Linear and nonlinear approximation of spherical radial basis function networks

Shaobo Lin

PII: S0885-064X(16)00015-7
DOI: http://dx.doi.org/10.1016/j.jco.2016.02.003
Reference: YJCOM 1286

To appear in: *Journal of Complexity*

Received date: 18 September 2015
Accepted date: 12 February 2016

Linear and nonlinear approximation of spherical radial basis function networks [☆]

Shaobo Lin[*]

Institute of Intelligent System and Decision, Wenzhou University, Wenzhou 325035, China

Abstract

In this paper, the center-selection strategy of spherical radial basis function networks (SRBFNs) is considered. To approximate functions in the Bessel-potential Sobolev classes, we provide two lower bounds of nonlinear SRBFN approximation. In the first one, we prove that, up to a logarithmic factor, the lower bound of SRBFN approximation coincides with the Kolmogorov n-width. In the other one, we prove that if a pseudo-dimension assumption is imposed on the activation function, then the logarithmic factor can even be omitted. These results together with the well known Jackson-type inequality of SRBFN approximation imply that the center-selection strategy doesn't affect the approximation capability of SRBFNs very much, provided the target function belongs to the Bessel-potential Sobolev classes. Thus, we can choose centers only for the algorithmic factor. Hence, a linear SRBFN approximant whose centers are specified before the training is recommended.

Keywords: Spherical radial basis function networks, linear approximation, nonlinear approximation, pseudo-dimension, Kolmogorov n-width.

1. Introduction

Fitting spherical data arising from sampling an unknown function defined on the sphere comes up frequently in applied problems. Examples include the study of seismic signals, gravitational phenomenon, solar corona and medical imaging of the brain. A

[☆] The research was supported by the National Natural Science Foundation of China (Grant Nos. 61502342, 11401462).

[*] Corresponding author: sblin1983@gmail.com

common procedure to fitting spherical data can boil down to two steps: choosing a specific class of functions to build up the candidates and selecting the final estimate from the candidates by using the spherical data (this process is also called as "training"). Therefore, the performance of the final estimate depends heavily on the quality of the candidates, which can be measured by the approximation capability of the selected class of functions.

The success of the radial basis function networks methodology in Euclidean space derives from its ability to generate approximants from data having arbitrary geometry. Thus, it is natural to introduce spherical radial basis function networks (SRBFNs) to tackle spherical data. This method has been extensively used in gravitational phenomenon [7], image processing [35] and learning theory [25]. An SRBFN can be mathematically represented as

$$S_n(x) := \sum_{i=1}^{n} c_i \phi(\xi_i \cdot x), \ x \in \mathbf{S}^d \tag{1.1}$$

where $c_i \in \mathbf{R}$ is the connection weight, ϕ is the activation function, $\{\xi_i\}_{i=1}^n \subset \mathbf{S}^d$ is the set of centers, and $\mathbf{S}^d$ is the unit sphere in $\mathbf{R}^{d+1}$.

Obviously, the approximation capability of SRBFNs depends on the activation function and centers. A seminal paper concerning the activation function selection is [34], in which Sun and Cheney deduced the sufficient and necessary conditions of the activation function to guarantee the universal approximation property of corresponding SRBFNs. Consequently, Mhaskar et al. [26] deduced a Jackson-type inequality of SRBFN approximation under the condition that the Fourier-Legendre coefficients of the activation function are not trivial. Later, Le Gia et al. [8] provided an upper bound error estimate for least-square SRBFN approximation with positive definition activation function by using the topological relation between $\mathbf{S}^d$ and the $d+1$-dimensional unit ball $\mathbf{B}^{d+1}$. For more details on this topic, the readers are referred to [5, 12, 14, 15, 28, 29, 30].

Compared with the activation function selection, the center-selection strategy of SRBFNs is more important and difficult, since it determines the computational burden of the training process. To be detailed, if centers are specified before the training, then solving a simple linear optimization problem can deduce the final estimate. If centers need to be tuned in the process of training, then we should tackle a nonlinear optimization problem that usually requires more computation. Up till now, there are roughly three categories of center-selection strategies. The first one is the spherical basis function (SBF) method (or zonal function networks method) that uses the linear combination of kernels located at points in a given scattered data. It follows from the definition that this type of networks is a linear approximant. Thus, we can use a linear algorithm to get the globally optimal solution. In particular, it was pointed out in [8] and [25] that solutions to the regularized least squares and support vector machine algorithms are SBFs. The second one is the minimal energy method that focuses on selecting centers by minimizing some quantities concerning the energy of the points. Examples include the Riesz minimal energy [10], ϕ-Riesz minimum energy [33] and other low discrepancy energies [32]. It should be highlighted that whether the SRBFN whose centers are minimal energy points is a linear approximant depends on the definition of energy. If the energy is independent of the data, such as the Riesz minimal energy, then the corresponding SRBFN is a linear approximant. The last one aims to select centers via training, which naturally results two types of parameters, the connection weights and centers, and makes the corresponding SRBFN be a nonlinear approximant. The main advantages of this approach is that the nonlinear SRBFN sometimes leads to better approximation capability [13, 24] and may circumvent the well known curse of dimensionality [1]. However, due to their nonlinearity, the implementation and training of the nonlinear SRBFN are much more difficult than its linear counterpart.

Our focus in this paper is not on selecting the most appropriate centers for a specified learning task, but on quantifying different approximation capabilities between linear and nonlinear SRBFNs. To this end, we should at first provide an answer to the following question: Is the approximation capability of the nonlinear SRBFN essentially better than that of linear SRBFN? Such a question is not new in the classical neural network approximation. For instance, in [17] and [21], Maiorov proved that the approximation capabilities of the nonlinear neural network and radial basis function network manifolds are better than that of arbitrary linear space with the same number of parameters, as the approximation errors of these nonlinear manifolds are essentially smaller than the Kolmogorov n-width [31]. A similar conclusion can also be found in [11]. The main novelty of this paper is to prove that similar conclusion is not valid for SRBFNs. In fact, we derive two lower bounds of nonlinear SRBFN approximation. The first one shows that, when the target function belongs to the well known Bessel-potential Sobolev class, up to a logarithmical factor, the lower bound of nonlinear SRBFN approximation coincides with the Kologorov n-width. The other one states that if a certain pseudo-dimension assumption is imposed on the activation function, then the logarithmical factor can even be omitted. These assertions together with the upper bound for linear SRBFN approximation [26, 27] imply that the approximation capability of the nonlinear SRBFN isn't essentially better than its linear counterpart, provided the target function belongs to the Bessel-potential Sobolev class. Based on this conclusion, the linear SRBFN method is recommended since the nonlinear SRBFN usually suffers from high computational burden. That is, if the target function is smooth, then we can choose a set of points with good geometrical distribution to built up the centers of SRBFN before the training process.

The rest of paper is organized as follows. In the next section, we present some preliminaries such as the spherical harmonics, Bessel-potential Sobolev class, and Kolmogorov n-width. In Section 3, we present the main results of this paper, where two lower bounds of nonlinear SRBFN approximation are presented, whose proofs are postponed to the last section.

