Accepted Manuscript

Nonparametric regression using needlet kernels for spherical data

Shao-Bo Lin

PII: S0885-064X(18)30074-8
DOI: https://doi.org/10.1016/j.jco.2018.09.003
Reference: YJCOM 1379

To appear in: *Journal of Complexity*

Received date : 27 March 2017
Accepted date : 14 September 2018

Please cite this article as: S.-B. Lin, Nonparametric regression using needlet kernels for spherical data, *Journal of Complexity* (2018), https://doi.org/10.1016/j.jco.2018.09.003

Nonparametric regression using needlet kernels for spherical data ☆

Shao-Bo Lin*

Department of Mathematics, Wenzhou University, Wenzhou 325035, China

Abstract

Needlets have been recognized as state-of-the-art tools to tackle spherical data, due to their excellent localization properties in both spacial and frequency domains. This paper considers developing kernel methods associated with the needlet kernel for nonparametric regression problems whose predictor variables are defined on a sphere. Due to the localization property in the frequency domain, we prove that the regularization parameter of the kernel ridge regression associated with the needlet kernel can decrease arbitrarily fast. A natural consequence is that the regularization term for the kernel ridge regression is not necessary in the sense of rate optimality. Based on the excellent localization property in the spacial domain further, we prove that all l^q $(0 < q \leq 2)$ kernel regularization estimates associated with the needlet kernel, including the kernel lasso estimate and kernel bridge estimate, possess almost same generalization capability for a large range of regularization parameters in the sense of rate optimality. This finding tentatively reveals that, if the needlet kernel is utilized, then the choice of q might not have a strong impact on the generalization capability in some modeling contexts. The above two properties reveal the theoretical advantages of the needlet kernel in kernel methods for spherical nonparametric regression problems.

Keywords: Nonparametric regression, needlet kernel, spherical data, kernel ridge regression, l^q regularization.

☆The research was supported by the National Natural Science Foundation of China (Grant Nos. 61876133, 61502342, 11771012)

*Corresponding author: sblin1983@gmail.com

1. Introduction

Contemporary scientific investigations frequently encounter a common issue of exploring the relationship between a response variable and a number of predictor variables whose domain is the surface of a sphere. Examples include the study of gravitational phenomenon [15], cosmic microwave background radiation [13], tectonic plat geology [7] and image rendering [47]. As the sphere is topologically a compact two-point homogeneous manifold, some widely used schemes for the Euclidean space such as neural networks [17] and support vector machines [39] are no more the most appropriate methods for tackling spherical data. Designing efficient and exclusive approaches to extract useful information from spherical data has been a recent focus in statistical learning [14, 26, 34, 38].

Recent years have witnessed considerable approaches on nonparametric regression for spherical data. A classical and long-standing technique is the orthogonal series methods associated with spherical harmonics [1], with which the local performance of the estimate are quite poor, since spherical harmonics are not well localized but spread out all over the sphere. Another widely used technique is the stereographic projection methods [14], in which the statistical problems on the sphere were formulated in the Euclidean space by use of a stereographic projection. A major problem is that the stereographic projection usually leads to a distorted theoretical analysis paradigm and a relatively sophisticate statistical behavior. Localization methods, such as the Nadaraya-Watson-like estimate [38], local polynomial estimate [3] and local linear estimate [26] are also interesting nonparametric approaches. Unfortunately, the manifold structure of the sphere is not well taken into account in these approaches. Mihn [32] also developed a general theory of reproducing kernel Hilbert space on the sphere and advocated to utilize the kernel methods to tackle spherical data. However, for some popular kernels such as the Gaussian [33] and polynomials [6], kernel methods suffer from either a similar problem as the localization methods, or a similar drawback as the orthogonal series methods. In fact, it remains open that whether there is an exclusive kernel for spherical data such that both the manifold structure of the sphere and the localization requirement are sufficiently considered.

Our focus in this paper is not on developing a novel technique to cope with spherical nonparametric regression problems, but on introducing an exclusive kernel for kernel

methods. To be detailed, we aim at finding a kernel that possesses excellent spacial localization property and makes fully use of the manifold structure of the sphere. Recalling that one of the most important factors to embody the manifold structure is the special frequency domain of the sphere, a kernel which can control the frequency domain freely is preferable. Thus, the kernel we need is required to possess excellent localization property, both in spacial and frequency domains. Under this circumstance, the needlet kernel comes into our sights. Needlets [29, 30, 36, 37] are new kinds of second-generation spherical wavelets, which can be shown to make up a tight frame with both perfect spacial and frequency localization properties. Furthermore, needlets have a clear statistical nature [2, 18], the most important of which is that in the Gaussian and isotropic random fields, the random spherical needlets behave asymptotically as an i.i.d. array [2]. It can be found in [36] that the spherical needlets correspond a needlet kernel, which is also well localized in spacial and frequency domains. Consequently, the needlet kernel is proved to possess the reproducing property [36, Lemma 3.8], compressible property [36, Theorem 3.7] and best approximation property [36, Corollary 3.10].

The aim of the present article is to pursue the theoretical advantages of the needlet kernel in kernel methods for spherical nonparametric regression problems. If the kernel ridge regression (KRR) associated with the needlet kernel is employed, the model selection then boils down to determining the frequency and regularization parameter. Due to the excellent localization in the frequency domain, we find that the regularization parameter of KRR can decrease arbitrarily fast for some suitable frequency. An extreme case is that the regularization term is not necessary for KRR in the sense of rate optimality. This attribution is totally different from other kernels without good localization property in the frequency domain [10], such as Gaussian [33] and Abel-Poisson [15] kernels. We attribute the above property as the first feature of the needlet kernel. Besides the good generalization capability, some real world applications also require the estimate to possess the smoothness, low computational complexity and sparsity [39]. This guides us to consider l^q $(0 < q \leq 2)$ kernel regularization schemes (KRS) associated with the needlet kernel, including the kernel bridge regression and kernel lasso estimate [48]. The first feature of the needlet kernel implies that the generalization capability of all l^q-KRS with

$0 < q \leq 2$ are almost the same, provided the regularization parameter is set to be small enough. However, such a setting makes all l^q-KRS with $0 < q \leq 2$ similar as least squares. To distinguish different behaviors of the l^q-KRS, we should establish similar results for large regularization parameters. By the aid of a probabilistic cubature formula and the excellent localization property in spacial domain of the needlet kernel, we find that all l^q-KRS with $0 < q \leq 2$ can attain the same almost optimal generalization error bounds, provided the regularization parameter is not larger than $\mathcal{O}(m^{q-1}\varepsilon)$. Here m is the number of samples and ε is the prediction accuracy. This implies that the choice of q does not have a strong impact in terms of the generalization capability for l^q-KRS, with relatively large regularization parameters depending on q. We consider it as the second feature of the needlet kernel.

2. Main Results

In spherical nonparametric regression problems with predictor variables $X \in \mathcal{X} = \mathbf{S}^d$ and response variables $Y \in \mathcal{Y} \subseteq \mathbf{R}$, we observe m i.i.d. samples $\mathbf{z}_m = \{(x_i, y_i)\}_{i=1}^m$ from an unknown distribution ρ, where $\mathbf{S}^d$ is the unit sphere embedded into $\mathbf{R}^{d+1}$. Throughout this paper, we assume $\mathcal{Y} \subseteq [-M, M]$ almost surely with $M > 0$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. One natural measurement of the estimate f is the generalization error $\mathcal{E}(f) := \int_{\mathcal{Z}} (f(X) - Y)^2 d\rho$, which is minimized by the regression function [17, P.2] defined by $f_\rho(x) := \int_{\mathcal{Y}} Y d\rho(Y|x)$. It is well known [9] that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho \tag{2.1}$$

for arbitrary $f \in L^2_{\rho_X}$, where $L^2_{\rho_X}$ denotes the Hilbert space of ρ_X-square integrable functions with norm $\|\cdot\|_\rho$ and ρ_X is the marginal distribution of ρ.

As it is impossible to obtain a nontrivial convergence rate without imposing any restriction on the distribution ρ [17, Theorem 3], we should introduce certain a-priori information. Let $r \geq 0$. Define the Bessel-potential Sobolev class W_r [31] to be all f such that

$$\|f\|_{W_r} := \left\| \sum_{k=0}^{\infty} (k + (d-1)/2)^r P_k f \right\|_2 \leq 1,$$

where

$$D_k^d := \begin{cases} \frac{2k+d-1}{k+d-1}\binom{k+d-1}{k}, & k \geq 1; \\ 1, & k = 0, \end{cases}$$

is the dimension of the family of spherical homogeneous harmonic polynomials of degree k, $P_k f := \sum_{j=1}^{D_k^d} \hat{f}_{k,j} Y_{k,j} := \langle f, Y_{k,j} \rangle Y_{k,j}$, $\{Y_{k,j} :\}_{j=1}^{D_k^d}$ is an arbitrary orthonormal basis of spherical homogeneous harmonic polynomials of degree k, $\langle f, g \rangle = \int_{\mathbf{S}^d} f(x)g(x)d\omega(x)$ and $d\omega$ denotes the surface area element on $\mathbf{S}^d$. It follows from the well known Sobolev embedding theorem that $W_r \subset C(\mathbf{S}^d)$, provided $r > d/2$. In our analysis, we assume $f_\rho \in W_r$ with $r > d/2$.

