Quantile Regression with ℓ_1-regularization and Gaussian Kernels[†]

Lei Shi[1,2], Xiaolin Huang[1], Zheng Tian[2] and Johan A.K. Suykens[1]

[1] Department of Electrical Engineering, KU Leuven,
ESAT-SCD-SISTA, B-3001 Leuven, Belgium
[2] Shanghai Key Laboratory for Contemporary Applied Mathematics,
School of Mathematical Sciences, Fudan University
Shanghai 200433, P. R. China

Abstract

The quantile regression problem is considered by learning schemes based on ℓ_1-regularization and Gaussian kernels. The purpose of this paper is to present an concentration estimates for the algorithms. Our analysis shows that the convergence behavior of ℓ_1-quantile regression with Gaussian kernels is almost the same as that of the RKHS-based learning schemes. Furthermore, the previous analysis for kernel-based quantile regression usually requires that the output sample values are uniformly bounded, which excludes the common case with Gaussian noise. Our error analysis presented in this paper can give satisfactory convergence rates even for unbounded sampling processes. Besides, numerical experiments are given which support the theoretical results.

Key words and phrases. Learning theory, Quantile regression, ℓ_1-regularization, Gaussian kernels, Unbounded sampling processes, Concentration estimate for error analysis

AMS Subject Classification Numbers: 68T05, 62J02

†The corresponding author is Lei Shi. Email addresses: leishi@fudan.edu.cn (L. Shi), huangxl06@mails.tsinghua.edu.cn (X. Huang), jerry.tianzheng@gmail.com (Z. Tian) and johan.suykens@esat.kuleuven.be (J. Suykens).

1 Introduction

In this paper, under the framework of learning theory, we study ℓ_1-regularized quantile regression with Gaussian kernels. Let X be a compact subset of $\mathbb{R}^n$ and $Y \subset \mathbb{R}$, the goal of quantile regression is to estimate the conditional quantile of a Borel probability measure ρ on $Z := X \times Y$. Denote by $\rho(\cdot|x)$ the conditional distribution of ρ at $x \in X$, the *conditional $\tau-$quantile* is a set-valued function defined by

$$F_\rho^\tau(x) = \{t \in \mathbb{R} : \rho((-\infty, t]|x) \geq \tau \text{ and } \rho([t, \infty)|x) \geq 1 - \tau\}, \quad x \in X, \tag{1.1}$$

where $\tau \in (0, 1)$ is a fixed constant specifying the desired quantile level. We suppose that $F_\rho^\tau(x)$ consists of singletons, i.e. there exists an $f_\rho^\tau : X \to \mathbb{R}$, called the *conditional $\tau-$quantile function*, such that $F_\rho^\tau(x) = \{f_\rho^\tau(x)\}$ for $x \in X$. In the setting of learning theory, the distribution ρ is unknown. All we have in hand is only a sample set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$, which is assumed to be independently distributed according to ρ. We additionally suppose that for some constant $M_\tau \geq 1$,

$$|f_\rho^\tau(x)| \leq M_\tau \text{ for almost every } x \in X \text{ with respect to } \rho_X, \tag{1.2}$$

where ρ_X denotes the marginal distribution of ρ on X. Throughout the paper, we will use these three assumptions without any further reference. We aim to approximate f_ρ^τ from the sample $\mathbf{z}$ through learning algorithms.

Relative to the classical least squares regression, quantile regression estimates are more robust against outliers in the response measurements and can provide richer information about the distributions of response variables such as stretching or compressing tails [12]. Due to its wide applications in data analysis, quantile regression attracts much attention in machine learning community and has been investigated in literature (e.g. [8, 20, 21, 33]). Define the *τ-pinball loss* $L_\tau : \mathbb{R} \to \mathbb{R}^+$ as

$$L_\tau(u) = \begin{cases} (1-\tau)u, & \text{if } u > 0, \\ -\tau u, & \text{if } u \leq 0. \end{cases}$$

Recall that f_ρ^τ minimizes $\int_{X\times Y} L_\tau(f(x) - y)d\rho$ over all measurable functions $f : X \to \mathbb{R}$, based on this observation, learning algorithms produce estimators of f_ρ^τ by minimizing $\frac{1}{m}\sum_{i=1}^m L_\tau(f(x_i) - y_i)$ when i.i.d. samples $\{(x_i, y_i)\}_{i=1}^m$ are given. In kernel-based machine learning, this minimization process usually takes place in a hypothesis space (a subset of continuous functions on X) generated by a kernel function $K : X \times X \to \mathbb{R}$. A popular choice is the Gaussian kernel with a variance $\sigma > 0$, which is given by

$$K^\sigma(x, y) = \exp\left\{-\frac{\|x - y\|^2}{2\sigma^2}\right\}.$$

Gaussian kernels are the most widely used kernels in practice because they are universal on every compact subset of $\mathbb{R}^n$ [17]. The variance σ is usually treated as a free parameter in training processes and can be chosen in a data-dependent way, e.g. by cross validation. It motivates the studies on the convergence behavior of algorithms with Gaussian kernels (e.g. [18, 32]). In particular, [33, 8] consider approximating f_ρ^τ by a solution of the optimization scheme

$$\arg\min_{f\in\mathscr{H}_\sigma}\left\{\frac{1}{m}\sum_{i=1}^{m}L_\tau(f(x_i)-y_i)+\lambda\|f\|_\sigma^2\right\}, \tag{1.3}$$

where $(\mathscr{H}_\sigma,\|\cdot\|_\sigma)$ is the Reproducing Kernel Hilbert Space (RKHS) [1] induced by K^σ. The positive constant λ is another tunable parameter and called the *regularization parameter*. Due to the Reprensenter Theorem [28], the solution of algorithm (1.3) belongs to a data-dependent hypothesis space

$$\mathscr{H}_{\mathbf{z},\sigma}=\left\{\sum_{i=1}^{m}\alpha_iK^\sigma(x,x_i):\alpha_i\in\mathbb{R}\right\}.$$

In this paper, for pursuing sparsity, we estimate f_ρ^τ by the ℓ_1-regularized learning algorithm. The algorithm is defined as the solution $\hat{f}_{\mathbf{z}}^\tau=f_{\mathbf{z},\lambda,\sigma}^\tau$ to the following minimization problem

$$\hat{f}_{\mathbf{z}}^\tau=\arg\min_{f\in\mathscr{H}_{\mathbf{z},\sigma}}\left\{\frac{1}{m}\sum_{i=1}^{m}L_\tau(f(x_i)-y_i)+\lambda\Omega(f)\right\}, \tag{1.4}$$

where the regularization term is given by

$$\Omega(f)=\sum_{i=1}^{m}|\alpha_i|\text{ for }f=\sum_{i=1}^{m}\alpha_iK^\sigma(x,x_i)\in\mathscr{H}_{\mathbf{z},\sigma},$$

i.e. the ℓ_1-norm of the coefficients in the kernel expansion of $f\in\mathscr{H}_{\mathbf{z},\sigma}$. The positive definiteness of K^σ ensures that the expression of $f\in\mathscr{H}_{\mathbf{z},\sigma}$ is unique. Thus the regularization term Ω as a functional on $\mathscr{H}_{\mathbf{z},\sigma}$ is well-defined. The regularization parameter λ controls the balance between the regularization term and the empirical error caused by data. And the parameters λ and σ are both free-determined in the algorithm.

The scheme with ℓ_1-regularization is often related to LASSO algorithm [24] in the linear regression model. And there have been extensive studies on the error analysis of ℓ_1-estimator for linear least square regression and linear quantile regression in statistics (e.g. see [2, 36]). In kernel-based machine learning, the ℓ_1-regularization was first introduced to design the linear programming support vector machine (e.g. [13, 26, 4]). Recently, a number of papers have begun to study the learning behavior of ℓ_1-regularized least square regression with a fixed kernel function (e.g. see [22, 16]).

The ℓ_1-regularizaiton is a very important regularization method as it may lead to sparse solutions. Particularly, the ℓ_1-regularized quantile regression has excellent computational properties. Since the loss function and the regularization term are both piecewise linear, the learning algorithm (1.4) is essentially a linear programming optimization problem and thus can be efficiently solved [34].

