NONPARAMETRIC REGRESSION USING NEEDLET KERNELS FOR SPHERICAL DATA

SHAOBO LIN

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 also prove that all the \(l^q\) (\(0 < q < \infty\)) kernel regularization estimates associated with the needlet kernel, including the kernel lasso estimate and the kernel bridge estimate, possess almost the 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 in terms of the generalization capability in some modeling contexts. From this perspective, \(q\) can be arbitrarily specified, or specified merely by other no generalization criteria like smoothness, computational complexity, sparsity, etc..

Key words. Nonparametric regression, needlet kernel, spherical data, kernel ridge regression.

AMS subject classifications. 68T05, 62J02.

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 [12], cosmic microwave background radiation [10], tectonic plate geology [6] and image rendering [31]. As the sphere is topologically a compact two-point homogeneous manifold, some widely used schemes for the Euclidean space such as the neural networks [13] and support vector machines [27] 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 [11, 18, 23, 26].

Recent years have witnessed a considerable approaches about 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 [11], 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 [26], local polynomial estimate [3] and local linear estimate [18] are also alternate and interesting nonparametric approaches. Unfortunately, the manifold structure of the sphere is not well taken into account in these approaches. Mihn [21] 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 [22] and polynomials [5], 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 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 to find 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 actually a function that possesses excellent localization properties, both in spacial and frequency domains. Under this circumstance, the needlet kernel comes into our sights. Needlets, introduced by Narcowich et al. [24, 25], are a new kind 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, 14], 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 [24] that the spherical needlets correspond a needlet kernel, which is also well localized in the spacial and frequency domains. Consequently, the needlet kernel is proved to possess the reproducing property [24, Lemma 3.8], compressible property [24, Theorem 3.7] and best approximation property [24, 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 a 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 [8], such as the Gaussian [22] and Abel-Poisson [12] 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 [27]. This guides us to consider the $l_q$ ($0 < q < \infty$) kernel regularization (KRS) schemes associated with the needlet kernel, including the kernel bridge regression and kernel lasso estimate [32]. The first feature of the needlet kernel implies that the regularization capability of all $l_q$-KRS with $0 < q < \infty$ are almost the same, provided the regularization parameter is set to be small enough. However, such a setting makes there be no difference among all $l_q$-KRS with $0 < q < \infty$, as each of them behaves similar as the least squares. To distinguish different behaviors of the $l_q$-KRS, we should establish a similar result for a large regularization parameter. By the aid of a probabilistic cubature formula and the excellent localization property in both frequency and spacial domain of the needlet kernel, we find that all $l^q$-KRS with $0 < q < \infty$ can attain almost the same optimal generalization error bounds, provided the regularization parameter is not larger than $O(m^{q-1} \varepsilon)$. Here $m$ is the number of samples and $\varepsilon$ 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 \( q \). From this perspective, \( q \) can be specified by other no generalization criteria like smoothness, computational complexity and sparsity. We consider it as the other feature of the needlet kernel.

The reminder of the paper is organized as follows. In the next section, the needlet kernel together with its important properties such as the reproducing property, compressible property and best approximation property is introduced. In Section 3, we study the generalization capability of the kernel ridge regression associated with the needlet kernel. In Section 4, we consider the generalization capability of the \( l^q \) kernel regularization schemes, including the kernel bridge regression and kernel lasso. In Section 5, we provide the proofs of the main results. We conclude the paper with some useful remarks in the last section.

2. The needlet kernel. Let \( S^d \) be the unit sphere embedded into \( \mathbb{R}^{d+1} \). For integer \( k \geq 0 \), the restriction to \( S^d \) of a homogeneous harmonic polynomial of degree \( k \) on the unit sphere is called a spherical harmonic of degree \( k \). The class of all spherical harmonics of degree \( k \) is denoted by \( H_k^d \), and the class of all spherical harmonics of degree \( k \leq n \) is denoted by \( \Pi_k^d \). Of course, \( \Pi_k^d = \bigoplus_{k=0}^n H_k^d \), and it comprises the restriction to \( S^d \) of all algebraic polynomials in \( d + 1 \) variables of total degree not exceeding \( n \). The dimension of \( H_k^d \) is given by

\[
D_k^d := \dim H_k^d = \begin{cases} 2k \frac{d-1}{k+d-1}, & k \geq 1; \\ 1, & k = 0, \end{cases}
\]

and that of \( \Pi_k^d \) is \( \sum_{k=0}^n D_k^d = D_{n+1}^d \sim n^d \).

The addition formula establishes a connection between spherical harmonics of degree \( k \) and the Legendre polynomial \( P_k^{d+1} \) [12]:

\[
\sum_{l=1}^{D_k^d} Y_{k,l}(x)Y_{k,l}(x') = \frac{D_k^d}{|S^d|} P_k^{d+1}(x \cdot x'),
\]

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

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

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

The following Funk-Hecke formula establishes a connection between spherical harmonics and function \( \phi \in L^1([-1,1]) \) [12]

\[
\int_{S^d} \phi(x \cdot x') H_k(x') d\omega(y) = B(\phi, k) H_k(x),
\]

where

\[
B(\phi, k) = |S^{d-1}| \int_{-1}^{1} P_k^{d+1}(t) \phi(t)(1-t^2)^{\frac{d-1}{2}} dt.
\]

A function \( \eta \) is said to be admissible [25] if \( \eta \in C^\infty[0,\infty) \) satisfies the following condition:

\[
\text{supp} \eta \subset [0, 2], \eta(t) = 1 \text{ on } [0, 1], \text{ and } 0 \leq \eta(t) \leq 1 \text{ on } [1, 2].
\]
The needlet kernel [24] is then defined to be

\[ K_n(x \cdot x') = \sum_{k=0}^{\infty} \eta \left( \frac{k}{n} \right) \frac{D_{d+k}^d}{|S^d|^k} P_{d+k}^{d+1}(x \cdot x'), \]

The needlets can be deduced from the needlet kernel and a spherical cubature formula [4]. We refer the readers to [2, 14, 24] for a detailed description of the needlets. According to the definition of the admissible function, it is easy to see that \( K_n \) possesses excellent localization property in the frequency domain. The following Lemma 2.1 that can be found in [24] and [4] yields that \( K_n \) also possesses perfect spatial localization property.

**Lemma 2.1.** Let \( \eta \) be admissible. Then for every \( k > 0 \) and \( r \geq 0 \) there exists a constant \( C \) depending only on \( k, r, d \) and \( \eta \) such that

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

For \( f \in L^1(S^d) \), we write

\[ K_n * f(\xi) := \int_{S^d} K_n(x \cdot x') f(x') d\omega(x'). \]

We also denote by \( E_N(f)_p \) the best approximation error of \( f \in L^p(S^d) \) (\( p \geq 1 \)) from \( \Pi_d^2 \), i.e.

\[ E_N(f)_p := \inf_{P \in \Pi_d^2} \| f - P \|_{L^p(S^d)}. \]

Then the needlet kernel \( K_n \) satisfies the following Lemma 2.2, which can be deduced from [24].

**Lemma 2.2.** \( K_n \) is a reproducing kernel for \( \Pi_d^2 \), that is \( K_n * P = P \) for \( P \in \Pi_d^2 \). Moreover, for any \( f \in L^p(S^d) \), \( 1 \leq p \leq \infty \), we have \( K_n * f \in \Pi_d^2 \), and

\[
\| K_n * f \|_{L^p(S^d)} \leq C \| f \|_{L^p(S^d)}, \quad \text{and} \quad \| f - K_n * f \|_{L^p(S^d)} \leq C E_n(f)_p,
\]

where \( C \) is a constant depending only on \( d, p \) and \( \eta \).

