Neurocomputing 175 (2016) 348–354

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Simultaneous approximation by spherical neural networks[☆]

Shaobo Lin [a,*], Feilong Cao [b]

[a] College of Mathematics and Information Science, Wenzhou University, Wenzhou, 325035, PR China
[b] Institute of Metrology and Computational Science, China Jiliang University, Hangzhou 310018, PR China

ARTICLE INFO

Article history:
Received 22 May 2015
Received in revised form
4 August 2015
Accepted 19 October 2015
Communicated by M. Bianchini
Available online 30 October 2015

Keywords:
Spherical neural networks
Simultaneous approximation
Laplace–Beltrami operator
Spherical polynomials

ABSTRACT

Approximation capabilities of the spherical neural networks (SNNs) are considered in this paper. Based on a known Taylor formula, we prove that, for non-polynomial target function, rates of simultaneously approximating the function itself and its (Laplace–Beltrami) derivatives by SNNs is not slower than those by the spherical polynomials (SPs). Then, the simultaneous approximation rates of SPs automatically derive the rates of SNNs.

1. Introduction

Spherical data abound in geodesy, meteorology, astrophysics, geophysics, and other areas [8,9]. For example, the mathematical models of some satellite missions such as the GOCE (Gravity Field and Steady-State Ocean Circulation Explorer) and the CHAMP (Challenging Mini-Satellite Payload for Geophysical Research and Application) which study the gravity potential of the earth are spherical Fredholm integral equations of the first kind [11]. Hence, finding a tool which can efficiently tackle spherical data becomes more and more important.

A basic and classical tool for fitting scattered data on the sphere is the spherical polynomial (SP). Up till now, the approximation capability of SPs has been widely studied [36]. For example, Ditzian [7] deduced a Jackson-type error estimate and its Stechkin-type inverse for SPs; Sloan [34] constructed a hyperinterpolation operator, which is an SP, and deduced the approximation error bound; Dai [5] provided a weighted Jackson inequality for SPs. The main algorithm to implement the SP approximation is the singular value decomposition approach [19]: Expand a function with respect to the orthonormal basis and estimate the corresponding Fourier coefficients. But it is well-known that the spherical harmonic basis is badly localized and incapable of representing local features of the target function, which is important in geophysics applications [10]. Therefore, one turns to find a tool which possesses nice localization performance. Consequently, spherical basis function (SBF) and spherical neural networks (SNNs) come into our sights.

SBF refers to a positive definite function on the $(d+1)$-dimensional unit sphere $\mathbf{S}^d$. Here a positive definite function on $\mathbf{S}^d$ means the matrix $A_\phi:=(\phi(\langle x_i,x_j\rangle))_{i,j=1}^M$ is positive definite [17, Def. 2.7]. The SBF method focuses on using the linear combination of shifts of SBF on the spherical data. Mathematically, the approximant is formed as

$$x\mapsto\sum_{i=1}^{M}c_i\phi(\langle x_i,x\rangle),\quad x\in\mathbf{S}^d,\quad c_i\in\mathbf{R},\tag{1.1}$$

where ϕ is an SBF, x_i is the spherical data and $\langle x,y\rangle$ denotes the inner product in $\mathbf{R}^{d+1}$. There are two topics on the SBF approximation. The first one is the density problem which concerns whether the approximant (1.1) can approximate arbitrary continuous function within arbitrary accuracy, provided the number of spherical data is sufficiently large and the coefficients $\{c_i\}_{i=1}^M$ are appropriately tuned. In the seminal paper [35], Sun and Cheney derived the sufficient and necessary conditions for the density of SBF approximation. The other one called the complexity problem is to determine how many samples are necessary to yield a prescribed degree of approximation by using the SBF approximant (1.1). For this problem, Mhaskar et al. [27] gave an upper bound of approximation by using the positive cubature and Marcinkiewicz–Zygmund inequality. They utilized the summation of the best approximation error of SPs and a residual depending on the smoothness of the SBF to bound the approximation error of the SBF approximant (1.1). Some studies for the SBF approximation can also be found in [4,14,16,21,30,32,33] and references therein.

[☆] The research was supported by the National Natural Science Foundation of China (Grant no. 61502342, 11401462, 91330118, 61272023).

[*] Corresponding author.
E-mail address: sblin1983@gmail.com (S. Lin).

http://dx.doi.org/10.1016/j.neucom.2015.10.067
0925-2312/© 2015 Elsevier B.V. All rights reserved.

A typical way to derive the SBF approximant is to use the least squares, which has already been proved to posses perfect approximation capability [12].

Besides the SBF method, [23] also proposed the SNN which formed as

$$\sum_{i=1}^{M} a_i \sigma(\langle w_i, x\rangle + b_i), \tag{1.2}$$

where, $w_i \in \mathbf{R}^{d+1}$, $a_i \in \mathbf{R}$, $b_i \in \mathbf{R}$ are the inner weight, outer weight, threshold of the SNN, respectively, and σ is named as the activation function in the terminology of neural networks [13]. It can be easily found that if σ is positive definite, w_i is the data point and $b_i{=}0$, then the SNN defined above coincides with the SBF approximant (1.1). In our previous paper [23], we theoretically proved that there exists an SNN which possesses essentially better approximation capability than the SP and SBF methods in the sense that SNN can deduce a similar approximation error by using much smaller M. However, the results in [23] do not present any hints in selecting the activation function and the related parameters. The aim of the present paper is to pursue the approximation capability of a large range of SNNs and provide a guidance about how to select the activation function and the corresponding parameters.

On the other hand, simultaneous approximation of a function and its derivatives are required in many science and engineering applications. There have been many studies on the problem of simultaneous approximation by neural networks on $\mathbf{R}^{d+1}$[1,22,37,38]. We also pursue the simultaneous approximation capability of SNN in this paper. By using a representation theorem and a Taylor formula, we firstly construct an SNN and use it and its (Laplace–Beltrami) derivatives to approximate SPs and their (Laplace–Beltrami) derivatives. After introducing the best approximation operator on the sphere and using the commutativity of the best approximation operator and the Laplace–Beltrami operator, we then derive the upper bound of simultaneous approximation error of SPs. Under this circumstance, a quantitative upper bound estimate on simultaneous approximation by SNNs can be derived. The obtained results reveal that the simultaneous approximation rate of the constructed SNN depends not only on the number of hidden units used, but also on the smoothness of functions to be approximated. Furthermore, it can be deduced that, for non-polynomial target functions the rate of simultaneous approximation by SNNs is not slower than that by SPs.

