, rather than neural networks. The basic idea behind the integral representation constructed in this work is the use of both the Radon and the wavelet transformations. First we observe that neural networks, as in (1), are characterized by constant values on hyper-planes in $\mathbb{R}^d$. This observation leads naturally to the Radon transform, which represents general multi-variate functions by a superposition of integrals over hyper-planes in $\mathbb{R}^d$. This property then permits the transformation of the problem into a one-dimensional one, for which the wavelet integral representation is well known and understood. One advantage in using the Radon transformation in this fashion is that, in contrast to the work of Delyon et al. [9], only one-dimensional wavelets need to be considered as opposed to d-dimensional wavelets in [9]. We observe that the Radon transform has been used previously in work related to approximation by neural networks. In particular, Chen et al. [3] have provided an elegant proof of the density property of neural networks using this transform. In this work, however, we go beyond density and establish convergence rates as well.

Let K be a compact set in the space $\mathbb{R}^d$. Consider the space $L_p(K, \mathbb{R}^d)$, $1 \leqslant p \leqslant \infty$, of functions defined on $\mathbb{R}^d$ with support on the set K and norm

$$\|f\|_p = \left(\int_{\mathbb{R}^d} \left|f(x)\right|^p \, \mathrm{d}x \right)^{1/p}.$$

We denote the ball of radius r in $\mathbb{R}^d$ by $B^d(r) = \{x = (x_1, \ldots, x_d)\colon \sum_{i=1}^d x_i^2 \leqslant r^2\}$. In the sequel we mainly consider the unit ball $B^d(1)$, and will simplify the notation somewhat by using $B^d = B^d(1)$ and $L_p = L_p(B^d, \mathbb{R}^d)$. The results can be immediately extended to general compact domains K by use of standard extension theorems, as in [1].

For any two sets $W, H \subseteq L_p$ we define a distance measure of W from H by

$$\mathrm{dist}(W, H, L_p) = \sup_{f \in W} \mathrm{dist}(f, H, L_p), \tag{2}$$

where $\mathrm{dist}(f, H, L_p) = \inf_{h \in H} \|f - h\|_{L_p}$. Furthermore, for any function $f \in L_1$ we denote by $\mathcal{F}(f)$ or $\hat{f}$ the Fourier transform of f

$$\hat{f}(u) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) \mathrm{e}^{\mathrm{i} u \cdot x} \, \mathrm{d}x,$$

where $u \in \mathbb{R}^d$ and $u \cdot x$ is the inner product of u and x. The inverse Fourier transform will be denoted by $\mathcal{F}^{-1}$.

In the space L_1 define the derivative of order $\rho \geqslant 0$ as

$$\mathcal{D}^\rho f = (-\Delta)^{\rho/2} f = \mathcal{F}^{-1}\left\{|u|^\rho \mathcal{F}(u)\right\}, \tag{3}$$

where $|u| = \sqrt{u_1^2 + \cdots + u_d^2}$, and the Fourier transform and derivatives are in the distributional sense. In the space L_p consider the Sobolev class of functions

$$W_p^{r,d} = \left\{ f\colon \max_{\rho \leqslant r} \left\|\mathcal{D}^\rho f\right\|_p \leqslant 1 \right\}.$$

As stressed above, a basic tool in the construction of this section is the wavelet transform (see, for example, [12]). We follow here the definitions and notation of Delyon et al. [9].

2.1. The wavelet transform

We introduce the class Φ_q^r of functions φ defining the neural network class (1), where $r > 0$, $1 < q < \infty$. Let φ be some function in the space $L_1(\mathbb{R})$. Using the function φ, construct a second function ψ on $\mathbb{R}$ satisfying the equality

$$\int_0^\infty a^{-1}\hat{\varphi}(aw)\overline{\hat{\psi}}(aw)\,\mathrm{d}a = 1, \tag{4}$$

for any w, where $\hat{\varphi}$ and $\hat{\psi}$ are the Fourier transforms of φ and ψ, respectively. The function class Φ_q^r consists of all functions $\varphi \in L_q(\mathbb{R}) \cap L_1(\mathbb{R})$ for which there exists a function ψ satisfying (4) and such that for all $\rho \in [0, r]$, $\mathcal{D}^\rho\varphi \in L_q(\mathbb{R})$ and $\mathcal{D}^{-\rho}\psi \in L_1(\mathbb{R})$. Observe that the functions φ are uni-variate, and the condition is imposed in $\mathbb{R}$ rather than $\mathbb{R}^d$. Further, set

$$M_\varphi = \max_{0\leqslant\rho\leqslant r}\left\{\left\|D^\rho\varphi\right\|_{L_q}, \left\|D^{-\rho}\psi\right\|_{L_1}\right\}. \tag{5}$$

We observe that in many neural network applications it is customary to use sigmoidal functions which approach a constant non-zero value at infinity. However, by taking suitable linear combinations of such functions, one can always obtain functions vanishing at infinity, which belong to $L_1(\mathbb{R})$. For example, for sigmoidal functions $\sigma(t)$, $t \in \mathbb{R}$, if $\lim_{t\to-\infty}\sigma(t) = 0$, $\lim_{t\to\infty}\sigma(t) = 1$ and $\sigma(t)$ is non-decreasing, then we require $\sigma(t+1) - \sigma(t) = \varphi(t) \in \Phi_q^r$.

Examples. One can easily verify that the functions

$$\varphi(t) = \sqrt{2}\mathrm{e}^{-t^2/2}, \quad \frac{1}{\sqrt{2}}\left(1-t^2\right)\mathrm{e}^{-t^2/2}, \quad \frac{(t+1)}{3}\chi_{[-1,0]}(t) + \frac{(1-t)}{3}\chi_{[0,1]}(t),$$

where χ_Δ is the characteristic function of the segment Δ, belong to the class Φ_q^r. The corresponding functions ψ are, respectively

$$\psi(t) = \sqrt{2}\left(1-t^2\right)\mathrm{e}^{-t^2/2}, \quad \frac{1}{\sqrt{2}}\left(1-t^2\right)\mathrm{e}^{-t^2/2}, \quad -\chi_{[-1,0]}(t) + \chi_{[0,1]}(t).$$

Using any legitimate pair of functions φ and ψ as above, and following [8], we construct the wavelet integral representation of functions on $\mathbb{R}$. Let $g(t)$ be any function from the space $L_1(\mathbb{R})$. Then

$$g(t) = \int_{\mathbb{R}^+\times\mathbb{R}} u(a,\tau)\varphi\bigl(a(t-\tau)\bigr)a\,\mathrm{d}a\,\mathrm{d}\tau, \tag{6}$$

where $a \in \mathbb{R}^+$, $\tau \in \mathbb{R}$, and the function u is defined as

$$u(a,\tau) = \int_{\mathbb{R}} g(t)\psi\bigl(a(t-\tau)\bigr)\,\mathrm{d}t. \tag{7}$$

In order to utilize this one-dimensional integral representation for neural networks, we first use the Radon transform to express a general multi-variate function in terms of projections onto hyper-planes.

2.2. The Radon transform

Let $S^{d-1} = \{x \in \mathbb{R}^d\colon \sum_{i=1}^d x_i^2 = 1\}$ be the unit sphere in $\mathbb{R}^d$. For a given $\omega \in S^{d-1}$ and $t \in \mathbb{R}^+$ consider the hyper-plane in $\mathbb{R}^d$ given by $\Pi_{\omega,t} = \{x \in \mathbb{R}^d\colon\ x\cdot\omega = t\}$. For any $f \in W_1^{d-1,d}$ define the Radon transform $\mathcal{R}f(\omega,t)$, defined on $S^{d-1}\times\mathbb{R}$,

$$\mathcal{R}f(\omega,t) = \int_{\Pi_{\omega,t}} f(x)\,\mathrm{d}m(x), \tag{8}$$

where $\mathrm{d}m(x)$ is the Lebesgue measure on the hyper-plane $\Pi_{\omega,t}$. For $t < 0$ define $\mathcal{R}f(\omega,t) = \mathcal{R}f(-\omega,-t)$.

Using the Radon transform, construct a function $g(\omega,t)$ on $S^{d-1}\times\mathbb{R}$. For odd d

$$g(\omega,t) = \left(\frac{\partial}{\partial t}\right)^{d-1}\bigl(\mathcal{R}f(\omega,t)\bigr)$$

and for even d

$$g(\omega,t) = \mathcal{H}_t\left(\frac{\partial}{\partial t}\right)^{d-1}\bigl(\mathcal{R}f(\omega,t)\bigr),$$

where $(\mathcal{H}p)(t) = \frac{1}{\pi}\,\mathrm{p.v.}\int_{\mathbb{R}}\frac{p(\tau)}{t-\tau}\,\mathrm{d}\tau$ is the Hilbert transform of the function $p(t)$, and the integral is evaluated as the Cauchy principal value, i.e.,

$$\mathrm{p.v.}\int_{\mathbb{R}}\frac{p(\tau)}{t-\tau}\,\mathrm{d}\tau = \lim_{\delta\to+0}\int_{\tau\colon\ |t-\tau|>\delta}\frac{p(\tau)}{t-\tau}\,\mathrm{d}\tau.$$

The function $f \in W_1^{d-1,d}$ is then given by the so-called inverse Radon transform (see [13])

$$f(x) = \int_{S^{d-1}} g(\omega, x\cdot\omega)\,\mathrm{d}\omega, \tag{9}$$

where $d\omega$ is the Lebesgue normed measure on S^{d-1}.

Two important properties of the Radon transform (see [13] and [26]) are the following. For any function $f \in L_1$ the following connection between the Radon and Fourier transforms

$$\mathcal{R}f(\omega,t) = \int_{\Pi_{\omega,t}} f(x)\,\mathrm{d}m(x) = (2\pi)^{-1}\int_{\mathbb{R}}\hat{f}(s\omega)\mathrm{e}^{\mathrm{i}st}\,\mathrm{d}s \tag{10}$$

holds. Moreover, the Hilbert transform with respect to t satisfies the equation $\widehat{\mathcal{H}g}(\omega, s) = \mathrm{sgn}(s)\hat{g}(\omega, s)$. Therefore, directly from the definition of the function $g(\cdot)$ and (10) it follows that

$$g(\omega, t) = (2\pi)^{-1}\mathrm{i}^{d-1}\int_{\mathbb{R}} |s|^{d-1}\hat{f}(s\omega)\mathrm{e}^{\mathrm{i}st}\,\mathrm{d}s. \tag{11}$$

At this stage we have all the tools needed in order to construct an integral representation combining the Radon and wavelet transforms. This representation will be used in section 3 to compute an upper bound for the approximation error by neural networks.