It is obvious that \( K_n \) is a semi-positive definite kernel, thus it follows from the known Mercer theorem [21] that \( K_n \) corresponds a reproducing kernel Hilbert space (RKHS), \( \mathcal{H}_K \).

**Lemma 2.3.** Let \( K_n \), be defined above, then the reproducing kernel Hilbert space associated with \( K_n \) is the space \( \Pi_d^2 \) with the inner product:

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

where \( \hat{f}_{k,j} = \int_{S^d} f(x) Y_{k,j}(x) d\omega(x) \).

**3. Kernel ridge regression associated with the needlet kernel.** In spherical nonparametric regression problems with predictor variables \( X \in \mathcal{X} = S^d \) and a real response variable \( Y \in \mathcal{Y} \subseteq \mathbb{R} \), we observe \( m \) i.i.d. samples \( z_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \) from an unknown distribution \( \rho \). Without loss of generality, it is always assumed that
\[ Y \subseteq [-M, M] \text{ almost surely, where } M \text{ is a positive constant. One natural measurement of the estimate } f \text{ is the generalization error,} \]

\[ \mathcal{E}(f) := \int_{\mathcal{Z}} (f(X) - Y)^2 d\rho, \]

which is minimized by the regression function \[13\] defined by

\[ f_{\rho}(x) := \int_{Y} Y d\rho(Y|x). \]

Let \( L^2_{\rho_X} \) be the Hilbert space of \( \rho_X \) square integrable functions, with norm \( \| \cdot \|_{\rho} \). In the setting of \( f_{\rho} \in L^2_{\rho_X} \), it is well known that, for every \( f \in L^2_{\rho_X} \), there holds

\[ (3.1) \quad \mathcal{E}(f) - \mathcal{E}(f_{\rho}) = \| f - f_{\rho} \|_{\rho}^2. \]

We formulate the learning problem in terms of probability estimates rather than expectation estimates. To this end, we present a formal way to measure the performance of learning schemes in probability. Let \( \Theta \subset L^2_{\rho_X} \) and \( M(\Theta) \) be the class of all Borel measures \( \rho \) such that \( f_{\rho} \in \Theta \). For each \( \varepsilon > 0 \), we enter into a competition over all estimators based on \( m \) samples \( \Phi_m : z \mapsto f_{z} \) by

\[ \text{AC}_m(\Theta, \varepsilon) := \inf_{f_{z} \in \Phi_m} \sup_{\rho \in M(\Theta)} \mathbb{P}^m\{z : \| f_{\rho} - f_{z} \|_{\rho}^2 > \varepsilon \}. \]

As it is impossible to obtain a nontrivial convergence rate without imposing any restriction on the distribution \( \rho \) \[13, \text{Chap.3}\], we should introduce certain prior information. Let \( \mu \geq 0 \). Denote the Bessel-potential Sobolev class \( W_r \) to be all \( f \) such that

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

where

\[ P_l f = \sum_{j=1}^{D^d_l} \langle f, Y_{k,j} \rangle Y_{k,j}. \]

It follows from the well known Sobolev embedding theorem that \( W_r \subset C(S^d) \), provided \( r > d/2 \). In our analysis, we assume \( f_{\rho} \in W_r \).

The learning scheme employed in this section is the following kernel ridge regression (KRR) associated with the needlet kernel

\[ (3.2) \quad f_{z,\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \| f \|_{K_m}^2 \right\}. \]

Since \( y \in [-M, M] \), it is easy to see that \( \mathcal{E}(\pi_M f) \leq \mathcal{E}(f) \) for arbitrary \( f \in L^2_{\rho_X} \), where \( \pi_M u := \min\{M, |u|\} \text{sgn}(u) \) is the truncation operator. As there isn’t any additional computation for employing the truncation operator, the truncation operator has been used in large amount of papers, to just name a few, \[5, 9, 13, 15, 21, 32, 33\]. The
following Theorem 3.1 illustrates the generalization capability of KRR associated with the needlet kernel and reveals the first feature of the needlet kernel.

**Theorem 3.1.** Let \( f_\rho \in W_r \) with \( r > d/2, m \in \mathbb{N}, \varepsilon > 0 \) be any real number, and \( n \sim \varepsilon^{-r/d} \). If \( f_{\rho, \lambda} \) is defined as in (3.2) with \( 0 \leq \lambda \leq M^{-2}\varepsilon \), then there exist positive constants \( C_i, i = 1, \ldots, 4 \), depending only on \( M, \rho, \) and \( d, \varepsilon_0 > 0 \) and \( \varepsilon_-, \varepsilon_+ \) satisfying

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

such that for any \( \varepsilon < \varepsilon_- \),

\[
\sup_{f_\rho \in W_r} \mathbb{P}^m\{ z : \| f_\rho - \pi_M f_{\rho, \lambda} \|_\rho^2 > \varepsilon \} \geq AC_m(W_r, \varepsilon) \geq \varepsilon_0,
\]

and for any \( \varepsilon \geq \varepsilon_+ \),

\[
e^{-C_3 m \varepsilon} \leq AC_m(W_r, \varepsilon) \leq \sup_{f_\rho \in W_r} \mathbb{P}^m\{ z : \| f_\rho - \pi_M f_{\rho, \lambda} \|_\rho^2 > \varepsilon \} \leq e^{-C_4 m \varepsilon}.
\]

We give several remarks on Theorem 3.1 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 \( \varepsilon \) and statisticians are required to assess the probability of success. It is obvious that the probability depends heavily on \( m \) and \( \varepsilon \). If \( m \) is too small, then there isn’t any estimate that can finish the learning task with small \( \varepsilon \). This fact is quantitatively verified by the inequality (3.4). More specifically, (3.4) shows that if the learning task is to yield an accuracy at most \( \varepsilon \leq \varepsilon_- \), and other than the prior knowledge, \( f_\rho \in W_r \), there are only \( m \leq \varepsilon_-(2r+d)/(2r) \) data available, then all learning schemes, including KRR associated with the needlet kernel, may fail with high probability. To circumvent it, the only way is to acquire more samples, just as inequalities (3.5) purport to show. (3.5) says that if the number of samples achieves \( \varepsilon_+ \), then the probability of success of KRR is at least \( 1 - e^{-C_4 m \varepsilon} \). The first inequality (lower bound) of (3.5) implies that this confidence can not be improved further. The values of \( \varepsilon_- \) and \( \varepsilon_+ \) thus are very critical since the smallest number of samples to finish the learning task lies in the interval \( [\varepsilon_-, \varepsilon_+] \). Inequalities (3.3) depicts that, for KRR, there holds

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

This implies that the interval \( [\varepsilon_-, \varepsilon_+] \) is almost the shortest one in the sense that up to a logarithmic factor, the upper bound and lower bound of the interval are asymptotically identical. Furthermore, Theorem 3.1 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 \( \varepsilon_0 \) to an exponentially small quantity. All the above assertions show that the learning performance of KRR is essentially revealed in Theorem 3.1.