We then introduce the spherical needlet kernel. A function η is said to be admissible [37] if $\eta \in C^\infty[0, \infty)$ satisfies the following condition:

$$\text{supp}\eta \subset [0, 2], \eta(t) = 1 \text{ on } [0, 1], \text{ and } 0 \leq \eta(t) \leq 1 \text{ on } [1, 2]. \tag{2.2}$$

The spherical needlet kernel [29, 36] is defined by

$$K_n(x \cdot x') = \sum_{k=0}^{\infty} \eta\left(\frac{k}{n}\right) \frac{D_k^d}{|\mathbf{S}^d|} P_k^{d+1}(x \cdot x'), \tag{2.3}$$

where $|\mathbf{S}^d|$ is the volume of the unit sphere $\mathbf{S}^d$, P_k^{d+1} is the normalized Legendre polynomial [35] with $P_k^{d+1}(1) = 1$ and

$$\int_{-1}^{1} P_k^{d+1}(t) P_j^{d+1}(t)(1-t^2)^{\frac{d-2}{2}} dt = \frac{|\mathbf{S}^d|}{|\mathbf{S}^{d-1}| D_k^d} \delta_{k,j}, \tag{2.4}$$

and $\delta_{k,j}$ is the usual Kronecker symbol. The property of the filtered kernel depends heavily on the filter $\eta(\cdot)$. By the assumption (2.2), the summation in (2.3) is finite, that is, one only needs to consider $k \leq 2n$. Thus, K_n is actually a polynomial-type kernel. We refer the readers to [43, P.101] for some concrete needlet kernels satisfying (2.3).

2.1. Kernel ridge regression associated with the needlet kernel

Our first result concerns the learning rate of kernel ridge regression (KRR) associated with the needlet kernel

$$f_{\mathbf{z},\lambda} := \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_{K_n}^2 \right\}, \tag{2.5}$$

where $\mathcal{H}_K$ is the reproducing kernel Hilbert space associated with the needlet kernel K_n. Since $y \in [-M, M]$, there holds $\mathcal{E}(\pi_M f) \leq \mathcal{E}(f)$ for arbitrary $f \in L^2_{\rho_X}$, where $\pi_M u := \min\{M, |u|\} sign(u)$ is the truncation operator. The following Theorem 2.1 illustrates the generalization capability of KRR associated with the needlet kernel.

Theorem 2.1. *Let $r > d/2$ and let c_1 and c_2 be absolute constants such that for each $\varepsilon > 0$, we pick an n_ε with $c_1\varepsilon^{-1/(2r)} \leq n_\varepsilon \leq c_2\varepsilon^{-1/2r}$. Then, there exist constants C_1, C_2 such that for all $m \geq 1$ there exists ε_- and ε_+ satisfying*

$$C_1 m^{-2r/(2r+d)} \leq \varepsilon_- \leq \varepsilon_+ \leq C_2 (m/\log m)^{-2r/(2r+d)}, \tag{2.6}$$

such that for all $0 \leq \lambda \leq M^{-2}|\mathbf{S}^d|^{-1}\varepsilon$ and $\varepsilon > 0$, the following holds for $f_{\mathbf{z},\lambda}$ defined by (2.5) with kernel K_{n_ε}:

1. If $\varepsilon < \varepsilon_-$, then

$$\sup_{f_\rho \in W_r} \mathbf{P}\{\mathbf{z} : \|f_\rho - \pi_M f_{\mathbf{z},\lambda}\|_\rho^2 > \varepsilon\} \geq C_0. \tag{2.7}$$

2. If $\varepsilon \geq \varepsilon_+$, then

$$e^{-C_3 m\varepsilon} \leq \sup_{f_\rho \in W_r} \mathbf{P}\{\mathbf{z} : \|f_\rho - \pi_M f_{\mathbf{z},\lambda}\|_\rho^2 > \varepsilon\} \leq e^{-C_4 m\varepsilon}. \tag{2.8}$$

Here, C_0, C_3, C_4 are constants depending only on M, r, ρ, and d.

We give several remarks on Theorem 2.1 below. In some real world applications, there are only m data available, and the purpose of learning is to produce an estimate with the prediction error at most ε and we are consulted to assess the probability of success. It is obvious that the probability depends on m and ε. The inequality (2.7) shows that if the learning task is to yield an estimate with accuracy $\varepsilon \leq \varepsilon_-$ with ε_- satisfying (2.6), then KRR associated with the needlet kernel fails with high probability. To circumvent it, the only way is to loose the requirement to the accuracy, just as inequalities (2.8) purport to show. (2.8) says that if the accuracy is larger than ε_+, then the probability of success of KRR is at least $1 - e^{-C_4 m\varepsilon}$. The first inequality (lower bound) of (2.8) implies that this confidence cannot be improved further. The values of ε_- and ε_+ thus are critical. Inequalities (2.6) depicts that, for KRR, there holds

$$[\varepsilon_-, \varepsilon_+] \subset [C_1 m^{-2r/(2r+d)}, C_2 (m/\log m)^{-2r/(2r+d)}].$$

This implies that the interval $[\varepsilon_-, \varepsilon_+]$ is almost the shortest one in the sense that up to a logarithmic factor, the lefthand and righthand of the interval are asymptotically identical.

Furthermore, Theorem 2.1 also presents a sharp phase transition phenomenon of KRR. The behavior of the confidence function changes dramatically within the critical interval $[\varepsilon_-, \varepsilon_+]$. It drops from a constant C_0 to an exponentially small quantity. All the above assertions show that the learning performance of KRR is actually revealed in Theorem 2.1.

To exhibit an explicit generalization error bound, we present the following estimates in terms of expectation. The following Corollary 2.2 can be directly deduced from Theorem 2.1 and [12, Chapter 3], if we notice the identity:

$$\mathbf{E}(\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)) = \int_0^\infty \mathbf{P}\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) > \varepsilon\} d\varepsilon.$$

Corollary 2.2. *Let $r > d/2$ and c_3, c_4 be constants such that for each $m \in \mathbf{N}$, we pick an n_m with $c_3 m^{1/(2r+d)} \le n_m \le c_4 m^{1/(2r+d)}$. Then, there exist constants C_5 and C_6 depending only on M, d, r and ρ such that for all $0 \le \lambda \le M^{-2} m^{-2r/(2r+d)}$, the following holds for $f_{\mathbf{z},\lambda}$ defined by (2.5) with kernel K_{n_m}:*

$$C_5 m^{-2r/(2r+d)} \le \sup_{f_\rho \in W_r} \mathbf{E}\left\{\|f_\rho - f_{\mathbf{z},\lambda}\|^2\right\} \le C_6 (m/\log m)^{-2r/(2r+d)}. \tag{2.9}$$

An interesting finding in Theorem 2.1 and Corollary 2.2 is that the regularization parameter of KRR can decrease arbitrarily fast, provided it is smaller than $M^{-2} m^{-2r/(2r+d)}$. The extreme case is that least-squares possess the same generalization performance as KRR. We attribute this as the first feature of needlet kernel in spherical nonparametric regression problems. This phenomenon is not a surprising discovery in the realm of nonparametric regression, due to the needlet kernel's localization property in the frequency domain. Via controlling the frequency of the needlet kernel, $\mathcal{H}_K$ is essentially a linear space with finite dimension. Thus, Theorems 3.2 and 11.3 in [17] together with Lemma 4.2 below automatically yields the same learning rates as (2.9) for least squares associated with the needlet kernel. Compared with Theorems 3.2 and 11.3 in [17], the novelty of our results is that we consider learning rate analysis in probability and study learning capability for KRR rather than least squares.

Due to Lemma 4.3 below, K_n is a Mercer kernel. Thus, the solution to (2.5) takes the form of $\sum_{i=1}^m a_i K_n(x_i \cdot x)$, where $(a_1, \ldots, a_m)^T := (A + m\lambda I)^{-1}(y_1, \ldots, y_m)^T$ and $A := (K_n(x_i \cdot x_j))_{i,j=1}^m$. Theorem 2.1 and Corollary 2.2 show that the purpose of introducing regularization term in KRR is only to conquer the singularity of the kernel matrix A, since

$m > D_n^{d+1}$ in our setting. Under this circumstance, a small λ leads to the ill-condition of the matrix $A + m\lambda I$ and a large λ conducts large approximation error. Corollary 2.2 illustrates that if the needlet kernel is employed, then we can set $\lambda = m^{-2r/(2r+d)}$ to guarantee both the small condition number of the kernel matrix and almost optimal generalization error bound. From Corollary 2.2, if we set $\lambda = m^{-2r/(2r+d)}$, it is easy to deduce that KRR possesses the optimal learning rate $m^{-2r/(2r+d)}$ [12] and the minimal eigenvalue of the matrix $A + m\lambda I$ is $m^{d/(2r+d)}$, which can guarantee that the matrix inverse technique is suitable to solve (2.5). Based on our theoretical analysis, it would be interesting to develop a feasible and efficient algorithm to solve (2.5) with $\lambda = 0$, since the above optimal regularization parameter depends on the a-priori knowledge of the data and needs cross-validation strategy to determine it.