The paper is organized as follows. In the next section, some preliminaries together with a representation theorem of SPs will be given. In Section 3, the upper bound of simultaneous approximation by SNNs will be established. In the last sections, we will give the proofs of the main results.

2. Spherical harmonics

Denote by $L^p(\mathbf{S}^d)(1 \le p \le \infty)$ the space of p-th Lebesgue integrable functions on $\mathbf{S}^d$ endowed with the norms

$$\|f\| := \|f(\cdot)\|_{L^\infty(\mathbf{S}^d)} := \operatorname*{ess\,sup}_{x \in \mathbf{S}^d} |f(x)|, \quad p = \infty,$$

and

$$\|f\|_p := \|f(\cdot)\|_{L^p(\mathbf{S}^d)} : = \left\{ \int_{\mathbf{S}^d} |f(x)|^p d\omega(x) \right\}^{1/p} < \infty, \quad 1 \le p < \infty.$$

We denote by $d\omega$ the surface area element on $\mathbf{S}^d$. The volume of $\mathbf{S}^d$ is denoted by Ω_d, and it is easy to deduce that

$$\Omega_d := \int_{\mathbf{S}^d} d\omega = \frac{2\pi^{(d+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}.$$

For integer $k \ge 0$, the restriction to $\mathbf{S}^d$ of a homogeneous harmonic polynomial of degree k on the unit sphere is called a spherical harmonic of degree k. The span of all spherical harmonics of degree k is denoted by $\mathbf{H}_k^d$, and the class of all SPs of degree $k \le n$ is denoted by ${\Pi_n}^d$. It is obvious that $\Pi_n^d = \oplus_{k=0}^n \mathbf{H}_k^d$. The dimension of $\mathbf{H}_k^d$ is given by

$$D_k^d := \dim \mathbf{H}_k^d = \begin{cases} \dfrac{2k+d-1}{k+d-1}\binom{k+d-1}{k}, & k \ge 1; \\ 1, & k = 0, \end{cases}$$

and that of ${\Pi_n}^d$ is $\sum_{k=0}^n D_k^d = D_n^{d+1} \sim n^d$, where $A \sim B$ denotes that there exist absolute constants C_1 and C_2 such that $C_1 A \le B \le C_2 A$.

Spherical harmonics have an intrinsic characterization. To describe this, we first introduce the Laplace–Beltrami operator Δ, which is defined by [31]:

$$\Delta f := \sum_{i=1}^{d+1} \frac{\partial^2 g(x)}{\partial x_i^2}\bigg|_{|x| := \sqrt{x_1^2 + \cdots + x_{d+1}^2} = 1}, \quad g(x) = f\left(\frac{x}{|x|}\right).$$

It is well-known that the operator Δ is an elliptic, (unbounded) selfadjoint operator on $L^2(\mathbf{S}^d)$, is invariant under arbitrary coordinate changes, and its spectrum comprises distinct eigenvalues $-\lambda_k := -k(k+d-1)$, $k = 0, 1, \ldots$, each having finite multiplicity. The space $\mathbf{H}_k^d$ can be characterized intrinsically as the eigenspace corresponding to λ_k, i.e.

$$\Delta H_k = -\lambda_k H_k, \quad H_k \in \mathbf{H}_k^d. \tag{2.1}$$

Since λ_ks are distinct, and the operator is selfadjoint, the spaces $\mathbf{H}_k^d$ are mutually orthonormal; also, $L^2(\mathbf{S}^d) = \text{closure}\{\oplus_k \mathbf{H}_k^d\}$. Hence, if we choose an orthonormal basis $\{Y_{k,l} : l = 1, \ldots, D_k^d\}$ for each $\mathbf{H}_k^d$, then the set $\{Y_{k,l} : k = 0, 1, \ldots, l = 1, \ldots, D_k^d\}$ is an orthonormal basis for $L^2(\mathbf{S}^d)$.

The well-known addition formula [36] is given by

$$\sum_{l=1}^{D_k^d} Y_{k,l}(x) Y_{k,l}(y) = \frac{D_k^d}{\Omega_d} P_k^{d+1}(\langle x, y\rangle), \tag{2.2}$$

where P_k^{d+1} is the Legendre polynomial with degree k and dimension $d+1$. The Legendre polynomial P_k^{d+1} can be normalized such that $P_k^{d+1}(1) = 1$, and satisfies the orthogonality relation

$$\int_{-1}^{1} P_k^{d+1}(t) P_j^{d+1}(t)(1-t^2)^{(d-2)/2}\, dt = \frac{\Omega_d}{\Omega_{d-1} D_k^d} \delta_{k,j},$$

where $\delta_{k,j}$ is the usual Kronecker symbol.

The following Funk–Hecke formula [36] plays an important role in computing the eigenvalues of the kernel $\phi \in L^1([-1,1])$.

$$\int_{\mathbf{S}^d} \phi(\langle x, y\rangle) H_k(y)\, d\omega(y) = B(\phi, k) H_k(x), \tag{2.3}$$

where

$$B(\phi, k) := \Omega_{d-1} \int_{-1}^{1} P_k^{d+1}(t) \phi(t)(1-t^2)^{(d-2)/2}\, dt.$$

In order to reveal the simultaneous approximation capability of SNNs, we need the following representation theorem, which was proven in [23].

Lemma 1. *Let $n \in \mathbf{N}$. Then for any $P_n \in \Pi_n^d$, there exists a set of points $\{a_k\}_{k=1}^{D_n^d} \subset \mathbf{S}^d$ and a set of univariate polynomials $\{g_k\}_{k=1}^{D_n^d}$ defined on $[-1,1]$ with degrees not larger than n such that*

$$P_n(x) = \sum_{k=1}^{D_n^d} g_k(\langle a_k, x\rangle), \quad x \in \mathbf{S}^d. \tag{2.4}$$

From the classical representation theorem (Theorem 3 of [31]), we know that every SP can be represented by a combination of ridge functions. But the number of parameters of both sides of the representation theorem is not asymptotically equal (the left side is $\mathcal{O}(n^d)$ and the right side is $\mathcal{O}(n^{d+1})$). It can be found in the present representation theorem (2.4) that the number of parameters of both sides are asymptotically equal to n^d. This properties will play a crucial role in studying the simultaneous approximation capacity of SNNs.

3. Simultaneous approximation capability of SNNs

It was mentioned above that the Laplace–Beltrami operator plays an important role in spherical harmonics. Thus, we consider the upper bound of approximation of SPs and their (Laplace–Beltrami) derivatives by SNNs and their (Laplace–Beltrami) derivatives. The following three theorems are main results of this paper. Theorem 1 shows that if the degrees of SPs and M satisfy some assumptions, then SPs can be simultaneously approximated by SNNs in any desired accuracy. Theorem 2 shows that as far as the simultaneous approximation capability is concerned, SNNs is somewhat better than SPs. Theorem 3 establishes a Jackson type error estimate for simultaneous approximation by SNNs.