An interesting finding in Theorem 3.1 is that the regularization parameter of KRR can decrease arbitrarily fast, provided it is smaller than \( M^{-2}\varepsilon \). The extreme case is that the least-squares possess the same generalization performance as KRR. It is not surprised 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, \( H_K \) is essentially a linear space with finite dimension. Thus, [13,
Th.3.2 and Th.11.3] together with Lemma 5.1 in the present paper automatically yields the optimal learning rate of the least squares associated with the needlet kernel in the sense of expectation. Differently, Theorem 3.1 presents an exponential confidence estimate for KRR, which together with (3.3) makes [13, Th.11.3] be a corollary of Theorem 3.1. Theorem 3.1 also shows that the purpose of introducing regularization term in KRR is only to conquer the singular problem of the kernel matrix, \( A := (K_n(x_i \cdot x_j))_{i,j=1}^m \), since \( m > D_n^{d+1} \) in our setting. Under this circumstance, a small \( \lambda \) leads to the ill-condition of the matrix \( A + m\lambda I \) and a large \( \lambda \) conducts large approximation error. Theorem 3.1 illustrates that if the needlet kernel is employed, then we can set \( \lambda = M^{-2}\varepsilon \) to guarantee both the small condition number of the kernel matrix and almost generalization error bound. From (3.3), it is easy to deduce that to attain the optimal learning rate \( m^{-2r/(2r+d)} \), 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 (3.2).

4. \( l^q \) kernel regularization schemes associated with the needlet kernel.

In the last section, we analyze the generalization capability of KRR associated with the needlet kernel. This section aims to study the learning capability of the \( l^q \) kernel regularization schemes (KRS) whose hypothesis space is the sample dependent hypothesis space [32] associated with \( K_n(\cdot, \cdot) \),

\[
\mathcal{H}_{K_n} := \left\{ \sum_{i=1}^m a_i K_n(x_i, \cdot) : a_i \in \mathbb{R} \right\}
\]

The corresponding \( l^q \)-KRS is defined by

\[
f_{z,\lambda,q} \in \arg \min_{f \in \mathcal{H}_{K_n}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \Omega^q_z(f) \right\},
\]

where

\[
\Omega^q_z(f) := \inf_{(a_1, \ldots, a_n) \in \mathbb{R}_n} \sum_{i=1}^m |a_i|^q, \text{ for } f = \sum_{i=1}^m a_i K_n(x_i, \cdot).
\]

With different choices of the order \( q \), (4.1) leads to various specific forms of the \( l_q \) regularizer. \( f_{z,\lambda,2} \) corresponds to the kernel ridge regression [27], which smoothly shrinks the coefficients toward zero and \( f_{z,\lambda,1} \) leads to the LASSO [30], which sets small coefficients exactly at zero and thereby also serves as a variable selection operator. The varying forms and properties of \( f_{z,\lambda,q} \) 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 4.1 shows that if the needlet kernel is utilized in \( l^q \)-KRS, then \( q \) may not have an important impact in the generalization capability for a large range of regularization parameter in the sense of rate optimality.

**Theorem 4.1.** Let \( f_{\rho} \in W_r \) with \( r > d/2 \), \( m \in \mathbb{N} \), \( \varepsilon > 0 \) be any real number, and \( n \sim \varepsilon^{-r/d} \). If \( f_{z,\lambda,q} \) is defined as in (4.1) with \( \lambda \leq m^{-1}\varepsilon \) and \( 0 < q < \infty \), then there exist positive constants \( C_i, i = 1, \ldots, 4 \), depending only on \( M, \rho, q \) and \( d \), \( \varepsilon_0 > 0 \) and \( \varepsilon_m, \varepsilon_+ \) satisfying

\[
C_1 m^{-2r/(2r+d)} \leq \varepsilon_m^{-1} \leq \varepsilon_+ \leq C_2 (m/ \log m)^{-2r/(2r+d)},
\]
such that for any $\varepsilon < \varepsilon_m$,
\begin{equation}
(4.3) \quad \sup_{f, \in W_r} \mathcal{P}^m \{ z : \| f - \pi_M f_{z, \lambda, q} \|_\rho^2 > \varepsilon \} \geq \mathcal{A} \mathcal{C}_m (W_r, \varepsilon) \geq \varepsilon_0,
\end{equation}
and for any $\varepsilon \geq \varepsilon_m$,
\begin{equation}
(4.4) \quad e^{-C_3 m \varepsilon} \leq \mathcal{A} \mathcal{C}_m (W_r, \varepsilon) \leq \sup_{f, \in W_r} \mathcal{P}^m \{ z : \| f - \pi_M f_{z, \lambda, q} \|_\rho^2 > \varepsilon \} \leq e^{-C_4 m \varepsilon}.
\end{equation}

Compared with KRR (3.2), a common consensus is that $l^q$-KRS (4.1) 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 always be true. This conclusion depends heavily on the value of the regularization parameter. If the regularization parameter is extremely small, then $l^q$-KRS for any $q \in (0, \infty)$ behave similar as the least squares. Under this circumstance, Theorem 4.1 obviously holds due to the conclusion of Theorem 3.1. To distinguish the character of $l^q$-KRS with different $q$, one should consider a relatively large regularization parameter. Theorem 4.1 shows that for a large range of regularization parameters, all the $l^q$-KRS associated with the needlet kernel can attain the same, almost optimal, generalization error bound. It should be highlighted that the quantity $m^{-1} \varepsilon$ is, to the best of knowledge, almost the largest value of the regularization parameter among all the existing results. We encourage the readers to compare our result with the results in [15, 28, 29, 32]. Furthermore, we find that $m^{-1} \varepsilon$ is sufficient to embody the feature of $l^q$ kernel regularization schemes. Taking the kernel lasso for example, the regularization parameter derived in Theorem 4.1 asymptotically equals to $\varepsilon$. It is to see that, to yield a prediction accuracy $\varepsilon$, we have
\[ f_{z, \lambda, 1} \in \arg \min_{f \in \mathcal{H}_{z, \lambda}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \Omega_1^z (f) \right\}, \]
and
\[ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 \leq \varepsilon. \]
According to the structural risk minimization principle and $\lambda = \varepsilon$, we obtain
\[ \Omega_1^z (f_{z, \lambda, 1}) \leq 1. \]
This implies naturally the sparsity of the kernel lasso estimate.

Intuitively, the generalization capability of $l^q$-KRS (4.1) with a large regularization parameter may depend on the choice of $q$. While from Theorem 4.1 it follows that the learning schemes defined by (4.1) can indeed achieve the same asymptotically optimal rates for all $q \in (0, \infty)$. 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. 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 $\mathcal{H}_{z, K}$ with a few $K_n(x, x)$’s whose centers $x_i$ are near
to \( x \). As \( f_p \) is bounded by \( M \), then the coefficient of these terms are also bounded. That is, we can construct, in \( \mathcal{H}_{R^d} \), a good approximant, whose \( l^q \) norm is bounded for arbitrary \( 0 < q < \infty \). Then, using the standard error decomposition technique in [7] that divide the generalization error into the approximation error and sample error, the approximation error of \( l^q \)-KR is independent of \( q \). For the sample error, we can tune \( \lambda \) that may depend on \( q \) to offset the effect of \( q \). Then, a generalization error estimate independent of \( q \) is natural.