2.3. *The Radon–wavelet integral representation*

Using (9), (5) and (6) we construct the neural network integral representation

$$f(x) = \int_{S^{d-1}}\int_{\mathbb{R}^+\times\mathbb{R}} u(\omega, a, \tau)\varphi\bigl(a(x\cdot\omega - \tau)\bigr)a\,\mathrm{d}a\,\mathrm{d}\tau\,\mathrm{d}\omega, \tag{12}$$

where

$$u(\omega, a, \tau) = \int_{\mathbb{R}} g(\omega, t)\psi\bigl(a(t-\tau)\bigr)\,\mathrm{d}t. \tag{13}$$

This expression may be simplified using the following notation. Let α and l be some positive numbers. Consider the set $\mathcal{Z} = S^{d-1}\times\mathbb{R}^+\times\mathbb{R}$, and denote elements of $\mathcal{Z}$ by $z = (\omega, a, \tau)$. Introduce a measure on $\mathcal{Z}$ by $\mathrm{d}z = \mathrm{d}\omega\,\mathrm{d}a\,\mathrm{d}\tau$. Then (12) may be re-written as

$$f(x) = Q\int_{\mathcal{Z}} F(x; z)\omega(z)\,\mathrm{d}z, \tag{14}$$

where we have used the notation

$$\begin{aligned}
F(x; z) \equiv F(x; z, \alpha, l) &= \frac{a^{1-\alpha}}{1+a^l}\varphi\bigl(a(x\cdot\omega-\tau)\bigr)\mathrm{sgn}\bigl(u(z)\bigr),\\
w(z) &= \frac{a^{\alpha}(1+a^l)|u(z)|}{Q(u,\alpha,l)},\\
Q \equiv Q(u,\alpha,l) &= \int a^{\alpha}\bigl(1+a^l\bigr)\bigl|u(z)\bigr|\,\mathrm{d}z.
\end{aligned} \tag{15}$$

From the definition of the function w it follows that $w(z) \geqslant 0$, $\forall z \in \mathcal{Z}$, and

$$\int_{\mathcal{Z}} w(z)\,\mathrm{d}z = 1. \tag{16}$$

Thus, $w(z)$ can be viewed as a probability density function on the set $\mathcal{Z}$. For any w-measurable function $G(z)$ we denote by

$$\mathrm{E}_w(G) = \int_{\mathcal{Z}} G(z)w(z)\,\mathrm{d}z \tag{17}$$

the average value of $G(z)$, where z is viewed as a random variable with probability density function $w(z)$. From (14) and (17) we have for all x

$$f(x) = Q\mathrm{E}_w\bigl(F(x,z)\bigr). \tag{18}$$

3. Upper bound

Starting from the integral representation (14) we construct an approximation to $f(x)$ by a finite sum, corresponding to a neural network. For this purpose use is made of the probabilistic structure inherent in the probability measure w. Fix a number n, and points $z_1,\dots,z_n \in \mathcal{Z}$, and let $\bar z = (z_1,\dots,z_n)$. Consider the function

$$f_n(x;\bar z) = \frac{Q}{n}\sum_{i=1}^{n} F(x;z_i), \tag{19}$$

as a function in the variable x.

Consider the direct product $\mathcal{Z}^n = \mathcal{Z}\times\cdots\times\mathcal{Z}$ of n copies of the set $\mathcal{Z}$, and define on $\mathcal{Z}^n$ the product measure $\bar w = w\times\cdots\times w$. For any function $h(\bar z)$ defined on $\mathcal{Z}^n$ we set

$$\mathrm{E}_{\bar w}(h) = \int_{\mathcal{Z}^n} h(\bar z)\bar w(\bar z)\,\mathrm{d}\bar z = \int_{\mathcal{Z}^n} h(z_1,\dots,z_n)w(z_1)\cdots w(z_n)\,\mathrm{d}z_1\cdots\mathrm{d}z_n.$$

Let $1<q<\infty$ be a fixed number. Below we will estimate the average value

$$\mathrm{E}_{\bar w}\bigl(\bigl\|f(x)-f_n(x;\bar z)\bigr\|_q^q\bigr) = \int_{B^d}\mathrm{E}_{\bar w}\bigl(\bigl|f(x)-f_n(x,\bar z)\bigr|^q\bigr)\,\mathrm{d}x. \tag{20}$$

From (14) and (19) we have

$$f(x)-f_n(x,\bar z) = \frac{Q}{n}\sum_{i=1}^{n}\bigl[\mathrm{E}_w\bigl(F(x;z)\bigr)-F(x;z_i)\bigr]. \tag{21}$$

The Burkholder inequality (see, for example, [11]) guarantees the existence of a constant c_0 depending only on q, such that for any $1<q<\infty$

$$\mathrm{E}_{\bar w}\left|\sum_{i=1}^{n}\bigl[\mathrm{E}_w\bigl(F(x;z)\bigr)-F(x;z_i)\bigr]\right|^q \leqslant c_0^q\mathrm{E}_{\bar w}\left[\left(\sum_{i=1}^{n}\bigl|\mathrm{E}_w\bigl(F(x;z)\bigr)-F(x;z_i)\bigr|^2\right)^{q/2}\right]. \tag{22}$$

From (20)–(22) using Hölder's inequality

$$\left(\sum_{i=1}^{n}|a_i|^2\right)^{1/2} \leqslant n^{(1/2-1/q)_+}\left(\sum_{i=1}^{n}|a_i|^q\right)^{1/q},$$

where $(\theta)_+ = \theta$ for $\theta \geqslant 0$, and zero otherwise, we obtain

$$\mathrm{E}_{\bar{w}}\big(\big\|f(x) - f_n(x;\bar{z})\big\|_q^q\big) \leqslant \left(\frac{c_0 Q n^{(1/2-1/q)_+}}{n}\right)^q \times \int_{B^d} \mathrm{E}_{\bar{w}}\left(\sum_{i=1}^n \big|\mathrm{E}_w F(x;z) - F(x;z_i)\big|^q\right) \mathrm{d}x$$

and hence

$$\mathrm{E}_{\bar{w}}\big(\big\|f(x) - f_n(x;\bar{z})\big\|_q^q\big) \leqslant \left(\frac{2c_0 Q}{n^\gamma}\right)^q \int_{B^d} \mathrm{E}_w\big|F(x,z)\big|^q \,\mathrm{d}x, \tag{23}$$

where $\gamma = \min\{1 - \frac{1}{q}, \frac{1}{2}\}$. We now estimate the integral.

Set $\alpha = 1 - 1/q$. Then we have

Lemma 1. For any $z \in \mathcal{Z}$ the inequality

$$\int_{B^d} \mathrm{E}_w\big|F(x,z)\big|^q \,\mathrm{d}x \leqslant v_{d-1}\|\varphi\|_{L_q(\mathbb{R})}^q$$

holds, where $v_{d-1} = \frac{2^{d-1}\Gamma(\frac{3}{2})^{d-1}}{\Gamma(\frac{d-1}{2}+1)}$ is the $(d-1)$-dimensional volume of the unit ball B^{d-1}.

Proof. We have from definition (15) of the function F

$$\begin{aligned}
\int_{B^d} \mathrm{E}_w\big(\big|F(x;z)\big|^q\big)\,\mathrm{d}x &= \int_{B^d\times\mathcal{Z}} \big|F(x;z)\big|^q w(z)\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{B^d\times S^{d-1}\times\mathbb{R}^+\times\mathbb{R}} \frac{a^{q(1-\alpha)}}{(1+a^l)^q}\big|\varphi(a(x\cdot\omega-\tau))\big|^q w(\omega,a,\tau)\,\mathrm{d}x\,\mathrm{d}\omega\,\mathrm{d}a\,\mathrm{d}\tau \\
&\leqslant \int_{B^d\times S^{d-1}\times\mathbb{R}^+\times\mathbb{R}} a\big|\varphi\big(a(x\cdot\omega-\tau)\big)\big|^q w(\omega,a,\tau)\,\mathrm{d}x\,\mathrm{d}\omega\,\mathrm{d}a\,\mathrm{d}\tau \\
&= \int_{S^{d-1}\times\mathbb{R}^+\times\mathbb{R}} w(\omega,a,\tau)\,\mathrm{d}\omega\,\mathrm{d}a\,\mathrm{d}\tau \int_{B^d} a\big|\varphi\big(a(x\cdot\omega-\tau)\big)\big|^q\,\mathrm{d}x.
\end{aligned} \tag{24}$$

Since for any $x \in \Pi_{\omega,s}$, $x\cdot\omega = s$, we have for any fixed ω, a and τ

$$\begin{aligned}
&\int_{B^d} a\big|\varphi\big(a(x\cdot\omega-\tau)\big)\big|^q\,\mathrm{d}x = \int_{-1}^{1}\mathrm{d}s\int_{\Pi_{\omega,s}\cap B^d} a\big|\varphi\big(a(x\cdot\omega-\tau)\big)\big|^q\,\mathrm{d}m(x) \\
&\quad= \int_{-1}^{1} a\big|\varphi\big(a(s-\tau)\big)\big|^q m\big(\Pi_{\omega,s}\cap B^d\big)\,\mathrm{d}s \leqslant v_{d-1}\int_{\mathbb{R}}\big|\varphi\big(a(s-\tau)\big)\big|^q\,\mathrm{d}(as) \\
&\quad= v_{d-1}\|\varphi\|_{L_q(\mathbb{R})}^q.
\end{aligned}$$

Hence, from (24) using (16) we obtain

$$\int_{B^d} \mathrm{E}_w \big|F(x;z)\big|^q \,\mathrm{d}q \leqslant v_{d-1}\|\varphi\|^q_{L_q(\mathbb{R})} \int_{S^{d-1}\times\mathbb{R}^+\times\mathbb{R}} w(\omega,a,\tau)\,\mathrm{d}\omega\,\mathrm{d}a\,\mathrm{d}\tau = v_{d-1}\|\varphi\|^q_{L_q(\mathbb{R})},$$

establishing the lemma. □

From inequality (23) and lemma 1 we have

$$\mathrm{E}_{\bar{w}}\big(\big\|f(x)-f_n(x;\bar{z})\big\|_q^q\big) \leqslant v_{d-1}\left(\frac{2c_0Q}{n^\gamma}\right)^q \|\varphi\|^q_{L_q(\mathbb{R})},$$

which, upon using Hölder's inequality, yields

$$\mathrm{E}_{\bar{w}}\big(\big\|f(x)-f_n(x;\bar{z})\big\|_q\big) \leqslant \big(\mathrm{E}_{\bar{w}}\big\|f(x)-f_n(x;\bar{z})\big\|_q^q\big)^{1/q} \leqslant v_{d-1}^{1/q}\frac{2c_0Q}{n^\gamma}\|\varphi\|_{L_q(\mathbb{R})}.$$

We therefore obtain the following statement.