At first we introduce a Sobolev space $W_p^{2k}(\mathbf{S}^d)$ whose properties can be found in [24],

$$W_p^{2k}(\mathbf{S}^d)=\left\{f: f\in L^p(\mathbf{S}^d), \Delta^j f\in L^p(\mathbf{S}^d), \quad 1\le j\le k\right\}$$

To describe the approximation error, we need the following modulus of smoothness [6,7].

$$\omega_p^r(f,t):=\sup_{\rho\in O_t}\left\|E_\rho^r f\right\|_p,$$

where

$$O_t:=\left\{\rho\in SO(d): \max_{x\in\mathbf{S}^d}\arccos\langle x,\rho x\rangle\le t\right\},$$
$$E_\rho^r f:=(I-T_\rho)^r f,\ T_\rho f(x):=f(\rho x),$$

$SO(d+1)$ denotes the set of all orthogonal matrix with determinant 1, and I denotes the identity operator.

Now we are in a position to give our main results.

Theorem 1. *Let $n\in\mathbf{N}$ and $M:=(n+1)D_n^d$. Suppose that ϕ is a univariate function with up to $(n+1)$ times bounded derivatives, and is not a polynomial of degree at mostn. Then for any $P_n\in\Pi_n^d$, and any $\varepsilon>0$, there is an SNN,*

$$N_{\phi,M}(x)=\sum_{j=0}^{M}a_j\phi(\langle w_j,x\rangle+b_j),\ x\in\mathbf{S}^d,\ w_j\in\mathbf{R}^{d+1},\ b_j,a_j\in\mathbf{R},$$

such that

$$|\Delta^k P_n(x)-\Delta^k N_{\phi,M}(x)|<\varepsilon,\ k=0,\dots,[n/2],x\in\mathbf{S}^d, \tag{3.1}$$

where $\Delta^k:=\Delta^{k-1}\Delta$ and $[t]$ denotes the largest integer not larger than t.

Theorem 2. *Let $1\le p\le\infty$, $n\in\mathbf{N}$ and $M:=(n+1)D_n^d$. Suppose that ϕ is a univariate function with up to $(n+1)$ times bounded derivatives, and is not a polynomial of degree at most n. Then for every non-polynomial function f, there holds*

$$E^*_{M,\phi,p}(\Delta^k f)\le 2E_{n,p}(\Delta^k f), \tag{3.2}$$

where

$$E^*_{M,\phi,p}(\Delta^k f):=\inf_{a_k,b_k\in\mathbf{R},w_k\in\mathbf{R}^{d+1}}\left\|\Delta^k_{\cdot}f(\cdot)-\Delta^k_{\cdot}\sum_{k=1}^{M}a_k\phi(\langle w_k,\cdot\rangle+b_k)\right\|_p$$

denotes the best simultaneous approximation error of SNNs,

$$E_{n,p}(\Delta^k f):=\inf_{P\in\Pi_n^d}\|\Delta^k_{\cdot}f(\cdot)-\Delta^k_{\cdot}P(\cdot)\|_p$$

is the best simultaneous approximation error of SPs and Δ_x denotes that the Laplace–Beltrami operator acts on x.

Theorem 3. *Let $1\le p\le\infty$, $n\in\mathbf{N}$, $r\in\mathbf{N}$ and $M:=(n+1)D_n^d$. Suppose that ϕ is a univariate function with up to $(n+1)$ times bounded derivatives, and is not a polynomial of degree at most n. Then for any $f\in W_p^{2k}(\mathbf{S}^d)$, there exists an SNN,*

$$N_{M,\phi}(x):=\sum_{j=1}^{M}a_j\phi(\langle w_j\cdot x\rangle+b_j),\ w_j\in\mathbf{R}^{d+1},a_j,b_j\in\mathbf{R}$$

such that

$$\|\Delta^k f-\Delta^k N_{M,\phi}\|_p\le C\omega_p^r\left(\Delta^k f,\frac{1}{n}\right),\quad k=0,\dots,[n/2]. \tag{3.3}$$

Since $D_n^d\sim n^{d-1}$, we have $M=(n+1)D_n^d\sim n^d\sim D_n^{d+1}$. This means that the number of parameters needed in the SNN approximant is asymptotically equals to that of SPs. The above theorems yield that as far as the simultaneous approximation capability is concerned, SNNs are somewhat superior to SPs.

Now, we compare our results with some related works. The relationship between neural networks and polynomial approximation has already studied in [23] on the sphere and [20,25,26,37] in Euclidean space. As shown in [23, Theorem31], Lin et al. proved that there exists an SNN formed as (1.2) with $\mathcal{O}(n^{d-1})$ parameters such that its approximation error can be upper bounded by SPs with degree at most n. Since $\mathcal{O}(n^{d-1})$ is essentially smaller than $D_n^{d+1}\sim n^d$, the authors concluded that the approximation capability of SNNs is essentially better than that of SPs. However, it should be highlighted that we do not known which activation function is appropriate to generate the SNN with the aforementioned property. Thus, the result in [23] is just a theoretical implication and incapable for applications. Differently, the results in this paper focus on a large range of activation functions. In fact, we prove that for arbitrary analytical and non-polynomial activation function, if the parameters are appropriately tuned, then the approximation capability of the corresponding SNN isn't worse than that of SPs. It can be found in the next section that our proof is constructive, that is, we propose a concrete strategy on selecting parameters of SNN to guarantee its approximation capability. Furthermore, another important advantage of the present paper is that we are concerned with the simultaneous approximation. The difference between our work with the results in [20,25,26,37] can be concluded as follows. The underlining space in our work is the unit sphere, which can be regarded as the simplest Remmian manifold. It should be noted that the methodology in spherical data analysis and Euclidean data analysis are quite different [17]. Especially, the definitions of the derivatives are different, which leads to a totally different analysis of the simultaneous approximation.

