Journal of Machine Learning Research 20 (2019) 1-44 Submitted 3/13; Revised 8/18; Published 10/19

Sparse Kernel Regression with Coefficient-based ℓ_q-regularization

Lei Shi MASTONE1983@GMAIL.COM
Shanghai Key Laboratory for Contemporary Applied Mathematics
School of Mathematical Sciences, Fudan University
Shanghai, P. R. China
Xiaolin Huang XIAOLINHUANG@SJTU.EDU.CN
Institute of Image Processing and Pattern Recognition
Institute of Medical Robotics, Shanghai Jiao Tong University
MOE Key Laboratory of System Control and Information Processing
Shanghai, P. R. China
Yunlong Feng YLFENG@ALBANY.EDU
Department of Mathematics and Statistics, State University of New York at Albany
New York, USA
Johan A.K. Suykens JOHAN.SUYKENS@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium

Editor: Mikhail Belkin

Abstract

In this paper, we consider the ℓ_q-regularized kernel regression with $0 < q \leq 1$. In form, the algorithm minimizes a least-square loss functional adding a coefficient-based ℓ_q-penalty term over a linear span of features generated by a kernel function. We study the asymptotic behavior of the algorithm under the framework of learning theory. The contribution of this paper is two-fold. First, we derive a tight bound on the ℓ_2-empirical covering numbers of the related function space involved in the error analysis. Based on this result, we obtain the convergence rates for the ℓ_1-regularized kernel regression which is the best so far. Second, for the case $0 < q < 1$, we show that the regularization parameter plays a role as a trade-off between sparsity and convergence rates. Under some mild conditions, the fraction of non-zero coefficients in a local minimizer of the algorithm will tend to 0 at a polynomial decay rate when the sample size m becomes large. As the concerned algorithm is non-convex, we also discuss how to generate a minimizing sequence iteratively, which can help us to search a local minimizer around any initial point.

Keywords: Learning Theory, Kernel Regression, Coefficient-based ℓ_q-regularization ($0 < q \leq 1$), Sparsity, ℓ_2-empirical Covering Number

1. Introduction and Main Results

The *regression* problem aims at estimating the function relations from random samples and occurs in various statistical inference applications. An output estimator of regression algorithms is usually expressed as a linear combination of features, i.e., a collection of candidate functions. As an important issue in learning theory and methodologies, *sparsity* focuses

on studying the sparse representations of such linear combinations resulted from the algorithms. It is widely known that an ideal way to obtain the sparsest representations is to penalize the combinatorial coefficients by the ℓ_0-norm. However, the algorithms based on ℓ_0-norm often lead to an NP-hard discrete optimization problem (see e.g., Natarajan (1995)), which motives the researchers to consider the ℓ_q-norm ($0 < q \leq 1$) as the substitution. In particular, the ℓ_1-norm constrained or penalized algorithms have achieved great success in a wide range of areas from signal recovery (Candès et al. (2006)) to variable selection in statistics (Tibshirani (1996)). Recently, several theoretical and experimental results (see e.g., Candès et al. (2008); Chartrand (2007); Fan and Li (2001); Saad and Ö. Yılmaz (2010); Rakotomamonjy et al. (2011); Xu et al. (2012)) suggest that ℓ_q-norm with $q \in (0, 1)$ yields sparser solutions than the ℓ_1-norm to produce accurate estimation. Due to the intensive study on compressed sensing (see e.g., Donoho (2006)), the algorithms involving the ℓ_q-norm ($0 < q \leq 1$) have drawn much attention in the last few years and been used for various applications, including image denoising, medical reconstruction and database updating.

In this paper, we focus on the ℓ_q-regularized kernel regression. In form, the algorithms minimize a least-square loss functional adding a coefficient-based ℓ_q-penalty term over a linear span of features generated by a kernel function. We shall establish a rigorous mathematical analysis on the asymptotic behavior of the algorithm under the framework of learning theory.

Let X be a compact subset of $\mathbb{R}^d$ and $Y \subset \mathbb{R}$, ρ be a Borel probability distribution on $Z = X \times Y$. For $f : X \to Y$ and $(x, y) \in Z$, the *least-square loss* $(f(x) - y)^2$ gives the error with f as a model for the process producing y from x. Then the resulting target function is called *regression function* and satisfies

$$f_\rho = \arg\min \left\{ \int_Z (f(x) - y)^2 d\rho \middle| f : X \to Y, \text{measurable} \right\}.$$

From Proposition 1.8 in Cucker and Zhou (2007), the regression function can be explicitly given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X, \tag{1}$$

where $\rho(\cdot|x)$ is the conditional probability measure induced by ρ at x. In the supervised learning framework, ρ is unknown and one estimates f_ρ based on a set of observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ which is assumed to be drawn independently according to ρ. We additionally suppose that $\rho(\cdot|x)$ is supported on $[-M, M]$, for some $M \geq 1$ and each $x \in X$. This uniform boundedness assumption for the output is standard in most literature in learning theory (see e.g., Zhang (2003); Smale and Zhou (2007); Mendelson and Neeman (2010); Wu et al. (2006)). Throughout the paper, we will use these assumptions without any further reference. Usually one may get an estimator of f_ρ by minimizing the empirical loss functional $\frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2$ over a *hypothesis space*, i.e., a pre-selected function set on X.

In kernel regression, the hypothesis space is generated by a kernel function $K : X \times X \to \mathbb{R}$. Recall that $\{x_i\}_{i=1}^m$ is the input data of the observations. The hypothesis space considered here is taken to be the linear span of the set $\{K_{x_i}\}_{i=1}^m$. For $t \in X$, we denote by K_t the

function

$$K_t: \quad X \to \mathbb{R} \\ x \mapsto K(x,t).$$

Let $0 < q \leq 1$, the output estimator of ℓ_q-regularized kernel regression is given by $\hat{f}_q = \sum_{i=1}^m c_{q,i}^{\mathbf{z}} K_{x_i}$, where its coefficient sequence $\mathbf{c}_q^{\mathbf{z}} = (c_{q,i}^{\mathbf{z}})_{i=1}^m$ satisfies

$$\mathbf{c}_q^{\mathbf{z}} = \arg\min_{\mathbf{c}\in\mathbb{R}^m} \left\{ \frac{1}{m}\sum_{j=1}^m \left(y_j - \sum_{i=1}^m c_i K(x_j, x_i) \right)^2 + \gamma \|\mathbf{c}\|_q^q \right\}. \tag{2}$$

Here $\gamma > 0$ is called a *regularization parameter* and $\|\mathbf{c}\|_q$ denotes the ℓ_q-norm of $\mathbf{c}$. Recall that for any $0 < q \leq 1$ and any sequence $\mathbf{w} = (w_n)_{n=1}^\infty$, the ℓ_0-norm and ℓ_q-norm are defined respectively as

$$\|\mathbf{w}\|_0 = \sum_{n=1}^\infty I(w_n \neq 0) \text{ and } \|\mathbf{w}\|_q = \left(\sum_{n\in \mathrm{supp}(\mathbf{w})} |w_n|^q \right)^{1/q},$$

where $I(\cdot)$ is the indicator function and $\mathrm{supp}(\mathbf{w}) := \{n \in \mathbb{N} : w_n \neq 0\}$ denotes the support set of $\mathbf{w}$. Strictly speaking, $\|\cdot\|_0$ is not a real norm and $\|\cdot\|_q$ merely defines a quasi-norm when $0 < q < 1$ (e.g., see Conway (2000)).

From an approximation viewpoint, we are in fact seeking for a function to approximate f_ρ from the function set spanned by the kernelized dictionary $\{K_{x_i}\}_{i=1}^m$. The kernelized dictionary together with its induced learning models has been previously considered in literature. In a supervised regression setting, to pursue a sparse nonlinear regression machine, Roth (2004) proposed the ℓ_1-norm regularized learning model induced by the kernelized dictionary, namely, the kernelized Lasso. It is indeed a basis pursuit method, the idea of which can be dated back to Chen and Donoho (1994); Girosi (1998). It was in Wu and Zhou (2008) that a framework of analyzing the generalization bounds for learning models induced by the kernelized dictionary was proposed. The idea behind is controlling the complexity of the hypothesis space and then investigating the approximation ability as well as the data-fitting risk of functions in this hypothesis space via approximation and concentration techniques, which is a typical learning theory approach. Following this line, a series of interesting studies have been expanded for various learning models induced by the kernelized dictionary. For example, probabilistic generalization bounds for different models were derived in Shi et al. (2011); Wang et al. (2012); Shi (2013); Lin et al. (2014); Feng et al. (2016) and many others. However, it is worth pointing out that one is not suggested to simply treat the kernelized dictionary as commonly used dictionaries. This is because learning models induced by the kernelized dictionary may possess more flexibility. For the kernelized dictionary, the positive semi-definite constraint on the kernel function is removed. The removing of the positive semi-definite constraints allows us to utilize some specific indefinite kernels to cope with real-world applications, see e.g., Schleif and Tino (2015). Moreover, $\{K_{x_i}\}_{i=1}^m$ is a data-dependent dictionary. In a nonlinear regression setting, comparing with models induced by basis functions, having seen enough observations, the data-dependent dictionary can provide adequate information. Consequently, the local information of the regression function can be captured with this redundant dictionary. An illustrative example

of this observation can be found in Ando et al. (2008). Recently, learning with indefinite kernels has drawn much attention. Most of work focused on the algorithmic study, e.g., Loosli et al. (2016); Huang et al. (2017). The algorithm under consideration can provide a simple scenario for regularized indefinite kernel regression. However, the theoretical work on this aspect is still limited so far.

Compared with the algorithms involving ℓ_0-penalty, the ℓ_1-regularized algorithms can be efficiently solved by convex programming methods. When $0 < q < 1$, the problem (2) is a non-convex optimization problem. Many efficient approaches have been developed to solve the ℓ_q-minimization problem of this type, e.g., Candès et al. (2008); Chen et al. (2010); Huang et al. (2008); Lai and Wang (2011); Xu et al. (2012); but there is no approach that guarantees to find a global minimizer. Most proposed approaches are *descent-iterative* in nature. To illustrate the principle of the minimization process, we define the objective functional of algorithm (2) as

$$\mathscr{T}_{\gamma,q}(\mathbf{c}) = \|\mathbf{y} - \mathbb{K}\mathbf{c}\|_2^2 + \gamma\|\mathbf{c}\|_q^q, \qquad \forall \mathbf{c} \in \mathbb{R}^m, \tag{3}$$

where $\mathbf{y} = \left(\frac{y_1}{\sqrt{m}}, \cdots, \frac{y_m}{\sqrt{m}}\right) \in \mathbb{R}^m$ and $\mathbb{K} \in \mathbb{R}^{m\times m}$ with entries $\mathbb{K}_{i,j} = \frac{K(x_i,x_j)}{\sqrt{m}}, 1 \le i,j \le m$. Given $\gamma > 0$ and an initial point $\mathbf{c}_0$, the descent-iterative minimization approach generates a minimizing sequence $\{\mathbf{c}_k\}_{k=1}^{\infty}$ such that $\mathscr{T}_{\gamma,q}(\mathbf{c}_k)$ are strictly decreasing along the sequence. Thus any local minimizer, including the global minimizer, that a descent approach may find must be in the level set $\{\mathbf{c} \in \mathbb{R}^m : \mathscr{T}_{\gamma,q}(\mathbf{c}) < \mathscr{T}_{\gamma,q}(\mathbf{c}_0)\}$. Therefore, in both theory and practice, one may be only interested in the local minimizer around some pre-given initial point. Specifically, for our problem, a reasonable choice of the initial point would be the solution in the case $q = 1$.

Assumption 1 *For $\gamma > 0$ and $0 < q < 1$, we assume that the coefficient sequence $\mathbf{c}_q^{\mathbf{z}}$ of the estimator $\hat{f}_q$ is a local minimizer of the problem (2) and satisfies $\mathscr{T}_{\gamma,q}(\mathbf{c}_q^{\mathbf{z}}) < \mathscr{T}_{\gamma,q}(\mathbf{c}_1^{\mathbf{z}})$, where $\mathbf{c}_1^{\mathbf{z}}$ is the global minimizer of the problem (2) at $q = 1$.*

In Section 2.1, we shall present a scheme for searching a local minimizer of problem (2) by constructing a descent-iterative minimization process. The previous theoretical analysis about least-square regression with a coefficient-based penalty term is valid only for a convex learning model, e.g., the ℓ_q-regularized regression with $q \ge 1$. To enhance the sparsity, one expects to use a non-convex ℓ_q penalty, i.e., $0 < q < 1$, but no optimization approach guarantees to find a global minimizer of the induced optimization problem. There is still a gap between the existing theoretical analysis and the optimization process: the estimator needs to be globally optimal in the theoretical analysis while the optimization method can not ensure the global optimality of its solutions. Up to our knowledge, due to the non-convexity of ℓ_q term, there still lacks a rigorous theoretical demonstration to support its efficiency in non-parametric regression. In this paper, we aim to fill this gap by developing an elegant theoretical analysis on the asymptotic performances of estimators $\hat{f}_q$ satisfying Assumption 1, where these estimators can be generated by the scheme in Section 2.1 or another descent-iterative minimization process. Here we would like to point out that the established convergence analysis of $\hat{f}_q$ only requires $\hat{f}_q$ to be a stationary point around $\hat{f}_1$.

One aim of this work is to discuss the sparseness of the algorithms (2), which is characterized by the upper bounds on the fraction of the non-zero coefficients in the expression

$\hat{f}_q = \sum_{i=1}^m c_{q,i}^{\mathbf{z}} K_{x_i}$. In general, the total number of coefficients is as large as the sample size m. But for some kind of kernels such as polynomial kernels, the representation of $\hat{f}_q$ is not unique whenever the sample size becomes large. For the sake of simplicity, we restrict our discussion on a special class of kernels.

Definition 1 *A function $K : X \times X \to \mathbb{R}$ is called an admissible kernel if it is continuous and for any $k \in \mathbb{N}$, $(c_1, \cdots, c_k) \in \mathbb{R}^k$ and distinct set $\{t_1, \cdots, t_k\} \subset X$, $\sum_{j=1}^k c_j K_{t_j} = 0$ for all $x \in X$ implies $c_j = 0$, $j = 1, \cdots, k$.*

It should be noticed that an admissible kernel here is not necessarily symmetric or positive semi-definite. Several widely used kernel functions from multi-variate approximation satisfy the condition of the admissible kernels, e.g., exponential kernels, Gaussian kernels, inverse multi-quadric kernels, B-spline kernels and compact supported radial basis function kernels including Wu's functions (Wu (1995)) and Wendland's functions (Wendland (1995)). One may see Wendland (2005) and references therein for more details of these kernels. For the same reason, the kernel function is required to be universal in Steinwart (2003) when discussing the sparseness of support vector machines. It is noticed that most universal kernels (see Micchelli et al. (2006)) are also admissible kernels. If K is admissible, as long as the input data are mutually different, the representation of $\hat{f}_q$ is unique and the number of non-zero coefficients is given by $\|\mathbf{c}_q^{\mathbf{z}}\|_0$. In the followings, when we refer to the linear combination of $\{K_{x_i}\}_{i=1}^m$, we always suppose that $\{x_i\}_{i=1}^m$ is pairwise distinct. In our setting, this assumption can be almost surely satisfied if the data generating distribution ρ is continuous.

Except sparsity, another purpose of this paper is to investigate how the estimator $\hat{f}_q$ given by (2) approximates the regression function f_ρ. Let ρ_X be the marginal distribution of ρ on X. With a suitable choice of γ depending on the sample size m, we show that the estimator $\hat{f}_q$ (or $\pi_M(\hat{f}_q)$ see Definition 2) converges to f_ρ in the function space $L^2_{\rho_X}(X)$ as m tends to infinity. Here, for a Borel measure Q on X, the space $L^2_Q(X)$ consists of all the square-integrable functions with respect to Q and the norm is given by $\|f\|_{L^2_Q} = \left(\int_X |f(x)|^2 dQ\right)^{1/2}$.

In order to state our results, we further recall some notations used in this paper. We say that K is a *Mercer kernel* if it is continuous, symmetric and positive semi-definite on $X \times X$. Such a kernel can generate a *reproducing kernel Hilbert space* (RKHS) $\mathscr{H}_K$ (e.g., see Aronszajn (1950)). For a continuous kernel function K, define

$$\widetilde{K}(u, v) = \int_X K(u, x) K(v, x) d\rho_X(x). \tag{4}$$

Then one can verify that $\widetilde{K}$ is a Mercer kernel.

Being an important convex approach for pursuing sparsity, regularizing with ℓ_1-norm deserves special attention in its own right. The following theorem illustrates our general error analysis for ℓ_1-regularized kernel regression. The result is stated in terms of properties of the input space X, the measure ρ and the kernel K.

Theorem 1 *Assume that X is a compact convex subset of $\mathbb{R}^d$ with Lipschitz boundary, $K \in \mathscr{C}^s(X \times X)$ with $s > 0$ is an admissible kernel and $f_\rho \in \mathscr{H}_{\widetilde{K}}$ with $\widetilde{K}$ defined by (4).*

Let $0 < \delta < 1$ *and*

$$\Theta = \begin{cases} \frac{d+2s}{2d+2s}, & \text{if } 0 < s < 1, \\ \frac{d+2\lfloor s \rfloor}{2d+2\lfloor s \rfloor}, & \text{otherwise}, \end{cases} \tag{5}$$

where $\lfloor s \rfloor$ *denotes the integral part of* s*. Take* $\gamma = m^{\epsilon-\Theta}$ *with* $0 < \epsilon \leq \Theta - \frac{1}{2}$*. Then with confidence* $1-\delta$*, there holds*

$$\|\hat{f}_1 - f_\rho\|^2_{L^2_{\rho_X}} \leq C_\epsilon \log(6/\delta) \left(\log(2/\delta) + 1\right)^6 m^{\epsilon-\Theta}, \tag{6}$$

where $C_\epsilon > 0$ *is a constant independent of* m *or* δ*.*

We shall prove Theorem 1 in Section 4.3 with the constant C_ϵ given explicitly. The convergence rate presented here improves the existing ones obtained in Wang et al. (2012); Shi et al. (2011); Guo and Shi (2012). In particular, for a $\mathscr{C}^\infty$ kernel (such as Gaussian kernels), the rate can be arbitrarily close to m^{-1}. To see the improvement, recall that the best convergence rates so far are given by Guo and Shi (2012). It was proved that if $f_\rho \in \mathscr{H}_{\widetilde{K}}$ and $K \in \mathscr{C}^\infty$, then with confidence $1-\delta$,

$$\|\hat{f}_1 - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{O}\left(\left(\log(24/\delta) + \log\log_2\left(\frac{16(1-\epsilon)}{\epsilon}\right)\right)^{\frac{8}{\epsilon}-7} m^{\epsilon-\frac{1}{2}}\right).$$

We improve this result in Theorem 1, as the convergence rate (6) is always faster than $m^{-1/2}$ even for Lipschitz continuous kernels. Next, we will further illustrate the optimality of the convergence rates obtained in Theorem 1 when f_ρ belongs to some specific function space. Let $X = [0,1]^d$, ρ_X be uniform distribution on $[0,1]^d$ and $K \in \mathscr{C}^s(X \times X)$ with $s := s' - d/2 > 0$ being an integer. Then the RKHS $\mathscr{H}_{\widetilde{K}} \subset \mathscr{C}^s([0,1]^d)$ and the Sobolev space $\mathscr{W}_2^{s'}([0,1]^d)$ consists of functions belonging to the Hölder space $\mathscr{C}^{s-1,\alpha}([0,1]^d)$ with an arbitrary $\alpha \in (0,1)$ (see e.g., Adams and Fournier (2003)). If $f_\rho \in \mathscr{H}_{\widetilde{K}} \cap \mathscr{W}_2^{s'}([0,1]^d)$, the claimed rate in (6) can be arbitrarily close to $\mathcal{O}(m^{-2s'/(2s'+d)})$ which is proven to be mini-max optimal in Fisher and Steinwart (2017).

Our refined result is mainly due to the following reasons. Firstly, when $K \in \mathscr{C}^s$ with $s \geq 2$ and the input space X satisfies some regularity condition, we obtain a tight upper bound on the empirical covering numbers of the related hypothesis space (see Theorem 11). Secondly, we apply the *projection operator* in the error analysis to get better estimates.

Definition 2 *For* $M \geq 1$*, the projection operator* π_M *on* $\mathbb{R}$ *is defined as*

$$\pi_M(t) = \begin{cases} -M & \text{if } t < -M, \\ t & \text{if } -M \leq t \leq M, \\ M & \text{if } t > M. \end{cases}$$

The projection of a function $f : X \to \mathbb{R}$ *is given by* $\pi_M(f)(x) = \pi_M(f(x)), \forall x \in X$*.*

The projection operator was introduced in the literature of Chen et al. (2004); Steinwart and Christmann (2008). It helps to improve the $\|\cdot\|_\infty-$bounds in the convergence analysis, which is very critical for sharp estimation.