Lemma 2. Let $\varphi \in \Phi_q^r$ where $1 < q < \infty$ and $r \geqslant 0$. Then

$$\mathrm{E}_{\bar{w}}\big(\big\|f(x)-f_n(x;\bar{z})\big\|_q\big) \leqslant \frac{c_1Q}{n^\gamma}\|\varphi\|_{L_q(\mathbb{R})},$$

where $c_1 = 2v_{d-1}^{1/q}c_0$.

A similar approximation result may be established for the derivative of the function f.

Lemma 3. For any $\varphi \in \Phi_q^r$, $1 < q < \infty$ and any $0 \leqslant l \leqslant r$,

$$\mathrm{E}_{\bar{w}}\big(\big\|\mathcal{D}_x^l f(x) - \mathcal{D}_x^l f_n(x;\bar{z})\big\|_q\big) \leqslant c_1\frac{Q}{n^\gamma}\big\|\mathcal{D}_t^l\varphi(t)\big\|_{L_q(\mathbb{R})}.$$

Proof. By analogy with (23) we have the estimate

$$\mathrm{E}_{\bar{w}}\big(\big\|\mathcal{D}_x^l f(x) - \mathcal{D}_x^l f_n(x;\bar{z})\big\|_q^q\big) \leqslant \left(\frac{c_0Q}{n^\gamma}\right)^q \int_{B^d} \mathrm{E}_{\bar{w}}\big(\big|\mathcal{D}_x^l F(x,\bar{z})\big|^q\big)\,\mathrm{d}x. \tag{25}$$

We estimate the last integral. From definition (15) of the function F and the relation $\alpha = 1 - 1/q$ we conclude that

$$\begin{aligned}\int_{B^d} \mathrm{E}_{\bar{w}}\big(\big|\mathcal{D}_x^l F(x,\bar{z})\big|^q\big)\,\mathrm{d}x &= \int_{B^d\times\mathcal{Z}} \frac{a^l}{(1+a^l)} a\big|\mathcal{D}_x^l\varphi\big(a(x\cdot\omega-\tau)\big)\big|^q w(z)\,\mathrm{d}z\,\mathrm{d}x \\ &\leqslant \int_{B^d\times\mathcal{Z}} \big|\mathcal{D}_x^l\varphi\big(a(x\cdot\omega-\tau)\big)\big|^q a\,\mathrm{d}x\,\mathrm{d}z.\end{aligned}$$

From definition (3) of the operator $\mathcal{D}^l$ it follows that $\mathcal{D}^l_x\varphi(a(x\cdot\omega-\tau)) = \mathcal{D}^l_t\varphi(a(t-\tau))|_{t=x\cdot\omega}$. Therefore

$$\begin{aligned}\int_{B^d}\bigl|\mathcal{D}^l_x\varphi\bigl(a(x\cdot\omega-\tau)\bigr)\bigr|^q a\,\mathrm{d}x &= \int_{B^d}\bigl|\bigl(\mathcal{D}^l_t\varphi\bigr)\bigl(a(x\cdot\omega-\tau)\bigr)\bigr|^q a\,\mathrm{d}x\\ &= \int_{-1}^{1}\mathrm{d}s\int_{\Pi_{\omega,s}\cap B^d}\bigl|\bigl(\mathcal{D}^l_t\varphi\bigr)\bigl(a(x\cdot\omega-\tau)\bigr)\bigr|^q a\,\mathrm{d}m(x)\\ &= \int_{-1}^{1}\bigl|\bigl(\mathcal{D}^l_s\varphi\bigr)\bigl(a(s-\tau)\bigr)\bigr|^q am\bigl(\Pi_{\omega,s}\cap B^d\bigr)\,\mathrm{d}s \leqslant v_{d-1}\bigl\|\mathcal{D}^l_t\varphi(t)\bigr\|^q_{L_q(\mathbb{R})}.\end{aligned}$$

Combining these results using Hölder's inequality the claim in the lemma follows. □

In the next lemma we estimate the function

$$Q = Q(u,\alpha,l) = \int_{\mathcal{Z}} a^{\alpha}\bigl(1+a^l\bigr)\bigl|u(z)\bigr|\,\mathrm{d}z,$$

used to define the approximant $f_n(x,\bar{z})$ in (19). The function $u(z)=u(\omega,a,t)$ was defined in (13).

Lemma 4. If $r \geqslant \rho > 1-\frac{1}{q}+l$, $l\geqslant 0$, $1<q<\infty$, are any numbers, $\alpha = 1-\frac{1}{q}$, and $\varphi\in\Phi^r_q$ then

$$Q(u,\alpha,l) \leqslant c_2\int_{S^{d-1}\times\mathbb{R}}\bigl(\bigl|g(\omega,t)\bigr| + \bigl|\mathcal{D}^{\rho}_t g(\omega,t)\bigr|\bigr)\,\mathrm{d}\omega\,\mathrm{d}t,$$

where $c_2 = M_\varphi \max\bigl\{\frac{2}{1-(1/q)}, \frac{2}{l\rho-l-1+(1/q)}\bigr\}$.

Proof. Denote $\int = \int_{\mathbb{R}}$. We have from (13)

$$\int\bigl|u(\omega,a,\tau)\bigr|\,\mathrm{d}\tau = \int\left|\int g(\omega,t)\psi\bigl(a(t-\tau)\bigr)\,\mathrm{d}t\right|\mathrm{d}\tau. \tag{26}$$

Set $\psi_\rho = D^{-\rho}\psi$. From Plancherel's formula it follows that for fixed ω, a and τ

$$\begin{aligned}\int g(\omega,t)\psi\bigl(a(t-\tau)\bigr)\,\mathrm{d}t &= a^{-1}\int\hat{g}(\omega,\theta)\hat{\psi}(-\theta/a)\mathrm{e}^{\mathrm{i}\tau\theta}\,\mathrm{d}\theta\\ &= a^{-1}\int|\theta|^{\rho}\hat{g}(\omega,\theta)|\theta|^{-\rho}\hat{\psi}(-\theta/a)\mathrm{e}^{\mathrm{i}\tau\theta}\,\mathrm{d}\theta\\ &= a^{-1-\rho}\int|\theta|^{\rho}\hat{g}(\omega,\theta)\hat{\psi}_\rho(-\theta/a)\mathrm{e}^{\mathrm{i}\tau\theta}\,\mathrm{d}\theta.\end{aligned}$$

Therefore from (26), using the inequality $\|\widehat{UV}\|_1 \leqslant \|\hat{U}\|_1\|\hat{V}\|_1$ we have

$$\begin{aligned}\int\bigl|u(\omega,a,\tau)\bigr|\,\mathrm{d}\tau &\leqslant a^{-\rho-1}\bigl\|\mathcal{F}^{-1}\bigl\{|\theta|^{\rho}\hat{g}(\omega,\theta)\bigr\}\bigr\|_{L_1(\mathbb{R})}\bigl\|\mathcal{F}^{-1}\bigl\{\hat{\psi}_\rho(-\theta/a)\bigr\}\bigr\|_{L_1(\mathbb{R})}\\ &= a^{-\rho-1}\bigl\|\mathcal{D}^{\rho}_t g(\omega,t)\bigr\|_{L_1(\mathbb{R})}\|\psi_\rho\|_{L_1(\mathbb{R})}.\end{aligned} \tag{27}$$

Represent the integral in $Q(u,\alpha,l)$ as

$$\begin{aligned} Q(u,\alpha,l) &= \int_{\mathcal{Z}} a^{\alpha}\left(1+a^{l}\right)\left|u(z)\right| \mathrm{d}z \\ &= \int_{S^{d-1}\times\mathbb{R}} \left[\int_0^1 a^{\alpha}\left(1+a^{l}\right)\left|u(\omega,a,\tau)\right| \mathrm{d}a + \int_1^{\infty} a^{\alpha}\left(1+a^{l}\right)\left|u(\omega,a,\tau)\right| \mathrm{d}a\right] \mathrm{d}\omega\, \mathrm{d}\tau, \end{aligned} \tag{28}$$

and separately estimate each summand. From Fubini's Theorem and the inequality (27) for $\rho = 0$ we have

$$\begin{aligned} &\int_{S^{d-1}\times\mathbb{R}} \int_0^1 a^{\alpha}\left(1+a^{l}\right)\left|u(\omega,a,\tau)\right| \mathrm{d}a\, \mathrm{d}\omega\, \mathrm{d}\tau \\ &\quad= \int_{S^{d-1}} \mathrm{d}\omega \int_0^1 \mathrm{d}a\, a^{\alpha}\left(1+a^{l}\right) \int_{\mathbb{R}} \left|u(\omega,a,\tau)\right| \mathrm{d}\tau \\ &\quad\leqslant \|\psi\|_{L_1(\mathbb{R})} \int_{S^{d-1}} \left\|g(\omega,t)\right\|_{L_1(\mathbb{R})} \mathrm{d}\omega \int_0^1 a^{-1+\alpha}\left(1+a^{l}\right) \mathrm{d}a \\ &\quad\leqslant c_3 \|\psi\|_{L_1(\mathbb{R})} \int_{S^{d-1}} \left\|g(\omega,t)\right\|_{L_1(\mathbb{R})} \mathrm{d}\omega, \end{aligned} \tag{29}$$

where $c_3 = 2/(1-(1/q))$. The second summand in (28) is estimated using Fubini's Theorem once more and the inequality (27)

$$\begin{aligned} &\int_{S^{d-1}\times\mathbb{R}} \int_1^{\infty} a^{\alpha}\left(1+a^{l}\right)\left|u(\omega,a,\tau)\right| \mathrm{d}a\, \mathrm{d}\omega\, \mathrm{d}\tau \\ &\quad\leqslant \|\psi_{\rho}\|_{L_1(\mathbb{R})} \int_{S^{d-1}} \left\|\mathcal{D}_t^{\rho} g(\omega,t)\right\|_{L_1(\mathbb{R})} \mathrm{d}\omega \int_1^{\infty} a^{\alpha-\rho-1}\left(1+a^{l}\right) \mathrm{d}a \\ &\quad\leqslant c_4 \|\psi_{\rho}\|_{L_1(\mathbb{R})} \int_{S^{d-1}} \left\|\mathcal{D}_t^{\rho} g(\omega,t)\right\|_{L_1(\mathbb{R})} \mathrm{d}\omega, \end{aligned} \tag{30}$$

where $c_4 = 2(1+l-\rho-(1/q))^{-1}$. From (29) and (30) we obtain the statement of lemma 4. □

In the following lemma we obtain an estimate for the approximation of a function f and its derivatives by neural networks.