2. Preliminaries

Denote by $L^2(\mathbf{S}^d)$ the space of square Lebesgue integrable functions on $\mathbf{S}^d$ endowed with the norm

$$\|f\|_2 := \|f(\cdot)\|_{L^2(\mathbf{S}^d)} := \left\{\int_{\mathbf{S}^d} |f(x)|^2 d\omega(x)\right\}^{1/2} < \infty,$$

where 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

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

For integer $k \geq 0$, the restriction to $\mathbf{S}^d$ of a homogeneous harmonic polynomial of degree k 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 spherical harmonics of degree $k \le s$ is denoted by Π_s^d. It is obvious that $\Pi_s^d = \bigoplus_{k=0}^s \mathbf{H}_k^d$. The dimension of $\mathbf{H}_k^d$ is given by

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

and that of Π_s^d is $\sum_{k=0}^s D_k^d = D_s^{d+1} \sim s^d$.

Spherical harmonics have an intrinsic characterization. To describe this, we first introduce the Laplace-Beltrami operator Δ [7], which is defined by

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

It is well known that Δ is an elliptic, (unbounded) selfadjoint operator on $L^2(\mathbf{S}^d)$, and is invariant under arbitrary coordinate changes. Furthermore, its spectrum comprises distinct eigenvalues $-\lambda_k := -k(k+d-1),\ k = 0, 1, \dots,$ 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, $L^2(\mathbf{S}^d) = \text{closure}\{\bigoplus_k \mathbf{H}_k^d\}$. Hence, if we choose an orthonormal basis $\{Y_{k,l} : l = 1, \dots, D_k^d\}$ of each $\mathbf{H}_k^d$, then the set $\{Y_{k,l} : k = 0, 1, \dots, l = 1, \dots, D_k^d\}$ is an orthonormal basis of $L^2(\mathbf{S}^d)$.

The following Funk-Hecke formula [7] plays an important role in computing the eigenvalues of the kernel $\phi \in L^1([-1,1])$

$$\int_{\mathbf{S}^d} \phi(x\cdot y) H_k(y) d\omega(y) = B(\phi, k) H_k(x), \tag{2.2}$$

where

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

and P_k^{d+1} is the Legendre polynomial [7] with degree k and dimension $d+1$.

The orthogonal projection P_l onto $\mathbf{H}_k^d$ is given by

$$P_l f := \sum_{j=1}^{D_k^d} \langle f, Y_{k,j}\rangle Y_{k,j}. \tag{2.3}$$

It can be found in (2.1) that $Y_{k,j}$ is an eigenfunction corresponding to the eigenvalue $-k(k+d-1) = ((d-1)/2)^2-(k+(d-1)/2)^2$ for the Laplace-Beltrami operator. It follows that $k+(d-1)/2$ is an eigenvalue corresponding to the eigen functions $Y_{k,j}, j = 1, \dots, D_k^d$, of the pseudo-differential operator

$$L_n := \sqrt{((d-1)/2)^2 - \Delta} = \sum_{k=1}^{\infty}(k+(d-1)/2)P_l. \tag{2.4}$$

Let $\mu \geq 0$. Denote the Bessel-potential Sobolev classes [27] H_μ to be all f such that

$$\|f\|_{H_\mu} := \left\| \sum_{k=0}^{\infty}(k+(d-1)/2)^{\mu}P_l f \right\|_2 \leq 1.$$

The Kolmogorov n-width for H_μ is then defined by

$$W_n(H_\mu, L^2) = \inf_{A_n} \sup_{f \in H_\mu} \inf_{g \in A_n} \|f-g\|_2,$$

where the left-most infimum is taken over all n-dimensional subspaces A_n of $L^2(\mathbf{S}^d)$. It can be easily deduced from [36] (see also [3, 31]) that there exist constants C_1 and C_2 independent of n such that

$$C_1 n^{-\mu/d} \leq W_n(H_\mu, L^2) \leq C_2 n^{-\mu/d}. \tag{2.5}$$

According to the definition of $W_n(H_\mu, L^2)$, (2.5) implies that if the target function belongs to H_μ, then the approximation rate of arbitrary linear space cann't be faster than $n^{-\mu/d}$.

3. Two lower bounds for nonlinear SRBFN approximation

In this section, we consider the approximation capability of the following set

$$\mathcal{N}_n(\phi) := \left\{ \sum_{i=1}^{n} c_i \phi(\xi_i \cdot x) : \xi_i \in \mathbf{S}^d, c_i \in \mathbf{R} \right\}, \tag{3.1}$$

where $\phi \in L^2([-1,1])$. $\mathcal{N}_n(\phi)$ is obviously a nonlinear manifold since the sum of two elements sometimes doesn't belong to $\mathcal{N}_n(\phi)$. For the nonlinear approximant, a commonly made statement in Euclidean space is that it can break the curse of dimensionality, in which the rate of approximation is faster than $n^{-1/2}$, independent of the dimension d of the variable space for f. To be detailed, Burger and Neubauer [4] proved that for arbitrary

$f(x)=\int_{\mathbf{B}^{d+1}} h(z)\phi(x\cdot z)dz$ with $x\in\mathbf{B}^{d+1}$, $h\in L^1(\mathbf{B}^{d+1})$ and $\phi\in C([-1,1])$, there exist a $g_n\in\mathcal{N}_n^*(\phi)$ and some constant $C_0>0$ independent of n such that

$$\|f-g_n\|\le C_0 n^{-1/2},$$

where

$$\mathcal{N}_n^*(\phi):=\left\{\sum_{i=1}^n c_i\phi(\xi_i\cdot x):\xi_i\in\mathbf{B}^{d+1},c_i\in\mathbf{R}\right\}.$$

Such a conclusion also holds for the nonlinear SRBFN manifold (See Appendix). However, it was pointed out in [2] that this is not exactly true since, in practice, the assumption on f becomes more and more stringent as d grows. Thus, to compare the approximation capabilities of the linear and nonlinear SRBFN, we should present the same restriction to the target functions.

In the following Theorem 3.1, we prove that if $f\in H_\mu$, then the nonlinear SRBFN approximation cann't essentially improve the approximation capability of its linear counterpart.

Theorem 3.1. *Let $\mu>0$ and $\mathcal{N}_n(\phi)$ be defined by (3.1). For arbitrary $n\in\mathbf{N}$, there exists a constant C independent of n such that*

$$dist(H_\mu,\mathcal{N}_n(\phi),L_2)\ge\frac{C}{(n\log n)^{\mu/d}},$$

where

$$dist(H_\mu,\mathcal{N}_n(\phi),L_2):=\sup_{f\in H_\mu}dist(f,\mathcal{N}_n(\phi),L_2):=\sup_{f\in H_\mu}\inf_{g\in\mathcal{N}_n(\phi)}\|f-g\|_2$$

denotes the distance between H_μ and $\mathcal{N}_n(\phi)$.