5. Proofs. In this section, we present the proof of Theorem 3.1 and Theorem 4.1, respectively.

5.1. Proof of Theorem 3.1. For the sake of brevity, we set \( f_n = R_n \ast f_p \). Let

\[
S(\lambda, m, n) := \{ \varepsilon(\pi_M f_n, \lambda) - \varepsilon_x(\pi_M f_n, \lambda) + \varepsilon_x(f_n) - \varepsilon(f_n) \}.
\]

Then it is easy to deduce that

\[
\varepsilon(\pi_M f_n, \lambda) - \varepsilon(f_n) \leq S(\lambda, m, n) + D_n(\lambda),
\]

where \( D_n(\lambda) := \| f_n - f_p \|^2 + \lambda \| f_n \|_{K^d}^2 \). If we set \( \xi_1 := (\pi_M f_n, \lambda)(x) - y^2 - (f_p(x) - y)^2 \), and \( \xi_2 := (f_n(x) - y)^2 - (f_p(x) - y)^2 \), then

\[
\varepsilon(\Pi_M f_n, \lambda)(x) = \varepsilon (f_n) - \varepsilon (f_p),
\]

Therefore, we can rewrite the sample error as

\[
S(\lambda, m, n) = \left\{ \varepsilon(\xi_1) - \frac{1}{m} \sum_{i=1}^{m} \xi_1(z_i) \right\} + \left\{ \frac{1}{m} \sum_{i=1}^{m} \xi_2(z_i) - \varepsilon(\xi_2) \right\} := S_1 + S_2.
\]

The aim of this subsection is to bound \( D_n(\lambda), S_1 \) and \( S_2 \), respectively. To bound \( D_n(\lambda) \), we need the following two lemmas. The first one is the Jackson-type inequality that can be deduced from [20, 24] and the second one describes the RKHS norm of \( f_n \).

**Lemma 5.1.** Let \( f \in W_r \). Then there exists a constant depending only on \( d \) and \( r \) such that

\[
\| f - f_n \| \leq Cn^{-2r},
\]

where \( \| \cdot \| \) denotes the uniform norm on the sphere.

**Lemma 5.2.** Let \( f_n \) be defined as above. Then we have

\[
\| f_n \|_{K^d}^2 \leq M^2.
\]

**Proof.** Due to the addition formula (2.1), we have

\[
K_n(x \cdot y) = \sum_{k=0}^{n} \eta \left( \frac{k}{n} \right) \left\{ \sum_{j=1}^{d} Y_{k,j}(x) Y_{k,j}(y) \right\} = \sum_{k=0}^{n} \eta \left( \frac{k}{n} \right) D_{k}^{d} F_{d+1}(x \cdot y).
\]

Since

\[
K_n \ast f(x) = \int_{S^d} K_n(x \cdot y) f(y) d\omega(y),
\]
it follows from the Funk-Hecke formula (2.2) that

\[
\hat{K}_n * f_{u,v} = \int_{S^d} K_n * f(x) Y_{u,v}(x) d\omega(x) = \int_{S^d} \int_{S^d} K_n(x \cdot x') f(x') d\omega(x') Y_{u,v}(x) d\omega(x)
\]

\[
= \int_{S^d} f(x') \int_{S^d} K_n(x \cdot x') Y_{u,v}(x) d\omega(x) d\omega(x')
\]

\[
= \int_{S^d} |S^{d-1}| \int_{-1}^1 K_n(t) P_{u,v}^{d+1}(t)(1 - t^2)^{\frac{d-2}{2}} dt Y_{u,v}(x') f(x') d\omega(x')
\]

\[
= |S^{d-1}| \hat{f}_{u,v} \int_{-1}^1 K_n(t) P_{u,v}^{d+1}(t)(1 - t^2)^{\frac{d-2}{2}} dt.
\]

Moreover,

\[
\int_{-1}^1 K_n(t) P_{u,v}^{d+1}(t)(1 - t^2)^{\frac{d-2}{2}} dt = \int_{-1}^1 \sum_{k=0}^n \eta \left( \frac{u}{n} \right) \frac{D_d^u}{|S^d|} P_{u,v}^{d+1}(t) P_{u,v}^{d+1}(1 - t^2)^{\frac{d-2}{2}} dt
\]

\[
= \int_{-1}^1 \eta \left( \frac{u}{n} \right) \frac{D_d^u}{|S^d|} P_{u,v}^{d+1}(t) P_{u,v}^{d+1}(1 - t^2)^{\frac{d-2}{2}} dt
\]

\[
= \eta \left( \frac{u}{n} \right) \frac{D_d^u}{|S^d|} \frac{|S^{d-1}|}{D_d^u} = \eta \left( \frac{u}{n} \right) \frac{1}{|S^{d-1}|}.
\]

Therefore,

\[
\hat{K}_n * f_{u,v} = \eta \left( \frac{u}{n} \right) \hat{f}_{u,v}.
\]

This implies

\[
\|K_n * f\|_{K_n}^2 = \sum_{u=0}^n \eta \left( \frac{u}{n} \right)^{-1} \sum_{v=1}^{D_d^u} (\hat{K}_n * f_{u,v})^2
\]

\[
\leq \sum_{u=0}^n \sum_{v=1}^{D_d^u} \hat{f}_{u,v}^2 \leq \|f\|_{L_2(S^d)}^2 \leq M^2.
\]

The proof of Lemma 5.2 is completed.

Based on the above two lemmas, it is easy to deduce an upper bound of \( \mathcal{D}_n(\lambda) \).

**Proposition 5.3.** Let \( f \in W_r \). There exists a positive constant \( C \) depending only on \( r \) and \( d \) such that

\[
\mathcal{D}_n(\lambda) \leq C n^{-2r} + M^2 \lambda
\]

In the rest of this subsection, we will bound \( \mathcal{S}_1 \) and \( \mathcal{S}_2 \) respectively. The approach used here is somewhat standard in learning theory. \( \mathcal{S}_2 \) is a typical quantity that can be estimated by probability inequalities. We shall bound it by the following one-side Bernstein inequality [7].

**Lemma 5.4.** Let \( \xi \) be a random variable on a probability space \( Z \) with mean \( \mathbf{E}(\xi) \), variance \( \sigma^2(\xi) = \sigma^2 \). If \( |\xi(z) - \mathbf{E}(\xi)| \leq M_\xi \) for almost all \( z \in Z \). then, for all \( \varepsilon > 0 \),

\[
P^m \left\{ \frac{1}{m} \sum_{i=1}^m \xi(z_i) - \mathbf{E}(\xi) \geq \varepsilon \right\} \leq \exp \left\{ -\frac{m \sigma^2 \varepsilon}{2 \left( \sigma^2 + \frac{1}{3} M_\xi \varepsilon \right)} \right\}.
\]
By the help of the above lemma, we can deduce the following bound of $S_2$.

**Proposition 5.5.** 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

$$1/m \sum_{i=1}^m \xi_2(z_i) - E(\xi_2) \leq \varepsilon.$$

**Proof.** It follows from Lemma 2.2 that $\|f_n\|_\infty \leq M$, which together with $|f_\rho(x)| \leq M$ yields that

$$|\xi_2| \leq (\|f_n\|_\infty + M)(\|f_n\|_\infty + M) \leq 4M^2.$$

Hence $|\xi_2 - E(\xi_2)| \leq 8M^2$. Moreover, we have

$$E(\xi_2^2) = E((f_n(X) - f_\rho(X))^2 \times (f_n(X) - Y) + (f_\rho(X) - Y))^2) \leq 16M^2\|f_n - f_\rho\|_\rho^2,$$

which implies that

$$\sigma^2(\xi_2) \leq E(\xi_2^2) \leq 16M^2\|f_n - f_\rho\|_\rho^2.$$

Now we apply Lemma 5.4 to $\xi_2$. It asserts that for any $t > 0$,

$$1/m \sum_{i=1}^m \xi_2(z_i) - E(\xi_2) \leq t$$

with confidence at least

$$1 - \exp\left(-\frac{mt^2}{2(\sigma^2(\xi_2) + \frac{\varepsilon}{3}M^2t)}\right) \geq 1 - \exp\left(-\frac{3mt^2}{48M^2 \left(2\|f_n - f_\rho\|_\rho^2 + t\right)}\right).$$

This implies the desired estimate. $\blacksquare$

It is more difficult to estimate $S_1$ because $\xi_1$ involves the sample $z$ through $f_{z,\lambda}$. We will use the idea of empirical risk minimization to bound this term by means of covering number [7]. The main tools are the following three lemmas.