Lemma 5. Assume $f \in W_2^{r,d}$ and $\varphi \in \Phi_q^r$ where $r = d/2 + 1/2 - 1/q + l + \varepsilon$, with $l \geqslant 0$, $1 < q < \infty$, $\varepsilon > n^{-\delta}$, $n > (q/(q-1))^{1/\delta}$ and $0 < \delta < 1$. Set $\gamma = \min\{1-\frac{1}{q}, \frac{1}{2}\}$. Then

$$\mathrm{E}_{\bar{w}}\left(\left\|\mathcal{D}^l f(x) - \mathcal{D}^l f_n(x;\bar{z})\right\|_q\right) \leqslant \frac{c_6}{n^{\gamma-\delta}}\left(\left\|\mathcal{D}^{\frac{d-1}{2}} f(x)\right\|_2 + \left\|\mathcal{D}^{\frac{d}{2}+\frac{1}{2}-\frac{1}{q}+l+\varepsilon} f(x)\right\|_2\right),$$

where $\mathcal{D} = \mathcal{D}_x$, and $c_6 = \frac{c_1}{2\pi}\frac{2^d\Gamma(3/2)^d}{\Gamma(d/2+1)} M_{\varphi}$.

Proof. Set $\rho = 1 - \frac{1}{q} + l + \varepsilon$. From lemmas 3 and 4 we conclude that

$$\begin{aligned}\mathrm{E}_{\bar{w}}\bigl(\bigl\|\mathcal{D}^l f(x) - \mathcal{D}^l f_n(x;\bar{z})\bigr\|_q\bigr) &\leqslant \frac{c_1 Q}{n^\gamma}\bigl\|\mathcal{D}_t^l \varphi(t)\bigr\|_{L_q(\mathbb{R})}\\ &\leqslant \frac{c_1 c_2}{n^\gamma}\int_{S^{d-1}\times\mathbb{R}}\bigl(\bigl|g(\omega,t)\bigr| + \bigl|\mathcal{D}_t^\rho g(\omega,t)\bigr|\bigr)\,\mathrm{d}\omega\,\mathrm{d}t.\end{aligned}$$

According to the conditions of the lemma we have

$$c_2 = M_\varphi \max\left\{\frac{2}{1-\frac{1}{q}}, \frac{2}{\rho - l - 1 + \frac{1}{q}}\right\} = M_\varphi \max\left\{\frac{2}{1-\frac{1}{q}}, \frac{2}{\varepsilon}\right\} \leqslant 2M_\varphi n^\delta.$$

Therefore

$$\mathrm{E}_{\bar{w}}\bigl(\bigl\|\mathcal{D}^l f(x) - \mathcal{D}^l f_n(x;\bar{z})\bigr\|_q\bigr) \leqslant \frac{2c_1 M_\varphi}{n^{\gamma-\delta}}\int_{S^{d-1}\times\mathbb{R}}\bigl(\bigl|g(\omega,t)\bigr| + \bigl|\mathcal{D}_t^\rho g(\omega,t)\bigr|\bigr)\,\mathrm{d}\omega\,\mathrm{d}t. \quad (31)$$

In the appendix we show that

$$\left(\int_{S^{d-1}\times\mathbb{R}}\bigl|\mathcal{D}_t^\rho g(\omega,t)\bigr|\,\mathrm{d}\omega\,\mathrm{d}t\right)^2 \leqslant c_5\int_{S^{d-1}\times\mathbb{R}}\bigl|\mathcal{D}_t^\rho g(\omega,t)\bigr|^2\,\mathrm{d}\omega\,\mathrm{d}t + c\|f\|^2_{W_2^{r,d}}, \quad (32)$$

where $c_5 = \frac{2^d\Gamma(3/2)^d}{\Gamma(d/2+1)}$. From the identity (11) and the definition of the operator $\mathcal{D}_t^\rho$ we have

$$\begin{aligned}\mathcal{D}_t^\rho g(\omega,t) &= (2\pi)^{-1}\mathrm{i}^{d-1}\mathcal{D}_t^\rho\left(\int_{\mathbb{R}}|s|^{d-1}f(s\omega)\mathrm{e}^{\mathrm{i}st}\,\mathrm{d}s\right)\\ &= (2\pi)^{-1}\mathrm{i}^{d-1}\int_{\mathbb{R}}|s|^{\rho+d-1}\hat{f}(s\omega)\mathrm{e}^{\mathrm{i}st}\,\mathrm{d}s.\end{aligned}$$

Hence by the Plancherel formula we obtain

$$\begin{aligned}\int_{\mathbb{R}}\bigl\|\mathcal{D}_t^\rho g(\omega,t)\bigr\|_2^2\,\mathrm{d}t &= (2\pi)^{-2}\int_{\mathbb{R}}\mathrm{d}t\left|\int_{\mathbb{R}}|s|^{\rho+d-1}\hat{f}(s\omega)\mathrm{e}^{\mathrm{i}st}\,\mathrm{d}s\right|^2\\ &= (2\pi)^{-2}\int_{\mathbb{R}}\bigl(|s|^{\rho+d-1}\bigl|\hat{f}(s\omega)\bigr|\bigr)^2\,\mathrm{d}s. \quad (33)\end{aligned}$$

Passing to polar coordinates we have

$$\begin{aligned}&\int_{S^{d-1}}\int_{\mathbb{R}}\bigl(|s|^{\rho+d-1}\bigl|\hat{f}(s\omega)\bigr|\bigr)^2\,\mathrm{d}s\,\mathrm{d}\omega = \int_{S^{d-1}}\mathrm{d}\omega\int_{\mathbb{R}}|s|^{2\rho+d-1}\bigl|\hat{f}(s\omega)\bigr|^2|s|^{d-1}\,\mathrm{d}s\\ &\quad= \int_{\mathbb{R}^d}|x|^{2\rho+d-1}\bigl|\hat{f}(x)\bigr|^2\,\mathrm{d}x = \bigl\|\mathcal{D}^{(\rho+\frac{d-1}{2})}f\bigr\|_2^2. \quad (34)\end{aligned}$$

Hence from (32)–(34) and $r = \rho + (d-1)/2$ we conclude

$$\left(\int_{S^{d-1}\times\mathbb{R}}\bigl|\mathcal{D}_t^\rho g(\omega,t)\bigr|\,\mathrm{d}t\,\mathrm{d}\omega\right)^2 \leqslant 2(2\pi)^{-2}c_5\bigl\|\mathcal{D}^{(\rho+\frac{d-1}{2})}f\bigr\|_2,$$

where an extra factor of 2 appears by assuming that the second term on the r.h.s. of (32) is smaller than the first.

From this and (31) we finally obtain the statement of lemma 5. □

We are now ready to prove the main result of this section, establishing an upper bound for the approximation of the Sobolev class by neural networks. The results are proved for the case $q = 2$. A simple corollary will then yield results for all $1 < q < \infty$. We state the results separately for small and large values of the smoothness parameters of the given class.

Theorem 1 (Small smoothness). Let $f \in W_2^{r,d}$, $r = d/2 + \varepsilon$, $\varepsilon > n^{-\delta}$ for some $0 < \delta < 1$. Furthermore, assume that $\varphi \in \Phi_2^r$. Then for any integers $d \geqslant 1$ and $n \geqslant 2$, there exists a function $h \in H_n(\varphi)$ such that

$$\|f - h\|_2 \leqslant \frac{c}{n^{1/2-\delta}} \|f\|_{W_2^{r,d}},$$

where c is a constant depending only on d and φ.

Proof. From lemma 5, using $q = 2$, we obtain

$$\begin{aligned} \mathrm{E}_{\bar{w}}\big(\big\|f(x) - f_n(x;\bar{z})\big\|_2\big) &\leqslant \frac{c_6}{n^{1/2-\delta}}\big(\big\|\mathcal{D}^{\frac{d-1}{2}} f(x)\big\|_2 + \big\|\mathcal{D}^r f(x)\big\|_2\big) \\ &\leqslant \frac{c_7}{n^{1/2-\delta}} \|f\|_{W_2^{r,d}}, \end{aligned}$$

where $c_7 = 2c_6 \frac{c_1}{\pi} \frac{2^d \Gamma(3/2)^d}{\Gamma(d/2+1)} M_\varphi$. Note that from the above result the expected value obeys the required bound. Therefore, clearly there exists a set of values for the parameters (ω, a, τ) which yield the desired result. □

Remark 1. Note that the conditions imposed on the functions φ in theorem 2, as well as theorem 2 below, pertain to the dimension $d = 1$, although the approximation is done in $\mathbb{R}^d$. The conditions on f, though, are given in terms of d, and become increasingly restrictive as the dimension increases. A similar situation occurs in [10].

Theorem 2 (Large smoothness). Let $f \in W_2^{r,d}$, $\varphi \in \Phi_2^r$, $d \geqslant 1$, and set $0 < \delta < 1$ and $\varepsilon > n^{-\delta}$. Let $r = (kd/2) + k\varepsilon$, $k \in \mathbb{N}$. Then for any $n \geqslant 1$ there exists a function $h \in H_n(\varphi)$ such that

$$\|f - h\|_2 \leqslant \frac{c}{n^{r/d-\delta'}} \|f\|_{W_2^{r,d}},$$

where $c = (4c_6)^k k^{r/d}$, $c_6 = \frac{c_1}{2\pi} \frac{2^d \Gamma(3/2)^d}{\Gamma(d/2+1)} M_\varphi$ and $\delta' = \delta + \frac{k\varepsilon}{d}$.

