Theory of Deep Convolutional Neural Networks II: Spherical Analysis

Zhiying Fang[a], Han Feng[b], Shuo Huang[b], Ding-Xuan Zhou[a,b]
School of Data Science[a], Department of Mathematics[b]
City University of Hong Kong, Kowloon, Hong Kong
Email: zyfang4-c@my.cityu.edu.hk, hanfeng@cityu.edu.hk
shuang56-c@my.cityu.edu.hk, mazhou@cityu.edu.hk

Abstract

Deep learning based on deep neural networks of various structures and architectures has been powerful in many practical applications, but it lacks enough theoretical verifications. In this paper, we consider a family of deep convolutional neural networks applied to approximate functions on the unit sphere $\mathbb{S}^{d-1}$ of $\mathbb{R}^d$. Our analysis presents rates of uniform approximation when the approximated function lies in the Sobolev space $W_\infty^r(\mathbb{S}^{d-1})$ with $r > 0$ or takes an additive ridge form. Our work verifies theoretically the modelling and approximation ability of deep convolutional neural networks followed by downsampling and one fully connected layer or two. The key idea of our spherical analysis is to use the inner product form of the reproducing kernels of the spaces of spherical harmonics and then to apply convolutional factorizations of filters to realize the generated linear features.

Keywords: deep learning, convolutional neural networks, approximation theory, spherical analysis, Sobolev spaces

1 Introduction

Deep learning has attracted tremendous attention from various fields of science and technology recently. Wide applications including those in image processing [9] and speech recognition [12] have received great successes. Based on deep neural network structures, it has a strong capability of obtaining data features and distinguishes itself from classical machine learning methods. Though it is successful in practical applications, theoretical assurances are still lacking and need to be investigated. Many attempts have been made trying to understand the practical power of deep neural networks [3, 18].

The classical (shallow) neural networks to approximate functions or process data on $\mathbb{R}^d$ take the form

$$f_N(x) = \sum_{k=1}^{N} c_k \sigma\left(\langle w_k, x\rangle - b_k\right), \qquad x \in \mathbb{R}^d, \tag{1.1}$$

where N is the number of neurons called width, $\{w_k \in \mathbb{R}^d, c_k \in \mathbb{R}, b_k \in \mathbb{R}\}_{k=1}^{N}$ are parameters corresponding to connection weights, biases, and coefficients, $\langle w_k, x\rangle = w_k \cdot x$ is the inner product in $\mathbb{R}^d$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function. Approximation of functions on subsets of $\mathbb{R}^d$ by shallow neural networks (1.1) was studied well around the late 1980s. See [19, 5] and references therein. As a natural extension of shallow nets, fully connected deep neural networks (DNNs) have been developed since the 1990s. A fully connected DNN of input $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ with J hidden layers of neurons $\{H^{(j)} : \mathbb{R}^d \to \mathbb{R}^{d_j}\}$ with width $\{d_j\}$ is defined iteratively by $H^{(0)}(x) = x$ with $d_0 = d$ and

$$\left(H^{(j)}(x)\right)_i = \sigma(\langle w_i^{(j)}, H^{(j-1)}(x)\rangle - b_i^{(j)}), \quad i = 1, 2, \ldots, d_j,$$

where $w_i^{(j)} \in \mathbb{R}^{d_{j-1}}$ and $b_i^{(j)} \in \mathbb{R}$ are connection weights and biases in the j-th layer. If we use $w_i^{(j)} \in \mathbb{R}^{d_{j-1}}$ with $i = 1, \ldots, d_j$ as rows to form a $d_j \times d_{j-1}$ matrix $F^{(j)}$ and $b_i^{(j)}$ to form a vector $b^{(j)} = (b_i^{(j)})_{i=1}^{d_j}$, then by acting σ componentwise on vectors, the DNN of depth J can be expressed as

$$H^{(j)}(x) = \sigma\left(F^{(j)} \; H^{(j-1)}(x) - b^{(j)}\right), \quad j = 1, 2, \ldots, J. \tag{1.2}$$

DNNs designed by convolutions are called **deep convolutional neural networks** (CNNs) and have been very successful in image classification and related applications [16]. Such a network is associated with a sequence of convolutional filters $\mathbf{w} = \{w^{(j)} : \mathbb{Z} \to \mathbb{R}\}_{j=1}^{J}$, where $w^{(j)}$ is supported in $\left\{0, \cdots, S^{(j)}\right\}$ for some $S^{(j)} \in \mathbb{N}$ called filter length. The convolution of such a filter w supported in $\{0, \cdots, S\}$ with another sequence $v = (v_1, \ldots, v_D)$ is a sequence $w * v$ given by

$$(w * v)_i = \sum_{k \in \mathbb{Z}} w_{i-k} v_k = \sum_{k=1}^{D} w_{i-k} v_k, \qquad i \in \mathbb{Z},$$

which is supported in $\{1, \cdots, D+S\}$. Then by restricting the convoluted sequence onto its support, for input $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, a deep CNNs with J hidden layers of neurons $\{h^{(j)} : \mathbb{R}^d \to \mathbb{R}^{d_j}\}$ and widths $\left\{d_j = d_{j-1} + S^{(j)}\right\}$ is defined iteratively by $h^{(0)}(x) = x$ and

$$h^{(j)}(x) = \sigma\left(\left(\sum_{k=1}^{d_{j-1}} w_{i-k}^{(j)} \left(h^{(j-1)}(x)\right)_k\right)_{i=1}^{d_j} - b^{(j)}\right). \tag{1.3}$$

By inducing a Toeplitz type **convolutional matrix**

$$T^w := (w_{i-k})_{i=1,\ldots,D+S,k=1,\ldots,D}$$

associated with a filter w of filter length S and $D \in \mathbb{N}$ given explicitly by

$$T^w = \begin{bmatrix} w_0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ w_1 & w_0 & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ w_S & w_{S-1} & \cdots & w_0 & 0 & \cdots & 0 \\ 0 & w_S & \cdots & w_1 & w_0 & 0\cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w_S & \cdots & w_1 & w_0 \\ 0 & \cdots & 0 & 0 & w_S & \cdots & w_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & w_S \end{bmatrix} \in \mathbb{R}^{(D+S)\times D}, \tag{1.4}$$

a deep CNN can be regarded as a special sparse form of a fully connected DNNs with the full connection matrices $F^{(j)}$ in (1.2) replaced by the sparse Toeplitz convolutional matrices $T^{(j)} := T^{w^{(j)}}$ with $D = d_{j-1}$ and $S = S^{(j)}$ for $j = 0, 1, \ldots, J$. That is, (1.3) becomes

$$h^{(j)}(x) = \sigma\left(T^{(j)}\, h^{(j-1)}(x) - b^{(j)}\right), \qquad j = 1, \ldots, J. \tag{1.5}$$

Compared with fully-connected DNNs, deep CNNs reduce the computational complexity by using at each layer a Toeplitz matrix due to the sparsity and convolutional nature. Throughout the paper we take an identical filter length $S^{(j)} \equiv S \in \mathbb{N}$ implying $\{d_j = d + jS\}$ and take the rectified linear unit (ReLU) activation function

$$\sigma(u) = \max\{u, 0\}, \qquad u \in \mathbb{R}.$$

It was shown in [33] that the output layer of any fully-connected DNN can be realized by a downsampled deep CNN with free parameters of the same order, and that deep CNNs can approximate ridge functions of the form $g(\xi \cdot x)$ with univariate functions g and unknown $\xi \in \mathbb{R}^d$ to the same accuracy with much smaller number of free parameters. Universality of approximation by deep CNNs was also established in [32] where rates of approximation were provided for restrictions of functions from the Sobolev space $H^r(\mathbb{R}^d)$ with a regularity index $r > 2 + d/2$. The regularity index r is large when d increases and is essentially needed in the analysis there due to the function regularity on the whole Euclidean space $\mathbb{R}^d$. Note that the issue of approximating non-smooth functions by fully connected DNNs has been studied in [28, 13]. Moreover, the Sobolev space $H^r(\mathbb{R}^d)$ requires derivatives of various orders to belong to the L_2 space, while the approximation considered in [32] is measured in the L_∞ norm.

In this paper, we overcome the difficulty in the large regularity index and the inconsistency of L_2 and L_∞ norms for the setting with data from the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$. With our novel analysis conducted with spherical harmonic expansions, we can present rates of approximating functions from the Sobolev space $W^r_\infty(\mathbb{S}^{d-1})$ (to be defined below) on $\mathbb{S}^{d-1}$, with any positive index r, by downsampled deep CNNs defined in [33] followed by two fully connected layers.

Definition 1. *The* **downsampling** *operator* $\mathfrak{D}_d : \mathbb{R}^D \to \mathbb{R}^{\lfloor D/d \rfloor}$ *with a scaling parameter* $d \leq D$ *is defined by*

$$\mathfrak{D}_d(v) = \left(v_{id}\right)_{i=1}^{\lfloor D/d \rfloor}, \qquad v = (v_i)_{i=1}^{D} \in \mathbb{R}^D, \tag{1.6}$$

where $\lfloor u \rfloor$ *denotes the integer part of* $u > 0$.