It follows from (2.5) that if $f\in H_\mu$, then the best approximation rate of the linear SRBFN behaves asymptomatically as $n^{-\mu/d}$. Let the set of centers $X:=\{x_i\}_{i=1}^n$ be specified. The SRBFN

$$L_{\phi,n}(x):=\sum_{i=1}^n c_i\phi(x_i\cdot x)\tag{3.2}$$

is a linear approximant. Denote by $\mathcal{L}_n(\phi)$ be the set of functions formed as (3.2). From the definition of linear and nonlinear SRBFN approximation, it follows that

$$dist(H_\mu,\mathcal{N}_n(\phi),L_2)\le dist(H_\mu,\mathcal{L}_n(\phi),L_2)$$

for arbitrary X. Then, the upper bound of $dist(H_\mu, \mathcal{L}_n(\phi), L_2)$ is automatical the upper bound of $dist(H_\mu, \mathcal{N}_n(\phi), L_2)$, which implies that the approximation capability of $\mathcal{N}_n(\phi)$ is intuitively better than that of $\mathcal{L}_n(\phi)$. However, in Theorem 3.1, we prove that except for a logarithmic factor, the nonlinear SRBFN cann't essentially improve the approximation rate of its linear counterpart when the target function belongs to H_μ.

In the realm of spherical approximation, it is a common consensus that $\mathcal{N}_n(\phi)$ is not much topologically larger than $\mathcal{L}_n(\phi)$. Thus, the lower bound deduced in Theorem 3.1 seems not very surprising at the first glance. We highlight the novelty of Theorem 3.1 in the following two directions. On one hand, admittedly, $\mathcal{N}_n(\phi)$ is not much larger than $\mathcal{L}_n(\phi)$, however, how to quantify the "not much larger than" is also very interesting. Our result in Theorem 3.1 shows that to approximate functions in H_μ, the words "not much larger than" can be mathematically reflected by the approximation rate. Here "larger than" means the approximation rate of the nonlinear SRBFN is faster than that of its linear counterpart and "not much" means that up to a logarithmic factor, their approximation rates are identical. On the other hand, Theorem 3.1 holds for $\mathcal{N}_n(\phi)$ with arbitrary square integrable activation function, which is, to the best of our knowledge, different from the previous studies concerning the SRBFN approximation.

At last, we should point out that Theorem 3.1 doesn't imply the uselessness of nonlinearity. In fact, [13] proved that there is a sigmoidal activation function σ such that

$$C_1 n^{-\mu/(d-1)} \leq dist(H_\mu, \mathcal{M}_n(\sigma), L_2) \leq C_2 n^{-\mu/(d-1)},$$

where

$$\mathcal{M}_n(\sigma) := \left\{ \sum_{i=1}^n c_i \sigma(w_i \cdot x - b_i) : w_i \in \mathbf{S}^d, b_i, c_i \in \mathbf{R} \right\}.$$

Similar result for $\mathbf{B}^{d+1}$ can be found in [19]. These assertions yield that nonlinearity can essentially improve the approximation capability. The result in Theorem 3.1 only shows that for SRBFNs, whether the centers are selected before the training cann't essentially affect the approximation capability, provided the target function belongs to the Bessel-potential Sobolev class.

In Theorem 3.1, we provided a lower bound of the nonlinear SRBFN approximation showing that up to a logarithmic factor, the approximation rate of the nonlinear SRBFN asymptomatically equals to the linear n-width. In the following, we will show that if some restrictions are imposed on the activation function, then the logarithmic factor can be omitted. To aid the description, we should introduce the pseudo-dimension [18, 23].

For any $t \in \mathbb{R}$, define

$$\operatorname{sgn}(t) := \begin{cases} 1, & \text{if } t \geq 0, \\ -1, & \text{if } t < 0. \end{cases}$$

If a vector $\mathbf{t} = (t_1, \dots, t_n)$ belongs to $\mathbf{R}^n$, then we denote by sgn($\mathbf{t}$) the vector (sgn(t_1), $\dots$,sgn(t_n)). The VC dimension [23] of a set V over D, denoted as $VCdim(V, D)$, is defined as the maximal natural number m such that there exists a collection $(\mu_1, \dots, \mu_m)$ in D such that the cardinality of the sgn-vectors set $S = \{(\operatorname{sgn}(v(\mu_1)), \dots, \operatorname{sgn}(v(\mu_m))) : v \in V\}$ equals to 2^m, that is, the set S coincides with the set of all vertexes of the unit cube in $\mathbf{R}^m$. The quantity

$$Pdim(V, D) := \max_g VCdim(V + g, D),$$

is called the pseudo-dimension of the set V over D, where g runs all functions defined on D and $V + g = \{v + g : v \in V\}$.

Theorem 3.2. *Let $\mu > 0$ and $\mathcal{Q}_n$ be a set of functions defined on $\mathbf{S}^d$. If $\mathcal{Q}_n$ satisfies $Pdim(\mathcal{Q}_n, L^2) \leq n$, then we have*

$$dist\{H_\mu, \mathcal{Q}_n, L_2\} \geq \frac{C}{n^{\mu/d}}$$

for some finite constant C independent of n.

It is obvious that the pseudo-dimension is an extension of the dimension, since the pseudo-dimension of an arbitrary n-dimensional vector space is n [18, Property 1]. Thus, Theorem 3.2 actually implies an extension of the Kolmogorov n-width. That is, if we define the Pseudo n-width for H_μ by

$$PS_n(H_\mu, L^2) = \inf_{Q_n} \sup_{f \in H_\mu} \inf_{g \in Q_n} \|f - g\|_2,$$

where the left-most infimum is taken over all sets $\mathcal{Q}_n$ whose pseudo-dimension is at most n, then there exist constants C_1 and C_2 independent of n such that

$$C_1 n^{-\mu/d} \leq PS_n(H_\mu, L^2) \leq C_2 n^{-\mu/d}. \tag{3.3}$$

According to (3.3), to present a lower bound of approximation rate, it suffices to deduce the pseudo-dimension of the set of approximants. Taking SRBFNs for example, its pseudo-dimension depends heavily on the the activation function. If the activation function is the well known polynomial kernel [37], i.e., $\phi(t) = (1+t)^s$, then $Pdim(\mathcal{N}_n(\phi)) = \min\{D_s^{d+1}, n\}$; If the activation function is the Gaussian, then it can be easily deduced from [20, 23] that there exist constants C_1 and C_2 independent of n such that

$$C_1 n \leq Pdim(\mathcal{N}_n(\phi)) \leq C_2 n \log n.$$

4. Proofs

In this section, we provide the proofs of Theorem 3.1 and Theorem 3.2, respectively.