Proof. The claim will be established using a variant of the iterative approach introduced in [9]. Define $\hat{\delta} = \delta/k$ and $\hat{\varepsilon} = n^{-\hat{\delta}}$. From lemma 5 with $q = 2$ it follows that for any function $f \in W_2^{r,d}$, and any numbers $l \geqslant 0$, the following inequalities hold:

$$\mathrm{E}_w\bigl(\bigl\|f(x) - f_n(x,\bar{z})\bigr\|_2\bigr) \leqslant \frac{c_6}{n^{1/2-\hat{\delta}}}\bigl(\bigl\|\mathcal{D}^{\frac{d-1}{2}} f(x)\bigr\|_2 + \bigl\|\mathcal{D}^{\frac{d}{2}+\hat{\varepsilon}} f(x)\bigr\|_2\bigr), \tag{35}$$

$$\mathrm{E}_w\bigl(\bigl\|\mathcal{D}^l(f(x) - f_n(x,\bar{z}))\bigr\|_2\bigr) \leqslant \frac{c_6}{n^{1/2-\hat{\delta}}}\bigl(\bigl\|\mathcal{D}^{\frac{d-1}{2}} f(x)\bigr\|_2 + \bigl\|\mathcal{D}^{\frac{d}{2}+l+\hat{\varepsilon}} f(x)\bigr\|_2\bigr). \tag{36}$$

Let $\{\bar{z}_i\}_{i=1}^k$ be k independent draws of the random variable z with corresponding densities $w_i = w$. Then from (35) we have

$$\begin{aligned}
&\mathrm{E}_{w_1}\mathrm{E}_{w_2}\cdots\mathrm{E}_{w_k}\bigl\|f(x) - f_n(x,\bar{z}_1) - \cdots - f_n(x,\bar{z}_k)\bigr\|_2 \\
&\quad\leqslant \frac{c_6}{n^{1/2-\hat{\delta}}}\bigl\{\mathrm{E}_{w_1}\cdots\mathrm{E}_{w_{k-1}}\bigl(\bigl\|\mathcal{D}^{\frac{d-1}{2}}\bigl(f(x) - f_n(x,\bar{z}_1) - \cdots - f_n(x,\bar{z}_{k-1})\bigr)\bigr\|_2\bigr) \\
&\qquad + \mathrm{E}_{w_1}\cdots\mathrm{E}_{w_{k-1}}\bigl(\mathcal{D}^{\frac{d}{2}+\hat{\varepsilon}}\bigl(f(x) - f_n(x,\bar{z}_1) - \cdots - f_n(x,\bar{z}_{k-1})\bigr)\bigr\|_2\bigr)\bigr\}.
\end{aligned}$$

Continuing this iterative inequality k times using (35) and (36) we obtain

$$\begin{aligned}
\mathrm{E} &\equiv \mathrm{E}_{w_1}\ldots\mathrm{E}_{w_k}\bigl(\bigl\|f(x) - f_n(x,\bar{z}_1) - \cdots - f_n(x,\bar{z}_k)\bigr\|_2\bigr) \\
&\leqslant \left(\frac{c_6}{n^{1/2-\hat{\delta}}}\right)^k \sum_{s=0}^{k}\bigl(\bigl\|\mathcal{D}^{\frac{d-1}{2}} f\bigr\|_2 + \bigl\|\mathcal{D}^{s(\frac{d}{2}+\hat{\varepsilon})} f\bigr\|_2\bigr).
\end{aligned}$$

Then for $k = (kd/2) + k\varepsilon$ we have

$$\mathrm{E} \leqslant \frac{2c_6^k k}{n^{k/2-\delta}} \max_{\rho \leqslant \frac{kd}{2}+k\varepsilon} \bigl\|\mathcal{D}^\rho f\bigr\|_2 \leqslant \frac{c}{n^{r/d-\delta}}\|f\|_{W_2^{r,d}},$$

which establishes the claim. □

The results of theorem 2 can easily be extended to general values of $1 < q < \infty$. The only modification to the proof will be the replacement of $n^{1/2}$ in (35), (36) and subsequent equations by n^γ in accordance with lemma 5, where $\gamma = \min(1 - 1/q, 1/2)$. Note that in this case, however, the condition on φ is $\varphi \in \Phi_q^r$. We summarize this observation in the following corollary.

Corollary 1. Let the conditions of theorem 2 hold with $\varphi \in \Phi_q^r$, $1 < q < \infty$. Then for any $n \geqslant 1$ there exists a function $h \in H_n(\varphi)$ such that

$$\|f - h\|_q \leqslant \frac{c}{n^{r/d-\delta}}\|f\|_{W_2^{r,d}},$$

where c depends on q.

Remark 2. Observe that theorems 1 and 2 apply only in situations where the degree of smoothness is large, namely $r > d/2$. In a recent paper [19], we have been able to extend these results to all values of r, using a somewhat different approach.

4. Lower bound – piecewise polynomial functions

In this section and the next we establish lower bounds on the approximation error by neural networks, focusing on two specific types of activation functions. In particular, we consider first piecewise polynomial activation functions, followed in section 5 by the study of the standard sigmoidal activation function.

The basic idea of the proof is based on transforming the problem into a finite-dimensional one by an appropriate linear transformation, as was done, for example, in [31]. Lower bounds for the latter problem can be more easily established and, due to the linearity of the transformation, serve as lower bounds for the original problem. We start with some notation. Denote the l_q^m-norm of vector $f = (f_1, ..., f_m) \in \mathbb{R}^m$ by $\|f\|_{l_q^m} = (\sum_{i=1}^m |f_i|^q)^{1/q}$, $q \geqslant 1$, and let the unit ball in the space l_q^m be denoted by $B_q^m = \{f \in \mathbb{R}^m\colon\ \|f\|_{l_q^m} \leqslant 1\}$. For any f define a vector $\operatorname{sgn} f = (\operatorname{sgn} f_1, ..., \operatorname{sgn} f_m)$, where $\operatorname{sgn} a = 1$ if $a > 0$, $\operatorname{sgn} a = -1$ if $a < 0$, and $\operatorname{sgn} a = 0$ if $a = 0$. For any set $G \subset \mathbb{R}^m$ let $\operatorname{sgn} G = \{\operatorname{sgn} f\colon\ f \in G\}$. Finally, if $A, B \subseteq l_q^m$, the distance of the set A from B is given by

$$\operatorname{dist}\bigl(A, B, l_q^m\bigr) = \sup_{a \in A} \operatorname{dist}\bigl(a, B, l_q^m\bigr),$$

where $\operatorname{dist}(a, B, l_q^m) = \inf_{b \in B} \|a - b\|_{l_q^m}$.

Let $\tilde{m}$ be any integer, and set $m = \tilde{m}^d$. Consider the cube $I^d = \bigl[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\bigr]^d$ contained in the unit ball B^d, and construct the uniform net

$$S_m = \left\{\frac{1}{\sqrt{d}}\left(\frac{2i_1}{\tilde{m}} - 1, \ldots, \frac{2i_d}{\tilde{m}} - 1\right)\colon\ 0 \leqslant i_1, \ldots, i_d \leqslant \tilde{m} - 1\right\}$$

consisting of the m points $\{\xi_1, \ldots, \xi_m\}$.

Introduce a manifold in $\mathbb{R}^m$

$$\bar{H}_{nm}(\varphi) = \bigl\{\bigl(h(\xi_1), \ldots, h(\xi_m)\bigr)\colon\ h \in H_n(\varphi)\bigr\},$$

which is the restriction of the manifold $H_n(\varphi)$ to S_m. Consider the finite set in $\mathbb{R}^m$

$$E = \bigl\{(\varepsilon_1, \ldots, \varepsilon_m)\colon\ \varepsilon_i \in \{-1, +1\}\ \forall\, i = 1, \ldots, m\bigr\}.$$

In the following theorem we establish a lower bound for the distance of E from $\bar{H}_{nm}(\varphi)$, specialized to the case where φ are piecewise polynomial functions.

Theorem 3. Let φ be a piecewise polynomial function on $\mathbb{R}$ with s breakpoints $\tau_1, \ldots, \tau_s$, such that on any interval $[\tau_i, \tau_{i+1})$, $i = 0, 1, \ldots, s$ (here $\tau_0 = -\infty$, and $\tau_s = +\infty$) the function φ is an algebraic polynomial of degree at most n. Let n and m be integers such that $m = [c(d+2)n \log_2 n]$. Then

$$\operatorname{dist}\bigl(E, \bar{H}_{nm}(\varphi), l_1^m\bigr) \geqslant cm,$$

where c depends only on d, s and r.

We first prove an auxiliary lemma.

Lemma 6. The cardinality of the set $\operatorname{sgn} \bar{H}_{nm}(\varphi)$ obeys, for any positive integers n and m,

$$\left|\operatorname{sgn} \bar{H}_{nm}(\varphi)\right| \leqslant \left(\frac{cm^2}{n}\right)^{(d+2)n},$$

where c depends only on d, s, and r.

Proof. The manifold $\bar{H}_{nm}(\varphi)$ is the set of all functions on S_m of the form

$$\xi \mapsto h(\xi; a, b, c) = \sum_{i=1}^{n} c_i \varphi(a_i \cdot \xi + b_i),$$

where $c_i, b_i \in \mathbb{R}$, $a_i \in \mathbb{R}^d$ for $1 \leqslant i \leqslant n$, and $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$, $c = (c_1, \ldots, c_n)$, and $a_i \cdot \xi$ is the inner product of the vectors a_i and ξ. Denote $\gamma = (a, b, c) \in \mathbb{R}^{(d+2)n}$ and $h(\xi; a, b, c) = h(\xi; \gamma)$.

The manifold $\bar{H}_{nm}(\varphi)$ can be expressed in terms of the variables γ,

$$\bar{H}_{nm}(\varphi) = \left\{ \left(h(\xi_1; \gamma), \ldots, h(\xi_m; \gamma)\right) \colon \gamma \in \mathbb{R}^{(d+2)n} \right\}.$$

The following statement is well known (see, for example, [31, theorem 3], or [28] for a more modern treatment).

Claim 1. If $p_1(\gamma), \ldots, p_M(\gamma)$ are algebraic polynomials of degree at most r in $N \leqslant M$ variables, $\gamma = (\gamma_1, \ldots, \gamma_N) \in \mathbb{R}^N$, then

$$\left|\left\{ \left(\operatorname{sgn} p_1(\gamma), \ldots, \operatorname{sgn} p_M(\gamma)\right) \colon \gamma \in \mathbb{R}^N \right\}\right| \leqslant \left(\frac{4\mathrm{e}Mr}{N}\right)^N.$$

Construct the partition of the space $\mathbb{R}^{(d+2)n}$ by hyper-planes of the form

$$\Gamma_{ijk} = \left\{ (a, b, c) \in \mathbb{R}^{(d+2)n} \colon a_i \cdot \xi_k - b_i = \tau_j \right\},$$

where $i = 1, \ldots, n$, $j = 1, \ldots, s$, and $k = 1, \ldots, m$. Set $M = nsm$, and $N = (d+2)n$.