After the last CNN layer, we add two fully connected layers $h^{(J+1)}, h^{(J+2)}$ with widths $\mathcal{D}_1, \mathcal{D}_2 > 0$, respectively, connection matrices $F^{(J+1)}, F^{(J+2)}$ and bias vectors $b^{(J+1)}, b^{(J+2)}$, to be determined. Precisely,

$$h^{(J+1)}(x) = \sigma\left(F^{(J+1)}\mathfrak{D}_d\left(h^{(J)}(x)\right) - b^{(J+1)}\right) \tag{1.7}$$

and

$$h^{(J+2)}(x) = \sigma\left(F^{(J+2)}h^{(J+1)}(x) - b^{(J+2)}\right). \tag{1.8}$$

Such a network with many convolutional layers followed by downsampling operations and very few fully connected layers is quite common in practical applications [16, 9]. The hypothesis space of functions on $\mathbb{S}^{d-1}$ induced by our network is given by

$$\mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S} = \left\{c^{(J+2)} \cdot h^{(J+2)}(x) - A : \ c^{(J+2)} \in \mathbb{R}^{\mathcal{D}_2}, A \in \mathbb{R}\right\}. \tag{1.9}$$

2 Main Results on Rates of Approximation

Our target is to establish rates of approximating functions in $W^r_\infty(\mathbb{S}^{d-1})$ by those from the hypothesis space $\mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S}$ defined by (1.9). Since the sums of the rows in the middle of the Toeplitz type matrix (1.4) are equal, we impose for the bias vectors $\{b^{(j)}\}_{j=1}^J$ of the convolutional layers a restriction

$$b^{(j)}_{S+1} = \ldots = b^{(j)}_{d_j - S}, \qquad j = 1, \ldots, J. \tag{2.1}$$

For the two fully connected layers we take the widths

$$\mathcal{D}_1 = (2N+3)\lfloor (d+JS)/d \rfloor, \qquad \mathcal{D}_2 = \lfloor (d+JS)/d \rfloor \tag{2.2}$$

for some positive integer $N \in \mathbb{N}$ and connection matrices

$$F^{(J+1)} = \Xi_{\mathcal{D}_2, \mathbf{1}_{2N+3}}, \qquad F^{(J+2)} = \Xi^T_{\mathcal{D}_2, \Theta_N} \tag{2.3}$$

with $\mathbf{1}_{2N+3} = (1,1,\ldots,1)^T \in \mathbb{R}^{2N+3}$ and $\Theta_N = (\theta_1, \ldots, \theta_{2N+3})^T \in \mathbb{R}^{2N+3}$. Here the matrix $\Xi_{\mathcal{D}_2,\vec{u}}$ takes a block form as

$$\Xi_{\mathcal{D}_2,\vec{u}} = \begin{bmatrix} \vec{u} & O & O \cdots & O \\ O & \vec{u} & O \cdots & O \\ \vdots & \ddots & \ddots & \vdots \\ O & \cdots & O & \vec{u} \end{bmatrix} \in \mathbb{R}^{(2N+3)\mathcal{D}_2 \times \mathcal{D}_2}, \ \vec{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_{2N+3} \end{bmatrix} \in \mathbb{R}^{2N+3}.$$

Our first main result, to be proved in Section 5, can be stated as follows. A positive parameter ε (which can be arbitrarily small) is needed due to the continuous embedding of the Sobolev space $W^s_2(\mathbb{S}^{d-1})$ with $s > \frac{d-1}{2}$ into $C(\mathbb{S}^{d-1})$ found in Proposition 1 below.

Theorem 1. *Let $2 \le S \le d$, $d \ge 3$, $J \ge \frac{d-1}{S-1}$, $0 < r \ne d-1$, and $\varepsilon > 0$ satisfy $\varepsilon < r - (d-1)$ when $r > d-1$. Take*

$$N = \begin{cases} \left\lfloor \left\lfloor \frac{(S-1)J+1}{d} \right\rfloor^{\frac{1}{2(d-1+\varepsilon)}} \right\rfloor^{d+1}, & \text{if } r < d-1, \\ \left\lfloor \left\lfloor \left\lfloor \frac{(S-1)J+1}{d} \right\rfloor^{\frac{1}{2r}} \right\rfloor^{2+r} \right\rfloor, & \text{if } r > d-1. \end{cases}$$

Then for any $f \in W^r_\infty(\mathbb{S}^{d-1})$, there exists a deep neural network consisting of J layers of CNNs with filters of length S and bias vectors satisfying (2.1) followed by downsampling and two fully connected layers with widths (2.2) and connection matrices (2.3) such that the hypothesis space $\mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S}$ contains a function $\hat{f}$ satisfying

$$\left\| f - \hat{f} \right\|_\infty \le C_{r,d,\varepsilon,S} J^{-\frac{r}{2(d-1+\varepsilon)}} \|f\|_{W^r_\infty(\mathbb{S}^{d-1})},$$

where $C_{r,d,\varepsilon,S}$ is a constant independent of J or f. Moreover, the total number of free parameters $\mathcal{N}$ in the network can be bounded as

$$\mathcal{N} \leq (3S+5)\, J + 4.$$

Remark 1. *To achieve the approximation accuracy $\left\| f - \hat{f} \right\|_\infty \leq \epsilon$, we only need to take*

$$J = \left\lceil \max \left\{ \left(C_{r,d,\varepsilon,S} \|f\|_{W^r_\infty(\mathbb{S}^{d-1})} / \epsilon \right)^{\max\left\{\frac{2(d-1+\varepsilon)}{r}, 2\right\}}, \frac{d-1}{S-1} \right\} \right\rceil,$$

where we denote the integer greater than or equal to $u > 0$ as $\lceil u \rceil$. When $0 < r < d-1$, we may set $\varepsilon = 1$ and see that the network achieving the approximation accuracy ϵ has depth $J = \mathcal{O}\left(\epsilon^{-\frac{2d}{r}}\right)$ and $\mathcal{N} = \mathcal{O}\left(\epsilon^{-\frac{2d}{r}}\right)$ free parameters. When $r > d-1$, we may set $0 < \varepsilon < r - (d-1)$ and find the depth and the number of free parameters of this network are both of orders $\mathcal{O}\left(\epsilon^{-2}\right)$, slightly better than that in [32].

Our second main result aims at demonstrating the superiority of deep CNNs observed empirically in many practical applications. Motivated by our earlier work [33] on approximating ridge functions of type $g(y \cdot x)$ with $y \in \mathbb{R}^d$, $g : \mathbb{R} \to \mathbb{R}$, and additive models (see, e.g., [4, 24]) in statistics of the form $f(x) = \sum_{j=1}^d g_j(x_j)$ with univariate functions $\{g_j\}_{j=1}^d$, we consider a family of **additive ridge functions** of the form

$$f(x) = \sum_{j=1}^{m} g_j(y_j \cdot x) \tag{2.4}$$

with $y_j \in \mathbb{S}^{d-1}, g_j : \mathbb{R} \to \mathbb{R}$ for $j \in \{1, \ldots, m\}$. The following theorem to be proved in Section 5 is about approximating additive ridge functions by deep CNNs followed by downsampling and one fully-connected layer. For $0 < \alpha \leq 1$, denote the space of Lipschitz-α functions on $[-1,1]$ as $W^\alpha_\infty([-1,1])$ with the semi-norm $|\cdot|_{W^\alpha_\infty}$ being the Lipschitz-α constant.

Theorem 2. *Let $m \in \mathbb{N}$, $d \geq 3$, $2 \leq S \leq d$, $J = \left\lceil \frac{md-1}{S-1} \right\rceil$, and $N \in \mathbb{N}$. If f is an additive ridge function (2.4) with unknown $\{y_1, \ldots, y_m\} \subset \mathbb{S}^{d-1}$, $\{g_1, \ldots, g_m\} \subset W^\alpha_\infty([-1,1])$ for some $0 < \alpha \leq 1$, then there exists a deep neural network consisting of J layers of CNNs with filters of length S and bias vectors satisfying (2.1) followed by downsampling and one fully connected layer $h^{(J+1)}$ with width $(2N+3)\lfloor (d+JS)/d \rfloor$ and connection matrix $F^{(J+1)} = \Xi_{\lfloor (d+JS)/d \rfloor, \mathbf{1}_{2N+3}}$ such that for some coefficient vector $c^{(J+1)} \in$*

$\mathbb{R}^{(2N+3)\lfloor (d+JS)/d \rfloor}$ *there holds*

$$\left\| f - c^{(J+1)} \cdot h^{(J+1)} \right\|_\infty \le \sum_{j=1}^{m} |g_j|_{W_\infty^\alpha} N^{-\alpha}.$$

The total number of free parameters $\mathcal{N}$ in the network can be bounded as

$$\mathcal{N} \le (3S+2) \left\lceil \frac{md-1}{S-1} \right\rceil + 2N + 2.$$

Remark 2. *To achieve an accuracy $\epsilon > 0$ for approximating an additive ridge function (2.4) on $\mathbb{S}^{d-1}$, we only need to take*

$$N = \left\lceil \left(\sum_{j=1}^{m} |g_j|_{W_\infty^\alpha} \right)^{1/\alpha} \epsilon^{-\frac{1}{\alpha}} \right\rceil = \mathcal{O}\left(\epsilon^{-\frac{1}{\alpha}} \right).$$

The total number of free parameters of the achieving network is of orders $\mathcal{O}\left(\epsilon^{-\frac{1}{\alpha}}\right)$. This is the complexity required by the classical literature on fully-connected networks to achieve an accuracy $\epsilon > 0$ for approximating univariate functions. This extends our earlier work [33] from the ridge case with $m = 1$ to an additive ridge case with $m \in \mathbb{N}$ and hints the superiority of deep CNNs in approximating multivariate functions of special structures.

3 Spherical Analysis for Deep CNNs

In this section, we introduce ideas of our analysis before proving our main results. We first give a brief review on relevant concepts from spherical harmonic analysis and introduce some classes of functions. More details can be found in [6, 29].