2.2. l^q kernel regularization schemes associated with the needlet kernel

Our second result is to study the learning capability of the l^q kernel regularization scheme (KRS) whose hypothesis space is the sample dependent hypothesis space [48] associated with $K_n(\cdot,\cdot)$, i.e.

$$\mathcal{H}_{K,\mathbf{z}} := \left\{ \sum_{i=1}^{m} a_i K_n(x_i, \cdot) : a_i \in \mathbf{R} \right\}.$$

The corresponding l^q-KRS is defined by

$$f_{\mathbf{z},\lambda,q} \in \arg\min_{f \in \mathcal{H}_{K,\mathbf{z}}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \Omega_{\mathbf{z}}^q(f) \right\}, \tag{2.10}$$

where

$$\Omega_{\mathbf{z}}^q(f) := \inf_{(a_1,\ldots,a_m)\in\mathbf{R}^m} \sum_{i=1}^{m} |a_i|^q, \qquad \text{for } f = \sum_{i=1}^{m} a_i K_n(x_i, \cdot).$$

With different choices of q, (2.10) leads to various specific forms of the l^q regularizer. $f_{\mathbf{z},\lambda,2}$ corresponds to the kernel ridge regression [39], which smoothly shrinks the coefficients toward zero and $f_{\mathbf{z},\lambda,1}$ leads to lasso [46], which sets small coefficients exactly at zero and thereby also serves as a variable selection operator. The varying forms and properties of $f_{\mathbf{z},\lambda,q}$ make the choice of order q crucial in applications. Apparently, an optimal q may depend on many factors such as the learning algorithms, the purposes of studies and so forth. The following Theorem 2.3 shows that if the needlet kernel is utilized in l^q-KRS,

then q may not have an important impact on the generalization capability for a large range of regularization parameters in the sense of rate optimality.

Theorem 2.3. *Let $r > d/2$, ρ_X is the uniform distribution and c_5, c_6 be absolute constants such that for each $\varepsilon > 0$, we pick an n_ε with $c_1\varepsilon^{-1/(2r)} \leq n_\varepsilon \leq c_2\varepsilon^{-1/2r}$. Then, there exist constants C_1', C_2' such that for all $m \geq 1$ there exists ε_- and ε_+ satisfying*

$$C_1' m^{-2r/(2r+d)} \leq \epsilon_m^- \leq \epsilon_m^+ \leq C_2'(m/\log m)^{-2r/(2r+d)}, \tag{2.11}$$

such that for all $0 \leq \lambda \leq M^{-2}|\mathbf{S}^d|^{-1}\varepsilon$, $0 < q \leq 2$ and $\varepsilon > 0$, the following holds for $f_{\mathbf{z},\lambda,q}$ defined by (2.10) with kernel K_{n_ε}:

1. *If $\varepsilon < \varepsilon_-$, then*

$$\sup_{f_\rho \in W_r} \mathbf{P}\{\mathbf{z} : \|f_\rho - \pi_M f_{\mathbf{z},\lambda,q}\|_\rho^2 > \varepsilon\} \geq C_0'. \tag{2.12}$$

2. *If $\varepsilon \geq \varepsilon_+$, then*

$$e^{-C_3' m\varepsilon} \leq \sup_{f_\rho \in W_r} \mathbf{P}\{\mathbf{z} : \|f_\rho - \pi_M f_{\mathbf{z},\lambda,q}\|_\rho^2 > \varepsilon\} \leq e^{-C_4' m\varepsilon}. \tag{2.13}$$

Here, C_0', C_3', C_4' are constants depending only on M, ρ, r, q and d.

Compared with KRR (2.5), a common consensus is that l^q-KRS (2.10) may bring a certain additional interest such as the sparsity for suitable choice of q. However, it should be noticed that this assertion may not be true, since it depends on the value of the regularization parameter. If the the regularization parameter is extremely small, then l^q-KRS for any $q \in (0, 2]$ behaves similar as least squares. Under this circumstance, Theorem 2.3 obviously holds according to Theorem 2.1. To distinguish the features of l^q-KRS with different q, one should consider a relatively large regularization parameter. Theorem 2.3 shows that for a large range of regularization parameters, all l^q-KRS associated with the needlet kernel can attain the same, almost optimal, generalization error bounds. It should be highlighted that the quantity $m^{q-1}\varepsilon$ is, to the best of knowledge, almost the largest value of the regularization parameter among all existing results. In fact , the regularization parameter adopted in the present paper is larger than that in [21, 41, 42, 45, 48]. Furthermore, we find that $m^{q-1}\varepsilon$ is sufficient to embody the feature of l^q kernel regularization schemes. Taking the kernel lasso for example, the regularization parameter derived in Theorem 2.3 asymptotically equals to ε. It is easy to see that, to yield a prediction accuracy ε, we have

$$f_{\mathbf{z},\lambda,1} \in \arg\min_{f \in \mathcal{H}_{K,\mathbf{z}}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \Omega_{\mathbf{z}}^1(f) \right\},$$

and

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2 \leq \varepsilon.$$

According to the structural risk minimization principle and $\lambda = \varepsilon$, it is easy to derive

$$\Omega_{\mathbf{z}}^1(f_{\mathbf{z},\lambda,1}) \leq C$$

with high probability.

Intuitively, the generalization capability of l^q-KRS (2.10) with a large regularization parameter may depend on the choice of q. While from Theorem 2.3 it follows that the learning schemes defined by (2.10) can indeed achieve the same asymptotically optimal rates for all $q \in (0, 2]$. In other words, on the premise of embodying the feature of l^q-KRS with different q, the choice of q has no influence on the generalization capability in the sense of rate optimality. Thus, we can determine q by taking other non-generalization considerations such as the smoothness, sparsity, and computational complexity into account. We attribute this phenomenon as the second feature of the needlet kernel in spherical nonparametric regression problems.

Finally, we explain the reason for this phenomenon by taking needlet kernel's perfect localization property in the spacial domain into account. To approximate $f_\rho(x)$, due to the localization property of K_n, we can construct an approximant in $\mathcal{H}_{\mathbf{z},K}$ with a few $K_n(x_i, \cdot)$ whose centers x_i are near to x. As f_ρ is bounded by M, then the coefficient of these terms are also bounded. That is, we can construct, in $\mathcal{H}_{\mathbf{z},K}$, a good approximant, whose l^q norm is bounded for arbitrary $0 < q \leq 2$. Then, using the standard error decomposition technique in [49] that divides the generalization error into the approximation error and sample error, the approximation error of l^q-KRS is independent of q. For the sample error, we can tune λ that may depend on q to offset the effect of q. Then, a generalization error estimate independent of q is natural.

3. Related Work and Discussion

Spherical regression, which was originally proposed in [23] in crystallography, is a regression model involving predictor and response variables that take values on spheres. Up to an asymptotically vanishing normalization task, the spherical regression problems

can be decomposed into d distinct regression problems, whose response variables are real numbers [26]. Roughly speaking, there are five approaches to tackle spherical nonparametric regression problems. The first method is spherical harmonics [1]. Due to the poor spacial localization property, this method do not allow learning efficiently high-resolution signals (or functions). The second method is the stereographic projection [14]. Although this method is intuitively feasible, there lacks concrete statistical behavior analysis. The third one is the localization method such as the Nadaraya-Watson-like estimate [38], local polynomial estimate [3] and local linear estimate [26]. The pros of this method is its perfect theoretical behavior, but the cons is that the manifold structure of the sphere may be neglected. The fourth one is spherical wavelets and needlets [15, 34]. The core of these methods is to present thresholds to the spherical wavelet coefficients or spherical needlet coefficients to avoid overfitting [34], which makes the approach possess a clearly statistical behavior. Unfortunately, some strong restrictions should be imposed to the distribution ρ to derive a concrete learning rates [34]. The last one is the spherical kernel methods, which was original proposed in [32]. This approach extends the popular support vector machines [39] in Euclidean space to tackle spherical data and provides concrete learning rates. However, it is still a question that which kernel is suitable for this purpose. Our aim in this paper is to present a theoretical analysis of kernel methods associated with the needlet kernel and reveal the advantages of needlet kernel. We compare our results with the following related work [32, 20, 6, 19, 34].