There are also many papers [21,29,30,32] focusing on the topic of simultaneous approximation by SBFs (or Sobolev error estimates of SBFs). In the nice paper [30], under some mild assumptions of SBFs, a Sobolev error estimate for approximation by SBFs was established. The geometric distribution of the scattered data such as the mesh norm and mesh ratio are utilized to bound the approximation error. The novelty of the results obtained in this paper is based on the following three points. The first one is that SNN is a non-linear approximant, and the parameters of SNN can be obtained by some algorithms such as the backprogramme (BP) [13] or greedy algorithm [3]. Then we use the number of parameters instead of the geometric distribution of scattered data to describe the approximation error. The second one is that our results are somewhat shaper than that of SBFs. For this purpose, we encourage the readers to compare Theorem 2 in this paper with Theorem 6.7 of [30]. It can be found in Theorem 2 that, for non-polynomial functions, the best simultaneous approximation error of SNNs is not larger than that of SPs, while in [30], there is another residual. The last one is that our results also hold for arbitrary subset of $\mathbf{S}^d$, but the approximation error deduced for SBFs sometimes only holds for the whole sphere $\mathbf{S}^d$. The main reason is that the tools we employed are the representation Theorem (2.4) and Taylor formula, which are different from the widely used tools in the SBF approximation such as the spherical harmonic analysis and cubature formula.

At last, as a nonlinear approximant, we should present the relation between our results and some existing papers concerning the tractability of approximation. A consensus on nonlinear approximation is that it can break the "curse of dimensionality". The results in [2,18,28] verified this statement by deducing approximation rates at least $M^{-1/2}$, which is independent of d. The most important difference is that in [2,18,28], the target functions depend heavily on the activation functions of the neural networks, which is the key reason why the deduced approximation error was independent of d. However, it was pointed out in [3, P68] that the restrictions to the target functions in [2,18,28] may become more and more strong as the dimension d grows. Thus, although the approximation error is independent of the dimension, the applicable target functions become more and more stringent as d grows. Different from these results, the approximation results in this paper are established for functions in the Sobolev space, which is obviously independent of the activation function. It was proved in [25] that, for such functions, the approximation error depends heavily on d, no matter the approximant is linear or nonlinear.

4. Proofs

In this section, we give the proofs of the main results of this paper. The following Lemma 2 will play a key role in our proofs.

Lemma 2. *Let $y \in \mathbf{S}^d$ be fixed, $n \in \mathbf{N}$. Suppose that h is a univariate function with up to n times continuous derivatives, then there exists a constant $C(d,n)$ depending only on n and d such that*

$$\max_{x \in \mathbf{S}^d} \left| \Delta_x^k h(\langle x, y\rangle) \right| \le C(d,n) \sum_{j=0}^{2k} \max_{t \in [-1,1]} |h^{(j)}(t)|, \quad k = 0, \ldots, [n/2]. \tag{4.1}$$

Proof. The following local coordinates are a slight modification of the usual spherical coordinates:

$$x_1 = \sqrt{1-t^2}\ \sin\,\theta_{d-2} \ldots \sin\,\theta_2\ \sin\,\theta_1\ \sin\,\phi$$
$$x_2 = \sqrt{1-t^2}\,\sin\theta_{d-2} \ldots \sin\,\theta_2\ \sin\,\theta_1\ \cos\,\phi$$
$$x_3 = \sqrt{1-t^2}\ \sin\,\theta_{d-2} \ldots \sin\,\theta_2\ \cos\,\theta_1$$
$$\ldots$$
$$\ldots$$
$$\ldots$$
$$x_{d-1} = \sqrt{1-t^2}\ \sin\,\theta_{d-2}\ \cos\,\theta_{d-3}$$
$$x_d = \sqrt{1-t^2}\ \cos\,\theta_{d-2}$$
$$x_{d+1} = t,$$

where $\theta_1, \ldots, \theta_{d-2} \in [0, \pi]$, $t \in [-1,1]$, and $\phi \in [0, 2\pi)$. Then the Laplace–Beltrami operator Δ has, in the local coordinates, the following representation for a function g on $\mathbf{S}^d$ with $g(x) = G(\theta_1, \ldots, \theta_{d-2}, t, \phi)$ [15]

$$\begin{aligned}\Delta g = {} & \left[(-d)t\frac{\partial G}{\partial t} + (1-t^2)\frac{\partial^2 G}{\partial t^2}\right] + (1-t^2)^{-1}(\ \sin\,\theta_{d-2})^{2-d}\frac{\partial}{\partial \theta_{d-2}} \\ & \times\left[(\sin\theta_{d-2})^{d-2}\frac{\partial G}{\partial \theta_{d-2}}\right] + (1-t^2)^{-1}(\sin\theta_{d-2})^{-2} \\ & \times(\sin\theta_{d-3})^{3-d}\frac{\partial}{\partial \theta_{d-3}}\left[(\sin\theta_{d-3})^{d-3}\frac{\partial G}{\partial \theta_{d-3}}\right] \\ & + \cdots + (1-t^2)^{-1}(\sin\theta_{d-2} \ldots \sin\theta_2)^{-2} \\ & \times(\sin\theta_1)^{-1}\frac{\partial}{\partial \theta_1}\left[(\sin\theta_1)^1\frac{\partial G}{\partial \theta_1}\right] \\ & + (1-t^2)^{-1}(\sin\theta_{d-2} \ldots\ \sin\theta_1)^{-2}\frac{\partial^2 G}{\partial \phi^2}.\end{aligned}$$

Without loss of generality, we only prove (4.1) for $d=2$. When $d \ge 3$, by using the above representation of the Laplace–Beltrami operator and the same method, we can deduce (4.1) easily. From the above representation with $d=2$, we have

$$\Delta_x\langle x, y\rangle = -2\langle x, y\rangle.$$

Hence a direct computation yields that

$$\Delta_x h(\langle x, y\rangle) = -2\langle x, y\rangle h'(\langle x, y\rangle) + (1-(\langle x, y\rangle)^2)h''(\langle x, y\rangle).$$

Then

$$\begin{aligned}\Delta_x^k h(\langle x, y\rangle) = \Delta_x^{k-1}\big(\Delta_x h(\langle x, y\rangle)\big) = \Delta_x^{k-1} \\ \times\Big(-2\langle x, y\rangle h'(\langle x, y\rangle) + (1-(\langle x, y\rangle)^2)h''(\langle x, y\rangle)\Big)\end{aligned}$$

If we set $h_1(t) := \Big(-2th'(t) + (1-t^2)h''(t)\Big)$, then

$$\begin{aligned}\Delta_x^k h(\langle x, y\rangle) = \Delta_x^{k-2}\big(\Delta_x h_1(\langle x, y\rangle)\big) = \Delta_x^{k-2} \\ \times\Big(-2\langle x, y\rangle h_1'(\langle x, y\rangle) + (1-(\langle x, y\rangle)^2)h_2''(\langle x, y\rangle)\Big).\end{aligned}$$