3.1 Spherical harmonics and Sobolev spaces on spheres

For $1 \le p \le \infty$, we denote by $L_p(\mathbb{S}^{d-1}) = L_p(\mathbb{S}^{d-1}, \sigma_d)$ the L_p-function space defined with respect to the normalized Lebesgue measure σ_d on $\mathbb{S}^{d-1}$, and $\|\cdot\|_p$ the norm of $L_p(\mathbb{S}^{d-1})$.

A spherical harmonic of degree $n \in \mathbb{Z}_+$ on $\mathbb{S}^{d-1}$ is the restriction to $\mathbb{S}^{d-1}$ of a homogeneous and harmonic polynomial of total degree n defined on $\mathbb{R}^d$. Let $\mathcal{H}_n^d$ denote the set of all spherical harmonics of degree n on $\mathbb{S}^{d-1}$. It can be found in [6] the dimension of the linear space $\mathcal{H}_n^d$ is

$$N(n,d) = \binom{n+d-1}{n} - \binom{n+d-3}{n-2} = \mathcal{O}_d(n^{d-2}). \tag{3.1}$$

Note that $L_2(\mathbb{S}^{d-1})$ is a Hilbert space with inner product $\langle f, g\rangle_2 := \int_{\mathbb{S}^{d-1}} f(x)g(x)d\sigma_d(x)$ for $f, g \in L_2\left(\mathbb{S}^{d-1}\right)$. The spaces $\mathcal{H}_n^d$, $n \in \mathbb{Z}_+$, of spherical harmonics are mutually orthogonal with respect to the inner product of $L_2(\mathbb{S}^{d-1})$. Since the space of spherical polynomials is dense in $L_2(\mathbb{S}^{d-1})$, each $f \in L_2(\mathbb{S}^{d-1})$ has a spherical harmonic expansion:

$$f = \sum_{n=0}^{\infty} \operatorname{Proj}_n f = \sum_{n=0}^{\infty} \sum_{l=1}^{N(n,d)} \widehat{f}_{n,l} Y_{n,l}$$

converging in the $L_2(\mathbb{S}^{d-1})$ norm. Here and elsewhere, $\{Y_{n,l}\}_{l=1}^{N(n,d)}$ is an orthonormal basis of $\mathcal{H}_n^d$, $\widehat{f}_{n,l}$ is the Fourier coefficients of f given by

$$\widehat{f}_{n,l} := \langle f, Y_{n,l}\rangle_{L_2(\mathbb{S}^{d-1})} := \int_{\mathbb{S}^{d-1}} f(x)Y_{n,l}(x)d\sigma_d(x), \tag{3.2}$$

and Proj_n is the orthogonal projection of $L_2(\mathbb{S}^{d-1})$ onto the subspace $\mathcal{H}_k^d$ of spherical harmonics, which has an integral representation:

$$\operatorname{Proj}_n f(x) = \int_{\mathbb{S}^{d-1}} f(y)Z_n(x, y)\, d\sigma_d(y), \qquad x \in \mathbb{S}^{d-1},$$

where

$$Z_n(x, y) = \sum_{i=1}^{N(n,d)} Y_{n,i}(x)Y_{n,i}(y), \qquad x, y \in \mathbb{S}^{d-1}.$$

It can be readily shown that $Z_n(x, y)$ is the reproducing kernel of $\mathcal{H}_n^d$ independent of the choice of $\{Y_{n,l}\}_{l=1}^{N(n,d)}$. Furthermore, with $\lambda = \frac{d-2}{2}$,

$$Z_n(x, y) = \frac{n+\lambda}{\lambda} C_n^{\lambda}\left(\langle x, y\rangle\right), \qquad x, y \in \mathbb{S}^{d-1}, \tag{3.3}$$

where $C_n^{\lambda}(t)$ is the Gegenbauer polynomial of degree n with parameter $\lambda > -1/2$, see, for instance, [6].

The spaces $\mathcal{H}_n^d$ of spherical harmonics can also be characterized as eigenfunction spaces of the Laplace-Beltrami operator Δ_0 on $\mathbb{S}^{d-1}$. Indeed,

$$\mathcal{H}_n^d = \{f \in C^2(\mathbb{S}^{d-1}) : \Delta_0 f = -\lambda_n f\},$$

where $\lambda_n = n(n + d - 2)$ and $C^2(\mathbb{S}^{d-1})$ denotes the space of all twice continuously differentiable functions on $\mathbb{S}^{d-1}$. As a matter of fact, we may define the fractional power $(-\Delta_0 + I)^{\alpha}$ of $-\Delta_0 + I$ for $\alpha \in \mathbb{R}$ in a distributional sense by

$$\operatorname{Proj}_n((-\Delta_0 + I)^{\alpha} f) = (\lambda_n + 1)^{\alpha} \operatorname{Proj}_n f,$$

which is a self-adjoint operator on $L_2(\mathbb{S}^{d-1})$ in the sense that

$$\Big\langle (-\Delta_0 + I)^\alpha f, g \Big\rangle_{L_2(\mathbb{S}^{d-1})} = \Big\langle f, (-\Delta_0 + I)^\alpha g \Big\rangle_{L_2(\mathbb{S}^{d-1})}, \quad \forall f, g \in L_2(\mathbb{S}^{d-1}).$$

Now we define the Sobolev space $W_p^r(\mathbb{S}^{d-1})$ to be a subspace of $L_p(\mathbb{S}^{d-1})$, $1 \le p \le \infty$, $r > 0$, with the finite norm

$$\begin{aligned} \|f\|_{W_p^r(\mathbb{S}^{d-1})} &:= \left\|(-\Delta_0 + I)^{r/2} f\right\|_p \qquad (3.4)\\ &= \left\| \sum_{n=0}^{\infty} (1+\lambda_n)^{\frac{r}{2}} \sum_{l=1}^{N(n,d)} \widehat{f}_{n,l} Y_{n,l} \right\|_p . \qquad (3.5) \end{aligned}$$

When $p = 2$ it is known that $W_2^r(\mathbb{S}^{d-1})$ is a Hilbert space with the inner product:

$$\langle f, g\rangle_{W_2^r(\mathbb{S}^{d-1})} = \sum_{n=0}^{\infty} (1+\lambda_n)^r \sum_{l=1}^{N(n,d)} \widehat{f}_{n,l} \widehat{g}_{n,l}.$$

In addition, we have the following continuous embedding lemma, see [11] and also [14, Eq. 14, p. 420]. By these lemmas, we know that the infinity norm can be bounded by the Sobolev norm when $r > \frac{d-1}{p}$.

Proposition 1. *For $d \ge 3$, $1 \le p \le \infty$, and $r > \frac{d-1}{p}$, the Sobolev space $W_p^r(\mathbb{S}^{d-1})$ is continuously embedded into $C(\mathbb{S}^{d-1})$, which implies*

$$\|f\|_\infty \le c_{r,d} \|f\|_{W_p^r(\mathbb{S}^{d-1})}, \qquad \forall f \in W_p^r(\mathbb{S}^{d-1}),$$

where $c_{r,d}$ is a constant independent of f.

3.2 Near-best approximation and discretization

The best approximation of a function by those from polynomial spaces of various degrees might be nonlinear. A useful tool in spherical harmonic analysis is a linear scheme L_n.

Definition 2. *Given a $C^\infty([0,\infty))$ function η with $\eta(t) = 1$ for $0 \le t \le 1$ and $\eta(t) = 0$ for $t \ge 2$, we define a sequence of linear operators L_n, $n \in \mathbb{Z}_+$, on $L_p(\mathbb{S}^{d-1})$ with $1 \le p \le \infty$ by*

$$L_n(f)(x) := \sum_{k=0}^{\infty} \eta\left(\frac{k}{n}\right) \mathrm{Proj}_k(f)(x) = \int_{\mathbb{S}^{d-1}} f(y) l_n(\langle x, y\rangle) d\sigma_d(y), \quad x \in \mathbb{S}^{d-1}, \qquad (3.6)$$

where with $\lambda = \frac{d-2}{2}$, l_n *is a kernel given by*

$$l_n(t) = l_{n,d}(t) := \sum_{k=0}^{2n} \eta\left(\frac{k}{n}\right) \frac{\lambda + k}{\lambda} C_k^{\lambda}(t), \qquad t \in [-1, 1]. \tag{3.7}$$

It can be found in [6, Chapter 4] that L_n is near best, achieving the order of best approximation for $f \in W_p^r(\mathbb{S}^{d-1})$.

Lemma 1. *For* $n \in \mathbb{N}$, $1 \le p \le \infty$ *and* $f \in W_p^r\left(\mathbb{S}^{d-1}\right)$, *there holds*

$$\|f - L_n(f)\|_p \le cn^{-r} \|f\|_{W_p^r(\mathbb{S}^{d-1})}, \tag{3.8}$$

where c *is a constant depending only on the function* η *in defining* L_n.

Note that since $(-\Delta_0 + I)^{-r/2}$ is self-adjoint, for $x \in \mathbb{S}^{d-1}$,

$$\begin{aligned} L_n(f)(x) =& \Big\langle f, l_n(\langle x, \cdot \rangle) \Big\rangle_{L_2(\mathbb{S}^{d-1})} \\ =& \Big\langle (-\Delta_0 + I)^{r/2} f, (-\Delta_0 + I)^{-r/2} l_n(\langle x, \cdot \rangle) \Big\rangle_{L_2(\mathbb{S}^{d-1})} \\ =:& \int_{\mathbb{S}^{d-1}} F_r(y) \zeta_{n,r}(\langle x, y \rangle) d\sigma(y), \end{aligned}$$

here and in what follows, we denote

$$F_r = (-\Delta_0 + I)^{r/2} f \quad \text{and} \quad \zeta_{n,r}(\langle x, \cdot \rangle) = (-\Delta_0 + I)^{-r/2} l_n(\langle x, \cdot \rangle).$$

The novelty here using a fractional power of $(-\Delta_0 + I)$ caused by the regularity of $f \in W_\infty^r(\mathbb{S}^{d-1})$ enables us to get an r-dependent error bound for discretizing $L_n(f)$: the larger the regularity index r, the smaller the bound.