4.1. Proof of Theorem 3.1

At first, we need to introduce the following Lemma 4.1, which can be found in [22]. Consider in $\mathbf{R}^m$ the polynomial manifold

$$\mathcal{P}_{m,p,s} = \{(p_1(u), \dots, p_m(u)) : u \in \mathbf{R}^p\},$$

where $p_1(u), \dots, p_m(u)$ are any algebraic polynomials with real coefficients each of total degree s. Let the vector set E^m consisting of all vectors $\varepsilon := (\varepsilon_1, \dots, \varepsilon_m)$, $m \in \mathbf{N}$ with coordinates $\varepsilon_1, \dots, \varepsilon_m = \pm 1$, i.e.,

$$E^m = \{\varepsilon = (\varepsilon_1, \dots, \varepsilon_m) : \varepsilon_i = \pm 1, i = 1, 2, \dots, m\}.$$

Lemma 4.1. *Let m, p and s be any natural numbers such that $m \geq 2p$ and $p \log(4ems/p) \leq m/4$. Then there exist a vector $\varepsilon \in E^m$ and some absolute constant $c > 0$ so that*

$$dist(\varepsilon, \mathcal{P}_{m,p,s}) \geq c m^{\frac{1}{2}}.$$

Now we proceed to the proof of Theorem 3.1.

Proof of Theorem 3.1. Let n and s be any natural numbers. Set $m = D_s^{d+1}$. Consider the set consisting of spherical harmonics

$$F_{s,d} := \left\{ h(x) = \sum_{k=0}^{s} \sum_{l=1}^{D_k^d} \varepsilon(k,l) Y_{k,l}(x) \right\},$$

where $\varepsilon(k,l)=\pm 1 : k=0,\dots,s, l=1,\dots,D_k^d$. It is obvious that there exists a constant C_3 depending only on d such that

$$\frac{1}{C_3 s^\mu m^{\frac{1}{2}}} h(x) \in H_\mu, \tag{4.1}$$

for any $h\in F_{s,d}$. Indeed, since $F_{s,d}$ is a subset of Π_s^d, it follows from the Parseval equality that

$$\left\| \frac{1}{C_3 s^\mu m^{\frac{1}{2}}} h(\cdot)\right\|_2^2 = \frac{\| \sum_{k=0}^s \sum_{l=1}^{D_k^d} \varepsilon(k,l) Y_{k,j}(\cdot)\|_2^2}{C_3^2 s^{2\mu} m} = \frac{1}{C_3^2 s^{2\mu}}.$$

The above equalities together with the well known Bernstein inequality [27] yield that

$$\left\| \frac{1}{C_3 s^\mu m^{\frac{1}{2}}} h(\cdot)\right\|_{H_\mu} \le C_3 s^\mu \left\| \frac{1}{C_3 s^\mu m^{\frac{1}{2}}} h(\cdot)\right\|_2 \le 1.$$

This yields (4.1).

Now we estimate the deviation of the set $F_{s,d}$ from the set $\mathcal{N}_n(\phi)$,

$$\mathrm{dist}(F_{s,d}, \mathcal{N}_n(\phi), L_2) = \sup_{h\in F_{s,d}} \inf_{g\in \mathcal{N}_n(\phi)} \|h(\cdot)-g(\cdot)\|_2.$$

Let

$$h(x) = \sum_{k=0}^s \sum_{l=1}^{D_k^d} \varepsilon(k,l) Y_{k,l}(x)$$

be an arbitrary function from $F_{s,d}$ and let

$$g(x) = \sum_{i=1}^n b_i \phi(\xi_i\cdot x), \;\; \xi_i \in \mathbf{S}^d$$

be an arbitrary function from $\mathcal{N}_n(\phi)$. Then by the Bessel inequality and Parseval equality we have

$$\begin{aligned}
\|h(\cdot)-g(\cdot)\|_2^2 &= \left\| \sum_{k=0}^s \sum_{l=1}^{D_k^d} \varepsilon(k,l) Y_{k,l}(\cdot) - g(\cdot)\right\|_2^2 \\
&\ge \left\| \sum_{k=0}^s \sum_{l=1}^{D_k^d} \varepsilon(k,l) Y_{k,l}(\cdot) - \sum_{k=0}^s \sum_{l=1}^{D_k^d} \langle g, Y_{k,l}\rangle Y_{k,l}(\cdot)\right\|_2^2 \\
&= \left\| \sum_{k=0}^s \sum_{l=1}^{D_k^d} (\varepsilon(k,l) - \langle g, Y_{k,l}\rangle) Y_{k,l}(\cdot)\right\|_2^2 \\
&= \sum_{k=0}^s \sum_{l=1}^{D_k^d} |\varepsilon(k,l) - \langle g, Y_{k,l}\rangle|^2.
\end{aligned}$$

Fix indices k, l, and consider the inner product $\langle g, Y_{k,l}\rangle$. It follows from the Funk-Hecke formula (2.2) that

$$\begin{aligned}\langle g, Y_{k,l}\rangle &= \sum_{i=1}^{n} \langle b_i\phi(\xi_i \cdot x), Y_{k,l}(x)\rangle \\ &= \sum_{i=1}^{n} b_i \int_{\mathbf{S}^d} \phi(\xi_i \cdot x) Y_{k,l}(x) d\omega(x) \\ &= \sum_{i=1}^{n} b_i B(\phi, k) Y_{k,l}(\xi_i).\end{aligned}$$

Set

$$\pi_{k,l}(u) := \pi_{k,l}(b_i, \xi_i) := \sum_{i=1}^{n} b_i B(\phi, k) Y_{k,l}(\xi_i),$$

then $\pi_{k,l}$ is an algebraic polynomial with degree at most s and variable $u \in \mathbf{R}^{n+dn}$. Therefore,we have

$$(\text{dist}(F_{s,d}, \mathcal{N}_n(\phi), L_2))^2 \geq \max_{\varepsilon(k,l)\in\{1,-1\}} \inf_{u\in\mathbf{R}^{(d+1)n}} \sum_{k=0}^{s} \sum_{l=1}^{D_k^d} |\varepsilon(k,l) - \pi_{k,l}(u)|^2.$$

Since $m = D_s^{d+1}$, we can rearrange the sequence $\{(k,l), k = 0, \dots, s, l = 1 \dots, D_k^d\}$ as $\{k'\}_{k'=1}^m$. Thus

$$(\text{dist}(F_{s,d}, \mathcal{N}_n(\phi), L_2))^2 \geq \max_{\varepsilon_{k'}\in\{1,-1\}} \inf_{u\in\mathbf{R}^{(d+1)n}} \sum_{k'=0}^{m} |\varepsilon(k') - \pi_{k'}(u)|^2. \tag{4.2}$$

Now, we use Lemma 4.1 to give a lower bound for (4.2). Since $C_1 s^d \leq m \leq C_2 s^d$, then there are a constant $c_0 \geq 2(d+1)^2/d$ and some n such that $m = 4c_0 n \log n$ and

$$\log n \leq \left(\frac{C_1(d+1)}{16ec_0^{(d+1)/d}4^{1/d}}\right)^{d/(d+1)} n.$$

Set $p = (d+1)n$, then

$$m = 4c_0 n \log n > 2(d+1)n = 2p.$$

and

$$\begin{aligned}p\log(4ems/p) &\leq (d+1)n \log \frac{16ec_0 n \log n \cdot 1/C_1(4c_0 n \log n)^{1/d}}{(d+1)n} \\ &\leq (d+1)n \log \frac{16ec_0 \log n \cdot 1/C_1 n^{1/d}(\log n)^{1/d} \cdot 4^{1/d} c_0^{1/d}}{d+1} \\ &\leq (d+1)n \log n^{(2d+2)/d} \\ &\leq \frac{2(d+1)^2}{d} n \log n \leq c_0 n \log n = \frac{m}{4}.\end{aligned}$$