Up to now, the kernel-based quantile regression mainly focuses on estimating f_ρ^τ by regularization schemes in RKHS. All results are stated under the boundedness assumption for the output, i.e. for some constant $M > 0$, $|y| \leq M$ almost surely. Our paper is devoted to establishing the convergence analysis for quantile regression based on ℓ_1-regularization and Gaussian kernels. Specifically, we investigate how the output function $\hat{f}_{\mathbf{z}}^\tau$ given in (1.4) approximates the quantile regression function f_ρ^τ with suitable chosen $\lambda = \lambda(m)$ and $\sigma = \sigma(m)$ as $m \to \infty$. We will show that the learning ability of algorithm (1.4) is almost the same as that of RKHS-based algorithm (1.3). Our error bounds are obtained under a weaker assumption: for some constants $M \geq 1$ and $c > 0$,

$$\int_Y |y|^\ell d\rho(y|x) \leq c\ell! M^\ell, \qquad \forall \ell \in \mathbb{N}, \quad x \in X. \tag{1.5}$$

Note that the boundedness assumption excludes the Gaussian noise while assumption (1.5) covers it. This assumption is well known in probability theory and was introduced in learning theory in [27, 10].

In the rest of this paper, we first present the main results in Section 2. After that, we give the framework of convergence analysis in Section 3 and prove the concerned theorems in Section 4. In Section 5, the results of numerical experiments are given to support the theoretical results.

2 Main Results

In order to illustrate our convergence analysis, we first state the definition of *projection operator* introduced in [6].

Definition 1. *For $B > 0$, the projection operator π_B on $\mathbb{R}$ is defined as*

$$\pi_B(t) = \begin{cases} -B & \text{if } t < -B, \\ t & \text{if } -B \leq t \leq B, \\ B & \text{if } t > B. \end{cases} \tag{2.1}$$

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

Since the target function f_ρ^τ takes value in $[-M_\tau, M_\tau]$ almost surely, it is natural to measure the approximation ability of $\hat{f}_{\mathbf{z}}^\tau$ by the error $\|\pi_{M_\tau}(\hat{f}_{\mathbf{z}}^\tau) - f_\rho^\tau\|_{L^r_{\rho_X}}$, where $L^r_{\rho_X}$ is a weighted L^r-space with the norm $\|f\|_{L^r_{\rho_X}} = \left(\int_X |f(x)|^r d\rho_X\right)^{1/r}$. Here the index $r > 0$ depends on the pair (ρ, τ) and takes the value $r = \frac{pq}{p+1}$ when the following noise condition on ρ is satisfied.

Definition 2. *Let $p \in (0, \infty]$ and $q \in [1, \infty)$. A distribution ρ on $X \times \mathbb{R}$ is said to have a $\tau-$quantile of $p-$average type q if for almost every $x \in X$ with respect to ρ_X, there exit a $\tau-$quantile $t^* \in \mathbb{R}$ and constants $0 < a_x \leq 1$, $b_x > 0$ such that for each $s \in [0, a_x]$,*

$$\rho((t^* - s, t^*)|x) \geq b_x s^{q-1} \text{ and } \rho((t^*, t^* + s)|x) \geq b_x s^{q-1}, \tag{2.2}$$

and that the function on X taking values $(b_x a_x^{q-1})^{-1}$ at $x \in X$ lies in $L^p_{\rho_X}$.

Condition (2.2) ensures the uniqueness of the conditional $\tau-$quantile function f_ρ^τ and the singleton assumption on F_ρ^τ. For more details and examples about this definition, see [20, 21] and references therein.

Denoted by $H^s(\mathbb{R}^n)$ the Sobolev space [15] with index $s > 0$ and for $p \in (0, \infty]$ and $q \in (1, \infty)$, we set

$$\theta = \min\left\{\frac{2}{q}, \frac{p}{p+1}\right\} \in (0, 1]. \tag{2.3}$$

Our main results are stated as follows.

Theorem 1. *Suppose that assumption (1.2) holds with $M_\tau \geq 1$, ρ has a $\tau-$quantile of $p-$average type q with some $p \in (0, \infty]$ and $q \in [1, \infty)$ and satisfies assumption (1.5). Assume that for some $s > 0$, f_ρ^τ is the restriction of some $\tilde{f}_\rho^\tau \in H^s(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ onto X and the density function $h = \frac{d\rho_X}{dx}$ exists and lies in $L^2(X)$. Take $\sigma = m^{-\alpha}$ with $0 < \alpha < \frac{1}{2(n+1)}$ and $\lambda = m^{-\beta}$ with $\beta > (n+s)\alpha$. Then with $r = \frac{pq}{p+1}$, for any $0 < \epsilon < \Theta/q$ and $0 < \delta < 1$, with confidence $1 - \delta$, we have*

$$\|\pi_{M_\tau}(\hat{f}_{\mathbf{z}}^\tau) - f_\rho^\tau\|_{L^r_{\rho_X}} \leq C^\epsilon_{X,\rho,\alpha,\beta} \left(\log\frac{5}{\delta}\right)^{1/q} m^{\epsilon - \frac{\Theta}{q}}, \tag{2.4}$$

where $C^\epsilon_{X,\rho,\alpha,\beta}$ is a constant independent of m or δ and

$$\Theta = \min\left\{\frac{1 - 2(n+1)\alpha}{2 - \theta}, \beta - (n+s)\alpha, \alpha s\right\}. \tag{2.5}$$

Let $\alpha = \frac{1}{2(n+1)+(2-\theta)s}$ and $\beta = \frac{n+2s}{2(n+1)+(2-\theta)s}$, the convergence rate given by (2.4) is $\mathcal{O}(m^{\epsilon - \frac{s}{q(2(n+1)+(2-\theta)s)}})$ with an arbitrarily small (but fixed) $\epsilon > 0$. Recall that, under the boundedness assumption for y, the convergence rate of algorithm (1.3) presented in [33] is $\mathcal{O}(m^{-\frac{s}{q(2(n+1)+(2-\theta)s)}})$. Actually when y is bounded, a tiny modification in our proof will yield the same learning rate. An improved bound can be achieved if ρ_X is supported in the closed unit ball of $\mathbb{R}^n$.

Theorem 2. *If X is contained in the closed unit ball of $\mathbb{R}^n$, under the same assumptions of Theorem 1, let $\sigma = m^{-\alpha}$ with $\alpha < \frac{1}{n}$, $\lambda = m^{-\beta}$ with $\beta > (n+s)\alpha$ and $r = \frac{pq}{p+1}$, then for any $0 < \epsilon < \Theta'/q$ and $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\|\pi_{M_\tau}(\hat{f}_{\mathbf{z}}^{\tau}) - f_\rho^\tau\|_{L^r_{\rho_X}} \leq \widetilde{C}^{\epsilon}_{X,\rho,\alpha,\beta} \left(\log \frac{5}{\delta}\right)^{1/q} m^{\epsilon - \frac{\Theta'}{q}}, \tag{2.6}$$

where $\widetilde{C}^{\epsilon}_{X,\rho,\alpha,\beta}$ is a constant independent of m or δ and

$$\Theta' = \min\left\{\frac{1-n\alpha}{2-\theta}, \beta - (n+s)\alpha, \alpha s\right\}. \tag{2.7}$$

In Theorem 2, we further set $\alpha = \frac{1}{n+(2-\theta)s}$ and $\beta = \frac{n+2s}{n+(2-\theta)s}$, and the convergence rate given by (2.6) is $\mathcal{O}(m^{\epsilon - \frac{s}{q(n+(2-\theta)s)}})$. This rate is exactly the same as that of algorithm (1.3) obtained in [8] for bounded output y. Based on these observations, we claim that the approximation ability of algorithm (1.4) is comparable with that of the RKHS-based algorithm (1.3). Considering that ℓ_1-regularized quantile regression is essentially a linear optimization problem and often leads to sparse solutions, the algorithm (1.4) may perform even better than RKHS-based algorithm (1.3) for large data sets. At the end of this section, we give an example to illustrate our main results.