To approximate the function $L_n(f)$ by a neural network, we need a stepping stone, discretizing the integral form (3.6) to an empirical version

$$\widehat{L}_{n,m}^{\mathbf{y}}(f)(x) = \frac{1}{m} \sum_{i=1}^{m} F_r(y_i) \zeta_{n,r}(\langle x, y_i \rangle), \qquad x \in \mathbb{S}^{d-1} \tag{3.9}$$

given in terms of a sample $\mathbf{y} = \{y_1, \ldots, y_m\} \subset \mathbb{S}^{d-1}$. The following estimate for the error $\widehat{L}_{n,m}^{\mathbf{y}}(f) - L_n(f)$ will be proved by a probability inequality in Section 5. Such a probabilistic argument has been applied in [15].

Lemma 2. *Let* $d \ge 3$, $r > 0$, *and* $\varepsilon > 0$. *If* $f \in W_\infty^r(\mathbb{S}^{d-1})$, *then for any* $n, m \in \mathbb{N}$, *there exist* $\mathbf{y} = \{y_1, y_2, \ldots, y_m\} \subset \mathbb{S}^{d-1}$ *such that*

$$\left\|\widehat{L}_{n,m}^{\mathbf{y}}(f) - L_n(f)\right\|_\infty \le C_{r,d,\varepsilon} \frac{\sqrt{\Lambda_{2(d-1-r+\varepsilon)}(n)}}{\sqrt{m}} \|f\|_{W_\infty^r(\mathbb{S}^{d-1})}, \tag{3.10}$$

where for $\tau \in \mathbb{R}$ and $n \in \mathbb{N}$ we denote

$$\Lambda_\tau(n) = \begin{cases} n^\tau, & \text{if } \tau > 0, \\ \log(n+1), & \text{if } \tau = 0, \\ 1, & \text{if } \tau < 0, \end{cases}$$

and $C_{r,d,\varepsilon}$ is a positive constant depending on r, d, ε but not on n, m or f.

3.3 Approximating ridge functions by deep CNNs

The last step in our spherical analysis of deep CNNs is to approximate the ridge function $\widehat{L}^{\mathbf{y}}_{n,m}(f)$ by functions from the network with a bound to be proved in Section 5.

Lemma 3. *Let $2 \leq S \leq d$, $d \geq 3$, $r > 0$, $m, n, N \in \mathbb{N}$, $f \in W^r_\infty(\mathbb{S}^{d-1})$, and $\mathbf{y} = \{y_1, \ldots, y_m\} \subset \mathbb{S}^{d-1}$. Let $J \geq \lceil \frac{md-1}{S-1} \rceil$. Then there exists a deep neural network consisting of J layers of CNNs with filters of length S and bias vectors satisfying (2.1) followed by downsampling and two fully connected layers with widths (2.2), connection matrices (2.3) and bias vectors involving two parameters $B^{(J)}, B^{(J+2)} \in \mathbb{R}$ given explicitly in (5.6), (5.8) below such that the hypothesis space $\mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S}$ contains a function $\hat{f}$ satisfying*

$$\left\| \widehat{L}^{\mathbf{y}}_{n,m}(f) - \hat{f} \right\|_\infty \leq c'_{r,d} \frac{n^2 \Lambda_{d-1-r}(n)}{N} \|f\|_{W^r_\infty(\mathbb{S}^{d-1})}, \tag{3.11}$$

where $c'_{r,d}$ is a constant depending only on r and d. The total number of free parameters $\mathcal{N}$ in the network can be bounded as

$$\mathcal{N} \leq J(3S+2) + m + 2N + 4.$$

4 Comparison with Related Work

In this section, we give a brief review of related work on rates of function approximation by neural networks.

The **fully connected** shallow nets (1.1) or multi-layer nets (1.2) have nice approximation properties due to the fully connected nature, which was well studied in a large literature around 30 years ago. When the activation function is a C^∞ sigmoidal type function, approximation rates were obtained by many authors. In particular, in [1] rates were given for functions in $f \in L_2(\mathbb{R}^d)$ whose Fourier transforms $\hat{f}$ satisfy a decay condition $\int_{\mathbb{R}^d} |w||\hat{f}(w)|dw < \infty$. Another typical result based on localized Taylor expansions asserts [19] that even for shallow nets (1.1), we have $\inf_{c_k, w_k, b_k} \|f_N - f\|_{C([-1,1]^d)} = O(N^{-r/d})$

for $f \in W^r_\infty([-1,1]^d)$, if for some $b \in \mathbb{R}$ and some integer $\ell \in \mathbb{N} \setminus \{1\}$, the C^∞ activation function σ satisfies $\sigma^{(k)}(b) \neq 0$ for all $k \in \mathbb{Z}_+$ and $\lim_{u \to -\infty} \sigma(u)/|u|^\ell = 0$ and $\lim_{u \to \infty} \sigma(u)/u^\ell = 1$. These conditions required by the localized Taylor expansion approach are not satisfied by ReLU, so the approximation theory developed 30 years ago does not apply to ReLU. The difficulty was overcome in the recent deep learning literature and approximation properties of ReLU nets were established in [15] for ReLU shallow nets and functions satisfying $\int_{\mathbb{R}^d} |w||\hat{f}(w)|dw < \infty$, in [30, 2, 22, 20] for deep nets and functions from $W^r_\infty([-1,1]^d)$ with $0 < r \leq 2$, and in [25] for approximation on manifolds. These results are obtained for fully connected nets.

Deep CNNs are different from fully connected nets. They have special sparse convolutional connection matrices (1.4), which leads to sparsity and reduces the computational complexity for structured data. Recently in [32], for functions f on $\Omega \subset [-1,1]^d$ satisfying $f = F|_\Omega$ with $F \in W^r_2\left(\mathbb{R}^d\right)$ and an integer index $r > 2 + d/2$, it was shown that the approximation accuracy $\|f - \hat{f}\|_\infty \leq \epsilon$ can be achieved by a deep CNN of depth $4\lceil \frac{1}{\epsilon^2} \log \frac{1}{\epsilon^2} \rceil$ and at most $\lceil \frac{75}{\epsilon^2} \log \frac{1}{\epsilon^2} \rceil d$ free parameters. The linear increment of the free parameter number with respect to d improves the bound in Theorem 1 of [30] which requires at least $2^d \epsilon^{-d/r}$ free parameters and $\frac{C_0 d}{4}(\log(1/\epsilon) + d)$ fully connected layers with $C_0 > 0$ to achieve the same approximation accuracy ϵ. Periodized deep CNNs with different architectures and connection matrices different from the Toeplitz convolutional ones (1.4) were shown in [21, 23] to be able to realize the output layer of any fully-connected DNN with free parameters of the same order. The same result was shown for deep CNNs (1.5) in [33].

The index $r > 2 + d/2$ required in [32] can be very large for processing high dimensional data. Hence approximated functions are required to possess high regularity which is not the usual case in applications. The technical difficulty is caused by embedding $W^r_2([-1,1]^d)$ into $W^s_\infty([-1,1]^d)$ which requires $s < r - d/2$ and makes the bound in [32] loose. To overcome the difficulty, we consider the case when the data is from the unit sphere $\mathbb{S}^{d-1}$. By applying the spherical harmonic expansions, we can approximate any functions $f \in W^r_\infty(\mathbb{S}^{d-1})$ by deep CNNs. In the literature, there have been some other harmonic analysis approaches in dealing with approximation by fully connected neural networks, using ridgelet transforms in [27], local Taylor expansions in [30], and B-spline functions in [28]. Our spherical harmonic analysis approach makes full use of the inner product nature (3.3) of the reproducing kernel of $\mathcal{H}^d_n$ which, after discretizing the polynomial approximation $L_n(f)$ to $\widehat{L}^{\mathbf{y}}_{n,m}(f)$, enables us to represent the linear transformations

$\{\langle y_i, x\rangle\}$ by deep CNNs with linearly increasing widths, an idea borrowed from our earlier work [32]. A key consequence of our approach is to allow the index r here to be an arbitrarily small positive number, which relaxes the restriction in [32] for the regularity of the approximated function. While the approximation of non-smooth functions is unified for $r > 0$ and the same order $\mathcal{O}\left(\epsilon^{-2}\right)$ for the number of network free parameters to achieve an approximation accuracy $\epsilon > 0$ is kept when $r > d-1$, a parameter number of order $\mathcal{O}\left(\epsilon^{-\frac{2d}{r}}\right)$ is required when $0 < r < d-1$. This is due to our approach of using a Hilbert space $W_2^s(\mathbb{S}^{d-1})$ in our probabilistic estimate for the discretization, which makes our rate suboptimal compared with [30, 20, 24] for $r < d-1$. It would be interesting to derive optimal rates of approximating by deep CNNs functions from $W_\infty^r(\mathbb{S}^{d-1})$ with small r.