Hence, it follows from Lemma 4.1 that

$$\mathrm{dist}(F_{s,d}, \mathcal{N}_n(\phi), L_2) \geq Cm^{\frac{1}{2}},$$

which together with (4.1) yields that

$$\begin{aligned} \mathrm{dist}(H_\mu, \mathcal{N}_n(\phi), L_2) &\geq \mathrm{dist}(C_3 s^{-\mu} m^{-\frac{1}{2}} F_{s,d}, \mathcal{N}_n(\phi), L_2) \\ &\geq Cs^{-\mu} m^{-\frac{1}{2}} \mathrm{dist}(F_{s,d}, \mathcal{N}_n(\phi), L_2) \geq Cs^{-\mu} m^{-\frac{1}{2}} \times m^{\frac{1}{2}} \\ &= Cs^{-\mu} \sim (n \log n)^{-\mu/d}. \end{aligned}$$

This completes the proof of Theorem 3.1. □

4.2. Proof of Theorem 3.2

To aid the proof, we need to introduce the following lemmas. The first one can be found in [16, P.489]. Let m be a fixed natural number. Consider the m-dimensional Hilbert space l_2^m consisting of vectors $t = (t_1, \dots, t_m) \in R^m$ with norm $|t| = (\sum_{i=1}^m t_i^2)^{1/2}$. Let H be some set in l_2^m. Define the distance of a point from the set H as $\mathrm{dist}(t, H) = \inf_{h \in H} |t - h|$.

Lemma 4.2. *There exists a set $G \subset E^m$ of cardinality at least $(4/3)^m$ such that there exists a constant $c_0 \leq 1$ satisfying*

$$\|u - v\|_{l_2^m} \geq c_0 m^{\frac{1}{2}}, \qquad \forall u, v \in G, u \neq v.$$

Lemma 4.3. *Let $s \in \mathbf{N}$, $m = D_s^{d+1}$ and $G \subset E^m$ be defined in Lemma 4.2. Denote by*

$$\mathcal{F}_m(G, c) = \left\{ h_a(x) : h_a(x) := \sum_{k=0}^{s} \sum_{l=1}^{D_k^d} \frac{1}{cs^\mu m^{1/2}} a_{k,l} Y_{k,l}(x) : a_{k,l} \in G \right\}. \tag{4.3}$$

Then for any $h_a \neq h_{a'} \in \mathcal{F}_m(G, c)$, there exists a constant C depending only on d, c and c_0 such that

$$\|h_a - h_{a'}\|_2 \geq \frac{C}{m^{\mu/d}}.$$

Proof. It follows from the Parseval equality and Lemma 4.2 that

$$\|h_a - h_{a'}\|_2^2 = \frac{1}{c^2 s^{2\mu} m} \sum_{k=0}^{s} \sum_{l=1}^{D_k^d} |a_{k,l} - a'_{k,l}|^2 \geq \frac{1}{c^2 s^{2\mu} m} \cdot c_0^2 m = \frac{C}{s^{2\mu}}.$$

Since $m = D_s^{d+1} \sim s^d$, we finishes the proof of Lemma 4.3. □

For a set of functions $\mathcal{F} \in L^p(\mathbf{S}^d)$, denote by

$$\mathcal{M}_\varepsilon(\mathcal{F})_p := \max\{s : \exists f_1, \dots, f_s \in L^p(\mathbf{S}^d), \|f_i - f_j\|_p \geq \varepsilon, \forall i \neq j\}$$

the ε-packing number for $\mathcal{F}$ in the L_p-norm. From the definition, it follows that for arbitrary $p \leq q$ and $\mathcal{F} \subset L^q(\mathbf{S}^d)$, there holds

$$\mathcal{M}_\varepsilon(\mathcal{F})_q \leq \mathcal{M}_\varepsilon(\mathcal{F})_p.$$

The following Lemma 4.4 can be easily deduced from the above inequality and [9, Corollary 3] (see also [18]).

Lemma 4.4. *Let $\mathcal{H}^n = \{h\}$ be a set of Lebesgue-measurable functions on $\mathbf{S}^d$ such that $\|h\|_\infty \leq \beta$ and $Pdim(\mathcal{H}^n, D) \leq n < \infty$. Then, for any $\varepsilon > 0$, the following upper bound on the ε-packing number holds:*

$$\mathcal{M}_\varepsilon(\mathcal{H}^n)_2 \leq e(n+1)\left(\frac{4e\beta}{\varepsilon}\right)^n. \tag{4.4}$$

Now we proceed to the proof of Theorem 3.2.

Proof of Theorem 3.2. Let n and s be any natural numbers. Set $m = D_s^{d+1}$. Then there exist constants C_1 and C_2 such that

$$C_1 s^d \leq m \leq C_2 s^d. \tag{4.5}$$

Let $\varepsilon > 0$ be any positive real number. Denote

$$\delta = \sup_{h\in\mathcal{F}_m(G,C_3)} \inf_{g\in\mathcal{Q}_n} \|h - g\|_2 + \varepsilon = dist(\mathcal{F}_m(G, C_3), \mathcal{Q}_n, L_2) + \varepsilon.$$

For any $h \in \mathcal{F}_m(G, C_3)$, define $Ph \in \mathcal{Q}_n$ by

$$\|h - Ph\|_2 \leq \delta.$$

Set $\beta = m^{-\mu/d}$. Introduce the clip operator

$$\pi_\beta h : \pi_\beta h(x) = \begin{cases} -\beta, & h(x) < \beta \\ h(x), & -\beta \leq h(x) \leq \beta, \\ \beta, & h(x) > \beta. \end{cases}$$

Consider the set of functions $S := \pi_\beta P(\mathcal{F}_m(G, C_3)) := \{\pi_\beta Ph : h \in \mathcal{F}_m(G, C_3)\}$. Let $h \neq h' \in \mathcal{F}_m(G, C_3)$, then

$$\|\pi_\beta Ph - \pi_\beta Ph'\|_2 = \|\pi_\beta Ph - h + h - h' + h' - \pi_\beta Ph'\|_2 \geq \|h - h'\|_2 - \|h - \pi_\beta Ph\|_2 - \|h' - \pi_\beta Ph'\|_2.$$

Since $|h(x)| \leq \beta$, it follows from the definition of the clip operator that

$$\|h - \pi_\beta Ph\|_2 \leq \|h - Ph\|_2.$$

Therefore, Lemma 4.3 implies that

$$\|\pi_\beta Ph - \pi_\beta Ph'\|_2 \geq \|h - h'\|_2 - \|h' - Ph'\|_2 - \|h - Ph\|_2 \geq \frac{C_4}{m^{\mu/d}} - 2\delta.$$