Repeating the above process k times, we obtain

$$\Delta_x^k h(\langle x, y\rangle) = \sum_{j=0}^{2k} s_j(\langle x, y\rangle)h^{(j)}(\langle x, y\rangle),$$

where $s_j (j = 0, \ldots, 2k)$ are algebraic polynomials defined on $[-1,1]$ with degrees not larger than $2k$, and the coefficients of s_js are independent of h. Since for fixed $y \in \mathbf{S}^d$, $|\langle x, y\rangle| \le 1$, a direct computation yields that there exists a constant $C(d,n)$ depending only on n and d such that $\max_{x \in [-1,1]} |s_j(\langle x, y\rangle)| \le C(n,d)$. Therefore,

$$\begin{aligned}\max_{x \in \mathbf{S}^d}\left|\Delta_x^k h(\langle x, y\rangle)\right| & \le \sum_{j=0}^{2k} \max_{x \in \mathbf{S}^d} |s_j(\langle x, y\rangle)| \max_{x \in \mathbf{S}^d} |h^{(j)}(\langle x, y\rangle)| \\ & \le \sum_{j=0}^{2k} \max_{t \in [-1,1]} |s_j(t)| \max_{t \in [-1,1]} |h^{(j)}(t)| \\ & \le C(d,n) \sum_{j=1}^{2k} \max_{t \in [-1,1]} |h^{(j)}(t)|.\end{aligned}$$

This finishes the proof of Lemma 2. □

By using the same method, we can obtain the following Lemma 3 directly.

Lemma 3. *Let $y \in \mathbf{S}^d$ be fixed, $n \in \mathbf{N}$, and $1 \le p \le \infty$. Suppose that h is a univariate function with up to n times continuous derivatives, then there exists a constant $C(d,n,p)$ dependent on n, d and p such that*

$$\|\Delta_\cdot^k h(\langle \cdot, y\rangle)\|_p \le C(d,n,p) \sum_{j=1}^{2k} \|h^{(j)}(\cdot)\|_{L^p([-1,1])}, \quad k = 0, \ldots, [n/2].$$

Now we proceed the proof of Theorem 1.

Proof of Theorem 1. It is obvious that there exists a $\theta \in (-1,1)$ such that

$$\phi^{(l)}(\theta) \ne 0, \quad l = 0, 1, \ldots, n.$$

In fact, from the assumption of the activation function ϕ, it follows that there exists a $\tau \in [-1,1]$ such that $\phi^{(n)}(\tau) \ne 0$. As $\phi^{(n)}$ is continuous on $[-1,1]$, there is a closed interval $[a_n, b_n] \subset [-1,1]$ such that $\phi^{(n)}(t) \ne 0$ for $t \in [a_n, b_n]$, where $b_n > a_n$. Similarly, there exists a closed interval $[a_{n-1}, b_{n-1}] \subset [a_n, b_n]$, where $b_{n-1} > a_{n-1}$, such that $\phi^{(n-1)}(t) \ne 0$ for $t \in [a_{n-1}, b_{n-1}]$. Applying this step repeatedly, we can show that there exists a nested set of closed intervals

$$[a_0,b_0]\subset[a_1,b_1]\subset\ldots\subset[a_n,b_n]$$

such that for $l=0,1,\ldots,n$, $\phi^{(l)}$ has no zero point on $[a_l,b_l]$.

Thus it is sufficient to prove that for any $\sigma:[-1,1]\to\mathbf{R}$ with $(n+1)$ times bounded derivatives, and $\sigma^{(l)}(0)\neq 0$ for $l=0,1,\ldots,n$, there exists an SNN,

$$N_{\sigma,M}(x):=\sum_{j=0}^{M}a'_j\sigma(\langle w'_j,x\rangle),\ \ a'_j\in\mathbf{R}^d,w'_j\in\mathbf{B}^{d+1},$$

such that

$$\max_{x\in\mathbf{S}^d}|\Delta_x^kf(x)-\Delta_x^kN_{\sigma,M}(x)|\le\varepsilon. \tag{4.2}$$

Indeed, if (4.2) holds, then for $\theta\in(a_0,b_0)$, set

$$\delta:=\min(\theta-a_0,b_0-\theta),$$

and it is obvious that $\delta>0$ and $-1\le\delta t+\theta\le 1$ for any $t\in[-1,1]$. If we set

$$\sigma(\delta t+\theta):=\phi(t),$$

then (3.1) can be deduced from (4.2) easily.

A main tool to prove (4.2) is the following Taylor formula. Suppose $-1\le\mu t\le 1$ and $1\le m\le n$,

$$\sigma(\mu t)=\sigma(0)+\frac{\sigma'(0)}{1!}\mu t+\cdots+\frac{\sigma^{(m)}(0)}{m!}(\mu t)^m+R_m(t), \tag{4.3}$$

where

$$R_m(t)=\frac{\mu^m}{(m-1)!}\int_0^t\left(\sigma^{(m)}(\mu u)-\sigma^{(m)}(0)\right)(t-u)^{m-1}\,du, \tag{4.4}$$

and

$$\sigma^{(m)}(\mu u)=\sigma^{(m)}(v)|_{v=\mu u}.$$

It follows from Lemma 1 that for any $P_n\in\Pi_n^d$,

$$P_n(x)=\sum_{j=0}^{D_n^d}g_j(\langle a_j,x\rangle)=\sum_{i=0}^{n}\sum_{j=0}^{D_n^d}C_{j,i}(\langle a_j,x\rangle)^i. \tag{4.5}$$

It follows from (4.3) that

$$(\langle a_j,x\rangle)^n=\frac{n!}{\sigma^{(n)}(0)\mu^n}\sigma(\mu\langle a_j,x\rangle)+p_{n-1}(\langle a_j,x\rangle)-\frac{n!R_n(\langle a_j,x\rangle)}{\mu^n\sigma^{(n)}(0)},$$

where p_{n-1} is a univariate polynomial with degree not exceeding $n-1$. Therefore,

$$\Delta_x^k(\langle a_j,x\rangle)^n=\frac{(n)!}{\sigma^{(n)}(0)\mu^n}\Delta_x^k\sigma(\mu\langle a_j,x\rangle)+\Delta_x^kp_{n-1}(\langle a_j,x\rangle)-\Delta_x^k\frac{n!R_n(\langle a_j,x\rangle)}{\mu^n\sigma^{(n)}(0)}. \tag{4.6}$$

Using (4.4) with $m=n$, we have

$$\frac{n!R_n(\langle a_j,x\rangle)}{\mu^n\sigma^{(n)}(0)}=\frac{n}{\sigma^{(n)}(0)}\int_0^{\langle a_j,x\rangle}\left(\sigma^{(n)}(\mu u)-\sigma^{(n)}(0)\right)(\langle a_j,x\rangle-u)^{n-1}\,du.$$