On the other hand, as stated in Remark 2, for an additive ridge function on $\mathbb{S}^{d-1}$ in the family (2.4), deep CNNs followed by a fully connected layer can extract linear features $\{y_j\}_{j=1}^m$ and then approximate the function efficiently, with the same order of network free parameters as that for approximating a univariate function by fully connected DNNs. This demonstrates the superiority of deep CNNs in approximating functions with structures. It would be of great interest to explore other structures of multivariate functions for which deep CNNs together with network architectures like pooling and parallel channels may have super performance in function approximations and representations. Applying deep CNNs to some practical or empirical problems involving additive ridge functions (2.4) would also help understand advantages of convolutional structures of deep learning in some practical domains.

5 Proof of the Main Results

This section is devoted to the proof of our main analysis. Our analysis for the error $f - \hat{f}$ is carried out by means of the bounds for $\|f - L_n(f)\|_\infty$ in Lemma 1, $\|L_n(f) - \widehat{L}_{n,m}^{\mathbf{y}}(f)\|_\infty$ in Lemma 2, and $\|\widehat{L}_{n,m}^{\mathbf{y}}(f) - \hat{f}\|_\infty$ in Lemma 3.

5.1 Proving the lemma on discretization

To complete our analysis, we first prove Lemma 2 and Lemma 3.

The proof of Lemma 2 is based on the following probability inequality for random variables with values in a Hilbert space which can be found in [26].

Lemma 4. *Let $(H, \|\cdot\|)$ be a Hilbert space and ξ be a random variable on (Y, ρ) with values in H. Assume $\|\xi\| \leq M < \infty$ almost surely. Denote*

$\sigma^2(\xi) = E\left(\|\xi\|^2\right)$. Let $\{y_i\}_{i=1}^m$ be independent random drawers of ρ. Then for any $0 < \delta < 1$, we have with confidence $1-\delta$,

$$\left\| \frac{1}{m}\sum_{i=1}^m \xi(y_i) - E(\xi) \right\|_H \le \frac{2M\log(\frac{2}{\delta})}{m} + \sqrt{\frac{2\sigma^2(\xi)\log(\frac{2}{\delta})}{m}}.$$

Proof of Lemma 2. Recall that $L_n(f)$ is defined by (3.6) with $f \in W_\infty^r(\mathbb{S}^{d-1})$ and $\varepsilon > 0$. In applying Lemma 4 we take the Sobolev space $W_2^s(\mathbb{S}^{d-1})$ with the smoothness index $s = \varepsilon + \frac{d-1}{2}$ to be the Hilbert space H and the random variable ξ on $\left(\mathbb{S}^{d-1}, \sigma_d\right)$ with values in H given by

$$\xi(y) = F_r(y)\sum_{k=0}^{2n}(1+\lambda_k)^{-r/2}\eta\left(\frac{k}{n}\right)Z_k(y,\cdot) \in H, \qquad y \in \mathbb{S}^{d-1}.$$

Then $E(\xi) = L_n(f)$ and $\frac{1}{m}\sum_{i=1}^m \xi(y_i) = \widehat{L}_{n,m}^{\mathbf{y}}(f)$.

To bound the norm $\|\xi\| = \|\xi\|_{W_2^s}$, we recall the norm of $W_2^s(\mathbb{S}^{d-1})$ given by (3.4) with $p=2$ and find for $y \in \mathbb{S}^{d-1}$,

$$\|\xi(y)\|_{W_2^s(\mathbb{S}^{d-1})} = \left\| F_r(y)\sum_{k=0}^{2n}(1+\lambda_k)^{\frac{s-r}{2}}\eta\left(\frac{k}{n}\right)Z_k(y,\cdot)\right\|_{L_2(\mathbb{S}^{d-1})},$$

where $\lambda_k = k(k+d-2)$. Then by the orthogonality and reproducing properties,

$$\begin{aligned}\|\xi(y)\|^2_{W_2^s(\mathbb{S}^{d-1})} &= (F_r(y))^2\sum_{k=0}^{2n}(1+\lambda_k)^{s-r}\eta^2\left(\frac{k}{n}\right)Z_k(y,y)\\ &= (F_r(y))^2\sum_{k=0}^{2n}(1+\lambda_k)^{s-r}\eta^2\left(\frac{k}{n}\right)N(k,d),\end{aligned}$$

where we have used the identity $Z_k(y,y) = N(k,d)$ found in [6] as Corollary 1.2.7 and $N(k,d)$ is the dimension of spherical harmonics $\mathcal{H}_k^d$. Notice that for $k \in \mathbb{N}$, $k^2 < 1+\lambda_k \le dk^2$. We find $(1+\lambda_k)^{s-r} \le d^{\max\{s-r,0\}}k^{2(s-r)}$ for either $s-r \ge 0$ or $s-r<0$. Since $0 \le \eta(t) \le 1$ for $t \in [0,2]$, we can apply (3.1) to estimate the summation as

$$\sum_{k=0}^{2n}(1+\lambda_k)^{s-r}\eta^2\left(\frac{k}{n}\right)N(k,d) \le 1 + \sum_{k=1}^{2n} d^{\max\{s-r,0\}}k^{2(s-r)}c_d' k^{d-2}$$

with a constant c_d' depending only on d, while

$$\sum_{k=1}^{2n} k^{2(s-r)+d-2} \le 1 + \begin{cases} \frac{3^{2(s-r)+d-1}}{2(s-r)+d-1}n^{2(s-r)+d-1}, & \text{if } 2(s-r)+d-2 > -1,\\ 1+\log(n+1), & \text{if } 2(s-r)+d-2 = -1,\\ \frac{1}{1-2(s-r)-d}, & \text{if } 2(s-r)+d-2 < -1.\end{cases}$$

Combining this with the definitions of the norm $\|f\|_{W^r_\infty(\mathbb{S}^{d-1})}$ and the function $\Lambda_\tau(n)$, we know that $\|\xi(y)\|^2_{W^s_2}$ can be bounded as

$$\|\xi(y)\|^2_{W^s_2(\mathbb{S}^{d-1})} \leq c^2_{s,r,d}\|f\|^2_{W^r_\infty(\mathbb{S}^{d-1})}\Lambda_{2s-2r+d-1}(n),$$

where $c_{s,r,d}$ is a positive constant independent of f or n. Thus the random variable ξ satisfies the condition $\|\xi\| \leq M < \infty$ in Lemma 4 with $M = c_{s,r,d}\|f\|_{W^r_\infty(\mathbb{S}^{d-1})}\sqrt{\Lambda_{2s-2r+d-1}(n)}$. So by Lemma 4 with $\delta = \frac{1}{2}$ and $\sigma^2(\xi) \leq M^2$, we know from the positive measure of the sample set that there exists a set of points $\mathbf{y} = \{y_i\}_{i=1}^m \in \mathbb{S}^{d-1}$ such that

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i=1}^m \xi(y_i) - E(\xi)\right\|_H &= \left\|\widehat{L}^{\mathbf{y}}_{n,m}(f) - L_n(f)\right\|_{W^s_2(\mathbb{S}^{d-1})} \\ &\leq \frac{6c_{s,r,d}\|f\|_{W^r_\infty(\mathbb{S}^{d-1})}\sqrt{\Lambda_{2s-2r+d-1}(n)}}{\sqrt{m}}.\end{aligned}$$

This verifies (3.10) by the embedding Proposition 1 with $p = 2$ and $s = \varepsilon + \frac{d-1}{2} > \frac{d-1}{2}$. □

5.2 Proving the lemma on ridge approximation

The proof of Lemma 3 about approximating the function $\widehat{L}^{\mathbf{y}}_{n,m}(f)$ is conducted by approximating the ridge functions $l_n(\langle y_i, x\rangle)$ in (3.9) with $y_i \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$ by deep CNNs. A key idea in our analysis is to use the inner product form (3.3) of the reproducing kernel of $\mathcal{H}^d_n$ and then to apply convolutional factorizations to realize the generated linear features, which enables us to conduct analysis after removing the restriction on large regularity index r. This idea might be applied to some other learning theory problems [8, 10, 17, 34].

We first apply the following two lemmas proved in [32] implying that the linear function $\langle y_i, x\rangle = y_i \cdot x$ can be realized by deep CNNs by factorizing y_i regarded as a sequence into convolutions of filters supported in $\{0, 1, \ldots, S\}$.

Lemma 5. *Let $S \geq 2$ and $W = (W_k)_{k=-\infty}^{\infty}$ be a sequence supported in $\{0, \cdots, \mathcal{M}\}$ with $\mathcal{M} \geq 0$. Then there exists a finite sequence of filters $\left\{w^{(j)}\right\}_{j=1}^p$ each supported in $\{0, \cdots, S\}$ with $p \leq \lceil \frac{\mathcal{M}}{S-1}\rceil$ such that the following convolutional factorization holds true*

$$W = w^{(p)} * w^{(p-1)} * \cdots * w^{(2)} * w^{(1)}.$$

Lemma 6. *Let $\{w^{(k)}\}_{k=1}^J$ be a set of sequences supported in $\{0, 1, \ldots, S\}$. Then*

$$T^{(J)} \cdots T^{(2)}T^{(1)} = T^{(J,1)} := (W_{i-k})_{i=1,\ldots,d+JS,k=1,\ldots,d} \in \mathbb{R}^{(d+JS)\times d} \qquad (5.1)$$

is a Toeplitz matrix associated with the filter $W = w^{(J)} * \cdots * w^{(2)} * w^{(1)}$ *supported in* $\{0, 1, \cdots, JS\}$*.*

We then construct a fully connected layer to approximate the univariate function l_n by continuous piecewise linear functions (splines) spanned by $\{\sigma(\cdot - t_i)\}_{i=1}^{N}$ with $t_i = -1 + \frac{i-2}{N}$, based on the following well known result in approximation by splines which can be found in [7] and [31, Lemma 6].