There are two types of polynomial kernels for spherical data learning: the localized kernels and non-localized kernels. For the non-localized kernels, there are three papers focused on its applications in nonparametric regression. [32, Theorem 1] is the first one to derive the learning rate of KRR associated with the polynomial kernel $(1 + x \cdot x')^n$. However their learning rates were built upon the assumption that f_ρ is a polynomial. [20] omitted this assumption by using the eigenvalue estimate of the polynomial kernel. But the derived learning rate of [20] is not optimal. [6] conducted a learning rate analysis for KRR associated with the reproducing kernel of the space $(\Pi_n^d, L_2(\mathbf{S}^d))$ and derived the similar learning rate as [20], where Π_n^d is the set of spherical polynomials of degrees at most n. In a nutshell, there is not almost optimal learning rate analysis for KRR

associated with non-localized kernels. Theorem 2.1 in the present paper shows that KRR associated with the needlet kernel can obtain the almost optimal learning rates. Furthermore, Theorem 2.3 shows that l^q-KRS with $0 < q \leq 2$ can also reach the almost optimal learning rates. It should be pointed out that when $y_i = f_\rho(x_i)$, the learning rate of least squares (KRR with $\lambda = 0$) associated with a localized kernel was derived in [19]. There are two difference between our paper and [19]. The first one is that we are faced with nonparametric regression problem (with noise), while [19] focused on the approximation problems (without noise). The other is that we are concerned with KRR and l^q-KRS rather than least squares. Compared with [34], our novelties are also two folds. The first one is that the learning schemes are different. To be detailed, in [34], Monnier constructed the nonparametric estimator by using a stochastic and deterministic shrinkage to the needlet coefficients. However, in the present paper, we are interested in KRR and l^q-KRS associated with the needlet kernel. The other one is that the theoretical analysis in [34] requires strong assumptions on the distribution ρ, ρ_X should be uniform distribution and the noise should be Gaussian. However, all these assumptions are removed (or weaken) in our analysis for KRR (or l^q-KRS).

The main features of the needlet kernel used in this paper are the localization property in both frequency and spacial domains. It should be noted that there are some other kernels possessing this property, such as the kernels proposed in [5, 16, 19, 30]. In fact, using the same methods in this paper, we can derive similar results for these kernels. Since needlets' popularity in statistics and real world applications, we only present the learning rate analysis for the needlet kernel.

4. Probabilistic Cubature Formula

In this section, we present some special properties of the needlet kernel and a probabilistic cubature formula, which are crucial to our proofs.

4.1. Special features of the needlet kernel

According to the definition of the admissible function (2.2), it is easy to see that K_n possesses excellent localization property in the frequency domain. The following

Lemma 4.1 which can be found in [36] and [5] yields that K_n also possesses perfect spacial localization property.

Lemma 4.1. *Let η be admissible. Then for every $k > 0$ and $s \geq 0$ there exists a constant $\tilde{C}_1$ depending only on k, s, d and η such that*

$$\left| \frac{d^s}{d\theta^s} K_n(\cos\theta) \right| \leq \tilde{C}_1 \frac{n^{d+2s}}{(1+n\theta)^k}, \ \theta \in [0, \pi].$$

For $f \in L^1(\mathbf{S}^d)$, we write

$$K_n * f(x) := \int_{\mathbf{S}^d} K_n(x \cdot x') f(x') d\omega(x'). \tag{4.1}$$

The following Lemma 4.2 can be deduced from [36, Theorem 3.7&Corollary 3.10], which is stemmed by the spacial localization property of K_n and will play a crucial role in proving Theorems 2.1 and 2.3.

Lemma 4.2. *Let $1 \leq p \leq \infty$. For any $f \in W_r$ with $r > d/2$, we have $K_n * f \in \Pi_{2n}^d$,*

$$\|K_n * f\|_{C(\mathbf{S}^d)} \leq \tilde{C}_2 \|f\|_{C(\mathbf{S}^d)}, \ \text{and} \ \|f - K_n * f\|_{L^p(\mathbf{S}^d)} \leq \tilde{C}_2 n^{-r},$$

where $\tilde{C}_2$ is a constant depending only on d, p, r and η.

It is obvious that K_n is continuous, symmetric and positive semi-definite, thus it follows from [32, Theorem 14] (or [40, Chapter 4]) that K_n is a reproducing kernel with corresponding reproducing kernel Hilbert space (RKHS).

Lemma 4.3. *The reproducing kernel Hilbert space associated with K_n is the space $(\Pi_{2n}^d, \langle \cdot, \cdot \rangle_{K_n})$ with*

$$\langle f, g \rangle_{K_n} := \sum_{k:\eta(k/n)\neq 0} \sum_{j=1}^{D_k^d} [\eta(k/n)]^{-1} \hat{f}_{k,j} \hat{g}_{k,j}.$$

The following proposition providing the boundedness of the RKHS norm of $K_n * f$ will play an important role in proving Theorem 2.1.

Proposition 4.4. *Let $K_n * f$ be defined by (4.1). Then we have*

$$\|K_n * f\|_{K_n}^2 \leq |\mathbf{S}^d| \|f\|_{C(\mathbf{S}^d)}.$$

Proof. The well known Funk-Hecke formula [32, Theorem 21] says

$$\int_{\mathbf{S}^d} K_n(x\cdot x')Y_{k,j}(x')d\omega(x') = B(K_n,k)Y_{k,j}(x), \tag{4.2}$$

where

$$B(K_n,k) = |\mathbf{S}^{d-1}|\int_{-1}^{1} P_k^{d+1}(t)K_n(t)(1-t^2)^{\frac{d-2}{2}}dt.$$

Then,

$$\begin{aligned}
\widehat{K_n * f}_{u,v} &= \int_{\mathbf{S}^d} K_n * f(x)Y_{u,v}(x)d\omega(x) = \int_{\mathbf{S}^d}\int_{\mathbf{S}^d} K_n(x\cdot x')f(x')d\omega(x')Y_{u,v}(x)d\omega(x)\\
&= \int_{\mathbf{S}^d} f(x')\int_{\mathbf{S}^d} K_n(x\cdot x')Y_{u,v}(x)d\omega(x)d\omega(x')\\
&= \int_{\mathbf{S}^d} |\mathbf{S}^{d-1}|\int_{-1}^{1} K_n(t)P_u^{d+1}(t)(1-t^2)^{\frac{d-2}{2}}dtY_{u,v}(x')f(x')d\omega(x')\\
&= |\mathbf{S}^{d-1}|\hat{f}_{u,v}\int_{-1}^{1} K_n(t)P_u^{d+1}(t)(1-t^2)^{\frac{d-2}{2}}dt.
\end{aligned}$$

But, (2.2) and (2.4) imply

$$\begin{aligned}
\int_{-1}^{1} K_n(t)P_u^{d+1}(t)(1-t^2)^{\frac{d-2}{2}}dt &= \int_{-1}^{1}\sum_{k=0}^{2n}\eta\left(\frac{u}{n}\right)\frac{D_k^d}{|\mathbf{S}^d|}P_k^{d+1}(t)P_u^{d+1}(t)(1-t^2)^{\frac{d-2}{2}}dt\\
&= \int_{-1}^{1}\eta\left(\frac{u}{n}\right)\frac{D_u^d}{|\mathbf{S}^d|}P_u^{d+1}(t)P_u^{d+1}(t)(1-t^2)^{\frac{d-2}{2}}dt\\
&= \eta\left(\frac{u}{n}\right)\frac{D_u^d}{|\mathbf{S}^d|}\frac{|\mathbf{S}^d|}{|\mathbf{S}^{d-1}|D_u^d} = \eta\left(\frac{u}{n}\right)\frac{1}{|\mathbf{S}^{d-1}|}.
\end{aligned}$$

We obtain

$$\widehat{K_n * f}_{u,v} = \eta\left(\frac{u}{n}\right)\hat{f}_{u,v}.$$

This together with Lemma 4.3 and (2.2) shows

$$\begin{aligned}
\|K_n * f\|_{K_n}^2 &= \sum_{u:\eta(u/n)\neq 0}\left[\eta\left(\frac{u}{n}\right)\right]^{-1}\sum_{v=1}^{D_u^d}(\widehat{K_n * f}_{u,v})^2\\
&\le \sum_{u:\eta(u/n)\neq 0}\sum_{v=1}^{D_u^d}\hat{f}_{u,v}^2 \le \|f\|_{L^2(\mathbf{S}^d)}^2 \le |\mathbf{S}^d|\|f\|_{C(\mathbf{S}^d)}^2.
\end{aligned}$$

The proof of Proposition 4.4 is completed. ■

4.2. Probabilistic cubature formula

To present the probabilistic cubature formula, we need the following three lemmas. The first one is the Nikolskii inequality for spherical polynomials [27].