Now, we assume $\delta \leq \frac{C_4}{4m^{\mu/d}}$. From the above inequality and Lemma 4.2, it follows that for any $f, f' \in S$ and $f \neq f'$,

$$\|f - f'\|_2 \geq \frac{C_4}{2m^{\mu/d}},$$

and the cardinality $|S| \geq \left(\frac{4}{3}\right)^m$. Fix $\alpha = \frac{C_4}{2m^{\mu/d}}$, then

$$\mathcal{M}_\alpha(S)_2 \geq \left(\frac{4}{3}\right)^m.$$

On the other hand, we have $|f(x)| \leq \beta$ for all $f \in S$ and $x \in \mathbf{S}^d$. It follows from the definition of the pseudo-dimension that $Pdim(\pi_\beta P(\mathcal{F}_m(G, C_3))) \leq Pdim(P(\mathcal{F}_m(G, C_3)))$. Since $P(\mathcal{F}_m(G, C_3)) \subset \mathcal{Q}_n$, we have $Pdim(P(\mathcal{F}_m(G, C_3))) \leq Pdim(\mathcal{Q}_n) = n$. Hence $Pdim(S) \leq n$. According to Lemma 4.4, we obtain

$$\mathcal{M}_\alpha(S)_2 \leq \mathcal{M}_\alpha(S)_1 \leq e(n+1)\left(\frac{4e\beta}{\alpha}\right)^n. \tag{4.6}$$

Since $C_1 s^d \leq m \leq C_2 s^d$, choose s ass the smallest natural number such that $s^d > n\left(\frac{8+\log_{4/3}\frac{4e\beta}{\alpha}}{C_1}\right)$. Then we have

$$\left(\frac{4}{3}\right)^m \geq \left(\frac{4}{3}\right)^{C_1 s^d} > \left(\frac{4}{3}\right)^{n\left(8+\log_{4/3}\left(\frac{4e\beta}{\alpha}\right)\right)} = \left(\frac{4e\beta}{\alpha}\right)^n \times \left(\frac{4}{3}\right)^{8n} > e(n+1)\left(\frac{4e\beta}{\alpha}\right)^n,$$

which contradicts (4.6). Thus, the assumption $\delta \leq \frac{C_4}{4m^{\mu/d}}$ is false. Recalling that $\beta = m^{-\mu/d}$ and $\alpha = \frac{C_4}{2}m^{-\mu/d}$, we obtain $m \sim n$. Therefore, we obtain

$$\delta > \frac{C_4}{4m^{\mu/d}} = Cn^{\frac{-\mu}{d}}.$$

According to the definition of δ and the arbitrariness of ε, there holds

$$dist(\mathcal{F}_m(G,C_3),\mathcal{L}_n,L_2) \geq Cn^{-\mu/d}.$$

Noticing $\mathcal{F}_m(G,C_3) \subset H_\mu$, we finish the proof of Theorem 3.2. □

Appendix: Fast approximation by nonlinear SRBFNs

In the Appendix, we give a fast rate of nonlinear SRBFNs approximation, which is motivated by [4]. At first we need to introduce the following modulus of smoothness of the activation function. If ϕ is a continuous univariate function defined on $[-1,1]$, then the modulus of smoothness $\omega_r(\phi,\tau)$ is given by (e.g. [6])

$$\omega_r(\phi,\tau) = \sup_{0<\alpha\leq\tau} \|\triangle_\alpha^r(\phi,\cdot)\|_{C[-1,1]},$$

where

$$\Delta_\alpha^r(\phi,\cdot) = \sum_{k=0}^{r}(-1)^k\binom{r}{k}\phi(\cdot+k\alpha).$$

If $r=1$, then we use $\omega(\phi,\tau)$ instead of $\omega_1(\phi,\tau)$ for brevity.

Theorem 4.5. *If ϕ is a continuous univariate function defined on $[-1,1]$ and f satisfies $f(x)=\int_{\mathbf{S}^d} h(z)\phi(x\cdot z)d\omega(z)$ for some positive function $h\in L^1(\mathbf{S}^d)$, then there exists an element of $\mathcal{N}_n(\phi)$, g_n, such that*

$$\|f-g_n\|_2 \leq Cn^{-\frac{1}{2}}\omega(\phi,n^{-1/d}) \tag{4.7}$$

holds for a constant C independent of n.

Proof. It is possible to find bounded measurable sets S_j such that

$$\mathbf{S}^d = \bigcup_{j=1}^{n} S_j,\ S_j\cap S_i=\varnothing,\ i\neq j,\ |S_j| = \mathcal{O}\left(\frac{1}{n}\right),\ \mathrm{diam}(S_j)=\mathcal{O}(n^{-\frac{1}{d}}). \tag{4.8}$$

We now define coefficients

$$c_j := \int_{S_j} h(t)d\omega(t),$$

and

$$\mu_j(t) := \begin{cases} \frac{1}{c_j}h(t), & t\in S_j,\ c_j\neq 0,\\ 0, & c_j\neq 0, t\in\mathbf{S}^d - S_j,\\ 0, & c_j=0.\end{cases}$$

Then we have

$$\int_{\mathbf{S}^d} \mu_j(t)d\omega(t) = \int_{S_j} \mu_j(t)d\omega(t) = 1, \tag{4.9}$$

and

$$\sum_{j=1}^{n} c_j\mu_j(t) = h(t). \tag{4.10}$$

In order to prove (4.7), it is sufficient to prove that

$$\int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} \left\| f - \sum_{j=1}^{n} c_j\phi(x \cdot t_j) \right\|_2^2 \mu_1(t_1)\cdots\mu_n(t_n)d\omega(t_1)\cdots d\omega(t_n) \leq C^2n^{-1}\omega(\phi, n^{-1/d})^2 \tag{4.11}$$

holds for all $\{t_j\}_{j=1}^n \subset \mathbf{S}^d$.

Indeed, if for all $\{t_j\}_{j=1}^n \subset \mathbf{S}^d$,

$$\left\| f - \sum_{i=1}^{n} c_j\phi(x \cdot t_j) \right\|_2 > C^2n^{-1}\omega(\phi, n^{-1/d})^2,$$

then it follows from (4.9) that

$$\begin{aligned} & \int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} \left\| f - \sum_{j=1}^{n} c_j\phi(x \cdot t_j) \right\|_2^2 \mu_1(t_1)\cdots\mu_n(t_n)d\omega(t_1)\cdots d\omega(t_n) \\ > \; & \int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} C^2n^{-1}\omega(\phi, n^{-1/d})^2\mu_1(t_1)\cdots\mu_n(t_n)d\omega(t_1)\cdots d\omega(t_n) \\ > \; & C^2n^{-1}\omega(\phi, n^{-1/d})^2, \end{aligned}$$

which contradicts (4.11). Therefore, their exists a set of points $\{t_j^*\}_{j=1}^n \subset \mathbf{S}^d$ such that (4.7) holds.