In fact, under the uniform boundedness assumption, the performance of the algorithm (2) can be measured by the error $\|\pi_M(\hat{f}) - f_\rho\|_{L^2_{\rho_X}}$, where $\hat{f}$ is a resulting estimator. To

explain the details, we recall the definition of regression function f_ρ given by (1). As the conditional distribution $\rho(\cdot|x)$ is supported on $[-M, M]$ for every $x \in X$, the target function f_ρ takes value in $[-M, M]$ on X. So to see how an estimator $\hat{f}$ approximates f_ρ, it is natural to project values of $\hat{f}$ onto the same interval by the projection operator $\pi_M(\cdot)$. Due to the analysis in this paper, one can always expect better estimates by projecting the output estimator onto the interval $[-M, M]$. So if we consider the estimator $\pi_M(\hat{f}_1)$ in Theorem 1, the obtained result can be further improved. However, in order to make comparisons with previous results, we just give the error analysis for $\hat{f}_1$. But for the case $0 < q < 1$, we shall only consider the error $\|\pi_M(\hat{f}_q) - f_\rho\|_{L^2_{\rho_X}}$. To illustrate the sparseness of the algorithm, we also derive an upper bound on the quantity $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$, where $\mathbf{c}_q^{\mathbf{z}}$ denotes the coefficient sequence of $\hat{f}_q$.

Theorem 3 *Assume that X is a compact convex subset of $\mathbb{R}^d$ with Lipschitz boundary, $K \in \mathscr{C}^\infty(X \times X)$ is an admissible kernel and $f_\rho \in \mathscr{H}_{\widetilde{K}}$ with $\widetilde{K}$ defined by (4). For $0 < q < 1$, the estimator $\hat{f}_q$ is given by algorithm (2) and satisfies Assumption 1. Let $0 < \delta < 1$ and $\gamma = m^{-\tau}$ with $1 - q < \tau < 1$. With confidence $1 - \delta$, there hold*

$$\|\pi_M(\hat{f}_q) - f_\rho\|^2_{L^2_{\rho_X}} \leq \widetilde{C} \left(\log(18/\delta) + \log\log \frac{8}{q(1-\tau)} \right)^3 m^{-(\tau-(1-q))}, \tag{7}$$

and

$$\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m} \leq \widetilde{C}' \left(q(1-q)\right)^{-q/(2-q)} \left(\log(2/\delta) + \log\log \frac{8}{q(1-\tau)} \right)^6 m^{-q(1-\frac{\tau}{2-q})}, \tag{8}$$

where $\widetilde{C}$ and $\widetilde{C}'$ are positive constants independent of m or δ.

From Theorem 3, one can see that under the restrictions on τ and q, the quantity $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$ converges to 0 at a polynomial rate when the sample size m becomes large. The regularization parameter γ plays an important role as a trade-off between sparsity and convergence rates. Thus one can obtain a sparser solution at the price of lower estimation accuracy. Due to Theorem 3, when $\frac{3-\sqrt{5}}{2} < q < 1$, we may take $\tau = (2-q)(1-q)$, then the quantity $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$ behaves like $\mathcal{O}(m^{-q^2})$ and the corresponding convergence rate is $\mathcal{O}(m^{-(1-q)^2})$. In our sparsity analysis (see Section 5), the regularization parameter γ also plays a role as a thresholding for the value of non-zero coefficient in $\hat{f}_q = \sum_{i=1}^m c_{q,i}^{\mathbf{z}} K_{x_i}$. Due to our analysis, a lower bound for non-zero coefficients is given by $O(\gamma^{2/(2-q)})$, which implies that a small q will lead to more zero coefficients in the kernel expansion for a fixed $\gamma < 1$. It should be mentioned that our sparsity analysis is only valid for $0 < q < 1$.

For the RKHS-based regularization algorithms, it is well known that a classical way to obtain sparsity is to introduce the $\epsilon-$insensitive zone in the loss function. A theoretical result for this approach in Steinwart and Christmann (2009) shows that the fraction of non-zero coefficients in the kernel expansion is asymptotically lower and upper bounded by constants. From this point of view, regularizing the combinatorial coefficients by the ℓ_q-norm is a more powerful way to produce sparse solutions. As our theoretical analysis only gives results for the worst case situations, one can expect better performance of the ℓ_q-regularized kernel regression in practice.

At the end of this section, we point out that the mathematical analysis for the case $0 < q < 1$ is far from optimal. It is mainly because of our analysis is based on Assumption 1. Under this assumption, one needs to bound $\|\mathbf{c}_1^{\mathbf{z}}\|_q^q$ which is critical in the error analysis, where $\mathbf{c}_1^{\mathbf{z}}$ denotes the coefficient sequence of the estimator $\hat{f}_1$. In fact, due to the discussion in Section 2.1, we can construct a minimizing sequence from any point. Besides the solution of ℓ_1-regularized kernel regression, we may consider some other choices of the starting point, e.g, the solution of RKHS-based regularized least-square regression. We believe that how to bound the $\|\cdot\|_q-$norm of these initial vectors is still a problem when one considers other possible starting points. In this paper, we use the reverse Hölder inequality to handle this term and the bound is too loose especially when q is small. Actually, even if $\hat{f}_q$ is a global minimizer of problem (2), we can not give an effective approach to conduct the error analysis. Additionally, we do not assume any sparsity condition on the target function. One possible condition that one may consider is that the regression function belongs to the closure of the linear span of $\{K_x|x \in X\}$ under the ℓ_q-constraint. Compared with the *hard sparsity* introduced by the ℓ_0-norm, such kind of sparsity assumption is referred as to *soft sparsity* (see Raskutti et al. (2011)), which is based on imposing a certain decay rate on the entries of the coefficient sequence. Developing the corresponding mathematical analysis under the soft sparsity assumption will help us to understand the role of the ℓ_q-regularization in feature selections in an infinite-dimensional hypothesis space. We shall consider this topic in future work. However, the sparsity analysis in this paper is still valid to derive the asymptotical bound for $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$ and will lead to better estimates if a more elaborate error analysis can be given.

The paper is organized as follows. The next section presents a descent-iterative minimization process for algorithm (2) and establishes the framework of error analysis. In Section 3, we derive a tight bound on empirical covering numbers of the hypothesis space under the ℓ_1-constraint. In Section 4 and Section 5, we derive the related results on the error analysis and sparseness of ℓ_q-regularized kernel regression.

2. Preliminaries

This section is devoted to generating the minimizing sequences and establishing the framework of mathematical analysis for ℓ_q-regularized kernel regression.

2.1. Minimizing sequences for ℓ_q-regularized kernel regression

In this part, we present a descent-iterative minimization process for algorithm (2), which can probably search a local minimizer starting from any initial point. Motivated by recent work on $\ell_{1/2}-$regularization in Xu et al. (2012), we generalize their strategy to the case $0 < q < 1$.

Let $\mathrm{sgn}(x)$ be given by $\mathrm{sgn}(x) = 1$ for $x \geq 0$ and $\mathrm{sgn}(x) = -1$ for $x < 0$. Define a function $\psi_{\eta,q}$ for $\eta > 0$ and $0 < q < 1$ as

$$\psi_{\eta,q}(x) = \begin{cases} \mathrm{sgn}(x)t_{\eta,q}(|x|), & |x| > a_q\eta^{1/(2-q)}, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $a_q = (1-\frac{q}{2})(1-q)^{\frac{q-1}{2-q}}$ and $t_{\eta,q}(|x|)$ denotes the solution of the equation

$$2t + \eta q t^{q-1} - 2|x| = 0 \tag{10}$$

on the interval $[(q(1-q)\eta/2)^{1/(2-q)}, \infty)$ with respect to the variable t. We further define a map $\Psi_{\eta,q} : \mathbb{R}^m \to \mathbb{R}^m$, which is given by

$$\Psi_{\eta,q}(\mathbf{d}) = (\psi_{\eta,q}(d_1), \cdots, \psi_{\eta,q}(d_m)), \qquad \forall \mathbf{d} = (d_1, \cdots, d_m) \in \mathbb{R}^m. \tag{11}$$

Then we have the following important lemma.

Lemma 4 *For any $\eta > 0$, $0 < q < 1$ and $\mathbf{d} = (d_1, \cdots, d_m) \in \mathbb{R}^m$, the map $\Psi_{\eta,q} : \mathbb{R}^m \to \mathbb{R}^m$ given by (11) is well-defined and $\Psi_{\eta,q}(\mathbf{d})$ is a global minimizer of the problem*

$$\min_{\mathbf{c}\in\mathbb{R}^m} \left\{ \|\mathbf{c} - \mathbf{d}\|_2^2 + \eta \|\mathbf{c}\|_q^q \right\}. \tag{12}$$

We shall leave the proof to the Appendix. The function $\psi_{\eta,q}$ defines the ℓ_q-*thresholding function* for $0 < q < 1$. According to the proof of Lemma 4, given a $x \in \mathbb{R}$ and $\eta > 0$, the value of $\psi_{\eta,q}(x)$ in equation (9) is essentially a global minimizer of the problem

$$\min_{t\in\mathbb{R}} \left\{ |t - x|^2 + \eta |t|^q \right\}.$$

When $q = 1/2$, the function $\psi_{\eta,1/2}$ is exactly the half thresholding function obtained in Xu et al. (2012). We also observe that though the analysis in Lemma 4 is based on the fact $0 < q < 1$, the expression of $\psi_{\eta,q}$ is coherent for $q \in [0, 1]$. Concretely, as $\lim_{q\to 1-} a_q = \frac{1}{2}$, by letting $q \to 1-$ in the definition of $\psi_{\eta,q}$, one may obtain the soft thresholding function for ℓ_1-regularization, which is given by (e.g., see Daubechies et al. (2004))

$$\psi_{\eta,1}(x) = \begin{cases} x - \operatorname{sgn}(x)\eta/2, & |x| > \eta/2, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, if taking $q = 0$ in the expression of $\psi_{\eta,q}$, one may also derive the hard thresholding function for ℓ_0-regularization, which is defined as (e.g., see Blumensath and Davies (2008))

$$\psi_{\eta,0}(x) = \begin{cases} x, & |x| > \eta^{1/2}, \\ 0, & \text{otherwise.} \end{cases}$$

The expression of $\psi_{\eta,q}$ is very useful as we can establish a descent-iterative minimization process for algorithm (2) based on the idea in Daubechies et al. (2004). Recall the definitions of $\mathbb{K}$ and $\mathbf{y}$ in the objective functional $\mathscr{T}_{\gamma,q}$ given by (3). For any $(\lambda, \gamma) \in (0, \infty)^2$ and $\mathbf{c}_0 \in \mathbb{R}^m$, we iteratively define a sequence $\{\mathbf{c}_n\}_{n=1}^\infty$ as

$$\mathbf{c}_{n+1} = \Psi_{\lambda\gamma,q}\left(\mathbf{c}_n + \lambda \mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}_n)\right). \tag{13}$$

Then $\{\mathbf{c}_n\}_{n=1}^\infty$ is a minimizing sequence with a suitable chosen $\lambda > 0$.

Proposition 5 *Let $0 < q < 1$, $\gamma > 0$ and $0 < \lambda \le \frac{q}{2}\|\mathbb{K}\|_2^{-2}$, where $\|\cdot\|_2$ denotes the spectral norm of the matrix. If the sequence $\{\mathbf{c}_n\}_{n=0}^\infty$ is generated by the iteration process (13), then the following statements are true.*

(i) *If* $\mathbf{c}^* \in \mathbb{R}^m$ *is a local minimizer of the objective functional* $\mathscr{T}_{\gamma,q}(\mathbf{c})$*, then* c^* *is a stationary point of the iteration process (13), i.e.,*

$$\mathbf{c}^* = \Psi_{\lambda\gamma,q}\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right).$$

(ii) *The sequence* $\{\mathbf{c}_n\}_{n=0}^{\infty}$ *is a minimizing sequence such that the sequence* $\{\mathscr{T}_{\gamma,q}(\mathbf{c}_n)\}_{n=0}^{\infty}$ *is monotonically decreasing.*

(iii) *The sequence* $\{\mathbf{c}_n\}_{n=0}^{\infty}$ *converges to a stationary point of the iteration process (13) whenever* λ *is sufficiently small.*

We also prove this proposition in the Appendix. The properties of $\psi_{\eta,q}$ play an important role in the proof. When $q = 1/2$ and $q = 2/3$, the equation $2t + \eta q t^{q-1} - 2|x| = 0$ can be analytically solved, i.e., the corresponding thresholding function can be explicit expressed as an analytical function. This motivated people to develop efficient algorithms based on (13) for these two special cases. In particular, the $\ell_{1/2}-$regularization problem has been intensively studied in some literature (see Xu et al. (2012) and references therein). Since a general formula for ℓ_q-thresholding function is given by (9), it is also interesting to develop corresponding iterative algorithms and compare their empirical performances for different values of q.

2.2. Framework of error analysis

In this subsection, we establish the framework of convergence analysis. Because of the least-square nature, one can see Cucker and Zhou (2007) that

$$\|f - f_\rho\|_{L^2_{\rho_X}}^2 = \mathscr{E}(f) - \mathscr{E}(f_\rho), \qquad \forall f : X \to \mathbb{R},$$

where $\mathscr{E}(f) = \int_{X\times Y}(f(x) - y)^2 d\rho$. Let $\hat{f}$ be the estimator produced by algorithm (2). Particularly, the estimator $\hat{f}$ under our consideration is $\hat{f}_1$ or $\pi_M(\hat{f}_q)$ with $0 < q \leq 1$. To estimate $\|\hat{f} - f_\rho\|_{L^2_{\rho_X}}^2$, we only need to bound $\mathscr{E}(\hat{f}) - \mathscr{E}(f_\rho)$. This will be done by applying the *error decomposition* approach which has been developed in the literature for regularization schemes (e.g., see Cucker and Zhou (2007); Steinwart and Christmann (2008)). In this paper, we establish the error decomposition formula based on the first author's previous work (Guo and Shi (2012)).

To this end, we still need to introduce some notations. For any continuous function $K : X \times X \to \mathbb{R}$, define an integral operator on $L^2_{\rho_X}(X)$ as

$$L_K f(x) = \int_X K(x,t) f(t) d\rho_X(t), \quad x \in X.$$

Since X is compact and K is continuous, L_K and its adjoint L_K^* are both compact operators. If K is a Mercer kernel, the corresponding integral operator L_K is a self-adjoint positive operator on $L^2_{\rho_X}$, and its $r-$th power L_K^r is well-defined for any $r > 0$. From Cucker and Zhou (2007), we know that the RKHS $\mathscr{H}_K$ is in the range of $L_K^{\frac{1}{2}}$. Recalling the Mercer kernel $\widetilde{K}$ defined as (4) for a continuous kernel K, it is easy to check that $L_{\widetilde{K}} = L_K L_K^*$.

Following the same idea in Guo and Shi (2012), we use the RKHS $\mathscr{H}_{\widetilde{K}}$ with the norm denoted by $\|\cdot\|_{\widetilde{K}}$ to approximate f_ρ. The approximation behavior is characterized by the *regularization error*

$$\mathcal{D}(\gamma) = \min_{f\in\mathscr{H}_{\widetilde{K}}} \left\{ \mathscr{E}(f) - \mathscr{E}(f_\rho) + \gamma \|f\|^2_{\widetilde{K}} \right\}.$$

The following assumption is standard in the literature of learning theory (e.g., see Cucker and Zhou (2007); Steinwart and Christmann (2008)).

Assumption 2 *For some $0 < \beta \le 1$ and $c_\beta > 0$, the regularization error satisfies*

$$\mathcal{D}(\gamma) \le c_\beta \gamma^\beta, \qquad \forall \gamma > 0. \tag{14}$$

The decay of $\mathcal{D}(\gamma)$ as $\gamma \to 0$ measures the approximation ability of the function space $\mathscr{H}_{\widetilde{K}}$. Next, for $\gamma > 0$, we define the *regularizing function* as

$$f_\gamma = \arg\min_{f\in\mathcal{H}_{\widetilde{K}}} \left\{ \mathscr{E}(f) - \mathscr{E}(f_\rho) + \gamma \|f\|^2_{\widetilde{K}} \right\}. \tag{15}$$

The regularizing function uniquely exists and is given by $f_\gamma = \left(\gamma I + L_{\widetilde{K}}\right)^{-1} L_{\widetilde{K}} f_\rho$ (e.g., see Proposition 8.6 in Cucker and Zhou (2007)), where I denotes the identity operator on $\mathscr{H}_{\widetilde{K}}$.

Now we are in a position to establish the error decomposition for algorithm (2). Recall that for $0 < q \le 1$, $\mathbf{c}_q^{\mathbf{z}}$ denotes the coefficient sequence of the estimator $\hat{f}_q$ and $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ is the sample set. The empirical loss functional $\mathcal{E}_{\mathbf{z}}(f)$ is defined for $f : X \to \mathbb{R}$ as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2.$$

Proposition 6 *For $\gamma > 0$, the regularization function f_γ is given by (15) and $f_{\mathbf{z},\gamma} = \frac{1}{m}\sum_{i=1}^m K_{x_i} g_\gamma(x_i)$ with*

$$g_\gamma = L_K^* \left(\gamma I + L_{\widetilde{K}}\right)^{-1} f_\rho.$$

Let $\hat{f}$ be an estimator under consideration, we define

$$\begin{aligned}
\mathscr{S}_1 &= \{\mathscr{E}(\hat{f}) - \mathscr{E}_{\mathbf{z}}(\hat{f})\} + \{\mathscr{E}_{\mathbf{z}}(f_{\mathbf{z},\gamma}) - \mathscr{E}(f_{\mathbf{z},\gamma})\},\\
\mathscr{S}_2 &= \{\mathscr{E}(f_{\mathbf{z},\gamma}) - \mathscr{E}(f_\gamma)\} + \gamma\{\frac{1}{m}\sum_{i=1}^m |g_\gamma(x_i)| - \|g_\gamma\|_{L^1_{\rho_X}}\},\\
\mathscr{S}_3 &= \mathscr{E}(f_\gamma) - \mathscr{E}(f_\rho) + \gamma\|g_\gamma\|_{L^2_{\rho_X}}.
\end{aligned}$$

If $\hat{f}_q$ satisfies Assumption 1 with $0 < q < 1$, for the estimator $\hat{f} = \pi_M(\hat{f}_q)$, there holds

$$\mathscr{E}(\hat{f}) - \mathscr{E}(f_\rho) + \gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \le \mathscr{S}_1 + \mathscr{S}_2 + \mathscr{S}_3 + \gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q. \tag{16}$$

When $q = 1$, for $\hat{f} = \hat{f}_1$ or $\pi_M(\hat{f}_1)$, there holds

$$\mathscr{E}(\hat{f}) - \mathscr{E}(f_\rho) + \gamma\|\mathbf{c}_1^{\mathbf{z}}\|_1 \le \mathscr{S}_1 + \mathscr{S}_2 + \mathscr{S}_3. \tag{17}$$

To save space, we shall leave the proof to the Appendix. In fact, Proposition 6 presents three error decomposition formulas. When $\hat{f} = \hat{f}_1$, the inequality (17) is exactly the error decomposition introduced in Guo and Shi (2012) for ℓ_1-regularized kernel regression. Note that the bound (16) involves an additional term $\gamma m^{1-q}\|\hat{\mathbf{c}}_1\|_1^q$. Therefore, the asymptotic behavior of the global estimator $\hat{f}_1$ plays a significant part in the convergence analysis of $\hat{f}_q$ with $0 < q < 1$.

With the help of Proposition 6, one can estimate the total error by bounding $\mathscr{S}_i$ $(i = 1, 2, 3)$ and $\gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q$ respectively. The terms $\mathscr{S}_2$ and $\mathscr{S}_3$ are well estimated in Guo and Shi (2012) by fully utilizing the structure of the functions $f_{\mathbf{z},\gamma}$ and g_γ. Here we directly quote the following bound for $\mathscr{S}_2 + \mathscr{S}_3$. One may see Guo and Shi (2012) for the detailed proof.

Lemma 7 *For any $(\gamma, \delta) \in (0, 1)^2$, with confidence $1 - \delta$, there holds*

$$\mathscr{S}_2 + \mathscr{S}_3 \leq 8\kappa^2 \left(2\kappa^2 + 1\right) \log^2(4/\delta) \left\{ \frac{\mathcal{D}(\gamma)}{\gamma^2 m^2} + \frac{\mathcal{D}(\gamma)}{\gamma m} \right\} + \frac{(2\kappa + 1)\sqrt{\mathcal{D}(\gamma)} \log(4/\delta)}{m} + \frac{3}{2}\sqrt{\gamma \mathcal{D}(\gamma)} + 2\mathcal{D}(\gamma), \tag{18}$$

where $\kappa = \|K\|_{\mathscr{C}(X \times X)}$.