Lemma 4.5. *Let $1 \le p < q \le \infty$, $n \ge 1$ be an integer. Then*

$$\|Q\|_{L^q(\mathbf{S}^d)} \le \tilde{C}_3 n^{\frac{d}{p}-\frac{d}{q}} \|Q\|_{L^p(\mathbf{S}^d)}, \ Q \in \Pi_n^d$$

where $\tilde{C}_3$ is a constant depending only on d.

The second one is the well known Bernstein inequality, (see [49] for example).

Lemma 4.6. *Let ξ be a random variable on a probability space Z with mean $\mathbf{E}(\xi)$, variance $\sigma^2(\xi) = \sigma^2$. If $|\xi(z) - \mathbf{E}(\xi)| \le M_\xi$ for almost all $\mathbf{z} \in Z$. then, for all $\varepsilon > 0$,*

$$\mathbf{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbf{E}(\xi)\right| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2 + \frac{1}{3}M_\xi\varepsilon\right)}\right\}.$$

To state the last lemma, we need introduce the following definitions. Let $\mathcal{V}$ be a finite dimensional vector space with norm $\|\cdot\|_{\mathcal{V}}$, and $\mathcal{U} \subset \mathcal{V}^*$ be a finite set. Here $\mathcal{V}^*$ denotes the dual space of $\mathcal{V}$. We say that $\mathcal{U}$ is a norm generating set for $\mathcal{V}$ if the mapping $T_{\mathcal{U}} : \mathcal{V} \to \mathbf{R}^{Card(\mathcal{U})}$ defined by $T_{\mathcal{U}}(x) = (u(x))_{u\in\mathcal{U}}$ is injective, where $Card(\mathcal{U})$ is the cardinality of the set $\mathcal{U}$ and $T_{\mathcal{U}}$ is named as the sampling operator. Let $\mathcal{W} := T_{\mathcal{U}}(\mathcal{V})$ be the range of $T_{\mathcal{U}}$, then the injectivity of $T_{\mathcal{U}}$ implies that $T_{\mathcal{U}}^{-1} : \mathcal{W} \to \mathcal{V}$ exists. Let $\mathbf{R}^{Card(\mathcal{U})}$ have a norm $\|\cdot\|_{\mathbf{R}^{Card(\mathcal{U})}}$, with $\|\cdot\|_{\mathbf{R}^{Card(\mathcal{U})^*}}$ being its dual norm on $\mathbf{R}^{Card(\mathcal{U})^*}$. Equipping $\mathcal{W}$ with the induced norm, and let $\|T_{\mathcal{U}}^{-1}\| := \|T_{\mathcal{U}}^{-1}\|_{\mathcal{W}\to\mathcal{V}}$. Then the following Lemma 4.7 can be found in [28].

Lemma 4.7. *Let $\mathcal{U}$ be a norm generating set for $\mathcal{V}$, with $T_{\mathcal{U}}$ being the corresponding sampling operator. If $v \in \mathcal{V}^*$ with $\|v\|_{\mathcal{V}^*} \le A$, then there exist real numbers $\{a_u\}_{u\in\mathcal{Z}}$, depending only on v such that for every $t \in \mathcal{V}$,*

$$v(t) = \sum_{u\in\mathcal{U}} a_u u(t),$$

and

$$\|(a_u)\|_{\mathbf{R}^{Card(\mathcal{U})^*}} \le A\|T_{\mathcal{U}}^{-1}\|.$$

By the help of the above lemmas, we can deduce the following probabilistic cubature formula. The established cubature formula is different from that in [4] since the result in [4] only holds with probability that is polynomial with respect to m.

Proposition 4.8. *Let $N \in \mathbf{N}$. If ρ_X is the uniform distribution on $\mathbf{S}^d$ and $\Lambda_N := \{t_i\}_{i=1}^N$ are i.i.d. random variables drawn according to ρ_X, then there exits a set of real numbers $\{a_i\}_{i=1}^N$ such that*

$$\int_{\mathbf{S}^d} P(x)d\omega(x) = \sum_{i=1}^N a_i P(t_i), \qquad \forall\ P \in \Pi_n^d$$

holds with confidence at least

$$1 - 2\exp\left\{-\tilde{C}_4\frac{N}{n^d} + \tilde{C}_4 n^d\right\},$$

subject to

$$\sum_{i=1}^N |a_i|^2 \le \tilde{C}_5 N^{-1},$$

where $\tilde{C}_4$ and $\tilde{C}_5$ are constants depending only on d.

Proof. In the proof only, we denote by $c_1, c_2, \ldots$ constants depending only on d. Without loss of generality, we assume $P \in \mathcal{P}^0 := \{f \in \Pi_n^d : \|f\|_\rho \le 1\}$. We denote the δ-net of all $f \in \mathcal{P}^0$, by $\mathcal{A}(\delta)$. It follows from [17, Chap.9] and the definition of the covering number that the smallest cardinality of $\mathcal{A}(\delta)$ is bounded by

$$\exp\{c_1 n^d \log 1/\delta\}.$$

Given $P \in \mathcal{P}^0$. Let P_j be the polynomial in $\mathcal{A}(2^{-j})$ which is closest to P in the uniform norm, with some convention for breaking ties. Since $\|Q_n - P_j\|_\infty \to 0$, with the denotation $\eta_i(P) = |P(t_i)|^2 - \|P\|_\rho^2$, we can write

$$\eta_i(P) = \eta_i(P_0) + \sum_{l=0}^\infty \eta_i(P_{l+1}) - \eta_i(P_l).$$

Furthermore,

$$|\eta_i(P)| \le \|P\|_\infty^2 + \|P\|_\rho^2.$$

It follows from Lemma 4.5 that

$$\|P\|_\infty \le c_2 n^{\frac{d}{2}} \|P\|_2.$$

But ρ_X is the uniform distribution, we then get

$$|\eta_i(P) - \mathbf{E}\eta_i(P)| \le c_3 n^d.$$

Moreover, using Lemma 4.5 again, there holds,

$$\sigma^2(\eta_i(P)) \le \mathbf{E}((\eta_i(P))^2) \le 2\|P\|_\infty^2\|P\|_\rho^2 + 2\|P\|_2^4 \le c_4 n^d.$$

Then, using Lemma 4.6 with $\varepsilon = 1/2$, $\sigma^2 = c_4 n^d$ and $M_\xi = c_3 n^d$, we have for fixed $P \in \mathcal{A}(1)$, with probability at most $2\exp\{-c_5 N/n^d\}$, there holds

$$\left|\frac{1}{N}\sum_{i=1}^N \eta_i(P)\right| \ge \frac{1}{4}.$$

Noting there are at $\exp\{c_1 n^d\}$ polynomials in $\mathcal{A}(1)$, we get

$$\mathbf{P}\left\{\left|\frac{1}{N}\sum_{i=1}^N \eta_i(P)\right| \ge \frac{1}{4} \text{ for some } P \in \mathcal{A}(1)\right\} \le 2\exp\left\{-\frac{c_6 N}{n^d} + c_6 n^d\right\}. \tag{4.3}$$

Now, we aim to bound the probability of the event:

(**e**1) for some $l \ge 1$, some $P \in \mathcal{A}(2^{-l})$ and some $Q \in \mathcal{A}(2^{-l+1})$ with $\|P-Q\|_\infty \le 3\times 2^{-l}$, there holds

$$|\eta_i(P) - \eta_i(Q)| \ge \frac{1}{4(l+1)^2}.$$

The main tool is also the Bernstein inequality in Lemma 4.6. To this end, we should bound $|\eta_i(P) - \eta_i(Q) - \mathbf{E}(\eta_i(P) - \eta_i(Q))|$ and the variance $\sigma^2(\eta_i(P) - \eta_i(Q))$. According to Lemma 4.5 and the uniformness of ρ_X, we have

$$|\eta_i(P) - \eta_i(Q)| \le c_7 n^d \|P - Q\|_\infty$$

almost surely and

$$\begin{aligned}\sigma^2(\eta_i(P) - \eta_i(Q)) &\le \mathbf{E}((\eta_i(P) - \eta_i(Q))^2) \\ &= \int_{\mathbf{S}^d} (|P(x)|^2 - |Q(x)|^2)^2 d\rho_X - (\|P\|_\rho^2 - \|Q\|_\rho^2)^2 \\ &\le c_8 n^d \|P - Q\|_\infty^2.\end{aligned}$$

If $P \in \mathcal{A}(2^{-l})$ and $Q \in \mathcal{A}(2^{-l+1})$ with $\|P - Q\|_\infty \le 3 \times 2^{-l}$, then it follows from Lemma 4.6 again that,

$$\begin{aligned}\mathbf{P}\left(\left|\sum_{i=1}^N \eta_i(P) - \eta_i(Q)\right| > \frac{1}{4(l+1)^2}\right) &\le 2\exp\left\{-\frac{N}{c_9 n^d(2^{-2l}l^4 + 2^{-l}l^2)}\right\} \\ &\le 2\exp\left\{-\frac{N}{c_{10} n^d 2^{-l/2}}\right\}\end{aligned}$$