Denote by $S_1, \ldots, S_P$ the connected components of the set $\mathbb{R}^{(d+2)n} \setminus \bigcup_{ijk} \Gamma_{ijk}$. Since all polynomials of the form $p(\gamma) = a_i \cdot \xi - b_i - \tau_j$ retain a single sign on any component S_l, then from claim 1 the following estimate for the number of components P follows:

$$P \leqslant \left(\frac{4\mathrm{e}M}{N}\right)^N = \left(\frac{4\mathrm{e}ms}{d+2}\right)^{(d+2)n}. \tag{37}$$

Using this result, we can now estimate the cardinality of the set $\operatorname{sgn} \bar{H}_{nm}(\varphi)$. From the construction of the hyper-planes $\{\Gamma_{ijk}\}$ it follows that the space $\mathbb{R}^{(d+2)n}$ can be divided into P polyhedral regions $S_1, \ldots, S_P$, and therefore we have

$$\begin{aligned}\bigl|\operatorname{sgn} \bar{H}_{nm}(\varphi)\bigr| &= \bigl|\bigl\{\bigl(\operatorname{sgn} h(\xi_1;\gamma), \ldots, \operatorname{sgn} h(\xi_m;\gamma)\bigr)\colon\ \gamma \in \mathbb{R}^{(d+2)n}\bigr\}\bigr| \\ &= \sum_{l=1}^{P} \bigl|\bigl\{\bigl(\operatorname{sgn} h(\xi_1;\gamma), \ldots, \operatorname{sgn} h(\xi_m;\gamma)\bigr)\colon\ \gamma \in S_l\bigr\}\bigr|. \qquad (38)\end{aligned}$$

Since for any l and i the function $\gamma \mapsto h(\xi_i;\gamma)$, $\gamma \in S_l$, is a polynomial of degree $\leqslant r$, then according to claim 1 we obtain

$$\bigl|\bigl\{\bigl(\operatorname{sgn} h(\xi_1;\gamma), \ldots, \operatorname{sgn} h(\xi_m;\gamma)\bigr)\colon\ \gamma \in S_l\bigr\}\bigr| \leqslant \left(\frac{4\mathrm{e}mr}{(d+2)n}\right)^{(d+2)n}. \qquad (39)$$

Hence from (38), (39) and (37) we have

$$\bigl|\operatorname{sgn} \bar{H}_{nm}(\varphi)\bigr| \leqslant \left(\frac{4\mathrm{e}ms}{d+2}\right)^{(d+2)n} \left(\frac{4\mathrm{e}mr}{(d+2)n}\right)^{(d+2)n} = \left(\frac{cm^2}{n}\right)^{(d+2)n},$$

where the constant c depends only on d, s, and r. □

Proof of theorem 3. Let a be an absolute constant satisfying the equation $2a^2 - 8a + 7 = 0$ (i.e., $a = 2 - 1/\sqrt{2}$) . Fix a vector $\gamma \in \mathbb{R}^{(d+2)n}$, and construct a subset in E

$$\mathrm{E}_\gamma = \left\{\varepsilon \in E\colon\ \sum_{i=1}^{m} \bigl|\varepsilon_i - \operatorname{sgn} h(\xi_i;\gamma)\bigr| \leqslant am\right\}.$$

The cardinality of the set E_γ can be estimated as follows:

$$\begin{aligned}|\mathrm{E}_\gamma| &= \left|\left\{\varepsilon \in E\colon\ \sum_{i=1}^{m}(\varepsilon_i + 1) \leqslant am\right\}\right| = \left|\left\{\varepsilon \in E\colon\ \sum_{i\colon\ \varepsilon_i=1} 1 \leqslant \frac{am}{2}\right\}\right| \\ &= \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{\lfloor \beta m\rfloor} \quad \left(\beta = \frac{a}{2}\right) \\ &\leqslant 2^m \mathrm{e}^{-2m(1/2 - \lfloor \beta m\rfloor/m)^2} \quad \text{(Chernoff bound)} \\ &\leqslant 2^m \mathrm{e}^{-2m(1/2-\beta)^2} \leqslant 2^{m(2-2a+a^2/2)} \leqslant 2^{m/4},\end{aligned}$$

where we have used $2a^2 - 8a + 7 = 0$.

Consider the set $E' = \bigcap_{\gamma \in \mathbb{R}^{(d+2)n}}(E \setminus \mathrm{E}_\gamma)$. From the definition of the set $\operatorname{sgn} \bar{H}_{nm}(\varphi)$ it follows that

$$\begin{aligned}|E'| &= \left|E \setminus \bigcup_{\gamma \in \mathbb{R}^{(d+2)n}} \mathrm{E}_\gamma\right| \\ &\geqslant 2^m - \bigl|\operatorname{sgn} \bar{H}_{nm}(\varphi)\bigr| \max_\gamma |\mathrm{E}_\gamma| \geqslant 2^m - \bigl|\operatorname{sgn} \bar{H}_{nm}(\varphi)\bigr| 2^{m/4}.\end{aligned}$$

Set $m = [(d+2)n\log_2 n]$. Then there exist constants $c_1, c_2 > 0$ such that $\frac{c_1 m}{\log_2 m} \leqslant n \leqslant \frac{c_2 m}{\log_2 m}$. Therefore it follows from lemma 6 that for some $0 < c < 1$

$$\bigl|\operatorname{sgn} \bar{H}_{nm}(\varphi)\bigr| \leqslant \left(\frac{cm^2}{n}\right)^{(d+2)n} \leqslant 2^{m/4}.$$

Thus $|E'| \geqslant 2^m - 2^{m/2} > 0$. Hence there exists in E' a vector $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)$ such that for all $\gamma \in \mathbb{R}^{(d+2)n}$

$$\sum_{i=1}^m \bigl|\varepsilon_i - h(\xi_i;\gamma)\bigr| \geqslant \frac{1}{2}\sum_{i=1}^m \bigl|\varepsilon_i - \operatorname{sgn} h(\xi_i;\gamma)\bigr| \geqslant \frac{am}{2}. \qquad \square$$

The following lemma yields a lower bound for the approximation error of the Sobolev space, with the help of the lower bound established for the finite-dimensional case in theorem 3.

Lemma 7. If $1 \leqslant p, q \leqslant \infty$ and $\frac{r}{d} > \left(\frac{1}{p} - \frac{1}{q}\right)_+$ then

$$\operatorname{dist}\bigl(W_p^{r,d}, H_n(\varphi), L_q\bigr) \geqslant \frac{c}{m^{\frac{r}{d}-\frac{1}{p}+\frac{1}{q}}} \operatorname{dist}\bigl(B_p^m, \bar{H}_{mn}(\varphi), l_q^m\bigr).$$

Proof. In order to derive a lower bound for $W_p^{r,d}$ it clearly suffices to find a lower bound holding for some set $F_m \subseteq W_p^{r,d}$, which will now be constructed. For $x \in \mathbb{R}^d$ let η be any function in $W_p^{r,d}$ which satisfies $\eta(x) = 1$ for $x \in \frac{1}{2}I^d$ and $\eta(x) = 0$ for $x \notin I^d$. The function in the remaining region may be computed by using spline functions.

Let m and $\tilde{m}$ be positive integers such that $m = \tilde{m}^d$. Consider the uniform grid of m points $S_m = \{\xi_i\}_{i=1}^m$, where $\xi_i \in I^d$. For each i define a function

$$\eta_i(x) = \eta\bigl(\tilde{m}(x - \xi_i)\bigr).$$

Consider the normed space l_p^m consisting of vectors $a = (a_1, \dots, a_m) \in \mathbb{R}^m$, and denote by B_p^m the unit ball in this space. Construct the functional subclass

$$F_m = \left\{ f_a(x) = \frac{1}{m^{r/d-1/p}} \sum_{i=1}^m a_i \eta_i(x)\colon\ \|a\|_{l_p^m} \leqslant 1 \right\}.$$

We first show that $F_m \subseteq W_p^{r,d}$. Let $\rho \in [0, r]$. Then for any $a \in B_p^m$

$$\begin{aligned}\|\mathcal{D}^\rho f_a\|_{L_p}^p &= \frac{1}{m^{rp/d-1}} \int_{I^d} \left|\sum_{i=1}^m a_i \mathcal{D}^\rho \eta_i(x)\right|^p \mathrm{d}x \\ &= \frac{1}{m^{rp/d-1}} \sum_{i=1}^m \int_{\frac{1}{\tilde{m}}I^d} \bigl|a_i \mathcal{D}^\rho \eta\bigl(\tilde{m}(x - \xi_i)\bigr)\bigr|^p \mathrm{d}x,\end{aligned}$$

where $\frac{1}{\tilde{m}}I^d$ is the cube I^d scaled down by a factor of $\tilde{m}$ for each side. Since

$$\mathcal{D}_x^\rho \eta\big(\tilde{m}(x-\xi_i)\big) = \tilde{m}^\rho \mathcal{D}_t^\rho \eta(t)\big|_{t=\tilde{m}(x-\xi_i)},$$

clearly

$$\big\|\mathcal{D}^\rho f_a\big\|_{L_p}^p = \frac{\tilde{m}^{\rho p-d}}{m^{rp/d-1}}\left(\sum_{i=1}^m |a_i|^p\right)\int_{I^d}\big|\mathcal{D}^\rho \eta(t)\big|^p\,\mathrm{d}t \leqslant \big\|\mathcal{D}^\rho \eta\big\|_{L_p}^p \leqslant 1.$$

We therefore conclude that $f_a \in W_p^{r,d}$.

We proceed with bounding $\operatorname{dist}(W_p^{r,d}, H_n(\varphi), L_q)$ from below. For all $f \in F_m$ and $h \in H_n(\varphi)$ we have

$$\begin{aligned}\|f-h\|_{L_q}^q &= \int_{I^d}\left|\frac{1}{m^{r/d-1/p}}\sum_{i=1}^m a_i\eta_i(x) - h(x)\right|^q \mathrm{d}x,\\ &= \sum_{i=1}^m \int_{\frac{1}{\tilde{m}}I^d}\left|\frac{1}{m^{r/d-1/p}}a_i\eta_i(y+\xi_i) - h(y+\xi_i)\right|^q \mathrm{d}y.\end{aligned}$$

Since $\eta_i(y+\xi_i) = \eta(\tilde{m}y)$, $i = 1,2,\ldots,m$, and $y \in \frac{1}{\tilde{m}}I^d$, we obtain

$$\begin{aligned}\operatorname{dist}\big(W_p^{r,d}, H_n(\varphi), L_q\big)^q &\geqslant \sup_{f\in F_m}\inf_{h\in H_n(\varphi)}\|f-h\|_{L_q}^q\\ &\geqslant \frac{1}{m^{(r/d-1/p)q}}\sup_{a\in B_p^m}\inf_{h\in H_n(\varphi)}\sum_{i=1}^m\int_{\frac{1}{\tilde{m}}I^d}\big|a_i\eta_i(y+\xi_i) - h(y+\xi_i)\big|^q\,\mathrm{d}y\\ &\geqslant \frac{1}{m^{(r/d-1/p)q}}\sup_{a\in B_p^m}\inf_{h\in H_n(\varphi)}\sum_{i=1}^m\frac{1}{\tilde{m}^d}\int_{I^d}\big|a_i\eta(y) - h(y/\tilde{m}+\xi_i)\big|^q\,\mathrm{d}y. \qquad (40)\end{aligned}$$