Therefore, our error analysis mainly focuses on bounding $\mathscr{S}_1$ and $\gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q$. The first term $\mathscr{S}_1$ can be estimated by uniform concentration equalities. These inequalities quantitatively characterize the convergence behavior of the empirical processes over a function set by various capacity measures, such as VC dimension, covering number, entropy integral and so on. For more details, one may refer to Van der Vaart and Wellner (1996) and references therein. In this paper, we apply the concentration technique involving the ℓ_2-*empirical covering numbers* to obtain bounds on $\mathscr{S}_1$. The ℓ_2-empirical covering number is defined by means of the normalized ℓ_2-metric d_2 on the Euclidian space $\mathbb{R}^l$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{l} \sum_{i=1}^{l} |a_i - b_i|^2 \right)^{1/2}, \qquad \mathbf{a} = (a_i)_{i=1}^l, \mathbf{b} = (b_i)_{i=1}^l \in \mathbb{R}^l.$$

Definition 8 *Let $(\mathcal{M}, d_{\mathcal{M}})$ be a pseudo-metric space and $S \subset \mathcal{M}$ a subset. For every $\epsilon > 0$, the covering number $\mathscr{N}(S, \epsilon, d_{\mathcal{M}})$ of S with respect to ϵ and the pseudo-metric $d_{\mathcal{M}}$ is defined to be the minimal number of balls of radius ϵ whose union covers S, that is,*

$$\mathscr{N}(S, \epsilon, d) = \min \left\{ \iota \in \mathbb{N} : \ S \subset \bigcup_{j=1}^{\iota} B(s_j, \epsilon) \text{ for some } \{s_j\}_{j=1}^{\iota} \subset \mathcal{M} \right\},$$

where $B(s_j, \epsilon) = \{s \in \mathcal{M} : d_{\mathcal{M}}(s, s_j) \leq \epsilon\}$ is a ball in $\mathcal{M}$. For a set $\mathscr{F}$ of functions on X and $\epsilon > 0$, the ℓ_2-empirical covering number of $\mathscr{F}$ is given by

$$\mathscr{N}_2(\mathscr{F}, \epsilon) = \sup_{l \in \mathbb{N}} \sup_{\mathbf{u} \in X^l} \mathscr{N}(\mathscr{F}|_{\mathbf{u}}, \epsilon, d_2),$$

where for $l \in \mathbb{N}$ and $\mathbf{u} = (u_i)_{i=1}^l \subset X^l$, we denote the covering number of the subset $\mathscr{F}|_{\mathbf{u}} = \{(f(u_i))_{i=1}^l : f \in \mathscr{F}\}$ of the metric space $(\mathbb{R}^l, d_2)$ as $\mathscr{N}(\mathscr{F}|_{\mathbf{u}}, \varepsilon, d_2)$.

As for the last term, we will derive a tighter bound for $\|\mathbf{c}_1^{\mathbf{z}}\|_1$ by an iteration technique based on the convergence analysis for the estimator $\pi_M(\hat{f}_1)$. The application of the projection operator will lead to better estimates.

3. Capacity of the hypothesis space under ℓ_1-constraint

In this section, by means of the ℓ_2-empirical covering numbers, we study the capacity of the hypothesis space generated by the kernel function. For $R > 0$, define $\mathscr{B}_R$ to be the linear combination of functions $\{K_x | x \in X\}$ under the ℓ_1-constraint

$$\mathscr{B}_R = \left\{ \sum_{i=1}^{k} \mu_i K_{u_i} : k \in \mathbb{N}, u_i \in X, \mu_i \in \mathbb{R} \text{ and } \sum_{i=1}^{k} |\mu_i| \leq R \right\}. \tag{19}$$

In this paper, we assume that the following capacity assumption for $\mathscr{B}_1$ comes into existence.

Assumption 3 *For a kernel function K, there exists an exponent p with $0 < p < 2$ and a constant $c_{K,p} > 0$ such that*

$$\log_2 \mathscr{N}_2(\mathscr{B}_1, \epsilon) \leq c_{K,p} \left(\frac{1}{\epsilon}\right)^p, \qquad \forall 0 < \epsilon \leq 1. \tag{20}$$

It is strictly proved in Shi et al. (2011) that, for a compact subset X of $\mathbb{R}^d$ and $K \in \mathscr{C}^s$ with some $s > 0$, the power index p can be given by

$$p = \begin{cases} 2d/(d+2s), & \text{when } 0 < s \leq 1, \\ 2d/(d+2), & \text{when } 1 < s \leq 1 + d/2, \\ d/s, & \text{when } s > 1 + d/2. \end{cases} \tag{21}$$

We will present a much tighter bound on the logarithmic ℓ_2-empirical covering numbers of $\mathscr{B}_1$. This bound holds for a general class of input space satisfying an *interior cone condition*.

Definition 9 *A subset X of $\mathbb{R}^d$ is said to satisfy an interior cone condition if there exist an angle $\theta \in (0, \pi/2)$, a radius $R_X > 0$, and a unit vector $\xi(x)$ for every $x \in X$ such that the cone*

$$C(x, \xi(x), \theta, R_X) = \left\{ x + ty : y \in \mathbb{R}^d, |y| = 1, y^T \xi(x) \geq \cos\theta, 0 \leq t \leq R_X \right\}$$

is contained in X.

Remark 10 *The interior cone condition excludes those sets X with cusps. It is valid for any convex subset of $\mathbb{R}^d$ with Lipschitz boundary (see e.g., Adams and Fournier (2003)).*

Now we are in a position to give our refined result on the capacity of $\mathscr{B}_1$.

Theorem 11 *Let X be a compact subset of $\mathbb{R}^d$. Suppose that X satisfies an interior cone condition and $K \in \mathscr{C}^s(X \times X)$ with $s \geq 2$ is an admissible kernel. Then there exists a constant $C_{X,K}$ that depends on X and K only, such that*

$$\log_2 \mathscr{N}_2(\mathscr{B}_1, \epsilon) \leq C_{X,K} \epsilon^{-\frac{2d}{d+2\lfloor s \rfloor}} \log_2 \left(\frac{2}{\epsilon}\right), \qquad \forall\ 0 < \epsilon \leq 1, \tag{22}$$

where $\lfloor s \rfloor$ denotes the integral part of s.

Recall the asymptotical bound obtained in Shi et al. (2011) with p given by (21). It asserts that for $K \in \mathscr{C}^s$ with $s \geq 2$, the quantity $\log_2 \mathscr{N}_2(\mathscr{B}_1, \epsilon)$ grows at most of the order $\epsilon^{-\min\{\frac{2d}{d+2}, \frac{d}{s}\}}$. Our stated result in Theorem 11 improves the previous bound a lot, as $\log_2 \mathscr{N}_2(\mathscr{B}_1, \epsilon)$ can be bounded by $\mathcal{O}(\epsilon^{-s_1})$ for any $s_1 > \frac{2d}{d+2\lfloor s \rfloor}$. Besides the ℓ_2-empirical covering number, another way to measure the capacity is the uniform covering number $\mathscr{N}(\mathscr{B}_1, \epsilon, \|\cdot\|_\infty)$ of $\mathscr{B}_1$ as a subset of the metric space $(\mathscr{C}(X), \|\cdot\|_\infty)$ of bounded continuous functions on X. From a classical result in function spaces (see Edmunds and Triebel (1996)), for $X = [0,1]^d$ and $K \in \mathscr{C}^s(X \times X)$, the unit ball $\mathscr{B}_1$ satisfies

$$c_s \left(\frac{1}{\epsilon}\right)^{d/s} \leq \log \mathscr{N}(\mathscr{B}_1, \epsilon, \|\cdot\|_\infty) \leq c_s' \left(\frac{1}{\epsilon}\right)^{d/s}, \qquad \forall \epsilon > 0.$$

When d is large or s is small, this estimate is rough. Moreover, the estimate above is asymptotic optimal and cannot be improved, which implies that the uniform covering number is not a suitable measurement for the capacity of $\mathscr{B}_1$. The ℓ_2-emprical covering number was first investigated in the field of empirical process (e.g., see Dudley (1987); Van der Vaart and Wellner (1996) and references therein). One usually assumes that, there exist $0 < p < 2$ and $c_p > 0$ such that

$$\log \mathscr{N}_2(\mathscr{F}, \epsilon) \leq c_p \epsilon^{-p}, \qquad \forall \epsilon > 0, \tag{23}$$

which guarantees the convergence of Dudley's entropy integral, i.e., $\int_0^1 \sqrt{\log \mathscr{N}_2(\mathscr{F}, \epsilon)} d\epsilon < \infty$. This fact is very important, since function class $\mathscr{F}$ with bounded entropy integrals satisfy the uniform central limit Theorem (see Dudley (1987) for more details). The classical result on ℓ_2-empirical covering number only asserts that if $\mathscr{N}_2(\mathscr{F}, \epsilon) = \mathcal{O}(\epsilon^{-p'})$ for some $p' > 0$, then the convex hull of $\mathscr{F}$ satisfies (23) with $p = \frac{2p'}{p'+2} < 2$. In this paper, we further clarify the relation between the smoothness of the kernel and the capacity of the hypothesis space. That is, we establish a more elaborate estimate for the power index p in Assumption 3 by using the prior information on the smoothness of the kernel. It should be pointed out that, from the relation $\mathscr{N}_2(\mathscr{F}, \epsilon) \leq \mathscr{N}(\mathscr{F}, \epsilon, \|\cdot\|_\infty)$, capacity assumption (23) for an RKHS generated by a positive semi-definite kernel K can be verified by bounding the uniform covering number. It is proved in Zhou (2003) that, when $\mathscr{F}$ is taken to be the unit ball in the RKHS $\mathscr{H}_K$, there holds $\log \mathscr{N}(\mathscr{F}, \epsilon, \|\cdot\|_\infty) = \mathcal{O}(\epsilon^{-2d/s})$ provided that $X \subset \mathbb{R}^d$ and $K \in \mathscr{C}^s(X \times X)$. Therefore, a sufficiently smooth kernel with $s > d$ can guarantee that capacity assumption (23) comes into existence. When X is a Euclidean ball in $\mathbb{R}^d$, the Sobolev space $\mathscr{W}_2^s(X)$ with $s > d/2$ is an RKHS. Birman and Solomyak (1967) proved that $\log \mathscr{N}_2(\mathscr{W}_2^s(X), \epsilon)$ is upper and lower bounded by $\mathcal{O}(\epsilon^{-p})$ with $p = d/s < 2$. However, up to our best knowledge, how to demonstrate capacity assumption (23) for a general RKHS is still widely open. From Theorem 11, one can see that even for positive semi-definite kernels,

the function space (19) is more suitable to be the hypothesis space than the classical RKHS, as its capacity can be well-estimated by the ℓ_2-empirical covering number.

We will prove Theorem 11 after a few lemmas. The improvement is mainly due to a local polynomial reproduction formula from the literature of multivariate approximation (see Wendland (2001); Jetter et al. (1999)). A point set $\Omega = \{\omega_1, \cdots, \omega_n\} \subset X$ is said to be Δ-*dense* if

$$\sup_{x \in X} \min_{\omega_j \in \Omega} |x - \omega_j| \leq \Delta,$$

where $|\cdot|$ is the standard Euclidean norm on $\mathbb{R}^d$. Denote the space of polynomials of degree at most s on $\mathbb{R}^d$ as $\mathcal{P}_s^d$. The following lemma is a formula for local polynomial reproduction (see Theorem 3.10 in Wendland (2001)).

Lemma 12 *Suppose X satisfies an interior cone condition with radius $R_X > 0$ and angle $\theta \in (0, \pi/2)$. Fix $s \in \mathbb{N}$ with $s \geq 2$. There exists a constant c_0 depending on θ, d and s such that for any Δ-dense point set $\Omega = \{\omega_1, \cdots, \omega_n\}$ in X with $\Delta \leq \frac{R_X}{c_0}$ and every $u \in X$, we can find real numbers $b_i(u)$, $1 \leq i \leq n$, satisfying*

(1) $\sum_{i=1}^n b_i(u) p(\omega_i) = p(u) \qquad \forall p \in \mathcal{P}_s^d$,

(2) $\sum_{i=1}^n |b_i(u)| \leq 2$,

(3) $b_i(u) = 0$ *provided that* $|u - \omega_i| > c_0 \Delta$.

This formula was first introduced by Wang et al. (2012) to learning theory for investigating the approximation property of the kernel-based hypothesis space. For any function set $\mathscr{F}$ on X, we use $absconv\mathscr{F}$ to denote the *absolutely convex hull* of $\mathscr{F}$, which is given by

$$absconv\mathscr{F} = \left\{ \sum_{i=1}^k \lambda_i f_i : k \in \mathbb{N}, f_i \in \mathscr{F} \text{ and } \sum_{i=1}^k |\lambda_i| \leq 1 \right\}.$$

In order to prove our result, we also need the following lemma.

Lemma 13 *Let Q be a probability measure on X and $\mathscr{F}$ be a class of n measurable functions of finite L_Q^2-diameter $diam\mathscr{F}$. Then for every $\epsilon > 0$,*

$$\mathscr{N}\left(absconv\mathscr{F}, \epsilon diam\mathscr{F}, d_{2,Q}\right) \leq \left(e + e(2n+1)\epsilon^2\right)^{2/\epsilon^2}, \tag{24}$$

where $d_{2,Q}$ is the metric induced by the norm $\|\cdot\|_{L_Q^2}$.

This lemma can be proved following the same idea of Lemma 2.6.11 in Van der Vaart and Wellner (1996). Now we can concentrate our efforts on deriving our conclusion in Theorem 11.

Proof [Proof of Theorem 11]. A set is called $\Delta-$*separated* if the distance between any two elements of the set is larger than Δ. We take $\{\Delta_n\}_{n=1}^\infty$ to be a positive sequence decreasing to 0 with Δ_n explicitly given later. Let the set $X_n = \left\{v_1, v_2, \cdots, v_{|X_n|}\right\}$ be an increasing family of finite sets, where $|X_n|$ denotes the cardinality of X_n. For every n, X_n is a maximal Δ_n-separated set in X with respect to inclusion, i.e., each X_n is

Δ_n-separated and if $X_n \subset W \subset X$ then W is not Δ_n-separated. Note that, if X_n is a maximal Δ_n-separated set in X, then it is Δ_n-dense in X. Based on $\{X_n\}_{n=1}^{\infty}$, we create a family of sets $\mathscr{A}_n = \{K_v | v \in X_n\}$. Similarly, let $\mathscr{A} = \{K_v | v \in X\}$, then $\mathscr{A}_1 \subset \mathscr{A}_2 \subset \cdots \subset \mathscr{A}$.

We first limit our discussion to the case that $s \geq 2$ is an integer. Motivated by the proof of Proposition 1 in Wang et al. (2012), we will show that for any $t_0 \in X$, the function K_{t_0} can be approximated by a linear combination of $\mathscr{A}_n$ whenever Δ_n is sufficiently small. For given $x \in X$, we let $g_x(t) = K(x,t)$. Then $g_x \in \mathscr{C}^s(X)$. We consider the Taylor expansion of g_x at a fixed point t_0 of degree less than s, which is given by

$$P_x(t) = K(x,t_0) + \sum_{|\alpha|=1}^{s-1} \frac{D^\alpha g_x(t_0)}{\alpha!}(t-t_0)^\alpha.$$

Here $\alpha = (\alpha^1, \cdots, \alpha^d) \in \mathbb{Z}_+^d$ is a multi-index with $|\alpha| = \sum_{j=1}^d |\alpha^j|$, $\alpha! = \prod_{j=1}^d (\alpha^j)!$, and $D^\alpha g_x(t_0)$ denotes the partial derivatives of g_x at the point t_0. Due to Lemma 12, if X satisfies an interior cone condition, we take a constant $\Delta_{X,s} := \frac{R_X}{c_0}$ depending only on X and s. Then for any $\Delta-$dense set given by $\{\omega_1, \cdots, \omega_n\}$ with $\Delta \leq \Delta_{X,s}$, we have

$$P_x(t_0) = \sum_{i \in I(t_0)} b_i(t_0) P_x(\omega_i),$$

where $\sum_{i\in I(t_0)} |b_i(t_0)| \leq 2$ and $I(t_0) = \{i \in \{1,\cdots,n\} : |\omega_i - t_0| \leq c_0\Delta\}$. Note that $K(x,t_0) = P_x(t_0)$ and $\max_{i\in I(t_0)} |K(x,\omega_i) - P_x(\omega_i)| \leq c_0^s \|K\|_{\mathscr{C}^s} \Delta^s$. It follows that

$$\begin{aligned} \left| K(x,t_0) - \sum_{i\in I(t_0)} b_i(t_0) K(x,\omega_i) \right| &= \left| \sum_{i\in I(t_0)} b_i(t_0) \left(P_x(\omega_i) - K(x,\omega_i)\right) \right| \\ &\leq 2c_0^s \|K\|_{\mathscr{C}^s} \Delta^s. \end{aligned} \tag{25}$$

It is important that the right hand side of the above inequality is independent of x or t_0.

For any function f belonging to $\mathscr{B}_1$, there exist $k \in \mathbb{N}$ and $\{u_i\}_{i=1}^k \subset X$, such that $f(x) = \sum_{i=1}^k \mu_i K(x,u_i)$ with $\sum_{i=1}^k |\mu_i| \leq 1$. Recall that $X_n = \left\{v_1, v_2, \cdots, v_{|X_n|}\right\}$ is Δ_n-dense in X. If $\Delta_n \leq \Delta_{X,s}$, we set

$$f_n(x) = \sum_{i=1}^k \mu_i \sum_{j\in I_n(u_i)} b_j(u_i) K(x,v_j),$$

where the index set $I_n(u)$ is defined for any $n \in \mathbb{N}$ and $u \in X$ as

$$I_n(u) = \{j \in \{1,\cdots,|X_n|\} : |v_j - u| \leq c_0\Delta_n\}.$$

Hence, we have

$$\|f - f_n\|_\infty = \sup_{x\in X} \left| \sum_{i=1}^k \mu_i \left(K(x,u_i) - \sum_{j\in I_n(u_i)} b_j(u_i) K(x,v_j) \right) \right| \leq 2c_0^s \|K\|_{\mathscr{C}^s} \Delta_n^s, \tag{26}$$

where the last inequality is from (25). Obviously, we can rewrite f_n as $f_n(x) = \sum_{j=1}^{|X_n|} \nu_j K(x, v_j)$ where

$$\sum_{j=1}^{|X_n|} |\nu_j| \leq \sum_{i=1}^{k} |\mu_i| \sum_{j \in I_n(u_i)} |b_j(u_i)| \leq 2.$$

Therefore, we have $f_n \in 2absconv\mathscr{A}_n$.

For a compact subset X of $\mathbb{R}^d$, there exists a constant $c_X > 0$ depending only on X, such that

$$\mathscr{N}(X, \epsilon, \tilde{d}_2) \leq c_X \left(\frac{1}{\epsilon}\right)^d, \qquad \forall\ 0 < \epsilon \leq 1,$$

where $\tilde{d}_2$ denotes the metric induced by the standard Euclidean norm on $\mathbb{R}^d$ (see Theorem 5.3 in Cucker and Zhou (2007)). Now we let $\Delta_n = 2c_X^{\frac{1}{d}} n^{-\frac{1}{d}}$. Recall that $|X_n|$ is the maximal cardinality of an Δ_n-separated set in X. Thus we have $|X_n| \leq \mathscr{N}(X, \Delta_n/2, \tilde{d}_2) \leq n$.

For any given $\epsilon > 0$, we choose $N := N(\epsilon)$ be to the smallest integer which is larger than $C'_{X,K} \left(\frac{1}{\epsilon}\right)^{\frac{d}{s}}$ with $C'_{X,K} = 2^{(s+2)d/s} c_0^d \|K\|_{\mathscr{C}^s}^{d/s} c_X$. Then the set $2absconv\mathscr{A}_N$ is $\frac{\epsilon}{2}-$dense in $\mathscr{B}_1$ due to (26). Recall that for any probability measure Q on X, $d_{2,Q}$ denotes the metric induced by the norm $\|\cdot\|_{L^2_Q}$. Therefore, we have

$$\mathscr{N}(\mathscr{B}_1, \epsilon, d_{2,Q}) \leq \mathscr{N}(2absconv\mathscr{A}_N, \epsilon/2, d_{2,Q}) = \mathscr{N}(absconv\mathscr{A}_N, \epsilon/4, d_{2,Q}).$$

To bound the latter, let $N_1 := N_1(\epsilon)$ be the integer part of $\epsilon^{-\frac{2d}{d+2s}}$. We choose a sufficiently small $\epsilon > 0$ satisfying

$$0 < \epsilon \leq \min\left\{ (C'_{X,K})^{s(d+2s)/d^2}, \left(2^{-1} c_X^{-1/d} \Delta_{X,s}\right)^{s+\frac{d}{2}}, \left(\frac{1}{2}\right)^{(d+2s)/2d} \right\} := \epsilon_{X,K}, \tag{27}$$

then $X_{N_1} \subset X_N$ and $\Delta_{N_1} \leq \Delta_{X,s}$.