Proposition 1. *Let X be a compact subset of $\mathbb{R}^n$ with Lipschitz boundary and ρ_X be the uniform distribution on X. For $x \in X$, the conditional distribution $\rho(\cdot|x)$ is a normal distribution with mean $f_\rho(x)$ and variance σ_x^2. If $\vartheta_1 := \sup_{x\in X} |f_\rho(x)| < \infty$, $\vartheta_2 := \sup_{x\in X} \sigma_x \leq 1$ and $f_\rho \in H^s(X)$ with $s > \frac{n}{2}$, let $\sigma = m^{-\frac{1}{2(n+1)+s}}$, $\lambda = m^{-\frac{n+2s}{2(n+1)+s}}$ and $\hat{f}_{\mathbf{z}}^{\frac{1}{2}}$ be given by algorithm (1.4) with $\tau = \frac{1}{2}$, then for $0 < \epsilon < \frac{s}{2s+4(n+1)}$ and $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\|\pi_{\vartheta_1}(\hat{f}_{\mathbf{z}}^{\frac{1}{2}}) - f_\rho\|_{L^2_{\rho_X}} \leq c_\epsilon \left(\log \frac{5}{\delta}\right)^{1/2} m^{\epsilon - \frac{s}{2s+4(n+1)}}, \tag{2.8}$$

where $c_\epsilon > 0$ is a constant independent of m or δ. Furthermore, if X is contained in the unit ball of $\mathbb{R}^n$, take $\sigma = m^{-\frac{1}{n+s}}$ and $\lambda = m^{-\frac{n+2s}{n+s}}$, then for $0 < \epsilon < \frac{s}{2s+2n}$, with confidence $1-\delta$, there holds

$$\|\pi_{\vartheta_1}(\hat{f}_{\mathbf{z}}^{\frac{1}{2}}) - f_\rho\|_{L^2_{\rho_X}} \leq \tilde{c}_\epsilon \left(\log \frac{5}{\delta}\right)^{1/2} m^{\epsilon - \frac{s}{2s+2n}}, \tag{2.9}$$

where $\tilde{c}_\epsilon > 0$ is a constant independent of m or δ.

Remark 1. *Although we evaluate the approximation ability of the estimator $\hat{f}_{\mathbf{z}}^{\tau}$ by its projection $\pi_{M_\tau}(\hat{f}_{\mathbf{z}}^{\tau})$, the error bounds still hold true for $\pi_B(\hat{f}_{\mathbf{z}}^{\tau})$ with some properly chosen $B := B(m) \geq M_\tau$. Actually, from the proofs of the main results, one can see that B will tend to infinity as the sample size increases. The analysis approach in this paper is also*

applicable to investigate the learning behavior of ℓ_1-regularized quantile regression with a fixed Mercer kernel. When $q = 2$ and the conditional τ-quantile function f_ρ^τ is smooth enough (meaning that the parameter s is large enough), the learning rates presented above can be arbitrarily close to $\frac{p+1}{2(p+2)}$. However, if one estimates f_ρ^τ by the same scheme associated with a fixed Mercer kernel, similar convergence rates can be achieved under a regularity condition that f_ρ^τ lies in the range of powers of an integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ defined by $L_K(f)(x) = \int_X K(x,y)f(y)d\rho_X(y)$. Specially, when applying the same algorithm with a single fixed Gaussian kernel, the same convergence behavior for approximating f_ρ^τ may actually require a very restrictive condition $f_\rho^\tau \in C^\infty$. Furthermore, the results of [23] indicate that, the approximation ability of a Gaussian kernel with a fixed variance is limited, one can not expect obtaining the polynomial decay rates for target functions of Sobolev smoothness.

3 Framework of Convergence Analysis

In this section, we establish the framework of convergence analysis for algorithm (1.4). Given $f : X \to \mathbb{R}$, the *generalization error* associated with the pinball loss L_τ is defined as

$$\mathcal{E}^\tau(f) = \int_{X\times Y} L_\tau(f(x) - y)d\rho.$$

We first state a result which plays an important role in our mathematical analysis.

Proposition 2. *Suppose that assumption (1.2) with $M_\tau \geq 1$ holds and ρ has a $\tau-$quantile of $p-$average type q. Then for any $f : X \to [-B, B]$, we have*

$$\|f - f_\rho^\tau\|_{L^r_{\rho_X}} \leq c_\rho \max\{B, M_\tau\}^{1-1/q} \left\{\mathcal{E}^\tau(f) - \mathcal{E}^\tau(f_\rho^\tau)\right\}^{1/q}, \tag{3.1}$$

where $r = \frac{pq}{p+1}$ and $c_\rho = 2^{1-1/q}q^{1/q}\|\{(b_x a_x^{q-1})^{-1}\}_{x\in X}\|^{1/q}_{L^p_{\rho_X}}$.

This proposition can be proved following the same idea in [21], and we move the proof to the Appendix just for completeness. By Proposition 2, in order to estimate error $\|\pi_B(\hat{f}_{\mathbf{z}}^\tau) - f_\rho^\tau\|$ in the $L^r_{\rho_X}-$space, we only need to bound the *excess generalization error* $\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau)$. This will be done by conducting an error decomposition which has been developed in the literature for RKHS-based regularization schemes (e.g. [6, 30]). A technical difficulty in our setting here is that the centers x_i of the basis functions in $\mathscr{H}_{\mathbf{z},\sigma}$ are determined by the sample $\mathbf{z}$ and cannot be freely chosen. One might consider regularization schemes in the infinite dimensional space of all linear combinations with $\{K^\sigma(x,t)|t \in X\}$. But due to the lack of a Reprensenter Theorem, the minimization

in such kind of space can not be reduced to a convex optimization problem in a finite dimensional space like (1.4).

In this paper, we shall overcome this difficulty by a stepping stone method [29]. We use $\hat{f}^{\tau}_{\mathbf{z},\gamma}$ to denote the solution of algorithm (1.3) with a regularization parameter γ, i.e.

$$\hat{f}^{\tau}_{\mathbf{z},\gamma} = \arg\min_{f\in\mathscr{H}_\sigma} \left\{ \frac{1}{m}\sum_{i=1}^{m} L_\tau(f(x_i)-y_i) + \gamma\|f\|^2_\sigma \right\}. \tag{3.2}$$

Note that $\hat{f}^{\tau}_{\mathbf{z},\gamma}$ belongs to $\mathscr{H}_{\mathbf{z},\sigma}$ and is a reasonable estimator for f^{τ}_{ρ}. We expect then that $\hat{f}^{\tau}_{\mathbf{z},\gamma}$ might play a stepping stone role in the analysis for the algorithm (1.4), which will establish a close relation between $\hat{f}^{\tau}_{\mathbf{z}}$ and f^{τ}_{ρ}. To this end, we need to estimate $\Omega(\hat{f}^{\tau}_{\mathbf{z},\gamma})$, the ℓ_1-norm of the coefficients in the kernel expression for $\hat{f}^{\tau}_{\mathbf{z},\gamma}$.

Lemma 1. *For every $\gamma>0$, the function $\hat{f}^{\tau}_{\mathbf{z},\gamma}$ defined by (3.2) satisfies*

$$\Omega(\hat{f}^{\tau}_{\mathbf{z},\gamma}) \le \frac{1}{2\gamma m}\sum_{i=1}^{m} L_\tau(\hat{f}^{\tau}_{\mathbf{z},\gamma}(x_i)-y_i) + \frac{1}{2\gamma} + \frac{1}{2}\|\hat{f}^{\tau}_{\mathbf{z},\gamma}\|^2_\sigma. \tag{3.3}$$

Proof. Setting $C=\frac{1}{2\gamma m}$ and introducing the slack variables, we can restate the optimization problem (3.2) as

$$\begin{array}{ll} \underset{f\in\mathscr{H}_\sigma,\xi_i\in\mathbb{R},\tilde{\xi}_i\in\mathbb{R}}{\text{minimize}} & \frac{1}{2}\|f\|^2_\sigma + C\sum_{i=1}^{m}\left\{(1-\tau)\xi_i+\tau\tilde{\xi}_i\right\} \\ \text{subject to} & f(x_i)-y_i\le\xi_i, \\ & y_i-f(x_i)\le\tilde{\xi}_i, \\ & \xi_i\ge 0,\ \tilde{\xi}_i\ge 0, \quad \text{for all } i=1,\cdots,m. \end{array} \tag{3.4}$$