Now we use (4.9) and (4.10) to prove (4.11).

$$
\begin{aligned}
& \int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} \left\| f - \sum_{j=1}^{n} c_j \phi(t_j \cdot) \right\|_2^2 \mu_1(t_1) \cdots \mu_n(t_n) d\omega(t_1) \cdots d\omega(t_n) \\
= & \int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} \left(\int_{\mathbf{S}^d} \left| f - \sum_{j=1}^{n} c_j \phi(t_j \cdot x) \right|^2 d\omega(x) \right) \mu_1(t_1) \cdots \mu_n(t_n) d\omega(t_1) \cdots d\omega(t_n) \\
= & \int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} \left\{ \int_{\mathbf{S}^d} \left[\left(\int_{\mathbf{S}^d} h(t)\phi(t \cdot x) d\omega(t) \right)^2 - 2 \sum_{j=1}^{n} c_j \phi(t_j \cdot x) \int_{\mathbf{S}^d} h(t)\phi(t \cdot x) d\omega(t) \right. \right. \\
+ & \left. \left. \left(\sum_{j=1}^{n} c_j \phi(t_j \cdot x) \right)^2 \right] d\omega(x) \right\} \mu_1(t_1) \cdots \mu_n(t_n) d\omega(t_1) \cdots d\omega(t_n) \\
= & \int_{\mathbf{S}^d} \left\{ \left[\int_{\mathbf{S}^d} h(t)\phi(t \cdot x) d\omega(t) \right]^2 - 2 \int_{\mathbf{S}^d} h(t)\phi(t \cdot x) d\omega(t) \sum_{j=1}^{n} c_j \int_{\mathbf{S}^d} \phi(t_j \cdot x) \mu_j(t_j) d\omega(t_j) \right. \\
+ & \left. \int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} \left[\sum_{j=1}^{n} c_j \phi(t_j \cdot x) \right]^2 \mu_1(t_1) \cdots \mu_n(t_n) d\omega(t_1) \cdots d\omega(t_n) \right\} d\omega(x)
\end{aligned}
$$

$$
\begin{aligned}
= & \int_{\mathbf{S}^d}\left\{\left[\int_{\mathbf{S}^d} h(t)\phi(t\cdot x)d\omega(t)\right]^2 - 2\int_{\mathbf{S}^d} h(t)\phi(t\cdot x)d\omega(t)\sum_{j=1}^n c_j\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right.\\
+ & \left.\left[\sum_{j=1}^n c_j\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]^2\right\}d\omega(x)\\
+ & \int_{\mathbf{S}^d}\left\{\int_{\mathbf{S}^d}\cdots\int_{\mathbf{S}^d}\left[\sum_{j=1}^n c_j\phi(t_j\cdot x)\right]^2\mu_1(t_1)\cdots\mu_n(t_n)d\omega(t_1)\cdots d\omega(t_n)\right.\\
- & \left.\left[\sum_{j=1}^n c_j\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]^2\right\}d\omega(x)\\
= & \int_{\mathbf{S}^d}\left\{\int_{\mathbf{S}^d} h(t)\phi(t\cdot x)d\omega(t) - \sum_{j=1}^n c_j\int_{\mathbf{S}^d}\phi(t\cdot x)\mu_j(t)d\omega(t)\right\}^2 d\omega(x)\\
+ & \int_{\mathbf{S}^d}\left\{\int_{\mathbf{S}^d}\cdots\int_{\mathbf{S}^d}\left[\sum_{j=1}^n c_j^2(\phi(t_j\cdot x))^2\right.\right.\\
+ & \left. 2\sum_{i\neq j=1}^n c_ic_j\phi(t_i\cdot x)\phi(t_j\cdot x)\right]\mu_1(t_1)\cdots\mu_n(t_n)d\omega(t_1)\cdots d\omega(t_n)\\
- & \sum_{j=1}^n c_j^2\left[\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]^2\\
- & \left. 2\sum_{i\neq j=1}^n c_ic_j\left[\int_{\mathbf{S}^d}\phi(t_i\cdot x)\mu_i(t_i)d\omega(t_i)\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]\right\}d\omega(x)\\
= & \int_{\mathbf{S}^d}\left\{\int_{\mathbf{S}^d}\phi(t\cdot x)\left[h(t)-\sum_{j=1}^n c_j\mu_j(t)\right]d\omega(t)\right\}^2 d\omega(x)\\
+ & \int_{\mathbf{S}^d}\left\{\sum_{j=1}^n c_j^2\int_{\mathbf{S}^d}(\phi(t_j\cdot x))^2 d\mu_j(t_j)d\omega(t_j)\right.\\
+ & 2\sum_{i\neq j=1}^n c_ic_j\left[\int_{\mathbf{S}^d}\phi(t_i\cdot x)\mu_i(t_i)d\omega(t_i)\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]\\
- & \sum_{j=1}^n c_j^2\left[\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]^2\\
- & \left. 2\sum_{i\neq j=1}^n c_ic_j\left[\int_{\mathbf{S}^d}\phi(t_i\cdot x)\mu_i(t_i)d\omega(t_i)\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]\right\}d\omega(x)\\
= & \int_{\mathbf{S}^d}\left\{\sum_{j=1}^n c_j^2\int_S(\phi_\sigma(x-t_j))^2 d\mu_j(t_j)dt_j - \sum_{j=1}^n c_j^2\left[\int_{\mathbf{S}^d}\phi(t_j\cdot x)\mu_j(t_j)d\omega(t_j)\right]^2\right\}d\omega(x),
\end{aligned}
$$

where the last equation is deduced by (4.10). Then we have

$$\begin{aligned}
&\int_{\mathbf{S}^d}\cdots\int_{\mathbf{S}^d}\left\|f-\sum_{j=1}^n c_j\phi(t_j\cdot)\right\|_2^2\mu_1(t_1)\cdots\mu_n(t_n)d\omega(t_1)\cdots d\omega(t_n)\\
= &\sum_{j=1}^n c_j^2\left\{\int_{\mathbf{S}^d}\left[\int_{\mathbf{S}^d}(\phi(t\cdot x))^2\mu_j(t)d\omega(t)-\left(\int_{\mathbf{S}^d}\phi(t\cdot x)\mu_j(t)d\omega(t)\right)^2\right]d\omega(x)\right\}\\
= &\sum_{j=1}^n c_j^2\left\{\int_{\mathbf{S}^d}\left[\int_{S_j}(\phi(t\cdot x))^2\mu_j(t)d\omega(t)-\left(\int_{S_j}\phi(t\cdot x)\mu_j(t)d\omega(t)\right)^2\right]d\omega(x)\right\}\\
= &\sum_{j=1}^n c_j^2\left\{\int_{\mathbf{S}^d}\left[\int_{S_j}(\phi(t\cdot x))^2\mu_j(t)d\omega(t)\right.\right.\\
- &\ 2\int_{S_j}(\phi(t\cdot x))\mu_j(t)d\omega(t)\int_{S_j}(\phi(x\cdot s))\mu_j(s)d\omega(s)\\
+ &\left.\left.\left(\int_{S_j}\phi(x\cdot s)\mu_j(s)d\omega(s)\right)^2\right]d\omega(x)\right\}\\
= &\sum_{j=1}^n c_j^2\left\{\int_{\mathbf{S}^d}\left[\int_{S_j}\mu_j(t)\left(\phi(t\cdot x)-\int_{S_j}\mu_j(s)\phi(s\cdot x)d\omega(s)\right)^2d\omega(t)\right]d\omega(x)\right\}\\
= &\sum_{j=1}^n c_j^2\left\{\int_{\mathbf{S}^d}\left[\int_{S_j}\mu_j(t)\left(\int_{S_j}\mu_j(s)\phi(t\cdot x)ds-\int_{S_j}\mu_j(s)\phi(s\cdot x)ds\right)^2d\omega(t)\right]d\omega(x)\right\}\\
= &\sum_{j=1}^n c_j^2\left\{\int_{\mathbf{S}^d}\left[\int_{S_j}\mu_j(t)\left(\int_{S_j}\mu_j(s)\left(\phi(t\cdot x)-\phi(s\cdot x)\right)d\omega(s)\right)^2d\omega(t)\right]d\omega(x)\right\}.
\end{aligned}$$