It is obvious that $n!R_n(\langle a_j,x\rangle)/\mu^n\sigma^{(n)}(0)$ is a univariate function with $\langle a_j,x\rangle$ being its variable. Then it follows from Lemma 2 that there exists a constant $C(d,n)$ depending only on d and n such that

$$\max_{x\in\mathbf{S}^d}\Delta_x\frac{n!R_n(\langle a_j,x\rangle)}{\mu^n\sigma^{(n)}(0)}\le C(d,n)\sum_{l=0}^{2k}\max_{t\in[-1,1]}\left|\frac{n!R_n^{(l)}(t)}{\mu^n\sigma^{(n)}(0)}\right|.$$

It was proved in [37] that for $l=0,1,\ldots,n$

$$\max_{t\in[-1,1]}|R_n^{(l)}(t)|\le\mu^n\omega^*\left(\sigma^{(n)},\mu\right),$$

where

$$\omega^*(\sigma,t)=\sup_{0\le h\le t}\max_{t,t+h\in[-1,1]}|\sigma(t)-\sigma(t+h)|.$$

Thus

$$\begin{aligned}\max_{x\in\mathbf{S}^d}\Delta_x\frac{n!R_n(\langle a_j,x\rangle)}{\mu^n\sigma^{(n)}(0)}&\le C(d,n)\frac{n!}{\mu^n\sigma^{(n)}(0)}\sum_{l=0}^{2k}\max_{t\in[-1,1]}|R_n^{(l)}(t)|\\&\le C(d,n)\frac{n!}{\sigma^{(n)}(0)}\omega^*\left(\sigma^{(n)},\mu\right)\end{aligned}$$

Define $\delta_n\in\mathbf{R}_+$ small enough such that

$$\sum_{j=1}^{D_n^d}C_{j,n}C(n,d)\frac{n!}{\sigma^{(n)}(0)}\omega^*\left(\sigma^{(n)},\delta_n\right)\le\frac{\varepsilon}{(n+1)}, \tag{4.7}$$

where $C_{j,n}$ is the constant defined in (4.5). Then by (4.5) and (4.6) (set $\mu=\delta_n$) and (4.7) we have

$$\Delta_x^kP_n(x)=\frac{n!}{\sigma^{(n)}(0)(\delta_n)^n}\Delta_x^k\sum_{j=1}^{D_n^d}\sigma(\delta_n\langle a_j,x\rangle)+\Delta_x^kQ_{n-1}(x)-r_n(x), \tag{4.8}$$

where

$$Q_{n-1}(x)=\sum_{i=0}^{n-1}\sum_{j=1}^{D_n^d}B_1(i,j)(\langle a_j,x\rangle)^i,$$

$$r_n(x):=\sum_{j=1}^{D_n^d}C_{j,n}\frac{n!\Delta_x^kR_n(\langle a_j,x\rangle)}{(\delta_n)^n\sigma^{(n)}(0)}, \tag{4.9}$$

and $B_1(i,j)$ is the constant independent of k and x. Thus

$$\max_{x\in\mathbf{S}^d}|r_n(x)|\le\left|\sum_{j=1}^{D_n^d}C_{j,n}C(n,d)\frac{n!}{\sigma^{(n)}(0)}\omega^*\left(\sigma^{(n)},\delta_n\right)\right|\le\frac{\varepsilon}{(n+1)}. \tag{4.10}$$

Using the same method above to deal with $(\langle a_j\cdot x\rangle)^{n-1}$, a similar result as (4.8), (4.9) and (4.10) can be obtained.

$$\begin{aligned}\Delta_x^kQ_{n-1}(x)=&\frac{(n-1)!}{\sigma^{(n-1)}(0)(\delta_{n-1})^{n-1}}\Delta_x^k\sum_{j=1}^{D_n^d}\sigma(\delta_{n-1}\langle a_j,x\rangle)\\&+\Delta_x^kQ_{n-2}^*(x)+r_{n-1}(x),\end{aligned}$$

where $\delta_{\text{n-1}}$ is defined similar as δ_n in (4.7): $\delta_{n-1}\in\mathbf{R}_+$ small enough such that

$$\sum_{j=1}^{D_n^d}C_{j,n-1}C(n-1,d)\frac{n!}{\sigma^{(n-1)}(0)}\omega^*\left(\sigma^{(n-1)},\delta_{n-1}\right)\le\frac{\varepsilon}{(n+1)},$$

$C_{j,n-1}$ is the constant defined in (4.5),

$$Q_{n-2}^*(x)=\sum_{i=1}^{n-2}\sum_{j=1}^{D_n^d}B_2(i,j)(\langle a_j,x\rangle)^i,$$

and $B_2(i,j)$ is a constant independent of x and k. Similarly, we have

$$\max_{x\in\mathbf{S}^d}|r_{n-1}(x)|\le\frac{\varepsilon}{(n+1)}.$$

After repeating the method above $(n+1)$ times, we finally obtain

$$\Delta_x^kP_n(x)=\Delta_x^k\sum_{i=0}^{n}\sum_{j=1}^{D_n^d}A(j,i)\sigma(\delta_i\langle a_j,x\rangle)+r(x),$$

where

$$r(x):=\sum_{i=0}^{n}r_i(x),$$

and $A(j,i)$ is a constant independent of x and k. Thus, we have

$$\max_{x\in\mathbf{S}^d}|r(x)|=\sum_{j=0}^{n}\max_{x\in\mathbf{S}^d}|r_i(x)|\le(n+1)\frac{\varepsilon}{(n+1)}=\varepsilon,$$

If we rewrite the index $\{(i,j): i=0,\dots,n, j=1,\dots,D_n^d\}$ as $\{j: j=1,\dots,M\}$ and set $w_j:=\delta_i a_j$ we obtain that there exists an SNN, $N_{\sigma,M}$, such that

$$\max_{x\in\mathbf{S}^d}\left|\Delta_x^k P_n(x)-\Delta_x^k\sum_{j=1}^{M}C(j)\sigma(\langle w_j,x\rangle)\right|\le\varepsilon.$$

This finishes the proof of Theorem 1. □

In order to prove Theorem 2, we need the following Lemma 4. The lemma shows the commutativity of the best approximation operator V_n and the Laplace–Beltrami operator Δ^k. Its proof is standard, we refer the reader to [24], where the commutativity of the Laplace–Beltrami operator and the well-known translation operator was established.We do believe that it has been given in some other papers, but for the sake of completeness, we give its proof. For preliminary, we need introduce the following spherical best approximation operator, which was studied in [7,5,36]. Let $\eta\in C^\infty([0,\infty))$ with the properties that $\eta(t)=1$ for $0\le t\le 1$ and $\eta(t)=0$ for $t\ge 2$. The best approximation kernel K_n is defined as

$$K_n(t):=\sum_{k=0}^{\infty}\eta\left(\frac{k}{n}\right)\frac{D_k^d}{\Omega_d}P_k^{d+1}(t),\quad t\in[-1,1].$$