The Lagrangian $\mathcal{L}$ associated with problem (3.4) is given by

$$\begin{aligned}\mathcal{L}(f,\xi,\tilde{\xi},\alpha,\tilde{\alpha},\beta,\tilde{\beta}) = \frac{1}{2}\|f\|^2_\sigma + C\sum_{i=1}^{m}\left\{(1-\tau)\xi_i+\tau\tilde{\xi}_i\right\} + \sum_{i=1}^{m}\alpha_i(f(x_i)-y_i-\xi_i) \\ + \sum_{i=1}^{m}\tilde{\alpha}_i(y_i-f(x_i)-\tilde{\xi}_i) - \sum_{i=1}^{m}\beta_i\xi_i - \sum_{i=1}^{m}\tilde{\beta}_i\tilde{\xi}_i.\end{aligned}$$

Denoting the inner product of $\mathscr{H}_\sigma$ as $\langle,\rangle_\sigma$, then for any $f\in\mathscr{H}_\sigma$, we have $\|f\|^2_\sigma=\langle f,f\rangle_\sigma$ and the reproducing property of $\mathscr{H}_\sigma$ [1] ensures that $f(x_i)=\langle f,K^\sigma(\cdot,x_i)\rangle_\sigma$. Considering $\mathcal{L}$ as a functional from $\mathscr{H}_\sigma$ to $\mathbb{R}$, the Fréchet derivative of $\mathcal{L}$ at $f\in\mathscr{H}_\sigma$ is written as $\frac{\partial\mathcal{L}}{\partial\mathscr{H}_\sigma}(f)$. We hence have $\frac{\partial\mathcal{L}}{\partial\mathscr{H}_\sigma}(f)=f+\sum_{i=1}^{m}\alpha_iK^\sigma(x,x_i)-\sum_{i=1}^{m}\tilde{\alpha}_iK^\sigma(x,x_i),\forall f\in\mathscr{H}_\sigma$. In order to derive the dual problem of (3.4), we first let

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathscr{H}_\sigma}(f)=0 &\to f+\sum_{i=1}^{m}\alpha_iK^\sigma(x,x_i)-\sum_{i=1}^{m}\tilde{\alpha}_iK^\sigma(x,x_i)=0,\\ \frac{\partial\mathcal{L}}{\partial\xi_i}=0 &\to C(1-\tau)-\alpha_i-\beta_i=0,\quad i=1,\cdots,m\\ \frac{\partial\mathcal{L}}{\partial\tilde{\xi}_i}=0 &\to C\tau-\tilde{\alpha}_i-\tilde{\beta}_i=0,\quad i=1,\cdots,m.\end{aligned}$$

From the above equations, we represent $(f, \xi, \tilde{\xi})$ by $(\alpha, \tilde{\alpha}, \beta, \tilde{\beta})$ and substitute them back into $\mathcal{L}$. Note that as $\alpha_i, \tilde{\alpha}_i, \beta_i, \tilde{\beta}_i \geq 0$, the equality constraints $C(1-\tau) - \alpha_i - \beta_i = 0$ and $C\tau - \tilde{\alpha}_i - \tilde{\beta}_i = 0$ amount to inequality constraints $0 \leq \alpha_i \leq C(1-\tau)$ and $0 \leq \tilde{\alpha}_i \leq C\tau$. Thus we can formulate the dual optimization problem of (3.4) as

$$\begin{array}{ll} \underset{\alpha_i \in \mathbb{R}, \tilde{\alpha}_i \in \mathbb{R}}{\text{maximize}} & \sum_{i=1}^m y_i(\tilde{\alpha}_i - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^m (\tilde{\alpha}_i - \alpha_i)(\tilde{\alpha}_j - \alpha_j) K^\sigma(x_i, x_j) \\ \text{subject to} & 0 \leq \alpha_i \leq C(1-\tau), \\ & 0 \leq \tilde{\alpha}_i \leq C\tau, \quad \text{for all } i = 1, \cdots, m. \end{array} \tag{3.5}$$

Here we also use the reproducing property to obtain that $\|f\|_\sigma^2 = \sum_{i,j=1}^m (\tilde{\alpha}_i - \alpha_i)(\tilde{\alpha}_j - \alpha_j) K^\sigma(x_i, x_j)$ for $f = \sum_{i=1}^m (\tilde{\alpha}_i - \alpha_i) K^\sigma(x, x_i)$. We denote the unique solution of (3.4) by $(f^*, \xi^*, \tilde{\xi}^*)$, then $f^* = \hat{f}^\tau_{\mathbf{z},\gamma}$. Furthermore, if $(\alpha_1^*, \tilde{\alpha}_1^*, \cdots, \alpha_m^*, \tilde{\alpha}_m^*)$ denotes the solution of (3.5), by the KKT conditions, we have

$$\begin{aligned} f^* &= \sum_{i=1}^m (\tilde{\alpha}_i^* - \alpha_i^*) K^\sigma(x_i, x), \\ \xi_i^* &= \max\{0, f^*(x_i) - y_i\}, \\ \tilde{\xi}_i^* &= \max\{0, y_i - f^*(x_i)\}, \end{aligned}$$

and

$$\begin{aligned} &\alpha_i^*(f^*(x_i) - y_i - \xi_i^*) = 0, \\ &\tilde{\alpha}_i^*(y_i - f^*(x_i) - \tilde{\xi}_i^*) = 0, \\ &(C(1-\tau) - \alpha_i^*)\xi_i^* = 0, \\ &(C\tau - \tilde{\alpha}_i^*)\tilde{\xi}_i^* = 0. \end{aligned}$$

By setting $\kappa_i^* = \tilde{\alpha}_i^* - \alpha_i^*$, then $\hat{f}^\tau_{\mathbf{z},\gamma} = \sum_{i=1}^m \kappa_i^* K^\sigma(x, x_i)$. From the definition of $\{(\alpha_i^*, \tilde{\alpha}_i^*)\}_{i=1}^m$, we have $\sum_{i=1}^m y_i \kappa_i^* - \frac{1}{2}\sum_{i,j=1}^m \kappa_i^* \kappa_j^* K^\sigma(x_i, x_j) \geq 0$, hence

$$\begin{aligned} \sum_{i=1}^m |\kappa_i^*| &\leq \sum_{i=1}^m \kappa_i^*(y_i + \operatorname{sgn}(\kappa_i^*)) - \frac{1}{2}\sum_{i,j=1}^m \kappa_i^* \kappa_j^* K^\sigma(x_i, x_j) \\ &= \sum_{i=1}^m \kappa_i^*(y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) + \operatorname{sgn}(\kappa_i^*)) + \frac{1}{2}\|\hat{f}^\tau_{\mathbf{z},\gamma}\|_\sigma^2, \end{aligned}$$

where $\operatorname{sgn}(\kappa_i^*)$ is defined by $\operatorname{sgn}(\kappa_i^*) = 1$ if $\kappa_i^* \geq 0$ and $\operatorname{sgn}(\kappa_i^*) = -1$ otherwise.

If $y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) > 0$, then $\tilde{\xi}_i^* > 0$ and $\xi_i^* = 0$, the KKT conditions imply that $\tilde{\alpha}_i^* = C\tau$ and $\alpha_i^* = 0$. Hence $\kappa_i^* = C\tau$ and

$$\kappa_i^*(y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) + \operatorname{sgn}(\kappa_i^*)) = C\tau(y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) + 1) \leq CL_\tau(\hat{f}^\tau_{\mathbf{z},\gamma}(x_i) - y_i) + C.$$

Similarly, if $y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) < 0$, we have $\kappa_i^* = -C(1-\tau)$ and

$$\kappa_i^*(y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) + \operatorname{sgn}(\kappa_i^*)) = -C(1-\tau)(y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) - 1) \leq CL_\tau(\hat{f}^\tau_{\mathbf{z},\gamma}(x_i) - y_i) + C.$$

When $y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) = 0$, it directly yields

$$\kappa_i^*(y_i - \hat{f}^\tau_{\mathbf{z},\gamma}(x_i) + \mathrm{sgn}(\kappa_i^*)) = |\kappa_i^*| \le |\tilde{\alpha}_i^*| + |\alpha_i^*| \le C.$$

Therefore,

$$\sum_{i=1}^m |\kappa_i^*| \le \sum_{i=1}^m C(1 + L_\tau(\hat{f}^\tau_{\mathbf{z},\gamma}(x_i) - y_i)) + \frac{1}{2}\|\hat{f}^\tau_{\mathbf{z},\gamma}\|_\sigma^2,$$

and the bound for $\Omega(\hat{f}^\tau_{\mathbf{z},\gamma})$ follows. □

Additionally, we need the following lemma to estimate the approximation performance of Gaussian kernels.