Lemma 7. *Given an integer* N*, let* $\boldsymbol{t} = \{t_i\}_{i=1}^{2N+3}$ *be the uniform mesh on* $\left[-1 - \frac{1}{N}, 1 + \frac{1}{N}\right]$ *with* $t_i = -1 + \frac{i-2}{N}$*. Construct a linear operator* $L_{\boldsymbol{t}}$ *on* $C[-1, 1]$ *by*

$$L_{\boldsymbol{t}}(f)(u) = \sum_{i=2}^{2N+2} f(t_i)\delta_i(u), \quad u \in [-1, 1], \, f \in C[-1, 1],$$

where $\delta_i \in C(\mathbb{R})$*,* $i = 2, \ldots, 2N + 2$*, is given by*

$$\delta_i(u) = N(\sigma\left(u - t_{i-1}\right) - 2\sigma\left(u - t_i\right) + \sigma\left(u - t_{i+1}\right)). \tag{5.2}$$

Then for $g \in C[-1, 1]$*,* $\|L_{\mathbf{t}}(g)\|_{C[-1,1]} \leq \|g\|_{C[-1,1]}$ *and*

$$\|L_{\mathbf{t}}(g) - g\|_{C[-1,1]} \leq 2\omega\left(g, 1/N\right)$$

where $\omega(g, \mu)$ *is the modulus of continuity of* g *given by*

$$\omega(g, \mu) = \sup_{|t| \leq \mu} \Big\{ |g(v) - g(v + t)| : v, v + t \in [-1, 1] \Big\}.$$

For the convenience of counting free parameter numbers, we introduce a linear operator $\mathcal{L}_N : \mathbb{R}^{2N+1} \to \mathbb{R}^{2N+3}$ given for $\zeta = (\zeta_i)_{i=1}^{2N+1} \in \mathbb{R}^{2N+1}$ by

$$\left(\mathcal{L}_N(\zeta)\right)_i = \begin{cases} \zeta_2, & \text{for } i = 1, \\ \zeta_3 - 2\zeta_2, & \text{for } i = 2, \\ \zeta_{i-1} - 2\zeta_i + \zeta_{i+1}, & \text{for } 3 \leq i \leq 2N + 1, \\ \zeta_{2N+1} - 2\zeta_{2N+2}, & \text{for } i = 2N + 2, \\ \zeta_{2N+2}, & \text{for } i = 2N + 3. \end{cases} \tag{5.3}$$

An important property of the operator $\mathcal{L}_N$ is to express the approximation operator $L_{\mathbf{t}}$ on $C[-1, 1]$ in terms of $\{\sigma\left(\cdot - t_j\right)\}_{j=1}^{2N+3}$ as

$$L_{\mathbf{t}}(f) = N \sum_{i=1}^{2N+3} \left(\mathcal{L}_N\left(\{f(t_k)\}_{k=2}^{2N+2}\right)\right)_i \sigma\left(\cdot - t_i\right), \quad \forall f \in C[-1, 1]. \tag{5.4}$$

Proof of Lemma 3. For $m \in \mathbb{N}$ and $\mathbf{y} = \{y_1, \ldots, y_m\} \subset \mathbb{S}^{d-1}$, we take W to be a sequence supported in $\{0, \cdots, md-1\}$ given by $W_{(j-1)d+(d-i)} = (y_j)_i$ where $j \in \{1, \cdots, m\}$ and $i \in \{1, \cdots, d\}$. By Lemma 5 with $\mathcal{M} = md - 1$, there exists a sequence of filters $\mathbf{w} = \left\{w^{(j)}\right\}_{j=1}^{J}$ supported in $\{0, \cdots, S\}$ with $J \geq \lceil \frac{\mathcal{M}}{S-1} \rceil$ satisfying the convolutional factorization $W = w^{(J)} * w^{(J-1)} * \cdots * w^{(2)} * w^{(1)}$. Here for $j = p+1, \ldots, J$, we have taken $w^{(j)}$ to be the delta sequence δ_0 given by $(\delta_0)_0 = 1$ and $(\delta_0)_k = 0$ for $k \in \mathbb{Z} \setminus \{0\}$. By Lemma 6, we have

$$T^{(J)}T^{(J-1)} \cdots T^{(1)} = T^{(J,1)} = (W_{i-k})_{i=1,\ldots,d+JS,k=1,\ldots,d} \in \mathbb{R}^{(d+JS)\times d},$$

where $T^{(j)}$ is the Toeplitz matrix with filter $w^{(j)}$ for $j = 1, 2, \ldots, J$.

Now we construct bias vectors in the neural networks. We denote $\|w\|_1 = \sum_{k=-\infty}^{\infty} |w_k|$. Take $b^{(1)} = -\left\|w^{(1)}\right\|_1 \mathbf{1}_{d_0}$ and

$$b^{(j)} = \left(\Pi_{p=1}^{j-1} \left\|w^{(p)}\right\|_1\right) T^{(j)} \mathbf{1}_{d_{j-1}} - \left(\Pi_{p=1}^{j} \left\|w^{(p)}\right\|_1\right) \mathbf{1}_{d_{j-1}+S}, \tag{5.5}$$

for $j = 2, \cdots, J$. The bias vectors satisfy $b_{S+1}^{(j)} = \ldots = b_{d_j-S}^{(j)}$. Observe that $\|x\|_\infty \leq 1$ for $x \in \mathbb{S}^{d-1}$. Denote $\|h\|_\infty = \max\{\|h_j\|_\infty : j = 1, \ldots, q\}$ for a vector of functions $h : \mathbb{S}^{d-1} \to \mathbb{R}^q$. We know that for $h : \mathbb{S}^{d-1} \to \mathbb{R}^{d_{j-1}}$,

$$\left\|T^{(j)} h\right\|_\infty \leq \left\|w^{(j)}\right\|_1 \|h\|_\infty.$$

Hence the components of $h^{(J)}(x)$ satisfy

$$\left(h^{(J)}(x)\right)_{kd} = \langle y_k, x\rangle + B^{(J)}, \qquad k = 1, \ldots, m,$$

where $B^{(J)} = \Pi_{p=1}^{J} \left\|w^{(p)}\right\|_1$. Applying the downsampling operator (1.6) leads to

$$\mathfrak{D}_d\left(h^{(J)}(x)\right) = \begin{bmatrix} \langle y_1, x\rangle \\ \vdots \\ \langle y_m, x\rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix} + B^{(J)} \mathbf{1}_{\lfloor (d+JS)/d \rfloor}.$$

Denote $\widehat{d} = \lfloor (d+JS)/d \rfloor$. Since $J \geq \lceil \frac{md-1}{S-1} \rceil$, we have

$$\frac{d+JS}{d} \geq 1 + \frac{md-1}{d} \frac{S}{S-1} > 1 + \frac{md-1}{d} \geq m.$$

Hence $\widehat{d} \geq m$.

We turn to expressing the last two fully connected layers. Of them, $h^{(J+1)}$ is given by

$$h^{(J+1)}(x) = \sigma(F^{(J+1)}\mathfrak{D}_d(h^{(J)}(x)) - b^{(J+1)})$$

with the connection matrix $F^{(J+1)} = \Xi_{\mathcal{D}_2,\mathbf{1}_{2N+3}}$ stated in (2.3) and the bias vector

$$b^{(J+1)}_{(j-1)(2N+3)+i} = \begin{cases} B^{(J)} + t_i, & \text{if } j = 1, \ldots, m,\ i = 1, \ldots, 2N+3, \\ B^{(J)} + 1, & \text{if } j > m, \end{cases} \tag{5.6}$$

where $\mathbf{t} := \{t_1 < \cdots < t_{2N+3}\}$ is given in Lemma 7. Then the first fully-connected layer $h^{(J+1)}(x) \in \mathbb{R}^{\hat{d}(2N+3)}$ of the deep network is

$$\left(h^{(J+1)}\right)_{(j-1)(2N+3)+i} = \left\{ \begin{array}{ll} \sigma\left(\langle y_j, \cdot\rangle - t_i\right), & \text{if } j \le m,\ 1 \le i \le 2N+3, \\ 0, & \text{if } j > m. \end{array}\right. \tag{5.7}$$

Write $h^{(J+1)}(x) \in \mathbb{R}^{\hat{d}(2N+3)}$ in a block form with $\hat{d}$ blocks of equal size $2N+3$, then the j-th block is $[\sigma\left(\langle y_j, x\rangle - t_i\right)]_{i=1}^{2N+3}$ for $j = 1, \ldots, m$, while the other blocks are zero vectors.