For any $g \in absconv\mathscr{A}_N$, there exists $\{\nu_i\}_{i=1}^{|X_N|} \subset \mathbb{R}^{|X_N|}$ with $\sum_{i=1}^{|X_N|} |\nu_i| \leq 1$, such that

$$g(x) = \sum_{i=1}^{|X_N|} \nu_i K(x, v_i) := g_1(x) + g_2(x),$$

where

$$\begin{aligned} g_1(x) &= \sum_{i=1}^{|X_{N_1}|} \nu_i K(x, v_i) + \sum_{i=|X_{N_1}|+1}^{|X_N|} \nu_i \left(\sum_{j \in I_{N_1}(v_i)} b_j(v_i) K(x, v_j) \right), \\ g_2(x) &= \sum_{i=|X_{N_1}|+1}^{|X_N|} \nu_i \left(K(x, v_i) - \sum_{j \in I_{N_1}(v_i)} b_j(v_i) K(x, v_j) \right). \end{aligned}$$

From the expression of g_1, we see that it is a linear combination of $\{K(x, v_i) | v_i \in X_{N_1}\}$. Similarly, one can check that the summation of the absolute value of the combinational

coefficients is still bounded by 2. Hence, we have $g_1 \in 2absconv\mathscr{A}_{N_1}$ and $g_2 \in absconv\mathscr{K}_{N_1,N}$ with

$$\mathscr{K}_{N_1,N} = \left\{ K(x,v_i) - \sum_{j\in I_{N_1}(v_i)} b_j(v_i)K(x,v_j) \right\}_{i=|X_{N_1}|+1}^{|X_N|}.$$

Therefore,

$$absconv\mathscr{A}_N \subset 2absconv\mathscr{A}_{N_1} + absconv\mathscr{K}_{N_1,N}.$$

And it follows that

$$\begin{aligned}&\mathscr{N}\left(absconv\mathscr{A}_N, \epsilon/4, d_{2,Q}\right)\\ &\quad\le \mathscr{N}\left(absconv\mathscr{A}_{N_1}, \epsilon/16, d_{2,Q}\right)\cdot \mathscr{N}\left(absconv\mathscr{K}_{N_1,N}, \epsilon/8, d_{2,Q}\right).\end{aligned} \tag{28}$$

To estimate the first term, we choose some suitable $\epsilon_1 > 0$ and $N_2 := N_2(\epsilon_1)$, such that the function set $\{f_j\}_{j=1}^{N_2}$ is the maximal ϵ_1-separated set in $absconv\mathscr{A}_{N_1}$ with respect to the distance $d_{2,Q}$. Hence, for any $f \in absconv\mathscr{A}_{N_1}$, f must belong to one of the N_2 balls of radius ϵ_1 centered at f_j. We denote these balls by $\mathscr{B}(f_j,\epsilon_1), j = 1,\cdots,N_2$. Moreover, as K is an admissible kernel, any function in $absconv\mathscr{A}_{N_1}$ has an unique expression as $f(x) = \sum_{i=1}^{|X_{N_1}|}\nu_i^f K(x,v_i)$ with $\sum_{i=1}^{|X_{N_1}|}|\nu_i^f| \le 1$. We define a mapping by $\Phi(f) = (\nu_1^f,\cdots,\nu_{|X_{N_1}|}^f) \in \mathbb{R}^{|X_{N_1}|}$. Under this mapping, the image of $\mathscr{B}(f_j,\epsilon_1)$ is given by

$$\begin{aligned}Im(\mathscr{B}(f_j,\epsilon_1)) &= \left\{ \{\nu_i\}_{i=1}^{|X_{N_1}|} \,\middle|\, \sum_{i=1}^{|X_{N_1}|}|\nu_i| \le 1 \text{ and } \sqrt{\sum_{i=1}^{|X_{N_1}|}(\nu_i - \nu_i^{f_j})^2} \le \epsilon_1 \right\}\\ &\subset \mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}(\nu^{f_j},\epsilon_1),\end{aligned} \tag{29}$$

where $\mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}(\nu^{f_j},\epsilon_1)$ denotes the ball in $\mathbb{R}^{|X_{N_1}|}$ of radius ϵ_1 centered at $\nu^{f_j} = (\nu_1^{f_j},\cdots,\nu_{|X_{N_1}|}^{f_j})$. Furthermore, for $\forall f,g \in absconv\mathscr{A}_{N_1}$, there holds

$$\|f-g\|_{L_Q^2} \le \kappa\sqrt{\sum_{i=1}^{|X_{N_1}|}(\nu_i^f - \nu_i^g)^2}, \tag{30}$$

where $\kappa = \|K\|_{\mathscr{C}(X\times X)}$.

Next for any $\epsilon_2 > 0$, from (30) and (29), we obtain

$$\begin{aligned}\mathscr{N}(\mathscr{B}(f_j,\epsilon_1),\epsilon_2,d_{2,Q}) &\le \mathscr{N}(Im(\mathscr{B}(f_j,\epsilon_1)),\epsilon_2/\kappa,\tilde{d}_2)\\ &\le \mathscr{N}(\mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}(\nu^{f_j},\epsilon_1),\epsilon_2/\kappa,\tilde{d}_2)\\ &= \mathscr{N}(\epsilon_1\mathscr{B}_{\mathbb{R}^{|X_{N_1}|}},\epsilon_2/\kappa,\tilde{d}_2),\end{aligned} \tag{31}$$

where $\mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}$ denotes the unit ball in $\mathbb{R}^{|X_{N_1}|}$. Recall that N_2 is the cardinality of the maximal ϵ_1-separated set in $absconv\mathscr{A}_{N_1}$. Then $N_2 \le \mathscr{N}\left(absconv\mathscr{A}_{N_1},\epsilon_1/2,d_{2,Q}\right)$. Hence,

for any positive ϵ_1 and ϵ_2, there holds

$$\begin{aligned}
&\mathscr{N}\left(absconv\mathscr{A}_{N_1}, \epsilon_2, d_{2,Q}\right) \\
&\leq \sum_{j=1}^{N_2} \mathscr{N}(\mathscr{B}(f_j, \epsilon_1), \epsilon_2, d_{2,Q}) \\
&\leq \mathscr{N}(\epsilon_1 \mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}, \epsilon_2/\kappa, \tilde{d}_2) \cdot \mathscr{N}\left(absconv\mathscr{A}_{N_1}, \epsilon_1/2, d_{2,Q}\right), \qquad (32)
\end{aligned}$$

where the second inequality is from (31).

Now we set $\epsilon_2 = \epsilon/16$ and $\epsilon_1 = 1/(16\kappa\sqrt{N_1})$, then by equation (32), there holds

$$\begin{aligned}
&\mathscr{N}\left(absconv\mathscr{A}_{N_1}, \epsilon/16, d_{2,Q}\right) \\
&\quad \leq \mathscr{N}(\mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}, \epsilon\sqrt{N_1}, \tilde{d}_2) \cdot \mathscr{N}\left(absconv\mathscr{A}_{N_1}, 1/(32\kappa\sqrt{N_1}), d_{2,Q}\right).
\end{aligned}$$

And from equation (28), we further find that

$$\begin{aligned}
&\mathscr{N}\left(absconv\mathscr{A}_N, \epsilon/4, d_{2,Q}\right) \leq \mathscr{N}(\mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}, \epsilon\sqrt{N_1}, \tilde{d}_2) \\
&\quad \cdot \mathscr{N}\left(absconv\mathscr{A}_{N_1}, 1/(32\kappa\sqrt{N_1}), d_{2,Q}\right) \cdot \mathscr{N}\left(absconv\mathscr{K}_{N_1,N}, \epsilon/8, d_{2,Q}\right). \qquad (33)
\end{aligned}$$

It is noticed that $\mathbb{R}^{|X_{N_1}|}$ is a finite dimensional space and $|X_{N_1}| \leq N_1$, then

$$\mathscr{N}(\mathscr{B}_{\mathbb{R}^{|X_{N_1}|}}, \epsilon\sqrt{N_1}, \tilde{d}_2) \leq (1 + 2N_1^{-1/2}\epsilon^{-1})^{|X_{N_1}|} \leq 3^{N_1}(N_1)^{-N_1/2}\epsilon^{-N_1}. \qquad (34)$$

Next, we use Lemma 13 to estimate the rest two terms. For the function set $\mathscr{A}_{N_1}$, it is easy to check that $diam\mathscr{A}_{N_1} \leq 2\kappa$. Then by (24), there holds

$$\mathscr{N}\left(absconv\mathscr{A}_{N_1}, 1/(32\kappa\sqrt{N_1}), d_{2,Q}\right) \leq \left(e + \frac{e}{1368\kappa^2}\right)^{8192\kappa^2 N_1}. \qquad (35)$$

Recall that $\Delta_{N_1} \leq \Delta_{X,s}$, then we have $Diam\mathscr{K}_{N_1,N} \leq 4c_0^s\|K\|_{\mathscr{C}^s}\Delta_{N_1}^s$ and the cardinality of $\mathscr{K}_{N_1,N}$ satisfies that $|\mathscr{K}_{N_1,N}| = |X_N| - |X_{N_1}| \leq N - 1$. Also by (24), there holds

$$\begin{aligned}
\mathscr{N}\left(absconv\mathscr{K}_{N_1,N}, \epsilon/8, d_{2,Q}\right) &= \mathscr{N}\left(absconv\mathscr{K}_{N_1,N}, \frac{\epsilon}{32c_0^s\|K\|_{\mathscr{C}^s}\Delta_{N_1}^s} \cdot 4c_0^s\|K\|_{\mathscr{C}^s}\Delta_{N_1}^s, L_Q^2\right) \\
&\leq \left(e + e(2N-1)\frac{\epsilon^2}{1024c_0^{2s}\|K\|_{\mathscr{C}^s}^2\Delta_{N_1}^{2s}}\right)^{\frac{2048c_0^{2s}\|K\|_{\mathscr{C}^s}^2\Delta_{N_1}^{2s}}{\epsilon^2}}.
\end{aligned}$$

Due to the choice of N_1, we have $\epsilon^{-\frac{2d}{d+2s}} - 1 < N_1 \leq \epsilon^{-\frac{2d}{d+2s}}$. Combining the restriction (27) for ϵ, we have $\Delta_{N_1}^{2s} = 2^{2s}c_X^{\frac{2s}{d}}(N_1)^{-\frac{2s}{d}} \leq 2^{4s}c_X^{\frac{2s}{d}}\epsilon^{\frac{4s}{d+2s}}$. It follows that $\Delta_{N_1}^{2s}\epsilon^{-2} \leq 2^{4s}c_X^{\frac{2s}{d}}\epsilon^{-\frac{2d}{d+2s}}$. On the other hand, one can bound $\epsilon^2\Delta_{N_1}^{-2s}$ by $2^{-2s}c_X^{-\frac{2s}{d}}\epsilon^{\frac{2d}{d+2s}}$. Recalling that $N \leq C'_{X,K}\left(\frac{1}{\epsilon}\right)^{\frac{d}{s}} + 1$, we then have

$$\begin{aligned}
&\mathscr{N}\left(absconv\mathscr{K}_{N_1,N}, \epsilon/8, d_{2,Q}\right) \\
&\leq \left(e + \frac{eC'_{X,s}\left(\frac{1}{\epsilon}\right)^{\frac{d^2}{s(d+2s)}}}{2^{2s+8}c_X^{2s/d}c_0^{2s}\|K\|_{\mathscr{C}^s}^2}\right)^{2^{4s+11}c_X^{2s/d}c_0^{2s}\|K\|_{\mathscr{C}^s}^2\epsilon^{-\frac{2d}{d+2s}}}. \qquad (36)
\end{aligned}$$

When $0 < \epsilon \leq \epsilon_{X,K}$, substituting the bounds (34), (35) and (36) into the inequality (33), we obtain

$$\log_2 \mathscr{N}(\mathscr{B}_1, \epsilon, d_{2,Q}) \leq C''_{X,s} \epsilon^{-\frac{2d}{d+2s}} \log_2 \left(\frac{1}{\epsilon}\right),$$

where

$$C''_{X,K} = 2 + d^{-1} + 8192\kappa^2 \log_2 \left(e + \frac{e}{1368\kappa^2}\right) + 2^{4s+11} c_X^{2s/d} d s^{-1} c_0^{2s} \|K\|^2_{\mathscr{C}^s} \log_2 \left(e + \frac{eC'_{X,s}}{2^{2s+8} c_X^{2s/d} c_0^{2s} \|K\|^2_{\mathscr{C}^s}}\right).$$

Note that $\mathscr{B}_1$ can be considered as a subset of $\mathscr{C}(X)$. We use d_∞ to denote the metric induced by $\|\cdot\|_\infty$. Then for $\epsilon_{X,s} < \epsilon \leq 1$, we find that

$$\log_2 \mathscr{N}(\mathscr{B}_1, \epsilon, d_{2,Q}) \leq \log_2 \mathscr{N}(\mathscr{B}_1, \epsilon_{X,s}/\kappa, d_\infty) \leq \log_2 \mathscr{N}(\mathscr{B}_1, \epsilon_{X,s}/\kappa, d_\infty) \epsilon^{-\frac{2d}{d+2s}} \log_2 \left(\frac{2}{\epsilon}\right).$$

We take $C_{X,K} = \max\{C''_{X,K}, \log_2 \mathscr{N}(\mathscr{B}_1, \epsilon_{X,s}/\kappa, d_\infty)\}$, then

$$\log_2 \mathscr{N}(\mathscr{B}_1, \epsilon, d_{2,Q}) \leq C_{X,K} \epsilon^{-\frac{2d}{d+2s}} \log_2 \left(\frac{2}{\epsilon}\right), \qquad \forall\ 0 < \epsilon \leq 1.$$

Notice that the above bound is independent of the distribution Q. Then for any sample $\mathbf{x} = \{x_i\}_{i=1}^\ell \in X^\ell$ with $\ell \in \mathbb{N}$, the above estimate holds true for $Q = \frac{1}{\ell}\sum_{i=1}^\ell \delta_{x_i}$. Consequently, we prove our conclusion if s is an integer. When s is not an integer, one can check that the proof is also valid if we replace s by its integral part $\lfloor s \rfloor$. Thus we complete the proof of Theorem 11. ■

4. Convergence analysis

In this section, we investigate the convergence behavior of ℓ_q-kernel regression based on concentration techniques involving ℓ_2-empirical covering numbers. Our error analysis presented in this part is under the capacity assumption (20). For an admissible kernel $K \in \mathscr{C}^s$ with $s > 0$, recall the previous result obtained in Shi et al. (2011) for $0 < s < 2$. Then under the assumption of Theorem 11, one can check that the assumption (20) can be satisfied with $0 < p < 2$. Concretely, when $0 < s < 2$, $p = \frac{2d}{d+2\min\{1,s\}}$; when $s \geq 2$, p can be chosen to be any constant satisfying $p > \frac{2d}{d+2\lfloor s \rfloor}$.

4.1. Deriving the estimator for the total error

Recall the quantity $\mathscr{S}_1$ defined for an estimator $\hat{f}$ in Proposition 6, we can rewrite it as

$$\mathscr{S}_1 = \mathcal{S}_{\mathbf{z}}(\hat{f}) - \mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma}),$$

where the quantity $\mathcal{S}_{\mathbf{z}}$ is defined for $f \in \mathscr{C}(X)$ by

$$\mathcal{S}_{\mathbf{z}}(f) = \{\mathscr{E}(f) - \mathscr{E}(f_\rho)\} - \{\mathscr{E}_{\mathbf{z}}(f) - \mathscr{E}_{\mathbf{z}}(f_\rho)\}.$$

We use the following proposition to estimate $\mathscr{S}_1$. Recall that $\mathscr{B}_R$ is defined as (19).

Proposition 14 *If $\mathscr{B}_1$ satisfies the capacity condition (20) with $0 < p < 2$, then for any $R \geq 1$ and $0 < \delta < 1$, with confidence $1 - \delta$, there hold*

$$\begin{aligned}\mathcal{S}_{\mathbf{z}}(\pi_M(f)) \leq \frac{1}{2}\left\{\mathscr{E}(\pi_M(f)) - \mathscr{E}(f_\rho)\right\} + \frac{176M^2\log(1/\delta)}{m} \\ +\tilde{c}_{K,p}M^2 m^{-\frac{2}{2+p}}R^{\frac{2p}{2+p}}, \qquad \forall f \in \mathscr{B}_R,\end{aligned} \tag{37}$$

and

$$\begin{aligned}\mathcal{S}_{\mathbf{z}}(f) \leq \frac{1}{2}\left\{\mathscr{E}(f) - \mathscr{E}(f_\rho)\right\} + \frac{20(3M+\kappa)^2R^2\log(1/\delta)}{m} \\ +\tilde{c}_{K,p}(3M+\kappa)^2 m^{-\frac{2}{2+p}}R^2, \qquad \forall f \in \mathscr{B}_R.\end{aligned} \tag{38}$$

Here $\tilde{c}_{K,p} > 0$ is a constant depending only on p and the kernel function K and $\kappa = \|K\|_{\mathscr{C}(X\times X)}$. The above bounds also hold true for $-\mathcal{S}_{\mathbf{z}}(\pi_M(f))$ and $-\mathcal{S}_{\mathbf{z}}(f)$.

We shall prove Proposition 14 in the Appendix by using the entropy integral based on ℓ_2-empirical covering numbers. Now we can derive the estimator for the total error. For $R \geq 1$ and $0 < q \leq 1$, denote

$$\mathcal{W}_q(R) = \left\{\mathbf{z} \in Z^m : \ \|\mathbf{c}_q^{\mathbf{z}}\|_q^q \leq R\right\}. \tag{39}$$

Proposition 15 *Assume that approximation assumption (14) with $0 < \beta \leq 1$ and capacity condition (20) with $0 < p < 2$ are valid, and the estimator $\hat{f}_q$ satisfies Assumption 1 with $0 < q < 1$. If $0 < \gamma \leq 1$, $0 < \delta < 1$ and $R \geq 1$, then there is a subset $V_R^{(q)}$ of Z^m with measure at most δ such that*

$$\begin{aligned}&\mathscr{E}(\pi_M(\hat{f}_q)) - \mathscr{E}(f_\rho) + \gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \\ &\leq 2\tilde{c}_{K,p}M^2 m^{-\frac{2}{2+p}}R^{\frac{2p}{q(2+p)}} + C_1\log^3(18/\delta)\max\left\{\gamma^{\beta-2}m^{-2}, \gamma^{\beta-1}m^{-\frac{2}{2+p}}\right\} \\ &\quad +(3\sqrt{c_\beta} + 6c_\beta)\gamma^\beta + 2\gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q, \qquad \forall \mathbf{z} \in \mathcal{W}_q(R)\backslash V_R^{(q)}.\end{aligned} \tag{40}$$

Moreover, when $q = 1$, there are subsets $V_R^{(1)}$ and $\widetilde{V}_R^{(1)}$ of Z^m with each measure at most δ such that

$$\begin{aligned}&\mathscr{E}(\pi_M(\hat{f}_1)) - \mathscr{E}(f_\rho) + \gamma\|\mathbf{c}_1^{\mathbf{z}}\|_1 \\ &\leq 2\tilde{c}_{K,p}M^2 m^{-\frac{2}{2+p}}R^{\frac{2p}{2+p}} + C_1\log^3(18/\delta)\max\left\{\gamma^{\beta-2}m^{-2}, \gamma^{\beta-1}m^{-\frac{2}{2+p}}\right\} \\ &\quad +(3\sqrt{c_\beta} + 6c_\beta)\gamma^\beta, \qquad \forall \mathbf{z} \in \mathcal{W}_1(R)\backslash V_R^{(1)},\end{aligned} \tag{41}$$

and

$$\begin{aligned}\mathscr{E}(\hat{f}_1) - \mathscr{E}(f_\rho) + \gamma\|\mathbf{c}_1^{\mathbf{z}}\|_1 \leq 2(20 + \tilde{c}_{K,p})(3M+\kappa)^2\log(3/\delta)m^{-\frac{2}{2+p}}R^2 \\ +\widetilde{C}_1\log^3(18/\delta)\max\left\{\gamma^{\beta-2}m^{-2}, \gamma^{\beta-1}m^{-\frac{2}{2+p}}, \gamma^\beta\right\}, \forall \mathbf{z} \in \mathcal{W}_1(R)\backslash\widetilde{V}_R^{(1)}.\end{aligned} \tag{42}$$

Here C_1 and $\widetilde{C}_1$ are positive constants depending on $\kappa, \tilde{c}_{K,p}$, M and c_β.