Lemma 2. *Let $s > 0$. Assume f_ρ^τ is the restriction of some $\tilde{f}_\rho^\tau \in H^s(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ onto X, and the density function $h = \frac{d\rho_X}{dx}$ exists and lies in $L^2(X)$. Then we can find $\{f^\tau_{\sigma,\gamma} \in \mathscr{H}_\sigma : 0 < \sigma \le 1, \gamma > 0\}$ such that*

$$\|f^\tau_{\sigma,\gamma}\|_{L^\infty(X)} \le \widetilde{B}, \tag{3.6}$$

and

$$\widetilde{\mathscr{D}}(\gamma,\sigma) := \mathcal{E}^\tau(f^\tau_{\sigma,\gamma}) - \mathcal{E}^\tau(f_\rho^\tau) + \gamma\|f^\tau_{\sigma,\gamma}\|_\sigma^2 \le \widetilde{B}(\sigma^s + \gamma\sigma^{-n}), \quad \forall 0 < \sigma \le 1, \gamma > 0, \tag{3.7}$$

where $\widetilde{B} \ge 1$ is a constant independent of σ or γ.

An early version of Lemma 2 associated with a general loss function was proved by [32] for regularized classification schemes. Since the pinball loss is Lipschitz continuous, the proof of Lemma 2 is exactly the same as [32]. The function sequence $\{f^\tau_{\sigma,\gamma}\}$ is constructed by means of a convolution type scheme with a Fourier analysis technique. Lemma 2 was firstly applied in [33] to analyze the conditional quantile regression algorithm (1.3). Recently, a more general version is presented by [8].

Define the *empirical error* associated with pinball loss as

$$\mathcal{E}^\tau_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m L_\tau(f(x_i) - y_i) \quad \text{for } f : X \to \mathbb{R}.$$

The error decomposition process is given by the following proposition.

Proposition 3. *Let $(\lambda,\sigma,\gamma) \in (0,1]^3$, $\hat{f}^\tau_{\mathbf{z}}$ be defined by (1.4) and $f^\tau_{\sigma,\gamma} \in \mathscr{H}_\sigma$ satisfying (3.6) and (3.7). Then for any $B > 0$, there holds*

$$\mathcal{E}^\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})) - \mathcal{E}^\tau(f_\rho^\tau) \le \mathscr{S}_1 + \mathscr{S}_2 + \mathscr{S}_3 + \mathscr{D}, \tag{3.8}$$

where

$$\mathscr{S}_1 = \left\{\mathcal{E}^\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})) - \mathcal{E}^\tau(f^\tau_\rho)\right\} - \left\{\mathcal{E}^\tau_{\mathbf{z}}(\pi_B(\hat{f}^\tau_{\mathbf{z}})) - \mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho)\right\},$$

$$\mathscr{S}_2 = \left(1+\frac{\lambda}{2\gamma}\right)\left(\left\{\mathcal{E}^\tau_{\mathbf{z}}(f^\tau_{\sigma,\gamma}) - \mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho)\right\} - \left\{\mathcal{E}^\tau(f^\tau_{\sigma,\gamma}) - \mathcal{E}^\tau(f^\tau_\rho)\right\}\right),$$

$$\mathscr{S}_3 = \frac{1}{m}\sum_{i=1}^m |\pi_B(y_i) - y_i| + \frac{\lambda}{2\gamma}\left(\mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho) - \mathcal{E}^\tau(f^\tau_\rho)\right),$$

$$\mathscr{D} = \left(1+\frac{\lambda}{2\gamma}\right)\widetilde{\mathscr{D}}(\gamma,\sigma) + \frac{\lambda}{2\gamma}(1+\mathcal{E}^\tau(f^\tau_\rho)).$$

Proof. Recall the definition of the projection operator π_B, for any given $a, b \in \mathbb{R}$, if $a \geq b$, simple calculation shows that

$$\pi_B(a) - \pi_B(b) = \begin{cases} 0 & \text{if } a \geq b \geq B \text{ or } -B \geq a \geq b \\ \min\{a, B\} + \min\{-b, B\} & \text{otherwise .} \end{cases}$$

Then we have $0 \leq \pi_B(a) - \pi_B(b) \leq a - b$ if $a \geq b$. Similarly, when $a \leq b$, we have $a - b \leq \pi_B(a) - \pi_B(b) \leq 0$. Hence for any $(x, y) \in Z$ and $f : X \to \mathbb{R}$, there holds

$$L_\tau(\pi_B(f)(x) - \pi_B(y)) \leq L_\tau(f(x) - y).$$

From the definition of $\hat{f}^\tau_{\mathbf{z}}$ (1.4), we have

$$\begin{aligned}
\mathcal{E}^\tau_{\mathbf{z}}(\pi_B(\hat{f}^\tau_{\mathbf{z}})) + \lambda\Omega(\hat{f}^\tau_{\mathbf{z}}) &= \frac{1}{m}\sum_{i=1}^m L_\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})(x_i) - y_i) + \lambda\Omega(\hat{f}^\tau_{\mathbf{z}}) \\
&\leq \frac{1}{m}\sum_{i=1}^m L_\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})(x_i) - \pi_B(y_i)) + \lambda\Omega(\hat{f}^\tau_{\mathbf{z}}) + \frac{1}{m}\sum_{i=1}^m |\pi_B(y_i) - y_i| \\
&\leq \mathcal{E}^\tau_{\mathbf{z}}(\hat{f}^\tau_{\mathbf{z}}) + \lambda\Omega(\hat{f}^\tau_{\mathbf{z}}) + \frac{1}{m}\sum_{i=1}^m |\pi_B(y_i) - y_i| \\
&\leq \mathcal{E}^\tau_{\mathbf{z}}(\hat{f}^\tau_{\mathbf{z},\gamma}) + \lambda\Omega(\hat{f}^\tau_{\mathbf{z},\gamma}) + \frac{1}{m}\sum_{i=1}^m |\pi_B(y_i) - y_i|,
\end{aligned}$$

where $\hat{f}^\tau_{\mathbf{z},\gamma}$ is defined by (3.2). Lemma 1 gives $\Omega(\hat{f}^\tau_{\mathbf{z},\gamma}) \leq \frac{1}{2\gamma}\mathcal{E}^\tau_{\mathbf{z}}(\hat{f}^\tau_{\mathbf{z},\gamma}) + \frac{1}{2\gamma} + \frac{1}{2}\|\hat{f}^\tau_{\mathbf{z},\gamma}\|^2_\sigma$, hence

$$\mathcal{E}^\tau_{\mathbf{z}}(\pi_B(\hat{f}^\tau_{\mathbf{z}})) + \lambda\Omega(\hat{f}^\tau_{\mathbf{z}}) \leq \left(1+\frac{\lambda}{2\gamma}\right)\left\{\mathcal{E}^\tau_{\mathbf{z}}(\hat{f}^\tau_{\mathbf{z},\gamma}) + \gamma\|\hat{f}^\tau_{\mathbf{z},\gamma}\|^2_\sigma\right\} + \frac{\lambda}{2\gamma} + \frac{1}{m}\sum_{i=1}^m |\pi_B(y_i) - y_i|.$$

This enables us to bound $\mathcal{E}^\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})) + \lambda\Omega(\hat{f}^\tau_{\mathbf{z}})$ by

$$\left\{\mathcal{E}^\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})) - \mathcal{E}^\tau_{\mathbf{z}}(\pi_B(\hat{f}^\tau_{\mathbf{z}}))\right\} + \left(1+\frac{\lambda}{2\gamma}\right)\left\{\mathcal{E}^\tau_{\mathbf{z}}(\hat{f}^\tau_{\mathbf{z},\gamma}) + \gamma\|\hat{f}^\tau_{\mathbf{z},\gamma}\|^2_\sigma\right\} + \frac{\lambda}{2\gamma} + \frac{1}{m}\sum_{i=1}^m |\pi_B(y_i) - y_i|.$$