Take the vector $\Theta_N \in \mathbb{R}^{2N+3}$ in the connection matrix $F^{(J+2)} = \Xi^T_{\mathcal{D}_2,\Theta_N}$ of the second fully-connected layer stated in (2.3) in terms of the linear operator $\mathcal{L}_N$ as

$$\Theta_N = \mathcal{L}_N\left(\{\zeta_{n,r}(t_i)\}_{i=2}^{2N+2}\right),$$

then by the identity (5.4), for $j = 1, \ldots, m$, the jth entry of the product $F^{(J+2)}h^{(J+1)}(x)$ equals

$$\begin{aligned} \mathbf{\Theta}_N^T\left[\sigma\left(\langle y_j, x\rangle - t_i\right)\right]_{i=1}^{2N+3} &= \sum_{i=1}^{2N+3}\left(\mathcal{L}_N\left(\{\zeta_{n,r}(t_i)\}_{i=2}^{2N+2}\right)\right)_i \sigma\left(\langle y_j, x\rangle - t_i\right) \\ &= \frac{1}{N} L_{\mathbf{t}}\left(\zeta_{n,r}\right)\left(\langle y_j, x\rangle\right). \end{aligned}$$

The other entries of the product $F^{(J+2)}h^{(J+1)}(x)$ vanish. Thus, by taking $B^{(J+2)} = \|\zeta_{n,r}\|_{C[-1,1]}$ and

$$b^{(J+2)} = \left[\begin{array}{c} -\frac{B^{(J+2)}}{N}\mathbf{1}_m \\ O \end{array}\right], \tag{5.8}$$

we see from the homogenous property $\sigma(u/N) = \sigma(u)/N$ that the last layer $h^{(J+2)}$ is given by

$$h^{(J+2)}(x) = \frac{1}{N}\left[\begin{array}{c} \left[L_{\mathbf{t}}\left(\zeta_{n,r}\right)\left(\langle y_j, x\rangle\right) + B^{(J+2)}\right]_{j=1}^m \\ O \end{array}\right].$$

For the coefficients we choose $c^{(J+2)} \in \mathbb{R}^{\hat{d}}$ as

$$c_j^{(J+2)} = \begin{cases} \frac{N}{m} F_r(y_j), & \text{if } j = 1, \dots, m, \\ 0, & \text{otherwise} \end{cases}$$

and $A = B^{(J+2)} \frac{1}{m} \sum_{j=1}^m F_r(y_j)$. Then we have that for $x \in \mathbb{S}^{d-1}$,

$$\begin{aligned} & \left| \widehat{L}_{n,m}^{\mathbf{y}}(f)(x) - c^{(J+2)} \cdot h^{(J+2)}(x) - A \right| \\ = & \left| \frac{1}{m} \sum_{j=1}^m F_r(y_j) \zeta_{n,r}(\langle y_j, x \rangle) - \frac{1}{m} \sum_{j=1}^m F_r(y_j) L_{\mathbf{t}} \left(\zeta_{n,r} \right) \left(\langle y_j, x \rangle \right) \right| \\ \leq & \|f\|_{W_\infty^r(\mathbb{S}^{d-1})} \|\zeta_{n,r} - L_{\mathbf{t}}(\zeta_{n,r})\|_{C[-1,1]}. \end{aligned} \tag{5.9}$$

Since $\zeta_{n,r}$ is an algebraic polynomial of degree at most $2n$, by Markov's inequality,

$$\|\zeta_{n,r}'\|_{C[-1,1]} \leq (2n)^2 \|\zeta_{n,r}\|_{C[-1,1]}.$$

Combining this with the bound $\|\zeta_{n,r}\|_{C[-1,1]} \leq \sum_{k=1}^{2n} k^{-r} N(k,d)$ followed from Corollary 1.2.7 of [6], we know that

$$\|\zeta_{n,r}'\|_{C([-1,1])} \leq c_d n^2 \sum_{k=1}^{2n} k^{d-2-r},$$

where c_d is a constant depending only on d. But

$$\sum_{k=1}^{2n} k^{d-2-r} \leq 1 + \begin{cases} \frac{3^{d-1-r}}{d-1-r} n^{d-1-r}, & \text{if } d-2-r > -1, \\ 1 + \log(n+1), & \text{if } d-2-r = -1, \\ \frac{1}{r+1-d}, & \text{if } d-2-r < -1, \end{cases}$$

which is bounded by $c_{r,d}'' \Lambda_{d-1-r}(n)$ with a positive constant $c_{r,d}''$ depending only on r and d. It follows that $\omega(\zeta_{n,r}, \frac{1}{N}) \leq c_d c_{r,d}'' n^2 \Lambda_{d-1-r}(n)/N$. Combining this with (5.9), Lemma 7 and the embedding Proposition 1 yields

$$\left\| \widehat{L}_{n,m}^{\mathbf{y}}(f)(x) - c^{(J+2)} \cdot h^{(J+2)}(x) - A \right\|_\infty \leq c_{r,d}' \frac{n^2 \Lambda_{d-1-r}(n)}{N} \|f\|_{W_\infty^r(\mathbb{S}^{d-1})},$$

where $c_{r,d}'$ is a constant depending only on r and d.

The total number of free parameters $\mathcal{N}$ in our network is the sum of $J(S+1)$ contributed by $\mathbf{w}$, $J(2S+1)$ by the bias vectors in the first J layers, $2N+1$ contributed by the vector $\{\zeta_{n,r}(t_i)\}_{i=2}^{2N+2}$ in choosing Θ_N, 2 by the parameters $B^{(J)}$, $B^{(J+2)}$ in the fully-connected layers, and at most $m+1$ by $c^{(J+2)}$ and A. So it can be bounded as

$$\mathcal{N} \leq J(S+1) + J(2S+1) + 2N + 1 + 2 + m + 1 \leq J(3S+2) + m + 2N + 4.$$

This proves Lemma 3. □

5.3 General error bounds

With the proved bounds for $\|f - L_n(f)\|_\infty$ in Lemma 1, $\|L_n(f) - \widehat{L}_{n,m}^{\mathbf{y}}(f)\|_\infty$ in Lemma 2, and $\|\widehat{L}_{n,m}^{\mathbf{y}}(f) - \hat{f}\|_\infty$ in Lemma 3, the following bounds for the error $f - \hat{f}$ follows immediately.

Theorem 3. *Let $2 \leq S \leq d$, $d \geq 3$, $r > 0$, $\varepsilon > 0$, $m, n, N \in \mathbb{N}$ and $f \in W_\infty^r(\mathbb{S}^{d-1})$. Let $J \geq \lceil \frac{md-1}{S-1} \rceil$, $\mathcal{D}_1 = (2N+3)\lfloor (d+JS)/d \rfloor$ and $\mathcal{D}_2 = \lfloor (d+JS)/d \rfloor$. Then for the network constructed in Lemma 3 there exists a function $\hat{f} \in \mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S}$ such that*

$$\left\| f - \hat{f} \right\|_\infty \leq C'_{r,d,\varepsilon} \left(n^{-r} + \frac{\sqrt{\Lambda_{2(d-1-r+\varepsilon)}(n)}}{\sqrt{m}} + \frac{n^2 \Lambda_{d-1-r}(n)}{N} \right) \|f\|_{W_\infty^r(\mathbb{S}^{d-1})}, \tag{5.10}$$

where $C'_{r,d,\varepsilon}$ is a constant depending only on r, d, ε. Moreover, the total number of free parameters $\mathcal{N}$ in the network can be bounded as

$$\mathcal{N} \leq J(3S+2) + m + 2N + 4.$$

5.4 Proving the main results

We are in a position to derive our main results from the general error bounds in Theorem 3.

Proof of Theorem 1. Since $J \geq \frac{d-1}{S-1}$, we know that $\frac{(S-1)J+1}{d} \geq 1$. Take $m = \lfloor \frac{(S-1)J+1}{d} \rfloor$. Then $m \in \mathbb{N}$ and $md - 1 \leq (S-1)J$. Hence the requirement $J \geq \lceil \frac{md-1}{S-1} \rceil$ in Theorem 3 is valid.

Now we take n, N as

$$\begin{cases} n = \lfloor m^{\frac{1}{2(d-1+\varepsilon)}} \rfloor \text{ and } N = n^{d+1}, & \text{if } 0 < r < d-1, \\ n = \lfloor m^{\frac{1}{2r}} \rfloor \text{ and } N = \lfloor n^{2+r} \rfloor, & \text{if } 0 < \varepsilon < r - (d-1). \end{cases}$$

Then we know by Theorem 3 that with $\mathcal{D}_1 = (2N+3)\lfloor (d+JS)/d \rfloor$ and $\mathcal{D}_2 = \lfloor (d+JS)/d \rfloor$, there exists a network constructed in Lemma 3 containing a function $\hat{f} \in \mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S}$ such that

$$\left\| f - \hat{f} \right\|_\infty \leq \left(2^{r+2} + 1 \right) C'_{r,d,\varepsilon} m^{-\min\left\{ \frac{r}{2(d-1+\varepsilon)}, \frac{1}{2} \right\}} \|f\|_{W_\infty^r(\mathbb{S}^{d-1})}.$$

But $m \geq \frac{(S-1)J+1}{2d} > \frac{(S-1)J}{2d}$. So we have

$$\left\| f - \hat{f} \right\|_\infty \leq C_{r,d,\varepsilon,S} J^{-\min\left\{ \frac{r}{2(d-1+\varepsilon)}, \frac{1}{2} \right\}} \|f\|_{W_\infty^r(\mathbb{S}^{d-1})}.$$

with the constant

$$C_{r,d,\varepsilon,S} := \left(2^{r+2}+1\right) C'_{r,d,\varepsilon} \left(2d/(S-1)\right)^{\min\left\{\frac{r}{2(d-1+\varepsilon)}, \frac{1}{2}\right\}}.$$

This yields the desired error bound.