Note that due to the infimum taken over $H_n(\varphi)$ it has been possible to rescale $h(\cdot)$ by the factor $m^{r/d-1/p}$. Define $a_y = (a_1,\ldots,a_m)\eta(y)$ and $h_y = (h(y/\tilde{m}+\xi_1),\ldots,h(y/\tilde{m}+\xi_m))$. Then moving the summation into the integral in (40) we have

$$\begin{aligned}\operatorname{dist}\big(W_p^{r,d}, H_n(\varphi), L_q\big)^q &\geqslant \frac{1}{m^{(r/d-1/p+1/q)q}}\sup_{a\in B_p^m}\inf_{h\in H_n(\varphi)}\int_{I^d}\|a_y - h_y\|_{l_q^m}^q\,\mathrm{d}y\\ &\geqslant \frac{1}{m^{(r/d-1/p+1/q)q}}\sup_{a\in B_p^m}\int_{I^d}\inf_{h\in H_n(\varphi)}\|a_y - h_y\|_{l_q^m}^q\,\mathrm{d}y.\end{aligned}$$

Using the affine invariance of $H_n(\varphi)$ mentioned in section 1 we conclude that for any y

$$\inf_{h\in H_n(\varphi)}\|a_y - h_y\|_{l_p^m} = \operatorname{dist}\big(a, \bar{H}_{mn}(\varphi), l_q^m\big)\eta(y).$$

Thus

$$\operatorname{dist}\big(W_p^{r,d}, H_n(\varphi), L_q\big)^q \geqslant \frac{1}{m^{(r/d-1/p+1/q)q}}\sup_{a\in B_p^m}\operatorname{dist}\big(a, \bar{H}_{mn}(\varphi), l_q^m\big)\int_{I^d}\big|\eta(y)\big|^q\,\mathrm{d}y.$$

The theorem follows upon taking $c = (\int_{I^d} |\eta(y)|^q \,\mathrm{d}y)^{1/q}$. □

We can now prove the main theorem.

Theorem 4. Let φ satisfy the conditions of theorem 3. Then for any $1 \leqslant p, q \leqslant \infty$ and $\frac{r}{d} > (\frac{1}{p} - \frac{1}{q})_+$ the inequality

$$\operatorname{dist}\bigl(W_p^{r,d}, H_n(\varphi), L_q\bigr) \geqslant \frac{c}{(n \log n)^{r/d}}$$

holds, where c depends on p, q, d and r.

Proof. Let ε be the vector constructed in lemma 6. Consider a vector $a = (a_1, \ldots, a_m)$, where $a_i = m^{-1/p}\varepsilon_i$. Clearly $a \in B_p^m$. Set $m = [cn(d+2)\log n]$. Then using lemma 7 and theorem 3 we obtain

$$\operatorname{dist}\bigl(W_p^{r,d}, H_n(\varphi), L_q\bigr) \geqslant \frac{m^{-1/p}(c_3 m)^{1/q}}{m^{\frac{r}{d} - \frac{1}{p} + \frac{1}{q}}} \geqslant \frac{c}{(n \log n)^{r/d}}. \qquad \square$$

5. Lower bound – the standard sigmoid

We extend the results of section 4 to cover the case of the widely studied sigmoidal function $\sigma(t) = \frac{1}{1+\mathrm{e}^{-t}}$. We derive estimates for the distance of $W_p^{r,d}$ from the manifold $H_n(\sigma)$. The main result is the following theorem.

Theorem 5. For any $1 \leqslant p, q \leqslant \infty$ and $\frac{r}{d} > (\frac{1}{p} - \frac{1}{q})_+$ the inequality

$$\operatorname{dist}\bigl(W_p^{r,d}, H_n(\sigma), L_q\bigr) \geqslant \frac{c}{(n \log n)^{r/d}}$$

holds.

First we need an auxiliary lemma. In the space $\mathbb{R}^m$ consider the manifold $\bar{H}_{n,m}(\sigma)$ defined in section 4, that is,

$$\bar{H}_{nm}(\sigma) = \bigl\{\bigl(h(\xi_1), \ldots, h(\xi_m)\bigr)\colon\ h \in H_n(\sigma)\bigr\}.$$

We make use of the following result.

Claim 2. Let $\rho_1(\gamma) = \frac{p_1(\gamma)}{q_1(\gamma)}, \ldots, \rho_M(\gamma) = \frac{p_M(\gamma)}{q_M(\gamma)}$ be rational functions of degree at most r, i.e., p_i and q_i, $i = 1, 2, \ldots, M$, are algebraic polynomials of degree at most r in $N \leqslant M$ variables $\gamma = (\gamma_1, \ldots, \gamma_N) \in \mathbb{R}^N$. Then

$$\bigl|\bigl\{\bigl(\operatorname{sgn}\rho_1(\gamma), \ldots, \operatorname{sgn}\rho_M(\gamma)\bigr)\colon\ \gamma \in \mathbb{R}^N\bigr\}\bigr| \leqslant \left(\frac{4\mathrm{e}Mr}{N}\right)^{2N}.$$

Proof. Clearly

$$\begin{aligned}&\left|\left\{\left(\operatorname{sgn}\rho_1(\gamma),\ldots,\operatorname{sgn}\rho_M(\gamma)\right):\ \gamma\in\mathbb{R}^N\right\}\right|\\&=\left|\left\{\left(\frac{\operatorname{sgn}p_1(\gamma)}{\operatorname{sgn}q_1(\gamma)},\ldots,\frac{\operatorname{sgn}p_M(\gamma)}{\operatorname{sgn}q_M(\gamma)}\right):\ \gamma\in\mathbb{R}^N\right\}\right|\\&\leqslant\left|\left\{\left(\operatorname{sgn}p_1(\gamma),\ldots,\operatorname{sgn}p_M(\gamma)\right):\ \gamma\in\mathbb{R}^N\right\}\right|\left|\left\{\left(\operatorname{sgn}q_1(\gamma),\ldots,\operatorname{sgn}q_M(\gamma)\right):\ \gamma\in\mathbb{R}^N\right\}\right|.\end{aligned}$$

The result then follows upon using claim 1. □

Lemma 8. The cardinality of the set $\operatorname{sgn}\bar{H}_{nm}(\sigma)$ is upper bounded as follows:

$$\left|\operatorname{sgn}\bar{H}_{nm}(\sigma)\right|\leqslant(cm)^{(1+1/d)(d+2)n}.$$

Proof. For any function $h(\xi;\gamma)\in\bar{H}_{nm}(\sigma)$ we have

$$h(\xi_k;\gamma)=\sum_{i=1}^{n}\frac{c_i}{1+\mathrm{e}^{-a_i\cdot\xi_k+b_i}},\tag{41}$$

where $\gamma=(a,b,c)$, $a\in\mathbb{R}^{dn}$, $c,b\in\mathbb{R}^n$, and the vector $\xi=(\xi_1,\ldots,\xi_m)$ has coordinates $\xi_i=\frac{k_i}{\tilde{m}}$, $0\leqslant k_i\leqslant\tilde{m}-1$, and $\tilde{m}=m^{1/d}$. Denote in (41) $\mathrm{e}^{-a_{ij}/\tilde{m}}=t_{ij}$ and $\mathrm{e}^{-b_i}=\tau_i$. Then

$$h(\xi_k;\gamma)=\sum_{i=1}^{n}\frac{c_i}{1+t_{i1}^{k_1}\cdots t_{id}^{k_d}\cdot\tau_i}.$$

Therefore, for each fixed k, the function $\gamma\mapsto h(\xi_k;\gamma)$ is a rational function in $(d+2)n$ variables c_i,τ_i and $t_{i1},\ldots,t_{id}$, $1\leqslant i\leqslant n$, of degree at most $((\tilde{m}-1)d+1)(n-1)+1\leqslant 4dnm^{1/d}$. Hence using claim 2 we conclude that

$$\left|\operatorname{sgn}\bar{H}_{nm}(\sigma)\right|\leqslant\left(\frac{4\mathrm{e}\cdot m\cdot 4ndm^{1/d}}{(d+2)n}\right)^{2(d+2)n}\leqslant(cm)^{2(1+1/d)(d+2)n}.\qquad\square$$

Similarly to the proof of theorem 4, theorem 5 is established using lemmas 6, 7 and 8.

The following result, by Mhaskar [20], establishes an upper bound for the error of approximation of the space $W_p^{r,d}$ by $H_n(\sigma)$. Let $r\geqslant 1$, $n\geqslant 1$ be integers, and $1\leqslant p\leqslant\infty$. Then

$$\operatorname{dist}\left(W_p^{r,d},H_n(\sigma),L_p\right)\leqslant cn^{-r/d}.\tag{42}$$

Combining (42) and the lower bound of theorem 5 we have

Corollary 2. If $r\geqslant 1$ is an integer, and $1\leqslant p\leqslant\infty$ then for any $n=1,2,\ldots$

$$\frac{c_1}{(n\log n)^{r/d}}\leqslant\operatorname{dist}\left(W_p^{r,d},H_n(\sigma),L_p\right)\leqslant\frac{c_2}{n^{r/d}},$$

where c_1 and c_2 depend only on r, d and p.