**Lemma 5.6.** Let $V_k$ be a $k$-dimensional function space defined on $S^d$. Denote by $\pi_M V_k = \{\pi_M f : f \in V_k\}$. Then

$$\log \mathcal{N}(\pi_M V_k, \eta) \leq ck \log \frac{M}{\eta},$$

where $c$ is a positive constant 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 $\eta$-net of $\pi_M V_k$.

Lemma 5.6 is a direct result through combining [16, Property 1] and [17, P.437]. It shows that the covering number of a bounded functional space can be also bounded properly. The following ratio probability inequality is a standard result in learning.
theory [7]. It deals with variances for a function class, since the Bernstein inequality
takes care of the variance well only for a single random variable.

**Lemma 5.7.** Let \( \mathcal{G} \) be a set of functions on \( \mathcal{Z} \) such that, for some \( c \geq 0 \), \( |g - E(g)| \leq B \) almost everywhere and \( E(g^2) \leq cE(g) \) for each \( g \in \mathcal{G} \). Then, for every \( \varepsilon > 0 \),

\[
P^m \left\{ \sup_{f \in \mathcal{G}} \frac{E(g) - \frac{1}{m} \sum_{i=1}^{m} g(z_i)}{\sqrt{E(g) + \varepsilon}} \geq \sqrt{\varepsilon} \right\} \leq \mathcal{N}(\mathcal{G}, \varepsilon) \exp \left\{ -\frac{m\varepsilon}{2c + 2B^2} \right\}.
\]

Now we are in a position to give an upper bound of \( S_2 \).

**Proposition 5.8.** For all \( \varepsilon > 0 \),

\[
S_1 \leq \frac{1}{2} \mathcal{E}(\pi_M f_{x, \lambda}) - \mathcal{E}(f_\rho) + \varepsilon
\]

holds with confidence at least

\[
1 - \exp \left\{ c n^d \log \frac{4M^2}{\varepsilon} - \frac{3m\varepsilon}{128M^2} \right\}.
\]

**Proof.** Set

\[
\mathcal{F} := \{(f(X) - Y)^2 - (f_\rho(X) - Y)^2 : f \in \pi_M \mathcal{H}_K\}.
\]

Then for \( g \in \mathcal{F} \), there exists \( f \in \mathcal{H}_K \) such that \( g(Z) = (\pi_M f(X) - Y)^2 - (f_\rho(X) - Y)^2 \).

Therefore,

\[
E(g) = \mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho) \geq 0, \quad \frac{1}{m} \sum_{i=1}^{m} g(z_i) = \mathcal{E}_\alpha(\pi_M f) - \mathcal{E}_\alpha(f_\rho).
\]

Since \( |\pi_M f| \leq M \) and \( |f_\rho(X)| \leq M \) almost everywhere, we find that

\[
|g(z)| = |(\pi_M f(X) - f_\rho(X))(\pi_M f(X) - Y) + (f_\rho(X) - Y)| \leq 8M^2
\]

almost everywhere. It follows that \( |g(Z) - E(g)| \leq 16M^2 \) almost everywhere and

\[
E(g^2) \leq 16M^2 \|\pi_M f - f_\rho\|_\rho^2 = 16M^2 E(g).
\]

Now we apply Lemma 5.7 with \( B = c = 16M^2 \) to the set of functions \( \mathcal{F} \) and obtain that

(5.3)

\[
\sup_{f \in \pi_M \mathcal{H}_K} \frac{\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}(\pi_M f) - \mathcal{E}_\alpha(f_\rho)\}}{\sqrt{\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} + \varepsilon}} \leq \sqrt{\varepsilon}
\]

with confidence at least

\[
1 - \mathcal{N}(\mathcal{F}, \varepsilon) \exp \left\{ -\frac{3m\varepsilon}{128M^2} \right\}.
\]

Observe that for \( g_1, g_2 \in \mathcal{F} \) there exist \( f_1, f_2 \in \pi_M \mathcal{H}_K \) such that

\[
g_j(Z) = (f_j(X) - Y)^2 - (f_\rho(X) - Y)^2, \quad j = 1, 2.
\]
In addition, for any \( f \in \pi_M \mathcal{H}_K \), there holds
\[
|g_1(Z) - g_2(Z)| = |(f_1(X) - Y)^2 - (f_2(X) - Y)^2| \leq 4M\|f_1 - f_2\|_\infty.
\]
We see that for any \( \varepsilon > 0 \), an \((\frac{\varepsilon}{4M})\)-covering of \( \pi_M \mathcal{H}_K \) provides an \( \varepsilon \)-covering of \( \mathcal{F} \). Therefore
\[
\mathcal{N}(\mathcal{F}, \varepsilon) \leq \mathcal{N}\left(\pi_M \mathcal{H}_K, \frac{\varepsilon}{4M}\right).
\]
Then the confidence is
\[
1 - \mathcal{N}(\mathcal{F}, \varepsilon) \exp \left\{ -\frac{3m\varepsilon}{128M^2} \right\} \geq 1 - \mathcal{N}\left(\pi_M \mathcal{H}_K, \frac{\varepsilon}{4M}\right) \exp \left\{ -\frac{3m\varepsilon}{128M^2} \right\}.
\]
Since
\[
\sqrt{\varepsilon}\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} + \varepsilon \leq \frac{1}{2}\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} + \varepsilon,
\]
it follows from (5.3) and Lemma 5.6 that
\[
S_2 \leq \frac{1}{2}\mathcal{E}(\pi_M f_{z, \lambda}) - \mathcal{E}(f_\rho) + \varepsilon
\]
holds with confidence at least
\[
1 - \exp \left\{ cn^d \log \frac{4M^2}{\varepsilon} - \frac{3m\varepsilon}{128M^2} \right\}.
\]
This finishes the proof. \( \square \)

Now we are in a position to deduce the final learning rate of the kernel ridge regression (3.2). Firstly, it follows from Propositions 5.3, 5.5 and 5.8 that
\[
\mathcal{E}(\pi_M f_{z, \lambda}) - \mathcal{E}(f_\rho) \leq D_n(\lambda) + S_1 + S_2 \leq C (n^{-2r} + \lambda M^2)
\]
\[
+ \frac{1}{2}(\mathcal{E}(\pi_M f_{z, \lambda}) - \mathcal{E}(f_\rho)) + 2\varepsilon
\]
holds with confidence at least
\[
1 - \exp \left\{ cn^d \log \frac{4M^2}{\varepsilon} - \frac{3m\varepsilon}{128M^2} \right\} \exp \left\{ -\frac{3m\varepsilon^2}{48M^2 (2\|f_n - f_\rho\|_\rho^2 + \varepsilon)} \right\}.
\]
Then, by setting \( \varepsilon \geq \varepsilon_+ \geq C(m/\log m)^{-2r/(2r + d)} \), \( n = \lceil eq^{-1/(2r)} \rceil \) and \( \lambda \leq M^{-2}\varepsilon \), we get, with confidence at least
\[
1 - \exp\{-Cm\varepsilon\},
\]
there holds
\[
\mathcal{E}(\pi_M f_{z, \lambda}) - \mathcal{E}(f_\rho) \leq 4\varepsilon.
\]