Since there are at most $2\exp\{-c_{11}n^d\log l\}$ polynomials in $\mathcal{A}(2^{-l})\cup\mathcal{A}(2^{-l+1})$, then the event (**e**1) holds with probability at most

$$\sum_{l=1}^{\infty}2\exp\left\{-\frac{c_{12}N}{n^d2^{-l/2}}+c_{10}n^d\log l\right\}\le\sum_{l=1}^{\infty}2\exp\left\{-2^{l/2}\left(\frac{c_{13}N}{n^d}-c_{13}n^d\right)\right\}.$$

Noting further $\sum_{i=1}^{\infty}e^{-a^lb}\le c_{12}e^{-b}$ for any $a>1$ and $b\ge1$, we then deduce that

$$\mathbf{P}\{\text{The event (}\mathbf{e}1\text{) holds}\}\le 2\exp\left\{\frac{c_{13}N}{n^d}-c_{13}n^d\right\}. \tag{4.4}$$

Thus, it follows from (4.3) and (4.4) that with confidence at least

$$1-2\exp\left\{\frac{c_{13}N}{n^d}-c_{13}n^d\right\}$$

there holds

$$\begin{aligned}\left|\sum_{i=1}^{n}\eta_i(P)\right| &\le \left|\sum_{i=1}^{n}\eta_i(P_0)\right|+\sum_{l=1}^{\infty}\left|\sum_{i=1}^{n}\eta_i(P_l)-\eta_i(P_l)\right|\\ &\le \frac14+\sum_{l=1}^{\infty}\frac{1}{4(l+1)^2}=\sum_{l=1}^{\infty}\frac{1}{4l^2}=\frac{\pi^2}{24}<\frac12.\end{aligned}$$

This means that with confidence at least

$$1-2\exp\left\{\frac{c_{13}N}{n^d}-c_{13}n^d\right\}$$

there holds

$$\frac12\|P\|_\rho^2\le\frac1N\sum_{i=1}^{N}|P(t_i)|^2\le\frac32\|P\|_\rho^2,\qquad\forall\ P\in\Pi_n^d.$$

Noting again that ρ_X is the uniform distribution, we then get

$$c_{14}\|P\|_2^2\le\frac1N\sum_{i=1}^{N}|P(t_i)|^2\le c_{15}\|P\|_2^2,\qquad\forall\ P\in\Pi_n^d. \tag{4.5}$$

Now, we use (4.5) and Lemma 4.7 to prove Lemma 4.8. In Lemma 4.7, we take $\mathcal{V}=\Pi_n^d$, $\|P\|_{\mathcal{V}}=\|P\|_2$, and $\mathcal{W}$ to be the set of point evaluation functionals $\{\delta_{t_i}\}_{i=1}^N$. The operator $T_{\mathcal{W}}$ is then the restriction map $P\mapsto P|_{\Lambda_N}$, with

$$\|f\|_{\Lambda_N,2}^2:=\frac1N\sum_{i=1}^{N}|f(t_i)|^2.$$

It follows from (4.5) that with confidence at least

$$1 - 2\exp\left\{\frac{c_{13}N}{n^d} - c_{13}n^d\right\}$$

there holds $\|T_{\mathcal{W}}^{-1}\| \leq c_{16}$. We now take u to be the functional

$$u : P \mapsto \int_{\mathbf{S}^d} P(x)d\omega(x).$$

By Hölder inequality, $\|y\|_{\mathcal{V}^*} \leq |\mathbb{S}^d|$. Therefore, Lemma 4.7 shows that

$$\int_{\mathbf{S}^d} P(x)d\omega(x) = \sum_{i=1}^{N} a_i P(t_i)$$

holds with confidence at least

$$1 - 2\exp\left\{\frac{c_{13}N}{n^d} - c_{13}n^d\right\}$$

subject to

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|a_i|}{1/N}\right)^2 \leq c_{17}.$$

This finishes the proof of Proposition 4.8. ■

5. Proofs

In this section, we present the proofs main results.

5.1. Proof of Theorem 2.1

We present the main ideas on the methodology before proceeding the detailed proof. The methodology we adopted in the proof of Theorem 2.1 is somewhat standard in the realm of non-parametric regression [17, Chapter 11] or learning theory [11, Section 1.4] that decomposes the generalization error into approximation and sample errors. The approximation error is independent of the samples while the sample error is reflected by the capacity of the hypothesis space [8, 22, 44]. In particular, we set $\mathcal{D}_n(\lambda) := \|f_n - f_\rho\|_\rho^2 + \lambda\|f_n\|_{K_n}^2$ and

$$\mathcal{S}(\lambda, m, n) := \mathcal{E}(\pi_M f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(\pi_M f_{\mathbf{z},\lambda}) + \mathcal{E}_{\mathbf{z}}(f_n) - \mathcal{E}(f_n)$$

as the approximation and sample errors respectively, where $f_n := K_n * f_\rho$. Then it is easy to deduce

$$\mathcal{E}(\pi_M f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \leq \mathcal{S}(\lambda, m, n) + \mathcal{D}_n(\lambda), \tag{5.1}$$

To bound the approximation error, it follows from Lemma 4.2, Proposition 4.4 and $|y_i| \leq M$ directly the following proposition.

Proposition 5.1. *Let $f_\rho \in W_r$ with $r > d/2$. There exists a positive constant C depending only on r and d such that*

$$\mathcal{D}_n(\lambda) \leq \bar{C}_1 n^{-2r} + |\mathbf{S}^d| M^2 \lambda.$$

To bound the sample error, we set $\xi_1 := (\pi_M(f_{\mathbf{z},\lambda})(x) - y)^2 - (f_\rho(x) - y)^2$, $\xi_2 := (f_n(x) - y)^2 - (f_\rho(x) - y)^2$, and then rewrite $\mathcal{S}(\lambda, m, n)$ as

$$\mathcal{S}(\lambda, m, n) = \left\{ \mathbf{E}(\xi_1) - \frac{1}{m} \sum_{i=1}^{m} \xi_1(z_i) \right\} + \left\{ \frac{1}{m} \sum_{i=1}^{m} \xi_2(z_i) - \mathbf{E}(\xi_2) \right\} =: \mathcal{S}_1 + \mathcal{S}_2, \tag{5.2}$$

where we used the relation

$$\mathbf{E}(\xi_1) = \int_Z \xi_1(x, y) d\rho = \mathcal{E}(\pi_M(f_{\mathbf{z},\lambda})(x)) - \mathcal{E}(f_\rho), \qquad \text{and} \quad \mathbf{E}(\xi_2) = \mathcal{E}(f_n) - \mathcal{E}(f_\rho).$$

Bounding $\mathcal{S}_2$ is standard, which can be easily deduced by using the same approach as that in [49, 42] based on the Bernstein inequality in Lemma 4.6. The following estimate describes the bound of $\mathcal{S}_2$. We omit the proof for the sake of brevity.

Proposition 5.2. *For every $0 < \delta < 1$, with confidence at least*

$$1 - \exp\left(-\frac{3m\varepsilon^2}{48M^2 \left(2\|f_n - f_\rho\|_\rho^2 + \varepsilon\right)} \right)$$

there holds

$$\mathcal{S}_2 \leq \varepsilon.$$

Since ξ_1 involves the sample $\mathbf{z}$ through $f_{\mathbf{z},\lambda}$, we should consider the capacity of the hypothesis space, which is measured by the covering number. The novelty of our proof, compared with [49] is a refined estimate for the covering number of the set $\pi_M V_k := \{\pi_M f : f \in V_k\}$, where V_k is a k-dimensional function space. This estimate can be directly deduced from two interesting papers on approximation theory, i.e., [24, Property 1] and [25, P.437]. The follow lemma is the main tool in our analysis and implies why the regularization parameter is not necessary in our analysis.

Lemma 5.3. *Let $k \in \mathbb{N}$. Then*

$$\log \mathcal{N}(\pi_M V_k, \eta) \leq \bar{C}_2 k \log \frac{M}{\eta},$$

where $\bar{C}_2$ is a positive constant depending only on d and M and $\mathcal{N}(\pi_M V_k, \eta)$ is the covering number associated with the uniform norm that denotes the number of elements in least η-net of $\pi_M V_k$.

Based on Lemma 5.3 and a standard ratio inequality in [49], we can derive the following estimate for $\mathcal{S}_1$ by using the same approach as that in [49].

Proposition 5.4. *For all $\varepsilon > 0$,*

$$\mathcal{S}_1 \leq \frac{1}{2}\left\{\mathcal{E}(\pi_M f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)\right\} + \varepsilon$$

holds with confidence at least

$$1 - \exp\left\{\bar{C}_3 n^d \log \frac{4M^2}{\varepsilon} - \frac{3m\varepsilon}{128M^2}\right\},$$

where $\bar{C}_3$ is a constant depending only on d and M.