Appendix

Proof of (32). We need to show that for any $\rho > 0$

$$\int_{S^{d-1}\times\mathbb{R}} \left|\mathcal{D}_t^\rho g(w,t)\right| \mathrm{d}\omega\,\mathrm{d}t \leqslant c\left(\int_{S^{d-1}\times\mathbb{R}} \left|\mathcal{D}_t^\rho g(w,t)\right|^2 \mathrm{d}\omega\,\mathrm{d}t\right)^{1/2} + c\|f\|_{W_2^r(K)}. \tag{43}$$

Note that Hölder's inequality *cannot* be used to show this, since the domain of the integral is infinite. Recall that the function f is supported over the unit ball $B^d(1) \subset \mathbb{R}^d$, namely $f \in W_2^{r,d}(B^d(1))$. We find it useful to extend the function to the Sobolev space of functions defined over the whole space $\mathbb{R}^d$. This can be done using standard results from the theory of function spaces, and is based on defining a function $\tilde{f}$ over $\mathbb{R}^d$ such that $\tilde{f} \in W_2^{r,d}(\mathbb{R}^d)$, $\tilde{f}(x) = 0$ for $x \notin B^d(2)$ and $\tilde{f}(x) = f(x)$ for $x \in B^d(1)$. Details of this procedure can be found in [27, section 4.2]. Let $J = [0,2]$ and fix $\omega \in S^{d-1}$. Then

$$\int_{\mathbb{R}^+} \left|\mathcal{D}_t^\rho g(w,t)\right| \mathrm{d}t = \left(\int_J + \int_{\mathbb{R}^+\setminus J}\right)\left|\mathcal{D}_t^\rho g(w,t)\right| \mathrm{d}t \equiv I_1 + I_2. \tag{44}$$

Applying Hölder's inequality to the first integral (defined over a compact domain) we have

$$\begin{aligned} I_1 &= \int_J \left|\mathcal{D}_t^\rho g(w,t)\right| \mathrm{d}t \leqslant |J|^{1/2}\left(\int_J \left|\mathcal{D}_t^\rho g(w,t)\right|^2 \mathrm{d}t\right)^{1/2} \\ &\leqslant c_8\left(\int_{\mathbb{R}^+} \left|\mathcal{D}_t^\rho g(w,t)\right|^2 \mathrm{d}t\right)^{1/2}. \end{aligned} \tag{45}$$

In order to estimate I_2 we use the definition of $g(\omega,t)$ from (11). Then we have

$$\begin{aligned} \mathcal{D}_t^\rho g(\omega,t) &= \frac{\mathrm{i}^{d-1}}{2\pi}\int_{\mathbb{R}} |s|^{\rho+d-1}\hat{f}(s\omega)\mathrm{e}^{\mathrm{i}st}\,\mathrm{d}s \\ &= \frac{\mathrm{i}^{d-1}}{2\pi}\int_{\mathbb{R}} |s|^{\rho+d-1}\left(\int_{\mathbb{R}^d} f(y)\mathrm{e}^{-\mathrm{i}sy\cdot\omega}\,\mathrm{d}y\right)\mathrm{e}^{\mathrm{i}st}\,\mathrm{d}s \\ &= \frac{\mathrm{i}^{d-1}}{2\pi}\int_{\mathbb{R}^d} f(y)\,\mathrm{d}y\int_{\mathbb{R}} |s|^{\rho+d-1}\mathrm{e}^{\mathrm{i}s(t-y\cdot\omega)}\,\mathrm{d}s. \end{aligned}$$

The Fourier transform implicit in the second integral should be understood in the sense of a Fourier transform of a generalized function (note that we only require the value of this quantity integrated over τ). From [23] we obtain for $\lambda \neq -1, -3, \ldots$

$$\mathcal{F}\left(|s|^\lambda\right)(\tau) = C(\lambda)|\tau|^{-\lambda-1},$$

where $C(\lambda) = -2\sin(\frac{\lambda\pi}{2})\Gamma(\lambda+1)$. Recall that the Fourier transform $\hat{f}$ of the generalized function F is defined by the relationship $(\hat{f},\tilde{\varphi}) = (F,\varphi)$ which holds for any compactly supported function $\varphi \in L_1(\mathbb{R})$.

Hence we obtain

$$\int_{\mathbb{R}^d} f(y)\,\mathrm{d}y \int_{\mathbb{R}} |s|^{\rho+d-1} \mathrm{e}^{\mathrm{i}s(t-y\cdot\omega)}\,\mathrm{d}s = C(\rho+d-1) \int_{\mathbb{R}^d} f(y)|t-y\cdot\omega|^{-\rho-d}\,\mathrm{d}y.$$

Set $A = C(\rho+d-1)/(2\pi)$. Then, since $\mathrm{supp}(f) \subseteq B_2^d(2)$,

$$I_2 = \int_{\mathbb{R}^+\setminus J} \left|D_t^\rho g(\omega,t)\right| \mathrm{d}t \leqslant A \int_{\mathbb{R}^+\setminus J} \left| \int_{\mathbb{R}^d} f(y)|t-y\cdot\omega|^{-\rho-d}\,\mathrm{d}y \right| \mathrm{d}t$$
$$= A \int_2^\infty \int_{|y|\leqslant 1} \left|f(y)\right| |t-y\cdot\omega|^{-\rho-d}\,\mathrm{d}y\,\mathrm{d}t = A \int_{|y|\leqslant 1} \left|f(y)\right| \mathrm{d}y \int_2^\infty \frac{\mathrm{d}t}{|t-y\cdot\omega|^{\rho+d}}.$$

Since $|2-y\cdot\omega| \geqslant 2-|y|^{1/2}|\omega|^{1/2} \geqslant 1$,

$$\int_2^\infty \frac{\mathrm{d}t}{|t-y\cdot\omega|^{\rho+d}} = \frac{\rho+d-1}{(2-y\cdot\omega)^{\rho+d-1}} \leqslant \rho+d-1.$$

Combining this result with Hölder's inequality we obtain

$$I_2 \leqslant c_9\|f\|_{L_2(K)}, \quad \text{where } c_9 = \frac{|B^d(1)|(\rho+d-1)C(\rho+d-1)}{2\pi}. \tag{46}$$

From (44), (45) and (46) we obtain

$$\int_{\mathbb{R}^+} \left|\mathcal{D}_t^\rho g(\omega,t)\right| \mathrm{d}t \leqslant c_8 \left(\int_{\mathbb{R}^+} \left|\mathcal{D}_t^\rho g(w,t)\right|^2 \mathrm{d}t \right)^{1/2} + c_9\|f\|_{L_2(K)}. \tag{47}$$

Integrating (47) over the unit sphere and using Hölder's inequality we then obtain

$$\int_{S^{d-1}\times\mathbb{R}} \left|\mathcal{D}_t^\rho g(\omega,t)\right| \mathrm{d}\omega\,\mathrm{d}t$$
$$\leqslant c_8 \int_{S^{d-1}} \left(\int_{\mathbb{R}^+} \left|\mathcal{D}^\rho g(\omega,t)\right|^2 dt \right)^{1/2} \mathrm{d}\omega + 2c_9\left|S^{d-1}\right| \|f\|_{W_2^r(K)}$$
$$\leqslant c_8\left|S^{d-1}\right|^{1/2} \left(\int_{S^{d-1}\times\mathbb{R}^+} \left|\mathcal{D}^\rho g(\omega,t)\right|^2 \mathrm{d}t\,\mathrm{d}\omega \right)^{1/2} + 2c_9\left|S^{d-1}\right| \|f\|_{W_2^r(K)},$$

which establishes (32) with $c = |S^{d-1}|^{1/2}\max(c_8, 2|S^{d-1}|^{1/2}c_9)$. □

Acknowledgements

We are grateful to Prof. Allan Pinkus for helpful discussions.

References

[1] R.A. Adams, *Sobolev Spaces* (Academic Press, New York, 1975).

[2] A.R. Barron, Universal approximation bounds for superposition of sigmoidal function, IEEE Trans. Inform. Theory 39 (1993) 930–945.

[3] T. Chen, H. Chen and R. Liu, Approximation capability in $C(\bar{R}_n)$ by multilayer feedforward networks and related problems, IEEE Trans. Neural Networks 6(1) (1995) 25–30.

[4] C.K. Chui, X. Li and H.N. Mhaskar, Some limitations of neural networks with one hidden layer, Adv. Comput. Math. 5 (1996) 233–244.

[5] G. Cybenko, Approximation by superposition of sigmoidal functions, Math. Control Signals Systems 2 (1989) 303–314.

[6] R.A. DeVore, R. Howard and C.A. Micchelli, Optimal nonlinear approximation, Manuscripta Math. 63 (1989) 469–478.

[7] R.A. DeVore, K. Oskolkov, P. Petrushev, Approximation by feed-forward neural networks, Ann. Numer. Math. 4 (1997) 261–287.

[8] I. Daubechies, *Ten Lectures on Wavelets* (SIAM Press, 1992).

[9] B. Delyon, A. Juditsky and A. Benveniste, Accuracy analysis for wavelet approximations, IEEE Trans. Neural Networks 6 (1995) 332–348.

[10] F. Girosi, Regularization theory, radial basis functions and networks, in: *From Statistics to Neural Networks*, eds. V. Cherkassy, J.H. Friedman and H. Wechsler (Springer, 1994) pp. 166–187.

[11] P. Hall and C.C. Heyde, *Martingale Limit Theory and Its Applications* (Academic Press, Orlando, 1980).

[12] E. Hernandez and G.L. Weiss, *First Course on Wavelets* (CRC Press, 1996).

[13] S. Helgason, *The Radon Transform* (Birkhäuser, Boston, 1980).

[14] M. Karpinski and A.J. Macintyre, Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks, J. Comput. System Sci. 54 (1997) 1600-176.

[15] V.Ya. Lin and A. Pinkus, Fundamentality of ridge functions, J. Approx. Theory 75 (1993) 295–311.

[16] V. Maiorov, On best approximation by ridge functions, J. Approx. Theory 99 (1999) 68–94.

[17] V. Maiorov and J. Ratsaby, On the degree of approximation using manifolds of finite pseudo-dimension, J. Construct. Approx. 15 (1999) 291–300.

[18] V. Maiorov and A. Pinkus, Lower bounds for approximation by MLP neural networks, Neurocomputing 25 (1998) 81–91.

[19] R. Meir and V. Maiorov, On the optimality of neural network approximation using incremental algorithms, Technical Report CC-257, Department of Electrical Engineering, Technion, Israel, October 1998. To appear in IEEE Trans. Neural Networks (2000).

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

[21] H.N. Mhaskar and C.A. Micchelli, Approximation by superposition of a sigmoidal function and radial basis functions, Adv. Appl. Math. 16 (1992) 350–373.

[22] H.N. Mhaskar and C.A. Micchelli, Dimension independent bounds on the degree of approximation by neural networks, IBM J. Res. Develop. 38 (1994) 277–284.

[23] O.P. Misra and J.L. Lavoine, *Transform Analysis of Generalized Functions*, North-Holland Mathematics Studies, Vol. 119 (North-Holland, Amsterdam, 1986).

[24] P.P. Petrushev, Approximation by ridge functions and neural networks, SIAM J. Math. Anal. 30 (1998) 155–189.

[25] A. Pinkus, *n-Widths in Approximation Theory* (Springer, Berlin, 1985).

[26] E. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces* (Princeton Univ. Press, Princeton, NJ, 1971).

[27] H. Triebel, *Interpolation Theory of Function Spaces and Differential Operators* (VEB Verlag, Berlin, 1978).

[28] M. Vidyasagar, *A Theory of Learning and Generalization* (Springer, London, 1997).

[29] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).

[30] A.G. Vitushkin, *Estimation of the Complexity of the Tabulation Problem* (Fizmatgiz, Moscow, 1959).

[31] H.E. Warren, Lower bounds for approximation by nonlinear manifold, Trans. Amer. Math. Soc. 133 (1968) 167–178.