Proof Let $\hat{f}$ be the estimator under consideration. We estimate $\mathscr{S}_1$ by bounding $\mathcal{S}_{\mathbf{z}}(\hat{f})$ and $-\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma})$ respectively. Due to Proposition 14, the quantities $-\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma})$ and $\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma})$ have the same upper bound. We thus turn to estimate $\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma})$. For $(\gamma,\delta)\in(0,1)^2$, let

$$R(\gamma,\delta):=\frac{2\kappa\sqrt{\mathcal{D}(\gamma)}\log(6/\delta)}{\gamma m}+\sqrt{\frac{2\mathcal{D}(\gamma)\log(6/\delta)}{\gamma m}}+\sqrt{\frac{\mathcal{D}(\gamma)}{\gamma}}.$$

From the proof of Proposition 5 in Guo and Shi (2012), there exist two subsets of Z^m denoted by U_1 and U_2 with $\rho(U_i)\leq\delta/3, i=1,2$, such that

$$f_{\mathbf{z},\gamma}\in\mathscr{B}_{R(\gamma,\delta)},\qquad\forall\ \mathbf{z}\in Z^m\backslash U_1, \tag{43}$$

and

$$\mathscr{E}(f_{\mathbf{z},\gamma})-\mathscr{E}(f_\rho)\leq 8\kappa^2\left(2\kappa^2+1\right)\log^2(6/\delta)\left\{\frac{\mathcal{D}(\gamma)}{\gamma^2m^2}+\frac{\mathcal{D}(\gamma)}{\gamma m}\right\}+2\mathcal{D}(\gamma),\forall\ \mathbf{z}\in Z^m\backslash U_2. \tag{44}$$

We use inequality (38) to bound $\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma})$. Now (43) and (44) are both valid for $f_{\mathbf{z},\gamma}$ with $\mathbf{z}\in Z^m\backslash(U_1\cup U_2)$. Then there exists a subset U_3 of Z^m with measure at most $\delta/3$ such that

$$\begin{aligned}\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma})\leq{}&4\kappa^2\left(2\kappa^2+1\right)\log^2(6/\delta)\left\{\frac{\mathcal{D}(\gamma)}{\gamma^2m^2}+\frac{\mathcal{D}(\gamma)}{\gamma m}\right\}\\&+\mathcal{D}(\gamma)+\frac{20(3M+\kappa)^2(R(\gamma,\delta))^2\log(3/\delta)}{m}\\&+\tilde{c}_{K,p}(3M+\kappa)^2m^{-\frac{2}{2+p}}(R(\gamma,\delta))^2,\forall\mathbf{z}\in Z^m\backslash(U_1\cup U_2\cup U_3).\end{aligned}$$

Note that $\rho(U_1\cup U_2\cup U_3)\leq\delta$ and

$$(R(\gamma,\delta))^2\leq(8\kappa^2+4)\log^2(6/\delta)\left(\frac{\mathcal{D}(\gamma)}{\gamma^2m^2}+\frac{\mathcal{D}(\gamma)}{\gamma m}+\frac{\mathcal{D}(\gamma)}{\gamma}\right).$$

Therefore, with confidence $1-\delta$, there holds

$$\begin{aligned}\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma})\leq{}&C_\kappa\log^2(6/\delta)\max\left\{\frac{\mathcal{D}(\gamma)}{\gamma^2m^2},\frac{\mathcal{D}(\gamma)}{\gamma m}\right\}+\mathcal{D}(\gamma)\\&+(20+\tilde{c}_{K,p})C_{\kappa,M}\log^3(6/\delta)m^{-\frac{2}{2+p}}\max\left\{\frac{\mathcal{D}(\gamma)}{\gamma^2m^2},\frac{\mathcal{D}(\gamma)}{\gamma}\right\},\end{aligned} \tag{45}$$

where $C_\kappa=8\kappa^2\left(2\kappa^2+1\right)$ and $C_{\kappa,M}=3(3M+\kappa)^2(8\kappa^2+4)$.

Next, for $\hat{f}=\hat{f}_1$ or $\pi_M(\hat{f}_q)$ with $0<q\leq1$, we consider bounding the quantity $\mathcal{S}_{\mathbf{z}}(\hat{f})$ with $\mathbf{z}\in\mathcal{W}(R)$. It is easy if we notice that $\|\mathbf{c}_q^{\mathbf{z}}\|_q^q\leq R$ implies the corresponding estimator $\hat{f}_q\in\mathscr{B}_{R^{1/q}}$. Thus we can bound $\mathcal{S}_{\mathbf{z}}(\hat{f})$ by directly applying Proposition 14.

When $\hat{f}=\pi_M(\hat{f}_q)$ with $0<q<1$, combining the bounds on $\mathscr{S}_1+\mathscr{S}_2$ given in Lemma 7, we are able to derive the estimator for the total error bound. Due to (37), there exist subsets $V_1^{(q)}$ with measure at most $\delta/3$ such that

$$\begin{aligned}\mathcal{S}_{\mathbf{z}}(\pi_M(\hat{f}_q))\leq{}&\frac{1}{2}\left\{\mathscr{E}(\pi_M(\hat{f}_q))-\mathscr{E}(f_\rho)\right\}+\frac{176M^2\log(3/\delta)}{m}\\&+\tilde{c}_{K,p}M^2m^{-\frac{2}{2+p}}R^{\frac{2p}{q(2+p)}},\qquad\forall\mathbf{z}\in\mathcal{W}_q(R)\backslash V_1^{(q)}.\end{aligned}$$

Similarly, from (45) and Lemma 7, we can find subsets V_2 and V_3 such that

$$-\mathcal{S}_{\mathbf{z}}(f_{\mathbf{z},\gamma}) \leq (C_\kappa + C_{K,p}) \log^3(18/\delta) \max\left\{\frac{\mathcal{D}(\gamma)}{\gamma^2 m^2}, \frac{\mathcal{D}(\gamma)}{\gamma m^{2/(2+p)}}\right\} + \mathcal{D}(\gamma), \forall \mathbf{z} \in Z^m \backslash V_2,$$

and

$$\begin{aligned}\mathscr{S}_2 + \mathscr{S}_3 \leq\ & 16\kappa^2 \left(2\kappa^2+1\right) \log^2(12/\delta) \max\left\{\frac{\mathcal{D}(\gamma)}{\gamma^2 m^2}, \frac{\mathcal{D}(\gamma)}{\gamma m}\right\} \\ & + \frac{(2\kappa+1)\sqrt{\mathcal{D}(\gamma)} \log(12/\delta)}{m} + \frac{3}{2}\sqrt{\gamma \mathcal{D}(\gamma)} + 2\mathcal{D}(\gamma), \forall \mathbf{z} \in Z^m \backslash V_3,\end{aligned}$$

where $C_{K,p} = (20 + \tilde{c}_{K,p}) C_{\kappa,M}$ and $\rho(V_i) \leq \delta/3$ for $i = 2, 3$.

Let $V_R^{(q)} = V_1^{(q)} \cup V_2 \cup V_3$, then $\rho(V_R^{(q)}) \leq \delta$. Recall the error decomposition formula (16). We combine the above three bounds and obtain

$$\begin{aligned}&\mathscr{E}(\pi_M(\hat{f}_q)) - \mathscr{E}(f_\rho) + \gamma \|\mathbf{c}_q^{\mathbf{z}}\|_q^q \\ &\leq \frac{1}{2}\left\{\mathscr{E}(\pi_M(\hat{f}_q)) - \mathscr{E}(f_\rho)\right\} + (176M^2 + 16\kappa^2(2\kappa^2+1)) \log^2(12/\delta) \max\left\{\frac{1}{m}, \frac{\mathcal{D}(\gamma)}{\gamma m}\right\} \\ &+ \tilde{c}_{K,p} M^2 m^{-\frac{2}{2+p}} R^{\frac{2p}{q(2+p)}} + (C_\kappa + C_{K,p} + 16\kappa^2\left(2\kappa^2+1\right)) \log^3(18/\delta) \max\left\{\frac{\mathcal{D}(\gamma)}{\gamma^2 m^2}, \frac{\mathcal{D}(\gamma)}{\gamma m^{2/(2+p)}}\right\} \\ &+ \frac{(2\kappa+1)\sqrt{\mathcal{D}(\gamma)} \log(12/\delta)}{m} + \frac{3}{2}\sqrt{\gamma \mathcal{D}(\gamma)} + 3\mathcal{D}(\gamma) + \gamma m^{1-q} \|\mathbf{c}_1^{\mathbf{z}}\|_1^q, \quad \forall \mathbf{z} \in \mathcal{W}_q(R) \backslash V_R^{(q)}.\end{aligned}$$

Due to this inequality, we use Approximation condition (14) and find that the inequality (40) holds true with

$$C_1 = 2((20 + \tilde{c}_{K,p}) C_{\kappa,M} + 176M^2 + 40\kappa^2\left(2\kappa^2+1\right)) c_\beta + (4\kappa+2)\sqrt{c_\beta}.$$

Following the same method, when $\hat{f} = \pi_M(\hat{f}_1)$, we can prove inequality (41) by using the error decomposition formula (17) .

We next focus on the case $\hat{f} = \hat{f}_1$. Due to inequality (38), for $R \geq 1$, with confidence $1 - \delta/3$, there holds

$$\mathcal{S}_{\mathbf{z}}(\hat{f}_1) \leq \frac{1}{2}\left\{\mathscr{E}(\hat{f}_1) - \mathscr{E}(f_\rho)\right\} + (20 + \tilde{c}_{K,p})(3M+\kappa)^2 \log(3/\delta) m^{-\frac{2}{2+p}} R^2, \forall \mathbf{z} \in \mathcal{W}_1(R).$$

Also by the same analysis above, we obtain inequality (42) with

$$\widetilde{C}_1 = 6((20 + \tilde{c}_{K,p})(3M+\kappa)^2(8\kappa^2+4) + 8\kappa^2\left(2\kappa^2+1\right) + 1) c_\beta + (4\kappa+5)\sqrt{c_\beta}.$$

Thus we complete our proof. ■

4.2. Bounding $\|\mathbf{c}_q^{\mathbf{z}}\|_q$ by iteration

To apply Proposition 15 for error analysis, we need to determine some $R \geq 1$ for the set $\mathcal{W}_q(R)$ given by (39). To this end, we shall apply an iteration technique to obtain a tight

bound for $\|\mathbf{c}_q^{\mathbf{z}}\|_q^q$. This technique can be found in Steinwart and Scovel (2007); Wu et al. (2006). Take the case $q = 1$ for example. Recall that $\mathbf{c}_1^{\mathbf{z}}$ is the global minimizer of the target functional $\mathscr{T}_{\gamma,1}$ defined by (3). Hence $\mathscr{T}_{\gamma,1}(\mathbf{c}_1^{\mathbf{z}}) \leq \mathscr{T}_{\gamma,1}(\mathbf{0})$. This inequality leads to a trivial bound on $\|\mathbf{c}_1^{\mathbf{z}}\|_1$, which is given by

$$\|\mathbf{c}_1^{\mathbf{z}}\|_1 \leq \frac{M^2}{\gamma}, \qquad \forall\ \mathbf{z} \in Z^m. \tag{46}$$

By applying inequality (41) iteratively, we are able to improve the above bound to the order $O(\gamma^{\beta-1})$ with some suitable choice of $\gamma = \gamma(m)$. Similarly, when $0 < q < 1$, a better estimates for $\|\mathbf{c}_q^{\mathbf{z}}\|_q^q$ can be derived based on the inequality (40). The following proposition illustrates the detailed process of iteration.

Proposition 16 *Under the assumptions of Proposition 15, let $(\delta, \tau) \in (0,1)^2$ and $J(\tau, p, q)$ be a constant given by*

$$J(\tau, p, q) = \max\left\{2, \frac{\log \frac{(2-(p+2)\tau)p}{(q-p\tau)(p+2)}}{\log \frac{2p}{q(p+2)}}\right\}. \tag{47}$$

For $q = 1$, take $\gamma = m^{-\tau}$ with $0 < \tau < \frac{2}{2+p}$. Then with confidence $1-\delta$, there holds

$$\|\mathbf{c}_1^{\mathbf{z}}\|_1 \leq C_2 \left(\log(1/\delta) + \log(J(\tau, p, 1))\right)^3 m^{(1-\beta)\tau}. \tag{48}$$

For $0 < q < 1$ satisfying $q > \max\left\{\frac{2p}{p+2}, \frac{p}{p+2\beta}\right\}$, take $\gamma = m^{-\tau}$ with $\frac{1-q}{1-q(1-\beta)} < \tau < \frac{2}{2+p}$. Then with confidence $1-\delta$, there holds

$$\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \leq \widetilde{C}_2 \left(\log J(\tau, p, 1)\right)^{3q} \left(\log(1/\delta) + \log(J(\tau, p, q) + 1)\right)^3 m^{1-q+q(1-\beta)\tau}. \tag{49}$$

Here C_2 and $\widetilde{C}_2$ are positive constants depending only on $\kappa, \tilde{c}_{K,p}$, M and c_β.

Proof Take $0 < \delta < 1$ and $\gamma = m^{-\tau}$ under the restriction $0 < \tau < \frac{2}{p+2}$. For $0 < q \leq 1$, we shall verify the following inequality derived from Proposition 15, that is

$$\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \leq \max\{a_m R^\theta, b_m\}, \qquad \forall \mathbf{z} \in \mathcal{W}_q(R) \setminus V_R^{(q)}, \tag{50}$$

where $V_R^{(q)}$ is a subset of Z^m with measure at most δ, $\theta = \frac{2}{q(p+2)}$, $a_m = 4\tilde{c}_{K,p} M^2 m^{\tau - \frac{2}{2+p}}$ and b_m will be given explicitly in accordance with specific situations. This inequality ensures that

$$\mathcal{W}_q(R) \subset \mathcal{W}_q\left(\max\{a_m R^\theta, b_m\}\right) \cup V_R^{(q)} \tag{51}$$

with $\rho(V_R^{(q)}) \leq \delta$. We then apply the inclusion (51) for a sequence of radiuses $\{R^{(j)}\}_{j\in\mathbb{N}}$ defined by $R^{(0)} = M_\delta m^\tau$ with a suitable chosen $M_\delta > 0$ and

$$R^{(j)} = \max\left\{a_m (R^{(j-1)})^\theta, b_m\right\}, \qquad j \in \mathbb{N}. \tag{52}$$

Then (51) holds for each $R^{(j)}$, which implies $\mathcal{W}_q(R^{(j-1)}) \subseteq \mathcal{W}_q(R^{(j)}) \cup V^{(q)}_{R^{(j-1)}}$ with $\rho(V^{(q)}_{R^{(j-1)}}) \leq \delta$. Applying this inclusion for $j = 1, 2, \ldots, J$ with J to be determined later, we see that

$$\mathcal{W}_q(R^{(0)}) \subseteq \mathcal{W}_q(R^{(1)}) \cup V^{(q)}_{R^{(0)}} \subseteq \cdots \subseteq \mathcal{W}_q(R^{(J)}) \cup \left(\cup_{j=0}^{J-1} V^{(q)}_{R^{(j)}} \right).$$

By the definition of the sequence $\{R^{(j)}\}_j$, we thus have

$$R^{(J)} = \max \left\{ a_m^{1+\theta+\theta^2+\ldots+\theta^{J-1}} (R^{(0)})^{\theta^J}, a_m^{1+\theta+\theta^2+\ldots+\theta^{J-2}} b_m^{\theta^{J-1}}, \ldots, a_m b_m^{\theta}, b_m \right\}. \tag{53}$$

Recall that $R^{(0)} = M_\delta m^\tau$. When $\theta \neq 1$, the first term on the right-hand side of equation (53) can be bounded as

$$\begin{aligned} a_m^{1+\theta+\theta^2+\ldots+\theta^{J-1}} (R^{(0)})^{\theta^J} &= a_m^{(\theta^J-1)/(\theta-1)} (R^{(0)})^{\theta^J} \\ &\leq (4\tilde{c}_{K,p} M^2)^{(\theta^J-1)/(\theta-1)} M_\delta^{\theta^J} m^{\frac{\tau(\theta^{J+1}-1)}{\theta-1} - \frac{2(\theta^J-1)}{(p+2)(\theta-1)}}. \end{aligned}$$

For the remaining terms, we find that

$$\begin{aligned} &\max \left\{ a_m^{1+\theta+\theta^2+\ldots+\theta^{J-2}} b_m^{\theta^{J-1}}, \ldots, a_m b_m^{\theta}, b_m \right\} \\ &= \max \left\{ (a_m)^{(\theta^{J-1}-1)/(\theta-1)} (b_m)^{\theta^{J-1}}, \ldots, a_m b_m^{\theta}, b_m \right\} \\ &= \max \left\{ a_m^{(\theta^{J-1}-1)/(\theta-1)} b_m^{\theta^{J-1}}, b_m \right\}. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} R^{(J)} \leq \max \Bigg\{ & A_{\theta,J} M_\delta^{\theta^J} m^{\frac{\tau(\theta^{J+1}-1)}{\theta-1} - \frac{2(\theta^J-1)}{(p+2)(\theta-1)}}, \\ & A_{\theta,J-1} m^{(\tau - \frac{2}{2+p})(\theta^{J-1}-1)/(\theta-1)} b_m^{\theta^{J-1}}, b_m \Bigg\}, \end{aligned} \tag{54}$$

where $A_{\theta,J}$ is defined as $A_{\theta,J} = (4\tilde{c}_{K,p} M^2)^{(\theta^J-1)/(\theta-1)}$ for any $J \in \mathbb{N}$ and $\theta \neq 1$.

Now we go back to our concrete examples. Due to the above analysis, we need to determine b_m, M_δ and J depending on the concerned cases.

When $q = 1$, then $\theta = \frac{2p}{p+2} < 1$ and we can take $b_m = C_1' \log^3(18/\delta) m^{(1-\beta)\tau}$ with $C_1' = 2C_1 + 6\sqrt{c_\beta} + 12 c_\beta$. One can check that the inequality (50) is valid due to (41). Recall the trivial bound (46) on $\|\mathbf{c}_1^{\mathbf{z}}\|_1$, which implies $\|\mathbf{c}_1^{\mathbf{z}}\|_1 \leq M^2/\gamma = M^2 m^\tau$ for $\forall \mathbf{z} \in Z^m$. We thus take $M_\delta = M^2$ and $R^{(0)} = M^2 m^\tau$. From (54), we further find that

$$R^{(J)} \leq \max \left\{ \left(4\tilde{c}_{K,p} M^2\right)^{\frac{2+p}{2-p}} M^2, C_1' \log^3(18/\delta) \left(4\tilde{c}_{K,p} M^2\right)^{\frac{2+p}{2-p}}, C_1' \log^3(18/\delta) \right\} m^{\theta_1},$$

where the power index θ_1 is given by

$$\begin{aligned} \theta_1 = \max \Bigg\{ & \left(\frac{2p}{p-2} \left(\frac{2p}{p+2} \right)^J - \frac{p+2}{p-2} \right) \tau - \frac{2}{p-2} \left(\left(\frac{2p}{p+2} \right)^J - 1 \right), \\ & (1-\beta)\tau, (1-\beta)\tau + \left(\left(\frac{2p}{p+2} \right)^{J-1} - 1 \right) \left((1-\beta)\tau + \frac{(p+2)\tau - 2}{p-2} \right) \Bigg\}. \end{aligned}$$

Let J be the smallest integer such that

$$\left(\frac{2p}{p-2}\left(\frac{2p}{p+2}\right)^J-\frac{p+2}{p-2}\right)\tau-\frac{2}{p-2}\left(\left(\frac{2p}{p+2}\right)^J-1\right)\leq 0$$

i.e., J is the integer satisfying

$$\max\left\{1,\frac{\log\frac{2-(p+2)\tau}{2-2p\tau}}{\log\frac{2p}{p+2}}\right\}\leq J<\max\left\{2,\frac{\log\frac{(2-(p+2)\tau)p}{(1-p\tau)(p+2)}}{\log\frac{2p}{p+2}}\right\}.$$

Then $\theta_1=(1-\beta)\tau$ and

$$R^{(J)}\leq\left(\left(4\tilde{c}_{K,p}M^2\right)^{\frac{2+p}{2-p}}M^2+C_1'\right)\log^3(18/\delta)m^{(1-\beta)\tau}.$$

Since $Z^m=\mathcal{W}_1(R^{(0)})$ and $\rho\left(\cup_{j=0}^{J-1}V_{R^{(j)}}^{(1)}\right)\leq J\delta$, the measure of the set $\mathcal{W}_1(R^{(J)})$ is at least $1-J\delta$. By scaling $J\delta$ to δ, we derive the bound (48) with

$$C_2=64\left(\left(4\tilde{c}_{K,p}M^2\right)^{\frac{2+p}{2-p}}M^2+C_1'\right).$$

Next we turn to the case $0<q<1$ with $q>\max\left\{\frac{2p}{p+2},\frac{p}{p+2\beta}\right\}$. The estimation in this case is more involved. In order to determine b_m and $R^{(0)}$ in this case, we need to use the result just obtained for $\|\mathbf{c}_1^{\mathbf{z}}\|_1$. In fact, for any $0<\delta<1$ and $0<\tau<\frac{2}{p+2}$, with confidence $1-\delta/2$ there holds

$$\|\mathbf{c}_1^{\mathbf{z}}\|_1\leq C_2\left(\log J(\tau,p,1)+\log(2/\delta)\right)^3m^{(1-\beta)\tau},\tag{55}$$

where $J(\tau,p,1)$ is given by (47) at $q=1$. The inequality (40) asserts that, with confidence $1-\delta/2$, there holds

$$\begin{aligned}\mathscr{E}(\pi_M(\hat{f}_q))-\mathscr{E}(f_\rho)+\gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q&\leq 2\tilde{c}_{K,p}M^2m^{-\frac{2}{2+p}}R^{\frac{2p}{q(2+p)}}\\&+\frac{1}{2}C_1'\log^3(36/\delta)\gamma^\beta+2\gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q,\qquad\forall\mathbf{z}\in\mathcal{W}_q(R).\end{aligned}\tag{56}$$

It is easy to check under the restriction $q>\max\left\{\frac{2p}{p+2},\frac{p}{p+2\beta}\right\}$, there holds $\frac{1-q}{1-q(1-\beta)}<\frac{2}{2+p}$ and $\frac{2p}{q(2+p)}<1$. Then we can take $\gamma=m^{-\tau}$ with $\frac{1-q}{1-q(1-\beta)}<\tau<\frac{2}{2+p}$. Combining (55) and (56), we find that (50) is satisfied with $b_m=B_\delta m^{\tilde{\alpha}}$, where

$$B_\delta=\left(C_1'+4C_2^q\right)\left(\log J(\tau,p,1)\right)^{3q}\log^3(36/\delta)\text{ and }\tilde{\alpha}=1-q+q(1-\beta)\tau.$$

We further define

$$R^{(0)}=M_\delta m^\tau\text{ and }M_\delta=M^2+C_2^q\left(\log J(\tau,p,1)+\log(1/\delta)\right)^{3q}.\tag{57}$$

Recall that the estimator $\hat{f}_q$ satisfies Assumption 1. Then we have

$$\gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q\leq\mathscr{T}_{\gamma,q}(\mathbf{c}_q^{\mathbf{z}})\leq\mathscr{T}_{\gamma,q}(\mathbf{c}_1^{\mathbf{z}})\leq\mathscr{T}_{\gamma,1}(\mathbf{c}_1^{\mathbf{z}})+\gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q\leq\mathscr{T}_{\gamma,1}(\mathbf{0})+\gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q.$$

Due to this inequality and the bound on $\|\mathbf{c}_1^{\mathbf{z}}\|_1$, one can easily check that with confidence $1-\delta$, there holds

$$\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \le M^2 m^\tau + m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q \le M_\delta m^\tau,$$

where the last inequality is due to $\tau > \frac{1-q}{1-q(1-\beta)}$. Hence the measure of the set $\mathcal{W}_q(R^{(0)})$ is at least $1-\delta$. We thus can iteratively define $R^{(j)}$ as (52) with $R^{(0)}$ given by (57). Now we can follow the same fashion to derive our desire bounds.