With these helps, we are in a position to prove Theorem 2.1.

Proof of Theorem 2.1. Firstly, it follows from Propositions 5.1, 5.2, 5.4, (5.1) and (5.2) that

$$\begin{aligned}\mathcal{E}(\pi_M f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)) \;&\leq\; \mathcal{D}_n(\lambda) + \mathcal{S}_1 + \mathcal{S}_2 \leq \bar{C}_1\left(n^{-2r} + \lambda|\mathbf{S}^d|M^2\right)\\ &+\; \frac{1}{2}(\mathcal{E}(\pi_M f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)) + 2\varepsilon\end{aligned}$$

holds with confidence at least

$$1 - \exp\left\{\bar{C}_3 n^d \log \frac{4M^2}{\varepsilon} - \frac{3m\varepsilon}{128M^2}\right\} - \exp\left(-\frac{3m\varepsilon^2}{48M^2\left(2\|f_n - f_\rho\|_\rho^2 + \varepsilon\right)}\right).$$

Then, by setting $\varepsilon \geq \varepsilon_+ := \bar{C}_1'(m/\log m)^{-2r/(2r+d)}$ for some $\bar{C}_1'$ depending only on M, d and $\bar{C}_3$, $n = \left[\varepsilon^{-1/(2r)}\right]$ and $\lambda \leq |\mathbf{S}^d|^{-1}M^{-2}\varepsilon$, we get, with confidence at least

$$1 - \exp\{-\bar{C}_4 m\varepsilon\},$$

there holds

$$\mathcal{E}(\pi_M f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \leq 4\varepsilon.$$

This provides the upper bound of (2.8).

The lower bound can be more easily deduced. Actually, it can be found in [12, Chap. 3] that for any estimator $f_{\mathbf{z}}$ based on samples $\mathbf{z}$, there holds

$$\sup_{f_\rho \in W_r} \mathbf{P}_m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_\rho^2 \geq \varepsilon\} \geq \begin{cases} \varepsilon_0, & \varepsilon < \varepsilon_-, \\ e^{-cm\varepsilon}, & \varepsilon \geq \varepsilon_-, \end{cases}$$

where $\varepsilon_0 = \frac{1}{2}$ and $\varepsilon_- = cm^{-2r/(2r+d)}$ for some universal constant c. With this, the proof of Theorem 2.1 is completed. ■

5.2. Proof of Theorem 2.3

Before proceeding the proof of Theorem 2.3, we present a simple description of the methodology. The methodology we adopted in the proof seems novel. Traditionally, the generalization error of learning schemes in the sample dependent hypothesis space (SDHS) is divided into the approximation, hypothesis and sample errors (three terms) [48]. All of the aforementioned results about coefficient regularization in SDHS fall into this style. According to [48], the hypothesis error has been regarded as the reflection of nature of data dependence of SDHS, and an indispensable part attributed to an essential characteristic of learning algorithms in SDHS, compared with the learning schemes in SIHS (sample independent hypothesis space). With the needlet kernel K_n, we will divide the generalization error of l^q kernel regularization into the approximation and sample errors (two terms) only. The core tool is needlet kernel's excellent localization properties in both the spacial and frequency domain, with which the reproducing property, compressible property and the best approximation property can be guaranteed. By Proposition 4.8, we can prove that all spherical polynomials can be represented by elements in SDHS. This helps us to deduce the approximation error. Since $\mathcal{H}_{\mathbf{z},K} \subseteq \mathcal{H}_K$, the bound of the sample error is as same as that in the previous subsection.

To be detailed, in order to estimate the upper bound of

$$\mathcal{E}(\pi_M f_{\mathbf{z},\lambda,q}) - \mathcal{E}(f_\rho),$$

we introduce a novel error decomposition strategy. It follows from the definition of $f_{\mathbf{z},\lambda,q}$

that, for arbitrary $f \in \mathcal{H}_{K,\mathbf{z}}$,

$$\begin{aligned}
\mathcal{E}(\pi_M f_{\mathbf{z},\lambda,q}) - \mathcal{E}(f_\rho) &\leq \mathcal{E}(\pi_M f_{\mathbf{z},\lambda,q}) - \mathcal{E}(f_\rho) + \lambda\Omega^q_{\mathbf{z}}(f_{\mathbf{z},\lambda,q}) \\
&\leq \mathcal{E}(\pi_M f_{\mathbf{z},\lambda,q}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda,q}) + \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) \\
&+ \mathcal{E}_{\mathbf{z}}(\pi_M f_{\mathbf{z},\lambda,q}) + \lambda\Omega^q_{\mathbf{z}}(\pi_M f_{\mathbf{z},\lambda,q}) - \mathcal{E}_{\mathbf{z}}(f) - \lambda\Omega^q_{\mathbf{z}}(f) \\
&+ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\Omega^q_{\mathbf{z}}(f) \\
&\leq \mathcal{E}(\pi_M f_{\mathbf{z},\lambda,q}) - \mathcal{E}_{\mathbf{z}}(\pi_M f_{\mathbf{z},\lambda,q}) + \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) \\
&+ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\Omega^q_{\mathbf{z}}(f).
\end{aligned}$$

Since $f_\rho \in W_r$ with $r > \frac{d}{2}$ and ρ_X is the uniform distribution, it follows from Jackson inequality [5] that there exists a $P_\rho \in \Pi^d_n$ such that

$$\|P_\rho\|_\rho \leq \bar{C}_5 \|f_\rho\|_\rho \quad \text{and} \quad \|f_\rho - P_\rho\|^2_\rho \leq \bar{C}_5 n^{-2r} \tag{5.3}$$

with $\bar{C}_5$ a constant depending only on d and r. Then,

$$\begin{aligned}
\mathcal{E}(f_{\mathbf{z},\lambda,q}) - \mathcal{E}(f_\rho) &\leq \{\mathcal{E}(P_\rho) - \mathcal{E}(f_\rho) + \lambda\Omega^q_{\mathbf{z}}(P_\rho)\} \\
&+ \{\mathcal{E}(f_{\mathbf{z},\lambda,q}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda,q}) + \mathcal{E}_{\mathbf{z}}(P_\rho) - \mathcal{E}(P_\rho)\} \\
&=: \mathcal{D}(\mathbf{z},\lambda,q) + \mathcal{S}(\mathbf{z},\lambda,q),
\end{aligned}$$

where $\mathcal{D}(\mathbf{z},\lambda,q)$ and $\mathcal{S}(\mathbf{z},\lambda,q)$ are called as the approximation error and sample error, respectively. The following Proposition 5.5, which is the main novelty of our proof, presents an upper bound for the approximation error.

Proposition 5.5. *Let $m, n \in \mathbf{N}$, $r > d/2$ and $f_\rho \in W_r$. Then, with confidence at least $1 - 2\exp\{-\bar{C}_6 m/n^d\}$, there holds*

$$\mathcal{D}(\mathbf{z},\lambda,q) \leq \bar{C}_7 \left(n^{-2r} + \lambda m^{1-q}\right),$$

where $\bar{C}_6$ and $\bar{C}_7$ are constants depending only on d, M and r.

Proof. From Lemma 4.2, it is easy to deduce that

$$P_\rho(x) = \int_{\mathbf{S}^d} P_\rho(x') K_n(x,x') d\omega(x').$$

Thus, Proposition 4.8, Hölder inequality and $r > d/2$ yield that with confidence at least $1 - 2\exp\{-\bar{C}_6 m/n^d\}$, there exists a set of real numbers $\{a_i\}_{i=1}^m$ satisfying $\sum_{i=1}^m |a_i|^q \leq \tilde{C}_5 m^{1-q}$ for $q > 0$ such that

$$P_\rho(x) = \sum_{i=1}^m a_i P_\rho(x_i) K_n(x_i, x).$$

The above observation together with (5.3) implies that with confidence at least $1 - 2\exp\{-\bar{C}_6 m/n^d\}$, P_ρ can be represented as

$$P_\rho(x) = \sum_{i=1}^m a_i P_\rho(x_i) K_n(x_i, x) \in \mathcal{H}_{K,\mathbf{z}}$$

such that for arbitrary $f_\rho \in W_r$, there holds

$$\|P_\rho - f_\rho\|_\rho^2 \leq \bar{C}_8 n^{-2r},$$

and

$$\Omega_{\mathbf{z}}^q(P_\rho) \leq \sum_{i=1}^m |a_i P_\rho(x_i)|^q \leq (\bar{C}_9 M)^q \sum_{i=1}^m |a_i|^q \leq \bar{C}_{10} m^{1-q},$$

where $\bar{C}_8, \bar{C}_9, \bar{C}_{10}$ are constants depending only on d, q and M. It thus implies that the inequalities

$$\mathcal{D}(\mathbf{z}, \lambda, q) \leq \|P_\rho - f_\rho\|_\rho^2 + \lambda \Omega_{\mathbf{z}}^q(P_\rho) \leq \bar{C}_7 \left(n^{-2r} + \lambda m^{1-q}\right) \tag{5.4}$$

holds with confidence at least $1 - 2\exp\{-\bar{C}_6 m/n^d\}$. ■

By the help of Proposition 5.5, we can prove Theorem 2.3 as follows.