The lower bound can be more easily deduced. Actually, it can be easily deduced from the Chapter 3 of [9] that for any estimator \( f_{z} \in \Phi_m \), there holds
\[
\sup_{f_\rho \in \mathcal{W}_r} \mathbb{P}_m \{ z : \|f_z - f_\rho\|_\rho \geq \varepsilon \} \geq \begin{cases} \varepsilon_0, & \varepsilon < \varepsilon_-, \\ e^{-cm\varepsilon}, & \varepsilon \geq \varepsilon_- \end{cases},
\]
where \( \varepsilon_0 = \frac{1}{2} \) and \( \varepsilon_- = cm^{-2r/(2r + d)} \) for some universal constant \( c \). With this, the proof of Theorem 3.1 is completed.
5.2. Proof of Theorem 4.1. Before we proceed the proof, we at first present a simple description of the methodology. The methodology we adopted in the proof of Theorem 4.1 seems of novelty. 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) [32]. All of the aforementioned results about coefficient regularization in SDHS fall into this style. According to [32], 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 $l^n$ kernel regularization into the approximation and sample errors (two terms) only. The core tool is needlet kernel’s excellent localization properties in both the spatial and frequency domain, with which the reproducing property, compressible property and the best approximation property can be guarantee. After presenting a probabilistic cubature formula for spherical polynomials, we can prove that all the polynomials can be represented by via the SDHS. This helps us to deduce the approximation error. Since all the polynomials can be represented by via the SDHS, this is an indispensable part attributed to an essential characteristic of learning algorithms in SDHS, compared with the learning schemes in SIHS.

To state the next lemma, we need introduce the following definitions. Let $\mathcal{U}$ be a finite dimensional vector space with norm $\| \cdot \|_V$, 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 $T_\mathcal{U} : \mathcal{V} \to \mathbb{R}^{\text{Card}(\mathcal{U})}$ defined by $T_\mathcal{U}(x) = (u(x))_{u \in \mathcal{U}}$ is injective, where $\text{Card}(\mathcal{U})$ 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 $\mathbb{R}_{\text{Card}(\mathcal{U})}^*$ have a norm $\| \cdot \|_{\mathbb{R}_{\text{Card}(\mathcal{U})}^*}$, with $\| \cdot \|_{\mathbb{R}_{\text{Card}(\mathcal{U})}^{\text{Card}(\mathcal{U})}}$ being its dual norm on $\mathbb{R}_{\text{Card}(\mathcal{U})}^*$. Equipping $\mathcal{W}$ with the induced norm, and let $\|T_\mathcal{U}^{-1}\| : = \|T_\mathcal{U}^{-1}\|_{\mathcal{W} \to \mathcal{V}}$. In addition, let $\mathcal{K}_+$ be the positive cone of $\mathbb{R}_{\text{Card}(\mathcal{U})}^*$ that is, all $(r_u) \in \mathbb{R}_{\text{Card}(\mathcal{U})}$ for which $r_u \geq 0$. Then the following Lemma 5.10 can be found in [19].

**Lemma 5.10.** 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 $\|v\|_{\mathcal{V}^*} \leq A$, then there exist real numbers $\{a_u\}_{u \in \mathcal{U}}$, depending only on $v$ such that for every $t \in \mathcal{V}$,

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

and

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

Also, if $\mathcal{W}$ contains an interior point $v_0 \in \mathcal{K}_+$ and if $v(T_\mathcal{U}^{-1} t) \geq 0$ when $t \in \mathcal{V} \cap \mathcal{K}_+$, then we may choose $a_u \geq 0$. 

To present the probabilistic cubature formula, we need the following two lemmas. 

**Lemma 5.10.** Let $1 \leq p < q \leq \infty$, $n \geq 1$ be an integer. Then

$$\|Q\|_{L_p(S^d)} \leq C n^{\frac{d}{p} - \frac{d}{q}} \|Q\|_{L_q(S^d)}, \quad Q \in \Pi_d^d$$

where the constant $C$ depends only on $d$.
By the help of Lemma 5.4, Lemma 5.9 and Lemma 5.10 we can deduce the following probabilistic cubature formula.

**Proposition 5.11.** Let $N$ be a positive integer and $1 \leq p \leq \infty$. If $\Lambda_N := \{t_i\}_{i=1}^N$ are i.i.d. random variables drawn according to arbitrary distribution $\mu$ on $\mathbb{S}^d$, then there exists a set of real numbers $\{a_i\}_{i=1}^N$ such that

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

holds with confidence at least

$$1 - 2 \exp \left\{ -CN^d \right\},$$

subject to

$$\sum_{i=1}^N |a_i|^p \leq \frac{|\mathbb{S}^d|}{1 - \varepsilon N^{1-p}}.$$ 

**Proof.** Without loss of generality, we assume $Q_n \in \mathcal{P}^0 := \{f \in \Pi^d_n : \|f\|_p \leq 1\}$. We denote the $\delta$-net of all $f \in \mathcal{P}^0$ by $A(\delta)$. It follows from [13, Chap.9] and the definition of the covering number that the smallest cardinality of $A(\delta)$ is bounded by

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

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

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

Since the sampling set $\Lambda_N$ consists of a sequence of i.i.d. random variables on $\mathbb{S}^d$, the sampling points are a sequence of functions $t_j = t_j(\omega)$ on some probability space $(\Omega, \mathcal{P})$. If we set $\xi_j^P(P) = |P(t_j)|^p$, then

$$\eta_i(P) = |P(t_i)|^p - \|P\|_p^p = |P(t_i)(\omega)|^p - \mathbb{E}\xi_j^P,$$

where we have used the equalities

$$\mathbb{E}\xi_j^P = \int_{\mathbb{S}^d} |P(x)|^p d\omega(x) = \|P\|_p^p = \int_{\Omega} |P(t(\omega))|^p d\omega.$$

Furthermore,

$$|\eta_i(P)| \leq \sup_{\omega \in \Omega} |P(t_i(\omega))|^p - \|P\|_p^p \leq \|P\|_{\infty}^p + \|P\|_p^p.$$ 

It follows from Lemma 5.9 that

$$\|P\|_{\infty} \leq C n^{\frac{d}{p}} \|P\|_p.$$
Hence
\[ |\eta_i(P) - \mathbb{E}\eta_i(P)| \leq Cn^d. \]

Moreover, using Lemma 5.9 again, there holds,
\[ \sigma^2(\eta_i(P)) \leq \mathbb{E}((\eta_i(P))^2) \leq \|P\|_{2p}^{2p} - \|P\|_p^{2p} \leq Cn^d. \]

Then, using Lemma 5.4 with \( \varepsilon = 1/2 \) and \( M_\xi = \sigma^2 = Cn^d \), we have for fixed \( P \in A(1) \), with probability at most \( 2 \exp\{ -CN/n^d \} \), there holds
\[ \left| \frac{1}{N} \sum_{i=1}^{N} \eta_i \right| \geq \frac{1}{4}. \]

Noting there are at \( \exp\{Cn^d\} \) polynomials in \( A(1) \), we get
\[ (5.4) \quad P^N \left\{ \frac{1}{N} \sum_{i=1}^{N} |\eta_i(N)| \geq \frac{1}{4} \text{ for some } P \in A(1) \right\} \leq 2 \exp \left\{ -\frac{CN}{n^d} + Cn^d \right\}. \]