Since $\theta = \frac{2p}{q(p+2)} < 1$, due to the inequality (54), we have

$$R^{(J)} \le \max\left\{A_{\theta,J}M_\delta^{\theta^J}, A_{\theta,J-1}B_\delta^{\theta^{J-1}}, B_\delta\right\}m^{\theta_2},$$

where θ_2 is given by

$$\max\left\{\left(\frac{2p}{2p-q(p+2)}\left(\frac{2p}{q(p+2)}\right)^J - \frac{q(p+2)}{2p-q(p+2)}\right)\tau - \frac{2q}{2p-q(p+2)}\left(\left(\frac{2p}{q(p+2)}\right)^J - 1\right),\right.$$
$$\left.\tilde{\alpha}, \tilde{\alpha} + \left(\left(\frac{2p}{q(p+2)}\right)^{J-1} - 1\right)\left(\tilde{\alpha} + \frac{q(p+2)}{2p-q(p+2)}\tau - \frac{2q}{2p-q(p+2)}\right)\right\}.$$

We chose J to be the smallest integer such that

$$\max\left\{1, \frac{\log\frac{2q-q(p+2)\tau}{2q-2p\tau}}{\log\frac{2p}{q(p+2)}}\right\} \le J < \max\left\{2, \frac{\log\frac{(2-(p+2)\tau)p}{(q-p\tau)(p+2)}}{\log\frac{2p}{q(p+2)}}\right\}.$$

Under the choice of J, we have $\theta_2 = \tilde{\alpha}$,

$$R^{(J)} \le (M^2 + C_1' + 4C_2^q)\left(4\tilde{c}_{K,p}M^2\right)^{\frac{2+p}{2-p}}(\log J(\tau,p,1))^{3q}\log^3(36/\delta)m^{\tilde{\alpha}}$$

and the measure of the set $\mathcal{W}_1(R^{(J)})$ is at least $1-(J+1)\delta$. We thus can derive the inequality (49) with

$$\widetilde{C}_2 = 64(M^2 + C_1' + 4C_2^q)\left(4\tilde{c}_{K,p}M^2\right)^{\frac{2+p}{2-p}}$$

by scaling $(J+1)\delta$ to δ. ■

4.3. Deriving the convergence rates

In this subsection, we estimate the total error for algorithm (2) and derive the explicit convergence rates. Recall that for $0 < q \le 1$, $\hat{f}_q$ denotes the estimator given by the algorithm (2).

Theorem 17 *Assume that approximation assumption (14) with $0 < \beta \le 1$ and capacity condition (20) with $0 < p < 2$ are valid. Let $(\delta,\tau,q) \in (0,1]^3$ and $J(\tau,p,q)$ be a constant*

given by (47). When $q = 1$, take $\gamma = m^{-\tau}$ with $0 < \tau < \frac{2}{2+p}$. Then with confidence $1-\delta$, there holds

$$\|\hat{f}_1 - f_\rho\|^2_{L^2_{\rho_X}} \leq C_3 \log(6/\delta) \left(\log(2/\delta) + \log(J(\tau, p, 1))\right)^6 m^{-\widetilde{\Theta}}, \tag{58}$$

where

$$\widetilde{\Theta} = \min\left\{\frac{2}{2+p} - 2(1-\beta)\tau, \beta\tau\right\}. \tag{59}$$

Additionally, if an estimator $\hat{f}_q$ satisfies Assumption 1 with $\max\left\{\frac{2p}{p+2}, \frac{p}{p+2\beta}\right\} < q < 1$. Take $\gamma = m^{-\tau}$ with $\frac{1-q}{1-q(1-\beta)} < \tau < \frac{2}{2+p}$. Then with confidence $1-\delta$, there holds

$$\|\pi_M(\hat{f}_q) - f_\rho\|^2_{L^2_{\rho_X}} \leq \widetilde{C}_3 \left(\log(9/\delta) + \log(J(\tau, p, 1)) + \log(J(\tau, p, q) + 1)\right)^3 m^{-(\tau - \tilde{\alpha})}, \tag{60}$$

where $\tilde{\alpha} = 1 - q + q(1-\beta)\tau$. Here C_3 and $\widetilde{C}_3$ are positive constants independent of m or δ.

Proof Recall that for any estimator $\hat{f}$ under consideration, there holds

$$\|\hat{f} - f_\rho\|^2_{L^2_{\rho_X}} = \mathscr{E}(\hat{f}) - \mathscr{E}(f_\rho).$$

Therefore, due to inequality (42) in Proposition 15, there is a subset $\widetilde{V}_R$ of Z^m with measure at most δ such that

$$\begin{aligned}\|\hat{f} - f_\rho\|^2_{L^2_{\rho_X}} &\leq 2(20 + \tilde{c}_{K,p})(3M + \kappa)^2 \log(3/\delta) m^{-\frac{2}{2+p}} R^2 \\ &+\widetilde{C}_1 \log^3(18/\delta) \max\left\{\gamma^{\beta-2} m^{-2}, \gamma^{\beta-1} m^{-\frac{2}{2+p}}, \gamma^\beta\right\}, \forall \mathbf{z} \in \mathcal{W}_1(R) \backslash \widetilde{V}_R.\end{aligned}$$

Now we choose R to be the right-hand side of inequality (48), i.e.,

$$R = C_2 \left(\log(1/\delta) + \log(J(\tau, p, 1))\right)^3 m^{(1-\beta)\tau}.$$

Then the measure of the set $\mathcal{W}_1(R) \setminus \widetilde{V}_R$ is at least $1 - 2\delta$. So with confidence at least $1 - 2\delta$, there holds

$$\|\hat{f}_1 - f_\rho\|^2_{L^2_{\rho_X}} \leq C_3 \max\left\{\log^3(3/\delta), \log(3/\delta) \left(\log(1/\delta) + \log(J(\tau, p, 1))\right)^6\right\} m^{-\widetilde{\Theta}},$$

where $\widetilde{\Theta}$ is given by (59) and

$$C_3 = 54(20 + \tilde{c}_{K,p})(3M + \kappa)^2 C_2^2 + 27\widetilde{C}_1.$$

Then we obtain the error bound (58) by scaling 2δ to δ.

Next we prove the error bound (60). Recalling the total error bound (40), for some $R \geq 1$, there is a subset $V_R^{(q)}$ with measure at most δ such that

$$\begin{aligned}&\|\pi_M(\hat{f}_q) - f_\rho\|^2_{L^2_{\rho_X}} \leq 2\tilde{c}_{K,p} M^2 m^{-\frac{2}{2+p}} R^{\frac{2p}{q(2+p)}} \\ &+\widetilde{C}_1' \log^3(18/\delta) \max\left\{\gamma^{\beta-2} m^{-2}, \gamma^{\beta-1} m^{-\frac{2}{2+p}}, \gamma^\beta\right\} + 2\gamma m^{1-q} \|\mathbf{c}_1^{\mathbf{z}}\|_1^q, \ \forall \mathbf{z} \in \mathcal{W}_q(R) \backslash V_R^{(q)},\end{aligned}$$

where $\widetilde{C}_1' = C_1 + 3\sqrt{c_\beta} + 6c_\beta$. Using the same argument, the inequality (48) ensures the existence of a subset V_1' with measure at most δ such that

$$\|\mathbf{c}_1^{\mathbf{z}}\|_1 \leq C_2 \left(\log(1/\delta) + \log(J(\tau,p,1))\right)^3 m^{(1-\beta)\tau}, \qquad \forall \mathbf{z} \in Z^m \backslash V_1'.$$

If we take R to be the right-hand side of the inequality (49), i.e.,

$$R = \widetilde{C}_2 \left(\log J(\tau,p,1)\right)^{3q} \left(\log(1/\delta) + \log(1 + J(\tau,p,q))\right)^3 m^{\tilde{\alpha}}$$

with $\tilde{\alpha} = 1 - q + q(1-\beta)\tau$. Then the measure of the set $\mathcal{W}_q(R)$ is at least $1-\delta$. Combining these bounds, for $\forall \mathbf{z} \in \mathcal{W}_q(R) \backslash (V_R^{(q)} \cup V_1')$, there holds

$$\begin{aligned}
&\|\pi_M(\hat{f}_q) - f_\rho\|_{L^2_{\rho_X}}^2 \\
&\quad \leq \widetilde{C}_3 \log^3(18/\delta) \left(\log(1/\delta) + \log(J(\tau,p,1)) + \log(1 + J(\tau,p,q))\right)^{\max\left\{\frac{6p}{q(2+p)}, 3q\right\}} m^{-\widetilde{\Theta}_1},
\end{aligned}$$

where $\widetilde{C}_3 = 2\tilde{c}_{K,p} M^2 \widetilde{C}_2^{\frac{2p}{q(2+p)}} + \widetilde{C}_1' + 2C_2^q$ and

$$\widetilde{\Theta}_1 = \min\left\{\frac{2}{2+p} - \frac{2p}{q(p+2)}\tilde{\alpha}, \frac{2}{2+p} - (1-\beta)\tau, \beta\tau, \tau - \tilde{\alpha}\right\}$$

Note that the measure of the set $\mathcal{W}_q(R) \backslash (V_R^{(q)} \cup V_1')$ is at least $1 - 3\delta$. And under the restrictions on τ, p and q, we have $\widetilde{\Theta}_1 = \tau - \tilde{\alpha}$ and $\max\left\{\frac{6p}{q(2+p)}, 3q\right\} \leq 3$. Then the error bound (60) follows by scaling 3δ to δ. Thus we complete the proof. ■

Now we can prove Theorem 1 based on the error bound (58).

Proof [Proof of Theorem 1]. Recall that the function space $\mathscr{H}_{\widetilde{K}}$ is in the range of the integral operator $L_{\widetilde{K}}^{1/2}$. Hence the assumption $f_\rho \in \mathscr{H}_{\widetilde{K}}$ implies the approximation condition (14) is valid with $\beta = 1$ (see Proposition 8.5 in Cucker and Zhou (2007)). The assumption on the input space X ensures that X satisfies an interior cone condition (see Definition 9). Then for an admissible kernel $K \in \mathscr{C}^s$ with $s > 0$, due to the previous result obtained in Shi et al. (2011) and Theorem 11, one can check that the capacity assumption (20) is achieved. Concretely, when $0 < s < 2$, $p = \frac{2d}{d+2\min\{1,s\}}$; when $s \geq 2$, p can be chosen to be any constant satisfying $p > \frac{2d}{d+2\lfloor s \rfloor}$.

We just focus on the case $s \geq 2$. For any $0 < \epsilon \leq \Theta - \frac{1}{2}$ with Θ given by (5), the capacity assumption is achieved for $p > \frac{2d}{d+2\lfloor s \rfloor}$ satisfying $\frac{2}{2+p} = \Theta - \frac{\epsilon}{2}$. We thus set $\tau = \Theta - \epsilon$. Next we need to bound the quantity $J(\tau,p,1)$. In fact, as $\frac{(2-(p+2)\tau)p}{(1-p\tau)(p+2)} = \frac{1-\frac{p+2}{2}\tau}{1-p\tau} \cdot \frac{2p}{p+2}$, we further have

$$\frac{\log \frac{(2-(p+2)\tau)p}{(1-p\tau)(p+2)}}{\log \frac{2p}{p+2}} = \frac{\log \frac{1-p\tau}{1-\frac{p+2}{2}\tau}}{\log \frac{p+2}{2p}} + 1.$$

Then we substitute $p = \frac{2}{\Theta - \epsilon/2} - 2$ and $\tau = \Theta - \epsilon$ into the above equation. Combining the restrictions for Θ and ϵ, we obtain

$$\frac{\log \frac{(2-(p+2)\tau)p}{(1-p\tau)(p+2)}}{\log \frac{2p}{p+2}} \leq \frac{\log \frac{4}{\epsilon(1-\epsilon)}}{\log \frac{1}{1-\epsilon}}$$

and

$$J(\tau,p,1)=\max\left\{2,\frac{\log\frac{(2-(p+2)\tau)p}{(1-p\tau)(p+2)}}{\log\frac{2p}{p+2}}\right\}\leq\frac{\log\frac{4}{\epsilon(1-\epsilon)}}{\log\frac{1}{1-\epsilon}}:=J_\epsilon.$$

Now recall the general bound (58). Since $\beta=1$, one can check that $\widetilde{\Theta}=\tau=\Theta-\epsilon$ and the bound (6) is valid with $C_\epsilon=C_3(J_\epsilon+1)^6$. As the error bound for the case $0<s<2$ can be derived following the same way, we thus complete the proof of Theorem 1. ■

5. Sparsity analysis

In this section, we shall derive an asymptotical upper bound on $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$, where $\mathbf{c}_q^{\mathbf{z}}$ denotes the coefficient sequence of the estimator $\hat{f}_q$ with $0<q<1$. In order to do so, we need a lower bound on the value of the non-zero elements in $\mathbf{c}_q^{\mathbf{z}}$. Recall that for any vector $\mathbf{c}=(c_1,\cdots,c_m)\in\mathbb{R}^m$, the support set of $\mathbf{c}$ is given by $\mathrm{supp}(\mathbf{c}):=\{j\in\{1,\cdots,m\}:c_j\neq 0\}$.

Proposition 18 *Let $I\subseteq\{1,2,\cdots,m\}$ be a non-empty index set, $0<q<1$ and $\mathbf{c}_q^*=(c_{q,1}^*,\cdots,c_{q,m}^*)$ be a local minimizer of the following optimization problem*

$$\min_{\mathbf{c}\in\mathbb{R}^m,supp(\mathbf{c})\subseteq I}\left\{\frac{1}{m}\sum_{j=1}^m\left(y_j-\sum_{i=1}^m c_iK(x_j,x_i)\right)^2+\gamma\|\mathbf{c}\|_q^q\right\}. \tag{61}$$

Then for $i\in supp(\mathbf{c}_q^*)$, there holds

$$|c_{q,i}^*|>\left(\frac{q(1-q)}{2\kappa^2}\right)^{1/(2-q)}\gamma^{1/(2-q)}, \tag{62}$$

where $\kappa=\|K\|_{\mathscr{C}(X\times X)}$. Moreover, if $\mathbf{c}_q^*$ is a global minimizer of the optimization problem (61), the above bound can be improved to

$$|c_{q,i}^*|\geq\left(\frac{1-q}{\kappa^2}\right)^{1/(2-q)}\gamma^{1/(2-q)}. \tag{63}$$

Proof Recall the target functional (3) to be optimized with $\mathbf{c}=(c_1,\cdots,c_m)\in\mathbb{R}^m$, which can be expressed as

$$\mathscr{T}_{\gamma,q}(\mathbf{c})=\frac{1}{m}\sum_{j=1}^m\left(y_j-\sum_{i=1}^m c_iK(x_j,x_i)\right)^2+\gamma\sum_{i=1}^m|c_i|^q.$$

For $i\in\mathrm{supp}(\mathbf{c}_q^*)$, we define an univariate function as

$$h_i(t)=\mathscr{T}_{\gamma,q}(\mathbf{c}_q^{*\backslash i}(t)),\qquad\forall t\in\mathbb{R},$$

where $\mathbf{c}_q^{*\backslash i}(t)=(c_{q,1}^*,\cdots,c_{q,i-1}^*,t,c_{q,i+1}^*,\cdots,c_{q,m}^*)$. As $\mathbf{c}_q^*$ is a local minimizer of the optimization problem (61), thus $c_{q,i}^*$ is a local minimizer of $h_i(t)$. We compute the first and the

second derivative of $h_i(t)$ on the interval $(c^*_{q,i} - \varrho, c^*_{q,i} + \varrho)$ for some $\varrho > 0$. Since $c^*_{q,i} \neq 0$, we can choose a sufficiently small ϱ such that 0 is not contained in the interval $(\hat{c}_i - \varrho, \hat{c}_i + \varrho)$. Then we obtain

$$h_i^{'}(t) = \frac{2}{m} \sum_{j=1}^{m} \left(\sum_{k \neq i} c^*_{q,k} K(x_j, x_k) + tK(x_j, x_i) - y_j \right) K(x_j, x_i) + \gamma q \mathrm{sgn}(t)|t|^{q-1}$$

and

$$h_i^{''}(t) = \frac{2}{m} \sum_{j=1}^{m} K^2(x_i, x_j) - \gamma q(1-q)|t|^{q-2}.$$

Recall that $c^*_{q,i}$ is a local minimizer of $h_i(t)$. Due to the optimality condition, we must have $h_i^{'}(c^*_{q,i}) = 0$ and $h_i^{''}(c^*_{q,i}) > 0$, i.e.,

$$\frac{2}{m} \sum_{j=1}^{m} \left(\sum_{k=1}^{m} c^*_{q,k} K(x_j, x_k) - y_j \right) K(x_j, x_i) + \gamma q \mathrm{sgn}(c^*_{q,i})|c^*_{q,i}|^{q-1} = 0 \tag{64}$$

and

$$\frac{2}{m} \sum_{j=1}^{m} K^2(x_j, x_i) - \gamma q(1-q)|c^*_{q,i}|^{q-2} > 0. \tag{65}$$

From the inequality (65), we have

$$\gamma q(1-q)|c^*_{q,i}|^{q-2} < \frac{2}{m} \sum_{j=1}^{m} K^2(x_j, x_i) \leq 2\kappa^2,$$

which leads to bound (62).

If $\mathbf{c}^*_q$ is the global minimizer of the optimization problem (61), the global optimality of $\mathbf{c}^*_q$ implies $c^*_{q,i}$ is a global minimizer of $h_i(t)$. We thus have $h_i^{'}(c^*_{q,i}) = 0$. Due to equation (64), there holds

$$\frac{2}{m} \sum_{j=1}^{m} \left(y_j - \sum_{k=1}^{m} c^*_{q,k} K(x_j, x_k) \right) K(x_j, x_i) = \gamma q \mathrm{sgn}(c^*_{q,i})|c^*_{q,i}|^{q-1}.$$

We multiply both sides of the above equation by $c^*_{q,i}$ and find that

$$\frac{2}{m} \sum_{j=1}^{m} \left(y_j - \sum_{k=1}^{m} c^*_{q,k} K(x_j, x_k) \right) K(x_j, x_i) c^*_{q,i} = \gamma q |c^*_{q,i}|^q. \tag{66}$$

Now we consider a new vector given by

$$\mathbf{c}_q^{*\setminus i}(0) = (c^*_{q,1}, \cdots, c^*_{q,i-1}, 0, c^*_{q,i+1}, \cdots, c^*_{q,m}).$$

We compare $\mathscr{T}_{\gamma,q}(\mathbf{c}_q^{*\backslash i}(0))$ with $\mathscr{T}_{\gamma,q}(\mathbf{c}_q^*)$ and find that

$$\begin{aligned}
&\mathscr{T}_{\gamma,q}(\mathbf{c}_q^*) - \mathscr{T}_{\gamma,q}(\mathbf{c}_q^{*\backslash i}(0)) \\
&= -\frac{1}{m}\sum_{j=1}^{m}\left\{2y_j - 2\sum_{k=1}^{m} c_{q,k}^* K(x_j,x_k) + c_{q,i}^* K(x_j,x_i)\right\} c_{q,i}^* K(x_j,x_i) + \gamma|c_{q,i}^*|^q \\
&= \frac{2}{m}\sum_{j=1}^{m}\left\{\sum_{k=1}^{m} c_{q,k}^* K(x_j,x_k) - y_j\right\} K(x_j,x_i)c_{q,i}^* - \frac{1}{m}\sum_{j=1}^{m} K^2(x_j,x_i)(c_{q,i}^*)^2 + \gamma|c_{q,i}^*|^q.
\end{aligned}$$

From the equality (66), it follows that

$$\begin{aligned}
\mathscr{T}_{\gamma,q}(\mathbf{c}_q^*) - \mathscr{T}_{\gamma,q}(\mathbf{c}_q^{*\backslash i}(0)) &= -\gamma q|c_{q,i}^*|^q - \frac{1}{m}\sum_{j=1}^{m} K^2(x_j,x_i)(c_{q,i}^*)^2 + \gamma|c_{q,i}^*|^q \\
&\geq \gamma(1-q)|c_{q,i}^*|^q - \kappa^2(c_{q,i}^*)^2.
\end{aligned}$$

Recall that $\mathbf{c}_q^*$ is a global minimizer of problem (61). The above inequality implies $\gamma(1-q)|c_{q,i}^*|^q - \kappa^2(c_{q,i}^*)^2 \leq 0$, which leads to the lower bound (63). Thus we complete our proof. ■

Remark 19 *We use the second order optimality condition to derive the lower bound on the non-zero coefficients of a local minimizer in (61). Based on the same idea, a lower bound estimation which is similar to (62) has been derived in Chen et al. (2010). However, our analysis gives another lower bound when the solution is a global minimizer, which indicates that the global optimality will help to improve the sparseness of the solutions.*

It should be noticed that for a given kernel function K, the lower bounds presented in Proposition 18 only depend on γ and q. Thus we can use these bounds to obtain some universal results. When $\mathbf{c}_q^{\mathbf{z}}$ is a local minimizer of optimization problem (2) satisfying Assumption 1, which implies that $\mathbf{c}_q^{\mathbf{z}}$ satisfies the condition of Proposition 18 for $I = \{1,\cdots,m\}$. We can derive the upper bound on $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$ based on the lower bound (62).