Observe that $m = \lfloor \frac{(S-1)J+1}{d} \rfloor \leq \frac{S}{d} J \leq J$ and

$$N \leq \begin{cases} m^{\frac{d+1}{2(d-1+\varepsilon)}} \leq J^{\frac{d+1}{2(d-1+\varepsilon)}}, & \text{if } 0 < r < d-1, \\ n^{2+r} \leq m^{\frac{2+r}{2r}} \leq J^{\frac{1}{2}+\frac{1}{r}}, & \text{if } 0 < \varepsilon < r-(d-1). \end{cases}$$

But $d \geq 3$ implies $\frac{d+1}{2(d-1+\varepsilon)} < 1$. In the case $0 < \varepsilon < r-(d-1)$ which implies $r > d-1+\varepsilon > 2$, we also have $\frac{1}{2}+\frac{1}{r} < 1$. So the total number of free parameters $\mathcal{N}$ in the network can be bounded as

$$\mathcal{N} \leq (3S+5)\, J + 4.$$

The proof of Theorem 1 is complete. □

Remark 3. *When $r = d-1$, from the above proof, we can see by taking $N = \lfloor n^{d+1} \log(n+1) \rfloor$ that the statement of Theorem 1 still holds except that the bound for the number of free parameters should be replaced by $\mathcal{N} \leq (3S+3)\, J + 2J^{\frac{d+1}{2(d-1+\varepsilon)}} \log(J+1) + 4$. Note that $\mathcal{N} = \mathcal{O}(J)$. So to achieve the approximation accuracy $\epsilon > 0$, the depth and the number of free parameters of the network are of orders $\mathcal{O}\left(\epsilon^{-2-\frac{2}{d-1}\varepsilon}\right)$.*

Proof of Theorem 2. We follow the proof of Lemma 3 and construct deep CNNs of depth $J = \left\lceil \frac{md-1}{S-1} \right\rceil$ with the m features $\{y_j \in \mathbb{S}^{d-1}\}_{j=1}^m$ in the additive ridge form (2.4) of the approximated function f, followed by downsampling and one fully-connected layer which produces $h^{(J+1)}(x) \in \mathbb{R}^{\hat{d}(2N+3)}$ expressed by (5.7). Then by making use of the univariate functions $\{g_j\}_{j=1}^m$ in the additive ridge form (2.4) of the approximated function f, we choose the coefficient vector $c^{(J+1)} \in \mathbb{R}^{\hat{d}(2N+3)}$ by means of the linear operator $\mathcal{L}_N$ as

$$\left\{ \left(c^{(J+1)}\right)_{(j-1)(2N+3)+i} \right\}_{i=1}^{2N+3} = N \mathcal{L}_N \left(\{g_j(t_i)\}_{i=2}^{2N+2} \right), \qquad j = 1, \ldots, m$$

and $\left(c^{(J+1)}\right)_{(j-1)(2N+3)+i} = 0$ for $j > m$. Then by the identity (5.4), we have

$$\begin{aligned} c^{(J+1)} \cdot h^{(J+1)}(x) &= N \sum_{j=1}^{m} \sum_{i=1}^{2N+3} \left(c^{(J+1)}\right)_{(j-1)(2N+3)+i} \sigma\left(\langle y_j, x\rangle - t_i\right) \\ &= \sum_{j=1}^{m} L_{\mathbf{t}}\left(g_j\right)\left(\langle y_j, x\rangle\right). \end{aligned}$$

Combining this with the additive ridge form (2.4) of f and Lemma 7, we know that for $x \in \mathbb{S}^{d-1}$,

$$\left|f(x) - c^{(J+1)} \cdot h^{(J+1)}(x)\right| = \left|\sum_{j=1}^{m} g_j(\langle y_j, x\rangle) - \sum_{j=1}^{m} L_{\mathbf{t}}\left(g_j\right)\left(\langle y_j, x\rangle\right)\right|$$

$$\leq \sum_{j=1}^{m} \|g_j - L_{\mathbf{t}}(g_j)\|_{C[-1,1]} \leq \sum_{j=1}^{m} |g_j|_{W_\infty^\alpha} N^{-\alpha}.$$

Then the desired error bound is verified.

The total number of free parameters $\mathcal{N}$ in the network is the sum of $J(S+1)$ contributed by $\mathbf{w}$, $J(2S+1)$ by the bias vectors in the first J layers, 1 by the parameter $B^{(J)}$ in the fully-connected layer, and $2N+1$ by the vector $\{g_j(t_i)\}_{i=2}^{2N+2}$ in choosing the coefficient vector $c^{(J+1)}$. So it can be bounded as

$$\mathcal{N} \leq J(S+1) + J(2S+1) + 1 + 2N + 1 \leq J(3S+2) + 2N + 2.$$

This proves Theorem 2. □

6 Conclusion and Discussion

In this paper spherical harmonic analysis is conducted rigorously for the approximation theory of deep CNNs followed by downsampling and one fully connected layer or two on spheres. Our analysis provides rates of uniformly approximating functions $f \in W_\infty^r(\mathbb{S}^{d-1})$ with $r > 0$ by deep CNNs followed by two fully connected layers. To approximate a Lipschitz function in a special additive ridge form, a network with one fully connected layer can be as fast as one for approximating a univariate Lipschitz function, which demonstrates the super power of deep CNNs in approximating or representing functions with special structures. Our spherical analysis relies on a special property of the reproducing kernel of $\mathcal{H}_n^d$ on the sphere. It would be interesting to extend our technique to approximation of non-smooth functions on $[-1,1]^d$ and to L_p approximation with $1 \leq p < \infty$.

Acknowledgments

The authors would like to thank the referees for their encouraging comments and constructive suggestions that led to an improved presentation of this paper. The second author is supported partially by the Research Grants

Council of Hong Kong [Project # CityU 21207019] and by City University of Hong Kong [Project # CityU 7200608]. The last author is supported partially by the Research Grants Council of Hong Kong [Project # CityU 11306318].

References

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

[2] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, Optimal approximation with sparsely connected deep neural networks, SIAM Journal on Mathematics of Data Science **1** (2019), 8–45.

[3] J. Bruna and S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013), 1872–1886.

[4] A. Christmann and D. X. Zhou, Learning rates for the risk of kernel-based quantile regression estimators in additive models, Anal. Appl. **14** (2016), 449–477.

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

[6] F. Dai and Y. Xu, Approximation Theory and Harmonic Analysis on Spheres and Balls, volume 27, Srpinger Monographs in Mathematics, Springer New York Heidelberg Dordrecht London, 2013.

[7] R. A. DeVore and G. G. Lorentz, Constructive Approximation, Springer-Verlag, Berlin, Heidelberg, 1993.

[8] J. Fan, T. Hu, Q. Wu and D. X. Zhou, Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmonic Anal. **41** (2016), 164-189.

[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[10] Z. C. Guo, D. H. Xiang, X. Guo, and D. X. Zhou, Thresholded spectral algorithms for sparse approximations, Anal. Appl. **15** (2017), 433–455.

[11] K. Hesse, A lower bound for the worst-case cubature error on spheres of arbitrary dimension, Numerische Mathematik **103** (2006), 413–433.

[12] G. E. Hinton, S. Osindero, Y. W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. **18** (2006), 1527-1554.

[13] M. Imaizumi and K. Fukumizu, Deep neural networks learn non-smooth functions effectively, in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.

[14] A. I. Kamzolov, The best approximation of the classes of functions $w_p^\alpha(S^n)$ by polynomials in spherical harmonics, Mathematical Notes of the Academy of Sciences of the USSR, **32** (1982), 622–626.

[15] J. Klusowski and A. Barron, Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls, IEEE Transactions on Information Theory **64** (2018), 7649–7656.

[16] A. Krizhevsky, I. Sutskever, and G. Hinton G, Imagenet classification with deep convolutional neural networks, *NIPS* (2012): 1097-1105.

[17] S. B. Lin and D. X. Zhou, Distributed kernel gradient descent algorithms, Constr. Approx. **47** (2018), 249-276.

[18] S. Mallat, Understanding deep convolutional networks, Phil. Trans. Royal Soc. A **374**:20150203.

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

[20] R. Nakada and M. Imaizumi, Adaptive approximation and estimation of deep neural network to intrinsic dimensionality, arXiv preprint arXiv: 1907.02177, 2019.

[21] K. Oono and T. Suzuki, Approximation and non-parametric estimation of ResNet-type convolutional neural networks, in Proceedings of the 36th International Conference on Machine Learning (PMLR) 97:4922-4931, 2019.

[22] P. Petersen and V. Voigtlaender, Optimal approximation of piecewise smooth functions using deep ReLU neural networks, Neural Networks **108** (2018), 296–330.

[23] P. Petersen and F. Voigtlaender, Equivalence of approximation by convolutional neural networks and fully-connected networks, Proceedings of the American Mathematical Society **148** (2020), 1567-1581.

[24] J. Schmidt-Hieber, Nonparametric regression using deep neural networks with ReLU activation function, arXiv preprint arXiv: 1708.06633, 2017.

[25] U. Shaham, A. Cloninger, and R. Coifman, Provable approximation properties for deep neural networks, Appl. Comput. Harmonic Anal. **44** (2018), 537–557.

[26] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. **26** (2007), 153–172.

[27] S. Sonoda and N. Murata, Neural network with unbounded activation functions is universal approximator, Applied and Computational Harmonic Analysis, **43** (2017), 233-268.

[28] T. Suzuki, Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality, in Proceedings of the International Conference on Learning Representations (ICLR), 2019.

[29] Y. G. Wang, Q. Le Gia, I. Sloan, and R. Womersley, Fully discrete needlet approximation on the sphere, Appl. Comput. Harmonic Anal. 43(2):292–316, 2017.

[30] D. Yarotsky, Error bounds for approximations with deep ReLU networks, Neural Networks **94** (2017), 103–114.

[31] D. X. Zhou, Deep distributed convolutional neural networks: universality, Anal. Appl. **16** (2018), 895–919.

[32] D. X. Zhou, Universality of deep convolutional neural networks, Appl. Comput. Harmonic Anal. **48** (2020), 787-794.

[33] D. X. Zhou, Theory of deep convolutional neural networks: Downsampling, Neural Networks **124** (2020), 319-327.

[34] D. X. Zhou, Distributed approximation with deep convolutional neural networks, preprint.