Now, we aim to bound the probability of the event:
\[ \{ \text{e1} \} \text{ for some } l \geq 1, \text{ some } P \in A(2^{-l}) \text{ and some } Q \in A(2^{-l+1}) \text{ with } \|P - Q\| \leq 3 \times 2^{-l}, \text{ there holds} \]
\[ |\eta_i(P) - \eta_i(Q)| \geq \frac{1}{4(l + 1)^2}. \]

The main tool is also the Bernstein inequality. To this end, we should bound
\[ |\eta_i(P) - \eta_i(Q) - \mathbb{E}(\eta_i(P) - \eta_i(Q))| \text{ and the variance } \sigma^2(\eta_i(P) - \eta_i(Q)). \]

According to the Taylor formula
\[ a^p = b^p + p\theta^{p-1}(a - b), \quad \theta \in (a, b), \]
and Lemma 5.9, we have
\[ \|\eta_i(P) - \eta_i(Q)\| \leq \sup_{\omega \in \Omega} \|P(t_i(\omega))|^p - |Q(t_i(\omega))|^p + \|P\|_p^p - \|P\|_p^p\| \leq Cn^d \|P - Q\|, \]

and
\[ \sigma^2(\eta_i(P) - \eta_i(Q)) \leq \mathbb{E}((\eta_i(P) - \eta_i(Q))^2) \leq \int_{S^d} (|P(x)|^p - |Q(x)|^p)^2 d\omega(x) - (\|P\|_p^p - \|Q\|_p^p)^2 \leq Cn^d \|P - Q\|^2. \]

If \( P \in A(2^{-l}) \) and \( Q \in A(2^{-l+1}) \) with \( \|P - Q\| \leq 3 \times 2^{-l} \), then it follows from Lemma 5.4 again that,
\[ P^N \left( \left| \sum_{i=1}^{N} \eta_i(P) - \eta_i(Q) \right| > \frac{1}{4(l + 1)^2} \right) \leq 2 \exp \left\{ -\frac{N}{Cn^d(2^{-2l^4} + (2^{-2l^2}))} \right\} \leq 2 \exp \left\{ -\frac{N}{Cn^d2^{-l^2}} \right\}.
Since there are at most \(2 \exp\{-Cn^d \log l\}\) polynomials in \(A(2^{-l}) \cup A(2^{-l+1})\), then the event (e1) holds with probability at most
\[
\sum_{i=1}^{\infty} 2 \exp\left\{-\frac{CN}{n^a 2^{-l/2}} + Cn^d \log l\right\} \leq \sum_{i=1}^{\infty} 2 \exp\left\{-2^{l/2} \left(\frac{CN}{n^a} - n^d\right)\right\}.
\]
Since \(\sum_{i=1}^{\infty} e^{-a^ib} \leq Ce^{-b}\) for any \(a > 1\) and \(b \geq 1\), we then deduce that
\[
\mathbb{P}_m\{\text{The event (e1) holds}\} \leq 2 \exp\left\{\frac{CN}{n^a} - Cn^d\right\}.
\]
Thus, it follows from (5.4) and (5.5) that with confidence at least
\[
1 - 2 \exp\left\{\frac{CN}{n^a} - Cn^d\right\}
\]
there holds
\[
\left|\sum_{i=1}^{n} \eta_i(Q_n)\right| \leq \left|\sum_{i=1}^{n} \eta_i(P_0)\right| + \sum_{i=1}^{\infty} \left|\sum_{i=1}^{n} \eta_i(P_i) - \eta_i(P_l)\right| \leq \frac{1}{4} + \sum_{l=1}^{\infty} \frac{1}{4(l+1)^2} = \sum_{l=1}^{\infty} \frac{1}{4l^2} = \frac{\pi^2}{24} < \frac{1}{2}.
\]
This means that with confidence at least
\[
1 - 2 \exp\left\{\frac{CN}{n^a} - Cn^d\right\}
\]
there holds
\[
\frac{1}{2} \|Q_n\|_p^p \leq \frac{1}{N} \sum_{i=1}^{N} |Q_n(\alpha_i)|^p \leq \frac{3}{2} \|Q_n\|_p^p \quad \forall Q_n \in \Pi_n^d.
\]
Now, we use (5.6) and Lemma 5.10 to prove Lemma 5.11. In Lemma 5.10, we take \(V = \Pi_n^d\), \(\|Q_n\|_V = \|Q_n\|_p\), and \(W\) to be the set of point evaluation functionals \(\{\delta_{t_i}\}_{i=1}^{N}\). The operator \(T_W\) is then the restriction map \(Q_n \mapsto Q_n|_{\Lambda}\), with
\[
\|f\|_{\Lambda,p} := \left\{\left(\frac{1}{N} \sum_{i=1}^{N} |f(t_i)|^p\right)^{\frac{1}{p}}, \quad 1 \leq p < \infty,\right.\]
\[\sup_{1 \leq i \leq N} \{\|f(t_i)\|_p\}, \quad p = \infty.\]
It follows from (5.6) that with confidence at least
\[
1 - 2 \exp\left\{\frac{CN}{n^a} - Cn^d\right\}
\]
there holds \(\|T_W^{-1}\| \leq 2\). We now take \(u\) to be the functional
\[
y : Q_n \mapsto \int_{S^d} Q_n(x) d\omega(x).\]
By Hölder inequality, \(\|y\|_{V^*} \leq |S^d|\). Therefore, Lemma 5.10 shows that

\[
\int_{S^d} Q_n(x) d\omega(x) = \sum_{i=1}^{N} a_i Q_n(t_i)
\]

holds with confidence at least

\[
1 - 2 \exp\left\{ \frac{CN}{n^d} - Cn^d \right\}
\]

subject to

\[
\frac{1}{N} \sum_{i=1}^{N} \left( \frac{|a_i|}{1/N} \right)^p \leq 2|S^d|.
\]

This finishes the proof of Proposition 5.11.

To estimate the upper bound of

\[\mathcal{E}(\pi_M f_{z,\lambda,q}) - \mathcal{E}(f_\rho),\]

we first introduce an error decomposition strategy. It follows from the definition of \(f_{z,\lambda,q}\) that, for arbitrary \(f \in \mathcal{H}_{K, z}\),

\[
\mathcal{E}(\pi_M f_{z,\lambda,q}) - \mathcal{E}(f_\rho) \leq \mathcal{E}(\pi_M f_{z,\lambda,q}) - \mathcal{E}(f_\rho) + \lambda \Omega_{f_\rho}^2(f_{z,\lambda,q})
\]

\[
\leq \mathcal{E}(\pi_M f_{z,\lambda,q}) - \mathcal{E}(f) + \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \Omega_{f_\rho}^2(f_{z,\lambda,q}) - \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \Omega_{f_\rho}^2(f)
\]

\[
\leq \mathcal{E}(\pi_M f_{z,\lambda,q}) - \mathcal{E}(f_{z,\lambda,q}) + \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \Omega_{f_\rho}^2(f).
\]

Since \(f_\rho \in W_r\) with \(r > d/2\), it follows from the Sobolev embedding theorem and Jackson inequality [4] that there exists a \(P_\rho \in \Pi_n^d\) such that

\[
\|P_\rho\| \leq c\|f_\rho\| \quad \text{and} \quad \|f_\rho - P_\rho\|^2 \leq Cn^{-2r}.
\]

Then we have

\[
\mathcal{E}(f_{z,\lambda,q}) - \mathcal{E}(f_\rho) \leq \{\mathcal{E}(P_\rho) - \mathcal{E}(f_\rho) + \lambda \Omega_{f_\rho}^2(P_\rho)\} + \{\mathcal{E}(f_{z,\lambda,q}) - \mathcal{E}(f_{z,\lambda,q}) + \mathcal{E}(f_{z,\lambda,q}) - \mathcal{E}(P_\rho)\}
\]

\[
= \mathcal{D}(z, \lambda, q) + \mathcal{S}(z, \lambda, q),
\]

where \(\mathcal{D}(z, \lambda, q)\) and \(\mathcal{S}(z, \lambda, q)\) are called as the approximation error and sample error, respectively. The following Proposition 5.12 presents an upper bound for the approximation error.