Theorem 20 *Assume that approximation assumption (14) with $0 < \beta \leq 1$ and capacity condition (20) with $0 < p < 2$ are valid, and the estimator $\hat{f}_q$ satisfies Assumption 1 with*

$$\max\left\{\frac{2p}{p+2}, \frac{p}{p+2\beta}\right\} < q < 1.$$

Let $(\delta,\tau) \in (0,1)^2$ and $J(\tau,p,q)$ be a constant given by (47). Take $\gamma = m^{-\tau}$ with $\frac{1-q}{1-q(1-\beta)} < \tau < \frac{2}{2+p}$. Then with confidence $1-\delta$, there holds

$$\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m} \leq C_q\left(\log J(\tau,p,1)\right)^{3q}\left(\log(1/\delta) + \log(1+J(\tau,p,q))\right)^3 m^{q\left(\frac{\tau}{2-q}-1+(1-\beta)\tau\right)}, \qquad (67)$$

where $\mathbf{c}_q^{\mathbf{z}}$ denotes the coefficient sequence of the estimator $\hat{f}_q$ and C_q is a constant positive constants depending on $q,\kappa,\tilde{c}_{K,p}$, M and c_β.

Proof From Proposition 18, if $\mathbf{c}_q^{\mathbf{z}}$ is a local minimizer of problem (2), then for $i \in \mathrm{supp}(\mathbf{c}_q^{\mathbf{z}})$, there holds

$$|c_{q,i}^{\mathbf{z}}| > \left(\frac{q(1-q)}{2\kappa^2}\right)^{1/(2-q)} \gamma^{1/(2-q)}.$$

Equivalently, we have

$$1 < \left(\frac{q(1-q)}{2\kappa^2}\right)^{-q/(2-q)} \gamma^{-q/(2-q)} |c_{q,i}^{\mathbf{z}}|^q.$$

Note that the above inequality holds true for every $c_{q,i}^{\mathbf{z}}$ with $i \in \mathrm{supp}(\mathbf{c}_q^{\mathbf{z}})$. Therefore, we take a summation on both sides of the inequality according to $i \in \mathrm{supp}(\mathbf{c}_q^{\mathbf{z}})$ and find that

$$\|\mathbf{c}_q^{\mathbf{z}}\|_0 < \left(\frac{q(1-q)}{2\kappa^2}\right)^{-q/(2-q)} \gamma^{-q/(2-q)} \|\mathbf{c}_q^{\mathbf{z}}\|_q^q.$$

Now taking $\gamma = m^{-\tau}$ with $\frac{1-q}{1-q(1-\beta)} < \tau < \frac{2}{2+p}$ and recalling the bound on $\|\mathbf{c}_q^{\mathbf{z}}\|_q^q$ given by (49), then the desired result is obtained with $C_q = \widetilde{C}_2 \left(\frac{q(1-q)}{2\kappa^2}\right)^{-q/(2-q)}$. Thus we complete the proof. ■

If $\mathbf{c}_q^{\mathbf{z}}$ is a global minimizer of algorithm (2), one can derive an upper bound on $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$ from (63) following the same method. From the upper bound (67), we can see that if $\tau < \frac{2-q}{1+(2-q)(1-\beta)}$, the quantity $\frac{\|\mathbf{c}_q^{\mathbf{z}}\|_0}{m}$ converges to 0 at a rate of polynomial decay as m tends to infinity. At the end of this section, we give the proof of Theorem 3.

Proof [Proof of Theorem 3]. From the proof of Theorem 1, we know that under the assumption of Theorem 3, the approximation condition (14) is valid with $\beta = 1$ and the capacity assumption (20) can be satisfied with an arbitrarily small $p > 0$. We derive our conclusions based on Theorem 17 and Theorem 20. We first check the restrictions for q and τ. As $\beta = 1$, we thus find $1 - q < \tau < \frac{2}{p+2}$ and $\frac{2p}{2+p} < q < 1$. Note that p can be arbitrarily closed to 0. Therefore, we can take τ and q to be any value in the interval $(0, 1)$ by choosing a small enough p. By the same argument, we can bound $J(\tau, p, q)$ by $\log \frac{4}{q(1-\tau)}$ for $1 - q < \tau < 1$ and $0 < q \leq 1$. Therefore, from bound (60), the inequality (7) holds with confidence $1 - \delta$ and $\widetilde{C} = 8\widetilde{C}_3$. Similarly, as a direct corollary of the bound (67), the inequality (8) holds with confidence $1 - \delta$ and $\widetilde{C}' = \widetilde{C}_2(1 + 2\kappa^2)$. Finally, if the two bounds hold together, we need to scale 2δ to δ. Thus we complete the proof. ■

6. Conclusion

In this paper, we investigate the sparse kernel regression with ℓ_q-regularization, where $0 < q \leq 1$. The data dependence nature of the kernel-based hypothesis space provides flexibility for this algorithm. The regularization scheme is essentially different from the standard one in an RKHS: the kernel is not necessarily symmetric or positive semi-definite and the regularizer is the ℓ_q-norm of a function expansion involving samples. When the underlying hypothesis space is finite-dimensional, the ℓ_q-regularization with $0 < q \leq 1$ is well understood in theory and widely used in various applications such as compressed sensing

and sparse recovery. However, its role in the non-parametric regression within an infinite-dimensional hypothesis space has not been fully understood yet. In this paper, we develop the mathematical analysis for the asymptotic convergence and sparseness of ℓ_q-regularized kernel regression. We first present a tight bound on the ℓ_2-empirical covering numbers of the kernel-based hypothesis space under ℓ_1-constrain, which is interesting on its own right. We thus demonstrate that, compared with classical RKHS, the hypothesis space involved in the error analysis induced by the non-symmetric kernel has nice behaviors in terms of the ℓ^2-empirical covering numbers of its unit ball. Moreover, the empirical covering number estimates developed in this paper can also be applied to obtain distribution-free error analysis for other sparse approximation schemes, for example Guo, Fan and Zhou (2016); Guo et al. (2017). Our theoretical analysis is based on the concentration estimates with ℓ^2-empirical covering numbers, a refined iteration technique for ℓ_q-regularization and a descent-iterative minimization process which can be realized by the ℓ_q-threshholding function. Based on our analysis, we show that the ℓ_q-regularization term plays a role as a trade-off between sparsity and convergence rates. We also prove that regularizing the combinatorial coefficients by the ℓ_q-norm can produce strong sparse solutions, i.e., the fraction of non-zero coefficients converges to 0 at a polynomial rate when the sample size m becomes large. Our mathematical analysis established in this paper can shed some lights on understanding the role of the ℓ_q-regularization in feature selections in an infinite-dimensional hypothesis space.

Acknowledgments

We thank the anonymous referees for their constructive suggestions. The work described in this paper is supported in part by the National Science Foundation of China (Grants No.11571078, 11631015, 61603248, 61977046); the Joint Research Fund by National Natural Science Foundation of China and Research Grants Council of Hong Kong (Project No. 11461161006 and Project No. CityU 104012); Program of Shanghai Subject Chief Scientist (Project No.18XD1400700). The research leading to these results has also received funding from the European Research Council ERC AdG A-DATADRIVE-B (290923) and ERC AdG E-DUALITY (787960) under the European Union's Horizon 2020 research and innovation programme. The corresponding author is Xiaolin Huang.

Appendix A.

This appendix includes some detailed proofs.

A.1. Proof of Lemma 4

In this subsection, we shall prove Lemma 4.

Proof [Proof of Lemma 4]. It is easy to verify that the vector $\Psi_{\eta,q}(\mathbf{d})$ is a global minimizer of the problem (12) if and only if $(\Psi_{\eta,q}(\mathbf{d}))_i$ is a global minimizer of $\min_{c\in\mathbb{R}}\left\{|c-d_i|^2+\eta|c|^q\right\}$, where $(\Psi_{\eta,q}(\mathbf{d}))_i$ denotes the $i-$th coordinate value of the vector $\Psi_{\eta,q}(\mathbf{d})$. In the following proof, we shall use x to denote a given constant. In order to prove our conclusion, we need to find the global minimizer of the univariate function $h(t)=t^2-2xt+\eta|t|^q$, i.e., the solution

of the minimization problem $\min_{t\in\mathbb{R}}\left\{|t-x|^2+\eta|t|^q\right\}$. Additionally, to ensure the well-definedness of the map $\Psi_{\eta,q}$, we also need to show that under the restriction $|x|>a_q\eta^{1/(2-q)}$, the equation (10) has only one solution on the interval $[(q(1-q)\eta/2)^{1/(2-q)},\infty)$.

When $x=0$, as $\eta>0$, the function $h(t)$ achieves its minimum at $t=0$.

We first limit our discussion to the case $x>0$. In this case, the function h is strictly decreasing on $(-\infty,0]$, hence all its possible minimizers are achieved on $[0,\infty)$. Let us consider the difference between $h(t)$ and $h(0)$ for $t>0$, i.e., $h(t)-h(0)=t(t-2x+\eta t^{q-1})$. It is noticed that the function $g(t)=t-2x+\eta t^{q-1}$ is continuously differentiable on $(0,\infty)$ and $\lim_{t\to 0+}g(t)=\lim_{t\to\infty}g(t)=\infty$. So we take its derivative on $(0,\infty)$ and find that its unique minimizer is given by $t_1=((1-q)\eta)^{1/(2-q)}$. It follows that $g(t)\geq g(t_1)=2a_q\eta^{\frac{1}{2-q}}-2x$ for all $t>0$.

When $0<x\leq a_q\eta^{1/(2-q)}$, then $\min_{t>0}g(t)=g(t_1)\geq 0$ which implies $h(t)-h(0)=tg(t)\geq 0$ for $t>0$. Hence, $t=0$ is a global minimizer of h in this case.

And then we consider the case $x>a_q\eta^{1/(2-q)}$. One may see that $t=0$ is not a global minimizer of h, because $g(t_1)<0$ in this case, which implies $h(t_1)<h(0)$. In addition, we have $\lim_{t\to\infty}h(t)=\infty$, thus the function $h(t)$ achieves its minimum on the interval $(0,\infty)$. We claim that this minimizer is given by t_2, where t_2 is the solution of the equation $h^{'}(t)=0$ on the interval $[(q(1-q)\eta/2)^{1/(2-q)},\infty)$. It should be noticed that $h^{'}(t)=0$ is exactly the equation given by (10) with $x>0$. We thus consider the second derivative of h given by $h^{''}(t)=2+\eta q(q-1)t^{q-2}$. We first prove the existence and uniqueness of t_2. In fact, a direct computation shows that

$$\begin{aligned}
h^{'}((q(1-q)\eta/2)^{1/(2-q)}) &= 2(\eta q)^{\frac{1}{2-q}}\left(\frac{1-q}{2}\right)^{\frac{1}{2-q}}+(\eta q)^{\frac{1}{2-q}}\left(\frac{1-q}{2}\right)^{\frac{q-1}{2-q}}-2x\\
&= (2-q)\left(\frac{1-q}{2}\right)^{\frac{q-1}{2-q}}(\eta q)^{\frac{1}{2-q}}-2x\\
&\leq (2-q)\left(\frac{1-q}{2}\right)^{\frac{q-1}{2-q}}q^{\frac{1}{2-q}}a_q^{-1}x-2x\\
&= \left(2^{\frac{3-2q}{2-q}}q^{\frac{1}{2-q}}-2\right)x<0,
\end{aligned}$$

the first inequality is from $x>a_q\eta^{1/(2-q)}$ and the last inequality holds as $q<2^{q-1}$ for $0<q<1$. We also observe that $h^{''}(t)\geq 0$ on $[(q(1-q)\eta/2)^{1/(2-q)},\infty)$, which implies $h^{'}(t)$ is strictly increasing on this interval. Since $h^{'}(t)$ is continuous on $(0,\infty)$ and $\lim_{t\to\infty}h^{'}(t)=\infty$, the equation $h^{'}(t)=0$ has a unique solution t_2 on $[(q(1-q)\eta/2)^{1/(2-q)},\infty)$. Because $h^{'}(t_2)=0$ and $h^{''}(t_2)>0$, we also conclude that t_2 is the only minimizer of $h(t)$ on $[(q(1-q)\eta/2)^{1/(2-q)},\infty)$. We further prove that t_2 is actually the minimizer of $h(t)$ on $(0,\infty)$. We just need to show $h(t)$ has no local minimizer on $(0,(q(1-q)\eta/2)^{1/(2-q)})$. This conclusion can be easily drawn from the fact that $h^{''}(t)<0$ on $(0,(q(1-q)\eta/2)^{1/(2-q)})$.

When $x<0$, one can easily find that $h(t)$ achieves its minimum on $(-\infty,0]$. Then for $t\in(-\infty,0]$, we can rewrite the function $h(t)$ as $(-t)^2+2x(-t)+\eta(-t)^q$. Due to the same analysis as above, we find that, when $-a_q\eta^{1/(2-q)}\leq x<0$, the global minimizer of $h(t)$ is $t=0$; when $x<-a_q\eta^{1/(2-q)}$, the minimizer of $h(t)$ is given by the unique solution of the equation $2(-t)+\eta q(-t)^{q-1}+2x=0$ on the interval $(-\infty,-(q(1-q)\eta/2)^{1/(2-q)}]$.

Finally, we combine all the cases together and find that for a given $x \in \mathbb{R}$, a global minimizer of $h(t)$ is $\psi_{\eta,q}(x)$ given by (9). Hence according to our analysis, if we evaluate $(\Psi_{\eta,q}(\mathbf{d}))_i$ as (11), the vector $\Psi_{\eta,q}(\mathbf{d})$ is a global minimizer of the optimization problem (12).

It is also easy to check that when $|x| = a_q\eta^{1/(2-q)}$, the univariate function $h(t)$ has two global minimizers given by 0 and $\frac{2-2q}{2-q}|x|$ respectively. Thus the vector $\Psi_{\eta,q}(\mathbf{d})$ indeed gives one global minimizer of problem (12). Finally we complete the proof. ■

Remark 21 *A similar discussion about the global minimizer of $h(t)$ can be found in Knight and Fu (2000).*

A.2. Proof of Proposition 5

In this subsection, we shall prove Proposition 5.

Proof [Proof of Proposition 5]. We first prove conclusion (i). Recall that $\mathbf{c}^* = (c_1^*, \cdots, c_m^*) \in \mathbb{R}^m$ is a local minimizer of the objective functional $\mathscr{T}_{\gamma,q}(\mathbf{c})$ defined by (3) and $0 < \lambda \leq \frac{q}{2}\|\mathbb{K}\|_2^{-2}$. It is sufficient to prove that for $i \in \mathrm{supp}(\mathbf{c}^*)$, i.e., $c_i^* \neq 0$, there holds

$$\left|\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right)_i\right| > a_q(\lambda\gamma)^{1/(2-q)}$$

and $c_i^* = \mathrm{sgn}\left(\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right)_i\right) t_i^*$, where t_i^* is the solution of the equation

$$2t + \lambda\gamma q t^{q-1} - 2\left|\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right)_i\right| = 0$$

on the interval $[(q(1-q)\lambda\gamma/2)^{1/(2-q)}, \infty)$. Here $(\cdot)_i$ denotes the $i-$th coordinate value of a vector.

By the same argument in the proof of Proposition 18, for $i \in \mathrm{supp}(\mathbf{c}^*)$, we have

$$2\left(\mathbb{K}^T(\mathbb{K}\mathbf{c}^* - \mathbf{y})\right)_i + \gamma q \mathrm{sgn}(c_i^*)|c_i^*|^{q-1} = 0 \tag{68}$$

and

$$2\sum_{j=1}^{m}\mathbb{K}_{j,i}^2 > \gamma q(1-q)|c_i^*|^{q-2}, \tag{69}$$

where $\mathbb{K}_{j,i}$ denotes the $(j,i)-$th entry of the matrix $\mathbb{K}$. We multiply both sides of the inequality (69) by a positive λ. Since $\lambda \leq \frac{q}{2}\|\mathbb{K}\|_2^{-2}$, we obtain

$$\lambda\gamma q(1-q)|c_i^*|^{q-2} < 2\lambda\sum_{j=1}^{m}\mathbb{K}_{j,i}^2 \leq q,$$

which implies $|c_i^*| > ((1-q)\lambda\gamma)^{1/(2-q)}$.

We consider the case $c_i^* > 0$. By equation (68), for $c_i^* > 0$, there holds

$$2c_i^* + \lambda\gamma q(c_i^*)^{q-1} = 2\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right)_i,$$

which verifies that $\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right)_i$ is positive and c_i^* is the zero point of

$$f(t) = 2t + \lambda\gamma q t^{q-1} - 2\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right)_i.$$

Note that $c_i^* > ((1-q)\lambda\gamma)^{1/(2-q)} > (q(1-q)\lambda\gamma/2)^{1/(2-q)}$. Recalling the analysis in Lemma 4, we know that $f(t)$ is monotonically increasing on the interval $[(q(1-q)\lambda\gamma/2)^{1/(2-q)}, \infty)$. We thus have $f((\lambda\gamma(1-q))^{1/(2-q)}) < 0$ which implies

$$\left(\mathbf{c}^* + \lambda\mathbb{K}^T(\mathbf{y} - \mathbb{K}\mathbf{c}^*)\right)_i > a_q(\lambda\gamma)^{1/(2-q)}.$$

Therefore, we verify conclusion (i) for $c_i^* > 0$. When $c_i^* < 0$, we can prove it following the same method.

Next, we shall prove the function $\psi_{\eta,q}(x)$ defined by (9) is Lipschitz continuous and strictly increasing for any $|x| > a_q\eta^{1/(2-q)}$. Then one can check that the proof of the rest two conclusions are almost the same as the proof of Theorem 3 in Xu et al. (2012). Since $\psi_{\eta,q}(x)$ is an odd function, we only need to prove our desired result for $x > a_q\eta^{1/(2-q)}$. From (9), when $x > a_q\eta^{1/(2-q)}$, $\psi_{\eta,q}(x)$ is defined to be the solution of the equation $2t+\eta q t^{q-1}-2x = 0$ on the interval $[(q(1-q)\eta/2)^{1/(2-q)}, \infty)$. Since the bivariate function $F(t,x) = 2t+\eta q t^{q-1} - 2x$ is continuously differentiable inside $[(q(1-q)\eta/2)^{1/(2-q)}, \infty) \times (a_q\eta^{1/(2-q)}, \infty)$, we have $\psi_{\eta,q}^{'}(x) = \frac{2}{2-\eta q(1-q)(\psi_{\eta,q}(x))^{q-2}}$ due to the implicit function theorem. Obviously $\psi_{\eta,q}^{'}(x) > 1$ which implies $\psi_{\eta,q}(x)$ is strictly increasing. In order to verify the Lipschitz continuity, we still need an upper bound for $\psi_{\eta,q}^{'}(x)$. To this end, we show that $\psi_{\eta,q}(x) > (q(1-q)\eta)^{1/(2-q)}$ for $x > a_q\eta^{1/(2-q)}$. Also following the analysis of Lemma 4, it is equivalent to check that when $x > a_q\eta^{1/(2-q)}$, there holds $2\left(q(1-q)\eta\right)^{\frac{1}{2-q}} + \eta q\left(q(1-q)\eta\right)^{\frac{q-1}{2-q}} - 2x < 0$. Since when $x > a_q\eta^{1/(2-q)}$, a simple calculation show that

$$2\left(q(1-q)\eta\right)^{\frac{1}{2-q}} + \eta q\left(q(1-q)\eta\right)^{\frac{q-1}{2-q}} - 2x < 2a_q\eta^{1/(2-q)}\left(\frac{3-2q}{2-q}q^{1/(2-q)} - 1\right).$$

The left hand side of the above inequality can be further bounded by 0 as $\frac{3-2q}{2-q}q^{1/(2-q)} < 1$ for $0 < q < 1$. Hence $\psi_{\eta,q}(x) > (q(1-q)\eta)^{1/(2-q)}$ for $x > a_q\eta^{1/(2-q)}$, which implies $\psi_{\eta,q}^{'}(x) < 2$. We thus verify the Lipschitz continuity of $\psi_{\eta,q}(x)$ for $x > a_q\eta^{1/(2-q)}$ and the proof is completed. ■

A.3. Proof of Proposition 6

In this subsection, we shall prove Proposition 6.