It is easy to deduce that

$$\sum_{j=1}^n c_j=\sum_{j=1}^n\int_{S_j}h(t)d\omega(t)=\int_{\mathbf{S}^d}h(t)d\omega(t)\le\|h\|_1$$

and

$$\begin{aligned}
\sum_{j=1}^n c_j^2 &= \sum_{j=1}^n\left(\int_{S_j}h(t)d\omega(t)\right)^2\le\sum_{j=1}^n\int_{S_j}1d\omega(t)\int_{S_j}h^2(t)d\omega(t)\\
&= \sum_{j=1}^n|S_j|\int_{S_j}h^2(t)d\omega(t)\le\frac{\|h\|_2^2}{n}.
\end{aligned}$$

Furthermore, it follows from the definition of $\omega(\phi,t)$ that

$$|\phi(x\cdot t)-\phi(x\cdot s)|\le C\omega(\phi,diam(S_j))\le C\omega(\phi,n^{-1/d}),$$

holds for arbitrary $x \in \mathbf{S}^d$ and $t, s \in S_j$. Thus, we have

$$\begin{aligned} & \int_{\mathbf{S}^d} \cdots \int_{\mathbf{S}^d} \left\| f - \sum_{j=1}^n c_j \phi(t_j \cdot) \right\|_2^2 \mu_1(t_1) \cdots \mu_n(t_n) d\omega(t_1) \cdots d\omega(t_n) \\ \leq & \sum_{j=1}^n c_j^2 \left\{ \int_{S_j} \mu_j(t) dt \left(\int_{S_j} \mu_j(s) ds \right)^2 \omega(\phi, n^{-1/d})^2 \right\} \\ \leq & \; Cn^{-1} \omega(\phi, n^{-1/d}). \end{aligned}$$

This completes the proof of Theorem 4.5. □

References

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

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

[3] G. Brown, F. Dai, Y. S. Sun, Kolmogorov width of classes of smooth functions on sphere, J. Complex., 18 (2002), 1001-1023.

[4] M. Burger, A. Neubauer, Error bounds for approximation with neural networks, J. Approx. Theory, 112 (2001), 235-250.

[5] D. Chen, V. Menegatto, X. Sun, A necessary and sufficient condition for strictly positive definite functions on spheres, Proc. Amer. Math. Soc., 131 (2003), 2733-2740.

[6] R. DeVore, G. Lorentz, Constructive Approximation, Springer-Verlag, Berlin, 1993.

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

[8] 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.

[9] D. Haussler, Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension, J. Combin. Theory, Ser A 69 (1995), 217-232.

[10] A. Kuijlaars, E. Saff, Asymptotics for minimal discrete energy on the sphere, Trans. Amer. Math. Soc., 350 (1998), 523-538.

[11] V. Kürková, M. Sanguineti, Comparision of worst case errors in linear and neural network apprxoimation, IEEE. Trans. Inform. Theory, 48 (2002), 264-275.

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

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

[14] S. Lin, F. Cao, X. Chang, Z. Xu, A general radial quasi-interpolation operator on the sphere, J. Approx. Theory, 164 (2012), 1402-1414.

[15] S. Lin, J. Zeng, L. Xu, Z. Xu, Jackson-type inequalities for spherical neural networks with doubling weights, Neural Networks, 63 (2015), 57-65.

[16] G. G. Lorentz, M. Z. Golitchek, Y. Makovoz, Constructive Approximation, Advanced Problems, Springer-Verlag, New York, 1996.

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

[18] V. Maiorov, J. Ratsaby, On the degree of approximation by manifolds of finite pseudo-dimension, Constr. Approx.,15 (1999), 291-300.

[19] V. Maiorov, A. Pinkus, Lower bounds for approximation by MLP neural networks, Neurocomputing, 25 (1999), 81-91.

[20] V. Maiorov, R. Meir, Lower bounds for multivariate approximation by affine invariant dictionaries, IEEE Trans. Inform. Theory, 47 (2001), 1569-1575.

[21] V. Maiorov, On best approximation of classes by radial functions, J. Approx. Theory, 120 (2003), 36-70.

[22] V. Maiorov, Almost optimal estimates for best approximation by translates on a torus, Constr. Approx., 21 (2005), 337-349.

[23] V. Maiorov, Pseudo-dimension and entropy of manifolds formed by affine invariant dictionary, Adv. Comput. Math., 25 (2006), 435-450.

[24] V. Michel, R. Telschow, A non-linear approximation method on the sphere, Int. J. Geomath., 5 (2014), 195-224.

[25] H. Minh, Some Properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory, Constr. Approx., 32 (2010), 307-338.

[26] 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.

[27] H. Mhaskar, F. Narcowich, J. Prestin, J. Ward, L^p Bernstein estimates and approximation by spherical basis functions, Math. Comput., 79 (2010), 1647-1679.

[28] T. Morton, M. Neamtu, Error bounds for solving pseudo-differential equations on spheres by colloctation with zonal kernels, J. Approx. Theory, 114 (2002), 242-268.

[29] F. Narcowich, J. Ward, Scattered data interpolation on spheres: Error estimates and locally supported basis functions, SIAM J. Math. Anal., 33 (2002), 1393-1410.

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

[31] A. Pinkus, n-Widths in Approximation Theory, Berlin Heidelberg, Germay: Springer-Verlag, 1985.

[32] E. Saff, A. Kuijlaars, Distributing many points on a sphere, Math. Intell., 19 (1997), 5-11.

[33] X. Sun, Z. Chen, Spherical basis functions and uniform distribution of points on spheres, J. Approx. Theory, 151 (2008), 186-207.

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

[35] Y. Tsai, Z. Shih, All-frequency precomputed radiance transfer using spherical radial basis functions and clustered tensor approximation, ACM Trans. Graph., 25 (2006), 967-976.

[36] H. Wang, Probabilistic and average widths of sobolev spaces on compact two-point homogeneous spaces equipped with a Gaussian measure, Constr. Approx., 39 (2014), 485-516.

[37] H. Wendland, Scattered Data Approximation, Cambridge University Press, Cambridge, 2005.