**Proposition 5.12.** Let \(m, n \in \mathbb{N}, r > d/2\) and \(f_\rho \in W_r\). Then, with confidence at least \(1 - 2 \exp\{-cm/n^d\}\), there holds

\[
\mathcal{D}(z, \lambda, q) \leq C \left( n^{-2r} + 2\lambda |S^d| m^{1-q} \right),
\]

where \(C\) and \(c\) are constants depending only on \(d\) and \(r\).
Proof. From Lemma 2.2, it is easy to deduce that
\[ P_\rho(x) = \int_{S^d} P_\rho(x') K_n(x, x') d\omega(x'). \]
Thus, Proposition 5.11, Hölder inequality and \( r > d/2 \) yield that with confidence at least \( 1 - 2 \exp\{-cm/n^q\} \), there exists a set of real numbers \( \{a_i\}_{i=1}^m \) satisfying
\[ \sum_{i=1}^m |a_i|^q \leq 2|S^d|m^{1-q} \] 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 (5.7) implies that with confidence at least \( 1 - 2 \exp\{-cm/n^d\} \), \( P_\rho \) can be represented as
\[ P_\rho(x) = \sum_{i=1}^m a_i P_\rho(x_i) K_n(x_i, x) \in H_{K, \pi} \]
such that for arbitrary \( f_\rho \in W_r \), there holds
\[ \|P_\rho - f_\rho\|^2 \leq \|P_\rho - f_\rho\|^2 \leq C n^{-2r}, \]
and
\[ \Omega_k^q(P_\rho) \leq \sum_{i=1}^m |a_i P_\rho(x_i)|^q \leq (cM)^q \sum_{i=1}^m |a_i|^q \leq 2|S^d|m^{1-q}, \]
where \( C \) is a constant depending only on \( d \) and \( M \). It thus implies that the inequalities
\[ (5.8) \quad D(z, \lambda, q) \leq \|P_\rho - f_\rho\|^2 + \lambda \Omega_k^q(g^*) \leq C \left( n^{-2r} + 2\lambda|S^d|m^{1-q} \right) \]
holds with confidence at least \( 1 - 2 \exp\{-cm/n^d\} \). \( \square \)

At last, we deduce the final learning rate of \( l^q \) kernel regularization schemes (4.1). Firstly, it follows from Propositions 5.12, 5.8 and 5.5 that
\[ \mathcal{E}(\pi_M f_{z, \lambda, q}) - \mathcal{E}(f_\rho) \leq D(z, \lambda, q) + S_1^q + S_2^q \leq C \left( n^{-2r} + \lambda m^{1-q} \right) \]
\[ + \frac{1}{2} \left( \mathcal{E}(f_{z, \lambda, q}) - \mathcal{E}(f_\rho) \right) + 2\varepsilon \]
holds with confidence at least
\[ 1 - 4 \exp\{-cm/n^d\} - \exp \left\{ \frac{cn^d \log 4M^2}{\varepsilon} - \frac{3m\varepsilon}{128M^2} \right\} - \exp \left( -\frac{3m\varepsilon^2}{48M^2(2n^{-2r} + \varepsilon)} \right). \]

Then, by setting \( \varepsilon \geq \varepsilon_m^+ = C(m/\log m)^{-2r/(2r+d)}, n = \lceil \varepsilon^{-1/(2r)} \rceil \) and \( \lambda \leq m^{q-1}\varepsilon \), it follows from \( r > d/2 \) that
\[ 1 - 5 \exp\{-Cm\varepsilon^{d/(2r)}\} \geq \exp\{-Cm\varepsilon\} \]
\[ - \exp \left\{ C\varepsilon^{-(d/(2r)} \left( \log 1/\varepsilon + \log m \right) - Cm\varepsilon \right\} \]
\[ \geq 1 - 6 \exp\{-Cm\varepsilon\}. \]
That is, for \( \varepsilon \geq \varepsilon_m^+ \),
\[ \mathcal{E}(f_{z, \lambda, q}) - \mathcal{E}(f_\rho) \leq 6\varepsilon \]
holds with confidence at least \( 1 - 6 \exp\{-Cm\varepsilon\} \). This finishes the proof of Theorem 4.1.
6. Conclusion and discussion. Since its inception in [24], needlets have become the most popular tools to tackle spherical data due to its perfect localization performance in both the frequency and spacial domains. The main novelty of the present paper is to suggest the usage of the needlet kernel in kernel methods to deal with spherical data. Our contributions can be summarized as follows. Firstly, the model selection problem of the kernel ridge regression boils down to choosing a suitable kernel and the corresponding regularization parameter. Namely, there are totally two types parameters in the kernel methods. This requires relatively large amount of computations when faced with large-scaled data sets. Due to needlet kernel’s excellent localization property in the frequency domain, we prove that, if a truncation operator is added to the final estimate, then as far as the model selection is concerned, the regularization parameter is not necessary in the sense of rate optimality. This means that there is only a discrete parameter, the frequency of the needlet kernel, needs tuning in the learning process, which presents a theoretically guidance to reduce the computation burden. Secondly, Compared with the kernel ridge regression, $l^q$ kernel regularization learning, including the kernel lasso estimate and kernel bridge estimate, may bring a certain additional attribution of the estimator, such as the sparsity. When utilized the $l^q$ kernel regularization learning, the focus is to judge whether it degrades the generalization capability of the kernel ridge regression. Due to needlet kernel’s excellent localization property in the spacial domain, we have proved in this paper that, on the premise of embodying the feature of the $l^q$ ($0 < q < \infty$) kernel regularization learning, the selection of $q$ doesn’t affect the generalization error in the sense of rate optimality. Both of them showed that the needlet kernel is a good choice of the kernel method to deal with spherical data.

REFERENCES

[1] Abrial, P., Moudden, Y., Starck, J., Delabrouille, J. and Nguyen, M. (2008). CMB data analysis and sparsity. Statist. 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] Brown, G. and Dai, F. (2005). Approximation of smooth functions on compact two-point homogeneous spaces. J. Funct. Anal. 220 401–423.
[5] 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.
[6] Chang, T., Ko, D., Royer, J. and Lu, J. (2000). Regression techniques in plate tectonics. Statist. Sci. 15 342–356.
[7] Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. Bull. Amer. Math. Soc. 39 1–49.
[8] 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.
[9] Devore, R. A., Kerkyacharian, G., Picard, D. and Temlyakov, V. (2006). Approximation methods for supervised learning. Found. Comput. Math. 6 3–58.
[10] Dodelson, S. (2003). Modern Cosmology. Academic Press, London.
[11] Downs, T. (2003). Spherical regression. Biometrika 90, 655–668.
[12] Freedman, W., Gervens, T. and Schreiner, M. (1998). Constructive Approximation on the Sphere. Oxford University Press Inc., New York.
[13] Györfy, L., Kohler, M., Krzyzak, A. and Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression. Springer, Berlin.
[14] 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.
[15] 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.

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

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

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

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

[20] 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.

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

[22] 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.

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

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

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

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

[27] 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.

[28] 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.

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

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

[31] 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.

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

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