Proof [Proof of Proposition 6]. We just prove the inequality (16), the bound (17) can be derived by the same approach. Recall the definition of the projection operator π_M (see Definition 2). As $\rho(\cdot|x)$ is supported on $[-M, M]$ at every $x \in X$, it obviously implies $|y_i| \le M$ for $i = 1, \cdots, m$ and $\mathscr{E}_{\mathbf{z}}(\pi_M(\hat{f}_q)) \le \mathscr{E}_{\mathbf{z}}(\hat{f}_q)$.

When $0 < q < 1$, since the estimator $\hat{f}_q$ satisfies Assumption 1, we have

$$\mathscr{E}_{\mathbf{z}}(\hat{f}_q) + \gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \le \mathscr{E}_{\mathbf{z}}(\hat{f}_1) + \gamma\|\mathbf{c}_1^{\mathbf{z}}\|_q^q.$$

Therefore,

$$\begin{aligned}
\mathscr{E}_{\mathbf{z}}(\pi_M(\hat{f}_q)) + \gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q &\le \mathscr{E}_{\mathbf{z}}(\hat{f}_q) + \gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \\
&\le \mathscr{E}_{\mathbf{z}}(\hat{f}_1) + \gamma\|\mathbf{c}_1^{\mathbf{z}}\|_q^q \\
&\le \mathscr{E}_{\mathbf{z}}(\hat{f}_1) + \gamma\|\mathbf{c}_1^{\mathbf{z}}\|_1 + \gamma\|\mathbf{c}_1^{\mathbf{z}}\|_q^q. \qquad (70)
\end{aligned}$$

Note that the coefficient sequence of $f_{\mathbf{z},\gamma}$ is given by $\{\frac{1}{m}g_\gamma(x_i)\}_{i=1}^m$ and $\hat{f}_1$ is the global minimizer of problem (2) at $q=1$. Hence, there holds

$$\mathscr{E}_{\mathbf{z}}(\hat{f}_1)+\gamma\|\mathbf{c}_1^{\mathbf{z}}\|_1 \le \mathscr{E}_{\mathbf{z}}(f_{\mathbf{z},\gamma})+\gamma m^{-1}\sum_{i=1}^m |g_\gamma(x_i)|. \tag{71}$$

A direct computation shows that

$$\begin{aligned}
&\mathscr{E}(\pi_M(\hat{f}_q))-\mathscr{E}(f_\rho)+\gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q \\
&=\{\mathscr{E}(\pi_M(\hat{f}_q))-\mathscr{E}_{\mathbf{z}}(\pi_M(\hat{f}_q))\}+\{\mathscr{E}_{\mathbf{z}}(\pi_M(\hat{f}_q))+\gamma\|\mathbf{c}_q^{\mathbf{z}}\|_q^q-\mathscr{E}_{\mathbf{z}}(f_{\mathbf{z},\gamma})-\gamma m^{-1}\sum_{i=1}^m|g_\gamma(x_i)|\} \\
&+\{\mathscr{E}_{\mathbf{z}}(f_{\mathbf{z},\gamma})-\mathscr{E}(f_{\mathbf{z},\gamma})\}+\{\gamma m^{-1}\sum_{i=1}^m|g_\gamma(x_i)|-\gamma\|g_\gamma\|_{L^1_{\rho_X}}\}+\{\gamma\|g_\gamma\|_{L^1_{\rho_X}}-\gamma\|g_\gamma\|_{L^2_{\rho_X}}\} \\
&+\{\mathscr{E}(f_{\mathbf{z},\gamma})-\mathscr{E}(f_\gamma)\}+\{\mathscr{E}(f_\gamma)-\mathscr{E}(f_\rho)+\gamma\|g_\gamma\|_{L^2_{\rho_X}}\}.
\end{aligned}$$

From inequalities (70) and (71), the second term on the right hand side of the above equation is at most $\gamma\|\mathbf{c}_1^{\mathbf{z}}\|_q^q$, which can be further bounded by $\gamma m^{1-q}\|\mathbf{c}_1^{\mathbf{z}}\|_1^q$ due to the reverse Hölder inequality. The inequality $\|g_\gamma\|_{L^1_{\rho_X}} \le \|g_\gamma\|_{L^2_{\rho_X}}$ implies the fifth term is at most zero, thus we obtain bound (16). Then the proof is completed. ■

A.4. Proof of Proposition 14

Now we concentrate our efforts on the proof of Proposition 14.

Definition 22 *A function $\psi:\mathbb{R}^+\to\mathbb{R}^+$ is sub-root if it is non-negative, non-decreasing, and if $\psi(r)/\sqrt{r}$ is non-increasing.*

It is easy to see for a sub-root function ψ and any $D>0$, the equation $\psi(r)=r/D$ has unique positive solution. Proposition 14 can be proved based on the following lemma, which is given in Blanchard et al. (2008).

Lemma 23 *Let $\mathcal{F}$ be a class of measurable, square-integrable functions such that $\mathbb{E}(f)-f\le b$ for all $f\in\mathcal{F}$. Let ψ be a sub-root function, D be some positive constant and r^* be the unique positive solution of $\psi(r)=r/D$. Assume that*

$$\forall r\ge r^*,\quad \mathbb{E}\left[\max\left\{0,\sup_{f\in\mathcal{F},\mathbb{E}f^2\le r}\left(\mathbb{E}(f)-\frac{1}{m}\sum_{i=1}^m f(z_i)\right)\right\}\right]\le\psi(r). \tag{72}$$

Then for all $x>0$ and all $T>D/7$, with probability at least $1-e^{-x}$ there holds

$$\mathbb{E}f-\frac{1}{m}\sum_{i=1}^n f(X_i)\le\frac{\mathbb{E}f^2}{D}+\frac{50T}{D^2}r^*+\frac{(T+9b)x}{m},\quad\forall f\in\mathcal{F}. \tag{73}$$

Proof [Proof of Proposition 14]. We assume that the function g in the concerned function set satisfies $\|g\|_\infty\le B$ and $\mathbb{E}g^2\le c\mathbb{E}g$ for some $c,B>0$. In order to prove the conclusion, we define the function set for $R\ge 1$ as

$$\mathcal{G}_R=\left\{g(z)=(\pi_M(f)(x)-y)^2-(f_\rho(x)-y)^2 : f\in\mathscr{B}_R\right\}.$$

We will apply Lemma 23 to $\mathcal{G}_R$ and find the sub-root function ψ in our setting. To this end, let $\sigma_1, \cdots, \sigma_n$ be the independent *Rademacher random variables*, that is, independent random variables for which $\text{Prob}(\sigma_i = 1) = \text{Prob}(\sigma_i = -1) = 1/2$. One can see Van der Vaart and Wellner (1996) that

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\left|\mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i)\right|\right] \le 2\mathbb{E}\left[\sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)\right|\right]. \tag{74}$$

Next, we will bound the right-hand side of the above inequality by using the ℓ_2-empirical covering numbers and entropy integral. Due to Van der Vaart and Wellner (1996), there is an universal absolute constant C such that

$$\frac{1}{\sqrt{m}}\mathbb{E}_\sigma \sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\left|\sum_{i=1}^m \sigma_i g(z_i)\right| \le C\int_0^{\sqrt{V}} \log_2^{\frac{1}{2}} \mathscr{N}_2(\mathcal{G}_R,\nu)d\nu, \tag{75}$$

where $V = \sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\frac{1}{m}\sum_{i=1}^m g^2(z_i)$ and $\mathbb{E}_\sigma$ denotes the expectation with respect to the random variables $\{\sigma_1,\cdots,\sigma_m\}$ conditioned on all of the other random variables. Now we use the capacity condition (20) for $\mathscr{B}_1$. It asserts that

$$\log_2 \mathscr{N}_2(\mathscr{B}_1,\epsilon) \le c_{K,p}\epsilon^{-p}, \qquad \forall\ 0<\epsilon\le 1.$$

For $g_1, g_2 \in \mathcal{G}_R$, we have

$$|g_1(z) - g_2(z)| = |\,(y-\pi_M(f_1)(x))^2 - (y-\pi_M(f_2)(x))^2\,| \le 4M|f_1(x) - f_2(x)|.$$

Therefore, it follows that

$$\mathscr{N}_2\left(\mathcal{G}_R,\epsilon\right) \le \mathscr{N}_2\left(\mathscr{B}_R,\frac{\epsilon}{4M}\right) \le \mathscr{N}_2\left(\mathscr{B}_1,\frac{\epsilon}{4MR}\right) \le \mathscr{N}_2\left(\mathscr{B}_1,\frac{\epsilon}{R_1}\right)$$

where $R_1 = \max\{B, 4MR\}$. Then

$$\log_2 \mathscr{N}_2\left(\mathcal{G}_R,\epsilon\right) \le c_{K,p}R_1^p\epsilon^{-p}, \qquad \forall\ 0<\epsilon\le R_1.$$

Using equation (75), since $V = \sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\frac{1}{m}\sum_{i=1}^m g^2(z_i)$ and $\sqrt{V}\le B\le R_1$, one easily gets

$$\begin{aligned}\frac{1}{\sqrt{m}}\mathbb{E}_\sigma \sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\left|\sum_{i=1}^m \sigma_i g(z_i)\right| &\le Cc_{K,p}^{1/2}R_1^{p/2}\int_0^{\sqrt{V}}\nu^{-p/2}d\nu \\ &= \tilde{c}_{K,p}R_1^{p/2}\left(\sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\frac{1}{m}\sum_{i=1}^m g^2(z_i)\right)^{\frac{1}{2}-\frac{p}{4}},\end{aligned} \tag{76}$$

where $\tilde{c}_{K,p} = 2Cc_{K,p}^{1/2}(2-p)^{-1}$ depending on p and the kernel function K. For simplicity, the constant $\tilde{c}_{K,p}$ may change from line to line in the following derivations. Due to Talagrand (1994), one see that

$$\mathbb{E}\sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\frac{1}{m}\sum_{i=1}^m g^2(z_i) \le \frac{8B}{m}\mathbb{E}\mathbb{E}_\sigma\sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\left|\sum_{i=1}^m \sigma_i g(z_i)\right| + r. \tag{77}$$

Therefore, from (76) and (77), we have

$$\mathcal{R}_m := \frac{1}{\sqrt{m}}\mathbb{E}\mathbb{E}_\sigma \sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\left|\sum_{i=1}^m \sigma_i g(z_i)\right| \le \tilde{c}_{K,p}\left(\frac{\mathcal{R}_m B}{\sqrt{m}}+r\right)^{\frac{1}{2}-\frac{p}{4}} R_1^{p/2}.$$

Solving the above inequality for $\mathcal{R}_m$ and substituting it to the equation (74), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{g\in\mathcal{G}_R,\mathbb{E}g^2\le r}\left|\mathbb{E}g-\frac{1}{m}\sum_{i=1}^m g(z_i)\right|\right] &\le \tilde{c}_{K,p}\max\left\{B^{\frac{2-p}{2+p}}m^{-\frac{2}{2+p}}R_1^{\frac{2p}{2+p}}, r^{\frac{1}{2}-\frac{p}{4}}m^{-\frac{1}{2}}R_1^{\frac{p}{2}}\right\}\\ &\le \tilde{c}_{K,p}R_1^{\frac{2p}{2+p}}\max\left\{B^{\frac{2-p}{2+p}}m^{-\frac{2}{2+p}}, r^{\frac{1}{2}-\frac{p}{4}}m^{-\frac{1}{2}}\right\},\end{aligned}$$

where the last inequality is due to $R_1 \ge 1$ and $0 < p < 2$. According to Lemma 23, one may take $\psi(r)$ to be the right-hand side of the above inequality. Then the solution r^* to the equation $\psi(r) = r/D$ satisfies

$$r^* \le \tilde{c}_{K,p}\max\left\{D^{\frac{4}{2+p}}, DB^{\frac{2-p}{2+p}}\right\}m^{-\frac{2}{2+p}}R_1^{\frac{2p}{2+p}}.$$

Recalling that $\mathbb{E}g^2 \le c\mathbb{E}g$, now we apply Lemma 23 to the function set $\mathcal{G}_R$ by let $T = D = 2c$, then with probability $1-e^{-x}$, there holds

$$\begin{aligned}\mathbb{E}g-\frac{1}{m}\sum_{i=1}^m g(z_i) \le \frac{1}{2}\mathbb{E}g+\frac{(2c+9b)x}{m}\\ +\tilde{c}_{K,p}\max\left\{c^{\frac{2-p}{2+p}}, B^{\frac{2-p}{2+p}}\right\}m^{-\frac{2}{2+p}}R_1^{\frac{2p}{2+p}}, \quad \forall g\in\mathcal{G}_R.\end{aligned}$$

For $g \in \mathcal{G}_R$, it is easy to verify that $b = c = 16M^2$, $B = 8M^2$ and $R_1 = \max\{4MR, 8M^2\} \le 8M^2R$. From the above inequality, we obtain

$$\mathbb{E}g-\frac{1}{m}\sum_{i=1}^m g(z_i) \le \frac{1}{2}\mathbb{E}g+\frac{176M^2x}{m}+\tilde{c}_{K,p}M^2m^{-\frac{2}{2+p}}R^{\frac{2p}{2+p}}, \forall g\in\mathcal{G}_R,$$

which is the exactly the inequality given by (37).

Next, we consider the function set defined as

$$\mathcal{G}'_R = \left\{g(z) = (f(x)-y)^2-(f_\rho(x)-y)^2 : f\in\mathscr{B}_R\right\}.$$

Note that $\mathscr{B}_R$ is a subset of $\mathscr{C}(X)$ and

$$\|f\|_{\mathscr{C}(X)} \le \kappa R, \qquad \forall f\in\mathscr{B}_R,$$

where $\kappa = \|K\|_{\mathscr{C}(X\times X)}$. Therefore, we have $B = c = (3M+\kappa R)^2$ and $b = 2B$ for $g \in \mathcal{G}'_R$. Moreover, $\forall g_1, g_2 \in \mathcal{G}'_R$, there holds

$$\begin{aligned}|g_1(z)-g_2(z)| &= |\,(y-f_1(x))^2-(y-f_2(x))^2\,|\\ &\le |2y-f_1(x)-f_2(x)||f_1(x)-f_2(x)|\\ &\le 2(M+\kappa R)|f_1(x)-f_2(x)|.\end{aligned}$$

Following the same analysis as above, we find that with probability $1 - e^{-x}$,

$$\mathbb{E}g - \frac{1}{m}\sum_{i=1}^{m} g(z_i) \leq \frac{1}{2}\mathbb{E}g + \frac{20(3M+\kappa R)^2 x}{m} + \tilde{c}_{K,p}(3M+\kappa)^{\frac{4-2p}{2+p}} R^{\frac{4-2p}{2+p}} m^{-\frac{2}{2+p}} R_2^{\frac{2p}{2+p}}, \quad \forall g \in \mathcal{G}'_R,$$

where $R_2 = \max\left\{(3M+\kappa R)^2, 2(M+\kappa R)R\right\} \leq (3M+\kappa)^2 R^2$. Hence, we obtain

$$\mathbb{E}g - \frac{1}{m}\sum_{i=1}^{m} g(z_i) \leq \frac{1}{2}\mathbb{E}g + \frac{20(3M+\kappa R)^2 x}{m} + \tilde{c}_{K,p}(3M+\kappa)^2 m^{-\frac{2}{2+p}} R^2, \quad \forall g \in \mathcal{G}'_R,$$

which leads to the inequality (38).

Finally, we can derive the same bounds for $-\mathcal{S}(\pi_M(f))$ and $-\mathcal{S}(f)$ by considering the function set $-\mathcal{G}_R$ and $-\mathcal{G}'_R$. Thus we complete the proof. ■

References

R. Adams and J. Fournier. *Sobolev Spaces*. Academic press, 2003.

T. Ando, S. Konishi, and S. Imoto. Nonlinear regression modeling via regularized radial basis function networks. *Journal of Statistical Planning and Inference*, 138:3616–3633, 2008.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36:489–531, 2008.

T. Blumensath and M. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14:629–654, 2008.

M. S. Birman and M. Z. Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α (Russian). *Matematicheskii Sbornik*, 73:331–355, 1967.

E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2006.

E. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.

R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *Signal Processing Letters, IEEE*, 14:707–710, 2007.

D. Chen, Q. Wu, Y. Ying, and D. X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

S. Chen and D. Donoho. Basis pursuit. In *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, volume 1, pages 41–44. IEEE, 1994.

X. Chen, F. Xu, and Y. Ye. Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization. *SIAM Journal on Scientific Computing*, 32:2832–2852, 2010.

J. B. Conway. *A Course in Operator Theory*. American Mathematical Society, 2000.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

R. M. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 15:1306–1326, 1987.

D. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, 1996.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

Y. Feng, S. G. Lv, H. Hang, and J. A. K. Suykens. Kernelized elastic net regularization: generalization bounds, and sparse recovery. *Neural Computation*, 28:525–562, 2016.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint*, arXiv:1702.07254, 2017.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480, 1998.

X. Guo, J. Fan, and D. X. Zhou. Sparsity and error analysis of empirical feature-based regularization schemes. *Journal of Machine Learning Research*, 17:3058–3091, 2017.

Z. C. Guo and L. Shi. Learning with coefficient-based regularization and ℓ_1-penalty. *Advances in Computational Mathematics*, 39:493–510, 2013.

Z. C. Guo, D. H. Xiang, X. Guo, and D. X. Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, 15:433–455, 2017.

J. Huang, J. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 36:587–613, 2008.

X. Huang, A. Maier, J. Hornegger, and J. A. K. Suykens. Indefinite kernels in least squares support vector machines and principal component analysis. *Applied and Computational Harmonic Analysis*, 43:162–172, 2017.

K. Jetter, J. Stöckler, and J. D. Ward. Error estimates for scattered data interpolation on spheres. *Mathematics of Computation*, 68:733–747, 1999.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.

M. J. Lai and J. Wang. An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems. *SIAM Journal on Optimization*, 21:82–101, 2011.

S. B. Lin, J. Zeng, J. Fang, and Z. Xu. Learning rates of ℓ^q-coefficient regularization learning with gaussian kernel. *Neural Computation*, 26:2350–2378, 2014.

G. Loosli, S. Canu, and C. S. Ong. Learning SVM in Krěin spaces. *IEEE transactions on pattern analysis and machine intelligence*, 38:1204–1216, 2016.

S. Mendelson and J. Neeman. Regularization in kernel learning. *Annals of Statistics*, 38:526–565, 2010.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

B. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu. $\ell_p - \ell_q$ penalty for sparse linear and sparse multiple kernel multi-task learning. *IEEE Transations on Neural Networks*, 22:1307–1320, 2011.

G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57:6976–6994, 2011.

V. Roth. The generalized Lasso. *IEEE transactions on Neural Networks*, 15:16–28, 2004.

R. Saad and Ö. Yılmaz. Sparse recovery by non-convex optimization–instance optimality. *Applied and Computational Harmonic Analysis*, 29:30–48, 2010.

F. Schleif and P. Tino. Indefinite proximity learning: a review. *Neural Computation*, 27:2039–2096, 2015.

L. Shi, Y. L. Feng, and D. X. Zhou. Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31:286-302, 2011.

L. Shi. Learning theory estimates for coefficient-based regularized regression. *Applied and Computational Harmonic Analysis*, 34:252–265, 2013.

S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

I. Steinwart and A. Christmann. *Support Vector Machines*. New York: Springer, 2008.

I. Steinwart and A. Christmann. Sparsity of SVMs that use the ϵ-insensitive loss. In *Advances in Neural Information Processing Systems 21* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), 1569–1576, 2009.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35:575–607, 2007.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

H. Y. Wang, Q. W. Xiao, and D. X. Zhou. An approximation theory approach to learning with ℓ^1 regularization. *Journal of Approximation Theory*, 167:240–258, 2013.

H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematic*, 4:389–396, 1995.

H. Wendland. Local polynomial reproduction and moving least squares approximation. *IMA Journal of Numercial Analysis*, 21:285–300, 2001.

H. Wendland. *Scattered data approximation*. Cambridge University Press Cambridge, Cambridge, 2005.

Q. Wu, Y. Ying, and D. X. Zhou. Learning rates of least-square regularized regressioin. *Foundations of Computational Mathematics*, 6:171–192, 2006.

Q. Wu and D. X. Zhou. Learning with sample dependent hypothesis spaces. *Computers & Mathematics with Applications*, 56:2896–2907, 2008.

Z. Wu. Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, 4:283–292, 1995.

A. Van der Vaart and J. Wellner. *Weak Convergence and Emprical Processes*. Springer-Verlag, New York, 1996.

Z. Xu, X. Chang, F. Xu, and H. Zhang. $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1013–1027, 2012.

T. Zhang. Leave-one-out bounds for kernel methods. *Neural Computation*, 15:1397–1437, 2003.

D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743–1752, 2003.