Proof of Theorem 2.3. Firstly, it follows from Propositions 5.5, 5.4 and 5.2 that

$$\begin{aligned} \mathcal{E}(\pi_M f_{\mathbf{z},\lambda,q}) - \mathcal{E}(f_\rho)) &\leq \mathcal{D}(\mathbf{z}, \lambda, q) + \mathcal{S}_1^q + \mathcal{S}_2^q \leq \bar{C}_7 \left(n^{-2r} + \lambda m^{1-q}\right) \\ &+ \frac{1}{2}(\mathcal{E}(f_{\mathbf{z},\lambda,q}) - \mathcal{E}(f_\rho)) + 2\varepsilon \end{aligned}$$

holds with confidence at least

$$1 - 2\exp\{-\bar{C}_6 m/n^d\} - \exp\left\{\bar{C}_3 n^d \log\frac{4M^2}{\varepsilon} - \frac{3m\varepsilon}{128M^2}\right\} - \exp\left(-\frac{3m\varepsilon^2}{48M^2\left(2n^{-2r} + \varepsilon\right)}\right).$$

Then, by setting $\varepsilon \geq \varepsilon_m^+ \geq \bar{C}_{11}(m/\log m)^{-2r/(2r+d)}$ for some $\bar{C}_{11}$ depending only on d and M, $n = \left[\varepsilon^{-1/(2r)}\right]$ and $\lambda \leq m^{q-1}\varepsilon$, it follows from $r > d/2$ that

$$\begin{aligned} & 1 - 2\exp\{-\bar{C}_{12} m\varepsilon^{d/(2r)}\} - \exp\{-\bar{C}_{12} m\varepsilon\} \\ - \; & \exp\left\{\bar{C}_{12}\varepsilon^{-d/(2r)}\left(\log 1/\varepsilon + \log m\right) - \bar{C}_{12} m\varepsilon)\right\} \\ \geq \; & 1 - 4\exp\{-\bar{C}_{13} m\varepsilon\}. \end{aligned}$$

That is, for $\varepsilon \geq \varepsilon_m^+$,

$$\mathcal{E}(f_{\mathbf{z},\lambda,q}) - \mathcal{E}(f_\rho) \leq 4\varepsilon$$

holds with confidence at least $1 - 4\exp\{-\bar{C}_{13} m\varepsilon\}$. The same method as [12, P.37] yields the lower bound of (2.13). This finishes the proof of Theorem 2.3. ■

References

[1] Abrial, P., Moudden, Y., Starck, J., Delabrouille, J. and Nguyen, M. (2008). CMB data analysis and sparsity. *Statis. Method.* **5** 289–298.

[2] Baldi, P., Kerkyacharian, G., Marinucci, D. and Picard, D. (2008). Asymptotics for spherical needlets. *Ann. Statist.* **37** 1150–1171.

[3] Bickel, P. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series* **54** 177–186.

[4] Böttcher, A., Kunis, S. and Potts, D (2009). Probabilistic spherical Marcinkiewicz-Zygmund inequalities. *J. Approx. Theory* **157** 113–126.

[5] Brown, G. and Dai, F. (2005). Approximation of smooth functions on compact two-point homogeneous spaces. *J. Funct. Anal.* **220** 401–423.

[6] Cao, F., Lin, S., Chang, X. and Xu, Z. (2013). Learning rates of regularized regression on the unit sphere. *Sci. China Math.* **56** 861–876.

[7] Chang, T., Ko, D., Royer, J. and Lu, J. (2000). Regression techniques in plate tectonics. *Statis. Sci.* **15** 342–356.

[8] Caponnetto, A. and DeVito, E. (2007). Optimal rates for the regularized least squares algorithm. *Found. Comput. Math.* **7** 331–368.

[9] Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. 39* 1–49.

[10] Cucker, F. and Smale, S. (2002). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Found. Comput. Math.* **2** 413–428.

[11] Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge.

[12] Devore, R. A., Kerkyacharian, G., Picard, D. and Temlyakov, V. (2006). Approximation methods for supervised learning. *Found. Comput. Math.* **6** 3–58.

[13] Dodelson, S. (2003). *Modern Cosmology.* Academic Press, London.

[14] Downs, T. (2003). Spherical regression. *Biometrika* **90**, 655-668.

[15] Freeden, W., Gervens, T. and Schreiner, M. (1998). *Constructive Approximation on the Sphere.* Oxford University Press Inc., New York.

[16] Filbir, F. and Themistoclakis, W. (2004). On the construction of de la vallèe poussin means for orthogonal polynomials using convolution structures. *J. Comput. Anal. Appl.* **6**, 297-312.

[17] Györfy, L., Kohler, M., Krzyzak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, Berlin.

[18] Kerkyacharian, G., Nickl, R. and Picard, D. (2011). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probability Theory and Related Fields* **153** 363–404.

[19] Le Gia, Q., and Mhaskar, H. (2008). Localized linear polynomial operators and quadrature formulas on the sphere. *SIAM J. Numer. Anal.* pages **47** 440–466.

[20] Li, L.(2009). Regularized least square regression with spaherical polynomial kernels. Inter. J. Wavelets, Multiresolution and Inform. Proces. *7* 781–801.

[21] Lin, S., Zeng, J., Fang, J. and Xu, Z. (2014). Learning rates of l^q coefficient regularization learning with Gaussian kernel. *Neural Comput.* **26** 2350–2378.

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

[23] Mackenzie, J. K. (1957). The estimation of an orientation relationship. *Acta Crystallog.* **10** 61–62.

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

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

[26] Marzio, M., Panzera, A. and Taylor, C. (2014). Nonparametric regression for spherical data. *J. Amer. Statis. Assoc.* **109** 748–763.

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

[28] Mhaskar, H. N., Narcowich, F. J. and Ward. J. D. (2000). Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature. *Math. Comput.* **70** 1113–1130.

[29] Mhaskar, H. (2004). Polynomial operators and local smoothness classes on the unit interval. *J. Approx. Theory 131* 243–267.

[30] Mhaskar, H. (2005). On the representation of smooth functions on the sphere using initely many bits. *Appl. Comput. Harmon. Anal. 18* 215–233.

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

[32] Minh, H. (2006). *Reproducing kernel Hilbert spaces in learning theory* Ph. D. Thesis in Mathematics, Brown University.

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

[34] Monnier, J. (2011). Nonparametric regression on the hyper-sphere with uniform design. *Test* **20** 412–446.

[35] Müller, C. (1966). *Spherical Harmonics, Lecture Notes in Mathematics.* Vol. 17, Springer, Berlin.

[36] Narcowich, F., Petrushev, V. and Ward, J. (2006). Localized tight frames on spheres. *SIAM J. Math. Anal. 38* 574–594.

[37] Narcowich, F., Petrushev, V. and Ward, J. (2006). Decomposition of Besov and Triebel-Lizorkin spaces on the sphere. *J. Funct. Anal. 238* 530–564.

[38] Pelletier, B. (2006). Non-parametric regression estimation on closed Riemannian manifolds. *J. Nonpar. Statis.* **18** 57-67.

[39] Schölkopf, B and Smola, A. J. (2001). *Learning with Kernel: Support Vector Machine, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning).* The MIT Press, Cambridge.

[40] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* Springer, New York.

[41] Steinwart, I., Hush, D. and Scovel, C. (2009). Optimal rates for regularized least squares regression. *In Proceedings of the 22nd Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.), pp. 79-93, 2009.

[42] Shi, L., Feng, Y. and Zhou, D. X. (2011). Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmon. Anal.* **31** 286–302.

[43] Sloan, I. H. and R. S. Womersley,R. S. (2012). Filtered hyperinterpolation: a constructive polynomial approximation on the sphere. *Int. J. Geomath.* **3** 95–117.

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

[45] Tong, H., Chen, D. and Yang, F. (2010). Least square regression with l^p-coefficient regularization. *Neural Comput.* **22** 3221–3235.

[46] Tibshirani, R. (1995). Regression shrinkage and selection via the LASSO. *J. ROY. Statist. Soc. Ser. B* **58** 267–288.

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

[48] Wu, Q and Zhou, D. X. (2008). Learning with sample dependent hypothesis space. *Comput. Math. Appl.* **56** 2896–2907.

[49] Zhou, D. X. and Jetter, K. (2006). Approximation with polynomial kernels and SVM classifiers. *Adv. Comput. Math.* **25** 323–344.