Then, the best approximation operator for $f\in L^1(\mathbf{S}^d)$ is defined by

$$V_nf(x):=\frac{1}{\Omega_d}\int_{\mathbf{S}^d}f(y)K_n(\langle x,y\rangle)\,d\omega(y).$$

We shall rely on the classical statement that every function $f\in L^p(\mathbf{S}^d),(1\le p\le\infty)$ admits a unique expansion of the form (see [24])

$$f(x)=\sum_{k=0}^{\infty}Y_kf(x).\tag{4.11}$$

with respect to the spherical harmonics, where the convergence of this series is meant in a (generalized) weak sense, and

$$Y_kf(x):=\frac{D_k^d}{\Omega_d}\int_{\mathbf{S}^d}f(y)P_k^{d+1}(\langle x,y\rangle)\,d\omega(y).$$

Therefore, an immediate computation yields

$$V_nf=\frac{1}{\Omega_d}\sum_{k=0}^{\infty}\eta\left(\frac{k}{n}\right)Y_k(f).\tag{4.12}$$

Since $\Pi_n^d=\oplus_{k=0}^{n}\mathbf{H}_k^d$, we have, for any $f\in L^1(\mathbf{S}^d)$, $V_nf\in\Pi_{2n}^d$ and for any $f\in\Pi_n^d$, $V_nf=f$.

Lemma 4. *Let $1\le p\le\infty$. Then for every $f\in W_p^{2k}(\mathbf{S}^d)$ there holds*

$$V_n\Delta^kf=\Delta^kV_nf.\tag{4.13}$$

Proof. From (2.1) and (4.11), for arbitrary $f\in W_p^{2k}(\mathbf{S}^d)$, we have

$$\Delta^kf=\Delta^k\sum_{j=0}^{\infty}Y_j(f)=(-1)^k\sum_{j=0}^{\infty}\lambda_j^kY_j(f).$$

The above equality together with (4.12) yields

$$\Delta^kV_nf=(-1)^k\sum_{j=0}^{\infty}\lambda_j^kY_j(V_nf)=(-1)^k\frac{1}{\Omega_d}\sum_{j=0}^{\infty}\lambda_j^kY_j\left(\sum_{l=0}^{\infty}\eta\left(\frac{l}{n}\right)Y_l(f)\right)$$

$$=(-1)^k\frac{1}{\Omega_d}\sum_{j=0}^{\infty}\eta\left(\frac{j}{n}\right)\lambda_j^kY_j(f).$$

On the other hand,

$$V_n\Delta^kf=\frac{1}{\Omega_d}\sum_{l=0}^{\infty}\eta\left(\frac{l}{n}\right)Y_l(\Delta^kf)$$

$$=\frac{1}{\Omega_d}\sum_{l=0}^{\infty}\eta\left(\frac{l}{n}\right)Y_l\left((-1)^k\sum_{j=0}^{\infty}\lambda_j^kY_j(f)\right)$$

$$=(-1)^k\frac{1}{\Omega_d}\sum_{j=0}^{\infty}\eta\left(\frac{j}{n}\right)\lambda_j^kY_j(f).$$

This finishes the proof of Lemma 4. □

Now we give the proof of Theorem 2.

Proof of Theorem 2. It follows from (4.13) that for arbitrary $\varepsilon>0$ there exists a $P_n\in\Pi_n^d$ such that

$$\|\Delta_\cdot^kf-\Delta_\cdot^kP_n(\cdot)\|_p<CE_{n,p}(\Delta^kf)+\frac{\varepsilon}{2}.$$

On the other hand, Theorem 1 requires that for $P_n\in\Pi_n^d$ there exists $N_{\phi,M}(x)$ such that

$$\|\Delta_\cdot^kP_n(\cdot)-\Delta_\cdot^kN_{\phi,M}(\cdot)\|_p<\frac{\varepsilon}{2}.$$

Thus

$$\|\Delta_\cdot^kf(\cdot)-\Delta_\cdot^kN_{\phi,M}(\cdot)\|_p\le\|\Delta_\cdot^kf(\cdot)-\Delta_\cdot^kP_n(\cdot)$$

$$\|_p+\|\Delta_\cdot^kP_n(\cdot)-\Delta_\cdot^kN_{\phi,M}(\cdot)\|_p\le E_{n,p}(\Delta^kf)+\varepsilon.$$

Then, setting $\varepsilon=E_{n,p}(\Delta^kf)$, we can deduce (3.2) directly. □

In order to give the explicit rate of simultaneous approximation by SNNs, we also need the rate of simultaneous approximation by SPs. The following Lemma 5 can be found in [5] or [7].

Lemma 5. *For every $f\in L^p(\mathbf{S}^d)$, $1\le p\le\infty$, there exists an absolute constant C such that*

$$\|V_nf-f\|_p\le CE_{n,p}(f),$$

where

$$E_{n,p}(f):=\inf_{P_n\in\Pi_n^d}\|f-P_n\|_p$$

is the best approximation error of SPs.

The following Jackson-type inequality can be deduced from [5].

Lemma 6. *Let $1\le p\le\infty$, $r\in\mathbf{N}$. Suppose that $f\in L^p(\mathbf{S}^d)$. Then*

$$\|V_nf-f\|_p\le C\omega_p^r\left(f,\frac{1}{n}\right).\tag{4.14}$$

Now, we are in a position to prove the last theorem of this paper.

Proof of Theorem 3. It can be easily deduced from Theorem 1 that there exists an SNN $N_{M,\phi}$ such that

$$\|\Delta^kf-\Delta^kN_{M,\phi}\|_p\le\|\Delta^kf-\Delta^kV_nf\|_p+$$

$$\|\Delta^kV_nf-\Delta^kN_{M,\phi}\|_p\le\|\Delta^kf-\Delta^kV_nf\|_p+\omega_p^r\left(\Delta^kf,\frac{1}{n}\right).$$

On the other hand, from (4.13) and (4.14) we obtain

$$\|\Delta^kf-\Delta^kV_nf\|_p=\|\Delta^kf-V_n(\Delta^kf)\|_p\le C\omega_p^r\left(\Delta^kf,\frac{1}{n}\right).$$

This completes the proof of Theorem 3. □

References

[1] G. Anastassiou, Rate of convergence of some multivariate neural network operators to the unit, Comput. Math. Appl. 40 (2000) 1–19.