Next, we further bound $\mathcal{E}^\tau_{\mathbf{z}}(\hat{f}^\tau_{\mathbf{z},\gamma}) + \gamma\|\hat{f}^\tau_{\mathbf{z},\gamma}\|^2_\sigma$ by Lemma 2. Let $f^\tau_{\sigma,\gamma} \in \mathscr{H}_\sigma$ be the functions constructed in Lemma 2, the definition of $\hat{f}^\tau_{\mathbf{z},\gamma}$ (3.2) tells us that

$$\mathcal{E}^\tau_{\mathbf{z}}(\hat{f}^\tau_{\mathbf{z},\gamma}) + \gamma\|\hat{f}^\tau_{\mathbf{z},\gamma}\|^2_\sigma \leq \mathcal{E}^\tau_{\mathbf{z}}(f^\tau_{\sigma,\gamma}) + \gamma\|f^\tau_{\sigma,\gamma}\|^2_\sigma = \left\{\mathcal{E}^\tau_{\mathbf{z}}(f^\tau_{\sigma,\gamma}) - \mathcal{E}^\tau(f^\tau_{\sigma,\gamma})\right\} + \mathcal{E}^\tau(f^\tau_{\sigma,\gamma}) + \gamma\|f^\tau_{\sigma,\gamma}\|^2_\sigma.$$

Combining the above two steps, we find that $\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau) + \lambda\Omega(\hat{f}_{\mathbf{z}}^\tau)$ is bounded by

$$\left\{\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}_{\mathbf{z}}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau))\right\} + \left(1+\frac{\lambda}{2\gamma}\right)\left\{\mathcal{E}_{\mathbf{z}}^\tau(f_{\sigma,\gamma}^\tau) - \mathcal{E}^\tau(f_{\sigma,\gamma}^\tau)\right\} + \frac{1}{m}\sum_{i=1}^m |\pi_B(y_i) - y_i|$$
$$+ \left(1+\frac{\lambda}{2\gamma}\right)\left\{\mathcal{E}^\tau(f_{\sigma,\gamma}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) + \gamma\|f_{\sigma,\gamma}^\tau\|_\sigma^2\right\} + \frac{\lambda}{2\gamma}(1+\mathcal{E}^\tau(f_\rho^\tau)).$$

Note that this bound is exactly $\mathscr{S}_1+\mathscr{S}_2+\mathscr{S}_3+\mathscr{D}$ and by the fact $\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau) \le \mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau) + \lambda\Omega(\hat{f}_{\mathbf{z}}^\tau)$, we draw our conclusion. □

With the help of Proposition 3, the excess generalization error is estimated by bounding $\mathscr{S}_i$ $(i = 1, 2, 3)$ and $\mathscr{D}$ respectively. Since the assumptions (1.2) and (1.5) imply that

$$\mathcal{E}^\tau(f_\rho^\tau) \le M_\tau + cM, \tag{3.9}$$

Lemma 2 immediately yields the estimates for $\mathscr{D}$. Our error analysis mainly focuses on how to estimate $\mathscr{S}_i$. We expect that $\mathscr{S}_i$ will tend to zero at a certain rate as the sample size tends to infinity. The asymptotical behaviors of $\mathscr{S}_i$ are usually illustrated by the convergence of the empirical mean $\frac{1}{m}\sum_{i=1}^m \xi_i$ to its expectation $\mathbb{E}\xi$, where $\{\xi_i\}_{i=1}^m$ are independent random variables on (Z, ρ). To be more concrete, in order to estimate $\mathscr{S}_1$ and $\mathscr{S}_2$, we define random variables as

$$\xi_i := \xi(z_i) = L_\tau(f(x_i) - y_i) - L_\tau(f_\rho^\tau(x_i) - y_i), \tag{3.10}$$

where f belongs to a bounded function set on X. Note that the Lipschitz property of the pinball loss guarantees the boundedness of ξ_i when f is bounded. So ξ_i defined by (3.10) are bounded random variables even if y_i is unbounded. When f is fixed, which is exactly the case as we estimate $\mathscr{S}_2$, the convergence is guaranteed by the following probability inequality [7].

Lemma 3. *Let ξ be a random variable on Z with mean $\mathbb{E}\xi$. Assume that $\mathbb{E}\xi \ge 0$, $|\xi - \mathbb{E}\xi| \le Q$ almost everywhere, and $\mathbb{E}\xi^2 \le c_1(\mathbb{E}\xi)^\theta$ for some $0 \le \theta \le 1$ and $c_1, Q \ge 0$. Then for every $\epsilon > 0$ there holds*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi}{\sqrt{(\mathbb{E}\xi)^\theta + \epsilon^\theta}} > \epsilon^{1-\frac{\theta}{2}}\right\} \le \exp\left\{-\frac{m\epsilon^{2-\theta}}{2c_1 + \frac{2}{3}Q\epsilon^{1-\theta}}\right\}. \tag{3.11}$$

When the random variables are given by (3.10), for a general distribution ρ, the variance-expectation condition $\mathbb{E}\xi^2 \le c_1(\mathbb{E}\xi)^\theta$ is satisfied with $\theta = 0$ and $c_1 = 1$. If ρ satisfies the noise condition (i.e. Definition 2), the following lemma provides an improved bound with $\theta > 0$.

Lemma 4. *Under the same assumptions of Proposition 2, for any $f : X \to [-B, B]$, there holds*

$$\mathbb{E}\left\{(L_\tau(f(x)-y) - L_\tau(f_\rho^\tau(x)-y))^2\right\} \le C_\theta \max\{B, M_\tau\}^{2-\theta} \left(\mathcal{E}^\tau(f) - \mathcal{E}^\tau(f_\rho^\tau)\right)^\theta, \quad (3.12)$$

where θ is given by (2.3) and $C_\theta = 2^{2-\theta} q^\theta \|\{(b_x a_x^{q-1})^{-1}\}_{x\in X}\|^\theta_{L^p_{\rho_X}}$.

This lemma is a direct corollary of Proposition 2 and has been proved in [21, 33] with $B = M_\tau = 1$. We shall omit the proof here. The positive θ will lead to sharper estimates and play an essential role in the convergence analysis.

The only difference between $\mathscr{S}_1$ and $\mathscr{S}_2$ is that the first term involves $\hat{f}_{\mathbf{z}}^\tau$ which varies with samples. Thus a uniform concentration inequality for a family of functions containing $\hat{f}_{\mathbf{z}}^\tau$ is needed to estimate $\mathscr{S}_1$. Since $\hat{f}_{\mathbf{z}}^\tau \in \mathscr{H}_\sigma$, let $\mathscr{B}_R^\sigma = \{f \in \mathscr{H}_\sigma : \|f\|_\sigma \le R\}$, we shall bound $\mathscr{S}_1$ by the following concentration inequality with a properly chosen R.

Lemma 5. *Under the same assumptions of Proposition 2 and θ is given by (2.3), for $R \ge 1$, $\Delta \ge 1$ and $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned}\left\{\mathcal{E}^\tau(\pi_B(f)) - \mathcal{E}^\tau(f_\rho^\tau)\right\} - \left\{\mathcal{E}_{\mathbf{z}}^\tau(\pi_B(f)) - \mathcal{E}_{\mathbf{z}}^\tau(f_\rho^\tau)\right\} \le \frac{1}{2}\left\{\mathcal{E}^\tau(\pi_B(f) - \mathcal{E}^\tau(f_\rho^\tau)\right\} \\ + C_{X,\rho} \max\{B, M_\tau\} \eta_\delta m^{-\frac{1}{2-\theta}} + 20Rm^{-\Delta}, \quad \forall f \in \mathscr{B}_R^\sigma,\end{aligned} \quad (3.13)$$

where

$$\eta_\delta = \log\frac{1}{\delta} + \sigma^{-\frac{2(n+1)}{2-\theta}} + (\Delta \log m)^{\frac{n+1}{2-\theta}} \quad (3.14)$$

and $C_{X,\rho} > 0$ is a constant only depending on X and ρ.

This lemma will be proved in the Appendix by applying a standard covering number argument. As an important measurement of the capacity of a function set, covering numbers have been well studied in the literature (see [25] and references therein). For the sake of completeness, we recall the definition of covering numbers.