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

[3] A. Barron, A. Cohen, W. Dahmen, R. DeVore, Approximation and learning by greedy algorithms, Ann. Stat. 36 (2008) 64–94.

[4] B. Baxter, S. Hubbert, Radial basis functions for the sphere, Recent Progr. Multivar. Approx. ISNM Int. Ser. Numer. Math. 137 (2001) 33–47.

[5] F. Dai, Jackson-type inequality for doubling weights on the sphere, Constr. Approx. 24 (2006) 91–112.

[6] Z. Ditzian, A modulus of smoothness on the unit sphere, J. d'Anal. Math. 79 (1999) 189–200.

[7] Z. Ditzian, Jackson-type inequality on the sphere, Acta Math. Hung. 102 (1–2) (2004) 1–35.

[8] G.E. Fasshauer, L.L. Schumaker, Scattered data fitting on the sphere. In: Mathematical Methods for Curves and Surfaces II, M. Dælen, T. Lyche, L.L. Schumaker(eds.), 1998. University Press Nashville, TN, pp. 117–166.

[9] W. Freeden, T. Gervens, M. Schreiner, Constructive Approximation on The Sphere, Calderon Press, Oxford, 1998.

[10] W. Freeden, V. Michel, Constructive approximation and numerical methods in geodetic research today—an attempt at a categorization based on an uncertainty principle, J. Geod. 73 (1999) 452–465.

[11] W. Freeden, V. Michel, H. Nutz, Satellite-to-satellite tracking and satellite gravity gradiometry (advanced techniques for high-resolution geopotential field determination), J. Eng. Math. 43 (2002) 19–56.

[12] Q. Le Gia, F. Narcowich, J. Ward, H. Wendland, Continuous and discrete least squares approximation by radial basis functions on spheres, J. Approx. Theory 143 (2006) 124–133.

[13] M. Hagan, M. Beale, H. Demuth, Neural Network Design, PWS Publishing Company, Boston, 1996.

[14] T. Hangelbroek, F. Narcowich, X. Sun, J. Ward, Kernel approximation on manifolds II: the L_∞ norm of the L_2 Projector, SIAM J. Math. Anal. 43 (2011) 662–684.

[15] K. Hesse, A lower bound for the worst-case cubature error on sheres of arbitrary dimension, Numer. Math. 103 (2006) 413–433.

[16] S. Hubbert, T. Morton, L_p error estimates for radial basis function interpolation on the sphere, J. Approx. Theory 129 (2004) 58–77.

[17] S. Hubbert, Q. Le Gia, T. Morton, Spherical Radial Basis Functions: Theory and Applications, Springer, New York, 2015.

[18] P. Kainen, V. Kürková, M. Sanguineti, Dependence of computational models on input dimension: tractability of approximation and optimization tasks, IEEE Trans. Inform. Theory 58 (2012) 1203–1214.

[19] P. Kim, J. Koo, Optimal spherical deconvolution, J. Multivariate Anal. 80 (2002) 21–42.

[20] V. Kürková, M. Sanguineti, Comparison of worst case errors in linear and neural network approxoimation, IEEE. Trans. Inform. Theory 48 (2002) 264–275.

[21] J. Levesley, X. Sun, Approximation in rough native spaces by shifts of smooth kernels on spheres, J. Approx. Theory 133 (2005) 269–283.

[22] X. Li, Simultaneous approximations of multivariate functions and their derivatives by neural networks with one hidden layer, Neurocomputing 12 (1996) 327–343.

[23] S. Lin, F. Cao, Z. Xu, Essential rate for approximation by spherical neural networks, Neural Netw. 24 (2011) 752–758.

[24] P. Lizorkin, S. Nikolskii, A theorem concerning approximation on the sphere, Anal. Math. 9 (1983) 207–221.

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

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

[27] H. Mhaskar, F. Narcowich, J. Ward, Approximation properties of zonal function networks using scattered data on the sphere, Adv. Comput. Math. 11 (1999) 121–137.

[28] H. Mhaskar, On the tractability of multivariate integration and approximation by neural networks, J. Complex. 20 (2004) 561–590.

[29] H. Mhaskar, Weighted quadrature formulas and approximation by zonal function networks on the sphere, J. Complex. 22 (2006) 348–370.

[30] H. Mhaskar, F. Narcowich, J. Prestin, J. Ward, L_p Bernstein estimates and approximation by spherical basis functions, Math. Comput. 79 (2010) 1647–1679.

[31] C. Müller, Spherical Harmonics, Lecture Notes in Mathematics, vol. 17, Springer, Berlin, 1966.

[32] F. Narcowich, X. Sun, J. Ward, H. Wendland, Direct and inverse sobolev error estimates for scattered data interpolation via spherical basis functions, Found. Comput. Math. 7 (2007) 369–370.

[33] F. Narcowich, X. Sun, J. Ward, Approximation power of RBFs and their associated SBFs: a connection, Adv. Comput. Math. 27 (2007) 107–124.

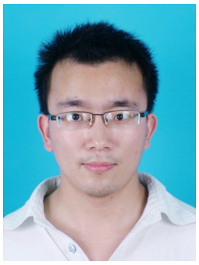
[34] I. Sloan, Polynomial interpolation and hyperinterpolation over general regions, J. Approx. Theory 83 (1995) 238–254.

[35] X. Sun, E. Cheney, Fundamental sets of continuous functions on spheres, Constr. Approx. 13 (1997) 245–250.

[36] K. Wang, L. Li, Harmonic Analysis and Approximation on The Unit Sphere, Science Press, Beijing, 2000.

[37] T. Xie, F. Cao, The errors in simultaneous approximation by feed-forward neural networks, Neurocomputing 73 (2010) 903–907.

[38] Z. Xu, F. Cao, Simultaneous L^p approximation order for neural networks, Neural Netw. 18 (2005) 914–923.



Shaobo Lin received the Ph.D. degree in Mathematics from Xi'an Jiaotong University in 2014. Now, he works in Wenzhou University. His main research interests include the neural networks and learning theory.

Feilong Cao, male, was born in Zhejiang Province, China, on August, 1965. He received the B.S. degree in Mathematics and the M.S. degree in Applied Mathematics from Ningxia University, China in 1987 and 1998, respectively. In 2003, he received the Ph.D. degree in Institute for Information and System Science, Xi'an Jiaotong University, China. From 1987 to 1992, he was an Assistant Professor. During 1992 to 2002, he was an Associate Professor. He now is a Professor in China Jiliang University. His current research interests include neural networks and approximation theory. He is the author or coauthor of more than 100 scientific papers.