Definition 3. *Let $(\mathscr{M}, d)$ be a pseudo-metric space and $S \subset \mathscr{M}$ a subset. For every $\epsilon > 0$, the covering number $\mathscr{N}(S, \epsilon, d)$ of S with respect to ϵ and d is defined as the minimal number of balls of radius ϵ of which the union covers S, that is,*

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

where $B(s_j, \epsilon) = \{s \in \mathscr{M} : d(s, s_j) \le \epsilon\}$ is a ball in $\mathscr{M}$.

When S is a subset of the metric space $(\mathscr{C}(X), \|\cdot\|_\infty)$ of bounded continuous functions on X, the uniform covering numbers $\mathscr{N}(S, \epsilon, \|\cdot\|_\infty)$ are defined to be the covering numbers

with respect to the uniform metric $\|\cdot\|_\infty$. Note that $\mathscr{B}_1^\sigma$ is a compact set of $\mathscr{C}(X)$. The proof of Lemma 5 is mainly based on the asymptotical behavior of $\mathscr{N}(\mathscr{B}_1^\sigma, \epsilon, \|\cdot\|_\infty)$ [35]: there exists a constant C_X depending only on X and n such that

$$\log \mathscr{N}(\mathscr{B}_1^\sigma, \epsilon, \|\cdot\|_\infty) \leq C_X \left(\left(\log \frac{1}{\epsilon} \right)^{n+1} + \frac{1}{\sigma^{2(n+1)}} \right), \quad \forall \epsilon > 0,\ \sigma > 0. \tag{3.15}$$

The upper bound appearing in the right hand side of (3.15) is dependent on ϵ and σ, which enables us to derive convergence rates even when σ varies with the sample size.

Besides the uniform covering number, empirical covering number is another choice to measure the capacity of $\mathscr{H}_\sigma$. Let $\mathscr{F}$ be a set of functions on X and $\omega = \{\omega_1, \cdots, \omega_k\} \subset X^k$. The metric $d_{2,\omega}$ is defined on $\mathscr{F}$ by $d_{2,\omega}(f,g) = \left\{ \frac{1}{k} \sum_{i=1}^k (f(\omega_k) - g(\omega_k))^2 \right\}^{1/2}, \forall f, g \in \mathscr{F}$. For every $\epsilon > 0$, the ℓ_2-empirical covering number of $\mathscr{F}$ is defined as

$$\mathscr{N}_2(\mathscr{F}, \epsilon) = \sup_{k \in \mathbb{N}} \sup_{\omega \in X^k} \mathscr{N}(\mathscr{F}, \epsilon, d_{2,\omega}).$$

It follows from Theorem 7.34 in [19] that, if X is contained in the closed unit ball of $\mathbb{R}^n$, then for any $\nu > 0$ and $0 < \mu < 1$, there exists a constant $c_{\nu,\mu} > 0$ such that

$$\log \mathscr{N}_2(\mathscr{B}_1^\sigma, \epsilon) \leq c_{\mu,\nu} \sigma^{-(1-\mu)(1+\nu)n} \left(\frac{1}{\epsilon} \right)^{2\mu}. \tag{3.16}$$

Although the term $\epsilon^{-2\mu}$ increases polynomially, as μ and ν can be chosen arbitrarily small, the bound (3.16) is tighter than the bound (3.15) and thus can lead to sharper estimates. We can bound $\mathscr{S}_1$ by the following lemma when bound (3.16) comes into existence.

Lemma 6. *If X is contained in a unit ball of $\mathbb{R}^n$, then under the same assumptions of Proposition 2, with θ given by (2.3) and C_θ given by Lemma 4, for any $0 < \mu < 1$ and $\nu > 0$, there exists a constant $c_\mu > 0$ depending only on μ and a constant $c_{\mu,\nu} > 0$ depending only on μ, ν such that for $R > 1$ and $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\begin{aligned} &\left\{ \mathcal{E}^\tau(\pi_B(f)) - \mathcal{E}^\tau(f_\rho^\tau) \right\} - \left\{ \mathcal{E}_{\mathbf{z}}^\tau(\pi_B(f)) - \mathcal{E}_{\mathbf{z}}^\tau(f_\rho^\tau) \right\} \\ &\quad \leq \frac{1}{2} \eta_R^{1-\theta} \left\{ \mathcal{E}^\tau(\pi_B(f)) - \mathcal{E}^\tau(f_\rho^\tau) \right\}^\theta + c_\mu \eta_R \\ &\quad + \left(2C_\theta^{\frac{1}{2-\theta}} + 36 \right) \log \frac{1}{\delta} \max\{B, M_\tau\} m^{-\frac{1}{2-\theta}}, \qquad \forall f \in \mathscr{B}_R^\sigma, \end{aligned} \tag{3.17}$$

where

$$\eta_R = c_{\mu,\nu,\rho} \max\{B, M_\tau\}^{\frac{(2-\theta)(1-\mu)}{2-\theta+\mu\theta}} R^{\frac{2\mu}{1+\mu}} \left(\frac{\sigma^{-(1-\mu)(1+\nu)n}}{m} \right)^{\frac{1}{2-\theta+\mu\theta}} \tag{3.18}$$

and $c_{\mu,\nu,\rho} = C_\theta^{\frac{1-\mu}{2-\theta+\mu\theta}} c_{\mu,\nu}^{\frac{1}{2-\theta+\mu\theta}} + 2^{\frac{1-\mu}{1+\mu}} c_{\mu,\nu}^{\frac{1}{1+\mu}}$.

We also leave the proof to the Appendix. Similarly, by considering suitable random variables on (Z,ρ), $\mathscr{S}_3$ can also be estimated by bounding the difference between the empirical mean and the expectation. Since y_i is unbounded, our error analysis relies on the following probability inequality for unbounded random variables [3].

Lemma 7. *Let $X_1, X_2 \cdots, X_m$ be independent random variables with $\mathbb{E}X_i = 0$. If for some constants $M_1, v_1 > 0$, the bound $\mathbb{E}|X_i|^\ell \le \frac{1}{2}\ell! M_1^{\ell-2} v_1$ holds for every $2 \le \ell \in \mathbb{N}$, then*

$$\text{Prob}\left\{\sum_{i=1}^m X_i \ge \epsilon\right\} \le \exp\left\{-\frac{\epsilon^2}{2(mv_1 + M_1\epsilon)}\right\}, \qquad \forall \epsilon > 0.$$

4 Concentration Estimates

This section is devoted to estimating $\mathscr{S}_i$ $(i = 1, 2, 3)$ and deriving convergence rates. This is conducted by using the concentration inequalities mentioned in Section 3. We shall give the proofs of the main results after the following proposition.

Proposition 4. *Under the same assumptions of Theorem 1, take $\sigma = m^{-\alpha}$ and $\lambda = m^{-\beta}$ with $0 < \alpha < \frac{1}{2(n+1)}$ and $\beta > (n+s)\alpha$. For $B \ge M_\tau$, $k \in \mathbb{N}$ and $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned}\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) \le 5C_{X,\rho,\alpha,\beta} \log\frac{5}{\delta}\left\{Bm^{-\frac{1-2(n+1)\alpha}{2-\theta}} + m^{-(\beta-(n+s)\alpha)} + m^{-\alpha s}\right\} \\ + 2c\left\{(k+1)! + 2^{k+2}k^k\right\} M^{k+1}B^{-k} + 6M2^{k+1}\log\frac{5}{\delta}m^{-1},\end{aligned} \tag{4.1}$$

where

$$\begin{aligned}C_{X,\rho,\alpha,\beta} = \max\Bigg\{24(C_\theta+1)(\widetilde{B}+M_\tau), 25(M+M_\tau)(c+1), \\ 2C_{X,\rho}\left(2 + \left(\frac{1+\beta}{2e\alpha}\right)^{\frac{n+1}{2-\theta}}\right), 40(3cM+4M), 9\widetilde{B}\Bigg\}.\end{aligned} \tag{4.2}$$

Proof. We first bound $\mathscr{S}_2$ by considering the random variable $\xi(z) = L_\tau(f_{\sigma,\gamma}^\tau(x) - y) - L_\tau(f_\rho^\tau(x) - y)$ on (Z,ρ). From Lemma 2, $|\xi(z)| \le |f_{\sigma,\gamma}^\tau(x) - f_\rho^\tau(x)| \le \widetilde{B} + M_\tau$ for almost every $z \in Z$. By Lemma 4, the variance-expectation condition of $\xi(z)$ is satisfied with θ given by (2.3) and $c_1 = C_\theta \max\{\widetilde{B}, M_\tau\}^{2-\theta}$. Applying Lemma 3, for any $0 < \delta < 1$, letting ϵ be the solution of the equation $\exp\left\{-\frac{m\epsilon^{2-\theta}}{2C_\theta \max\{\widetilde{B},M_\tau\}^{2-\theta} + \frac{2}{3}(\widetilde{B}+M_\tau)\epsilon^{1-\theta}}\right\} = \delta/5$, with confidence $1-\delta/5$, we have

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi \le \sqrt{(\mathbb{E}\xi)^\theta + \epsilon^\theta}\epsilon^{1-\frac{\theta}{2}} \le (\mathbb{E}\xi)^{\frac{\theta}{2}}\epsilon^{1-\frac{\theta}{2}} + \epsilon \le \frac{\theta}{2}\mathbb{E}\xi + \left(2 - \frac{\theta}{2}\right)\epsilon. \tag{4.3}$$

Here the last inequality is from Young's inequality. Since ϵ satisfies

$$\epsilon^{2-\theta} - \frac{2(\widetilde{B}+M_\tau)\log\frac{5}{\delta}}{3m}\epsilon^{1-\theta} - \frac{2C_\theta \max\{\widetilde{B},M_\tau\}^{2-\theta}\log\frac{5}{\delta}}{m} = 0,$$

using Lemma 7.2 in [7], we find

$$\epsilon \le \max\left\{\frac{4(\widetilde{B}+M_\tau)\log\frac{5}{\delta}}{3m}, \left(\frac{4C_\theta \max\{\widetilde{B},M_\tau\}^{2-\theta}}{m}\right)^{\frac{1}{2-\theta}}\right\}.$$

Substituting the above bound to (4.3), we obtain

$$\begin{aligned}
&\left\{\mathcal{E}^\tau_{\mathbf{z}}(f^\tau_{\sigma,\gamma}) - \mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho)\right\} - \left\{\mathcal{E}^\tau(f^\tau_{\sigma,\gamma}) - \mathcal{E}^\tau(f^\tau_\rho)\right\} \\
&\le \frac{\theta}{2}\left\{\mathcal{E}^\tau(f^\tau_{\sigma,\gamma}) - \mathcal{E}^\tau(f^\tau_\rho)\right\} + \left(2-\frac{\theta}{2}\right)4(C_\theta+1)(\widetilde{B}+M_\tau)\log\frac{5}{\delta}m^{-\frac{1}{2-\theta}} \\
&\le \frac{1}{2}\widetilde{\mathscr{D}}(\sigma,\gamma) + 8(C_\theta+1)(\widetilde{B}+M_\tau)\log\frac{5}{\delta}m^{-\frac{1}{2-\theta}}.
\end{aligned}$$

Therefore, there exists a subset of Z_1 of Z^m with measure at least $1-\delta/5$, such that

$$\mathscr{S}_2 \le \left(1+\frac{\lambda}{2\gamma}\right)\left(\frac{1}{2}\widetilde{\mathscr{D}}(\sigma,\gamma) + 8(C_\theta+1)(\widetilde{B}+M_\tau)\log\frac{5}{\delta}m^{-\frac{1}{2-\theta}}\right), \quad \forall \mathbf{z}\in Z_1. \tag{4.4}$$

Next we use Lemma 7 to estimate $\mathscr{S}_3$. Set a random variable ζ on (Z,ρ) as $\zeta(z) = |y-\pi_B(y)|$. Denote the indicator function of a set $\{|y|\ge B\}$ as $I_{|y|\ge B}$. It follows from assumption (1.5) and the inequalities $I_{|y|\ge B} \le B^{-k}|y|^k$, $|\zeta - \mathbb{E}\zeta|^\ell \le 2^\ell(|\zeta|^\ell + \mathbb{E}|\zeta|^\ell)$ and $(k+\ell)! \le \ell! k^k 2^{\ell k}$ that

$$\begin{aligned}
&\mathbb{E}|\zeta-\mathbb{E}\zeta|^\ell \le 2^{\ell+1}\mathbb{E}|\zeta|^\ell = 2^{\ell+1}\int_Z |y-\pi_B(y)|^\ell d\rho \le 2^{\ell+1}\int_Z |y|^\ell I_{|y|\ge B}d\rho \\
&\le 2^{\ell+1}B^{-k}\int_Z |y|^{\ell+k}d\rho \le c2^{\ell+1}(\ell+k)!M^{\ell+k}B^{-k} \le c2^{\ell+1}k^k 2^{\ell k}\ell! M^{\ell+k}B^{-k} \le \frac{1}{2}\ell! M_1^{\ell-2}v_1,
\end{aligned}$$

where $M_1 = M2^{k+1}$ and $v_1 = 4M_1^2 B^{-k}ck^kM^k$. Then we apply Lemma 7 to the random variables $\{X_i = \zeta(z_i) - \mathbb{E}\zeta\}$, and see that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^m \zeta(z_i) - \mathbb{E}\zeta \ge \frac{\epsilon}{m}\right\} \le \exp\left\{-\frac{\epsilon^2}{2(mv_1 + M_1\epsilon)}\right\}.$$

Setting the right-hand side to be $\delta/5$, we find that the positive solution to the corresponding quadratic equation $\epsilon^2 = 2M_1\epsilon\log\frac{5}{\delta} + 2mv_1\log\frac{5}{\delta}$ is

$$\epsilon = M_1\log\frac{5}{\delta} + \sqrt{M_1^2\log^2\frac{5}{\delta} + 2mv_1\log\frac{5}{\delta}} \le 3M_1\log\frac{5}{\delta} + 2^{k+2}mB^{-k}ck^kM^{k+1}.$$

Thus with confidence $1-\delta/5$, there holds

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}|\pi_B(y_i)-y_i| &\le \int_Z |y-\pi_B(y)|d\rho + \frac{3M2^{k+1}}{m}\log\frac{5}{\delta} + 2^{k+2}B^{-k}Ck^kM^{k+1}\\ &\le \int_Z |y|I_{|y|\ge B}d\rho + \frac{3M2^{k+1}}{m}\log\frac{5}{\delta} + 2^{k+2}B^{-k}ck^kM^{k+1}\\ &\le B^{-k}\int_Z |y|^{k+1}d\rho + \frac{3M2^{k+1}}{m}\log\frac{5}{\delta} + 2^{k+2}B^{-k}ck^kM^{k+1}\\ &\le c\left\{(k+1)!+2^{k+2}k^k\right\}M^{k+1}B^{-k} + \frac{3M2^{k+1}}{m}\log\frac{5}{\delta}.\end{aligned}$$

Similarly, we can estimate $\mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho)-\mathcal{E}^\tau(f^\tau_\rho)$ by considering a random variable $\zeta(z)=L_\tau(f^\tau_\rho(x)-y)$ defined on (Z,ρ). It follows from the assumptions (1.5) and (1.2) that

$$\mathbb{E}|\zeta-\mathbb{E}\zeta|^\ell \le 2^{\ell+1}\mathbb{E}|\zeta|^\ell \le 2^{2\ell+1}\left(\int_Z |y|^\ell d\rho + M_\tau^\ell\right) \le 2^{2\ell+1}\left(c\ell!M^\ell + M_\tau^\ell\right).$$

Then we use Lemma 7 with $M_1=4(M+M_\tau)$ and $v_1=64(c+1)(M+M_\tau)^2$ and find that with confidence $1-\delta/5$, there holds

$$\mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho)-\mathcal{E}^\tau(f^\tau_\rho) \le \frac{24\log\frac{5}{\delta}(M+M_\tau)(c+1)}{\sqrt{m}}.$$

Therefore, there exists a subset of Z_2 of Z^m with measure at least $1-2\delta/5$, such that

$$\begin{aligned}\mathscr{S}_3 \le c\left\{(k+1)!+2^{k+2}k^k\right\}M^{k+1}B^{-k} + \frac{3M2^{k+1}\log\frac{5}{\delta}}{m}\\ + \frac{12\lambda\log\frac{5}{\delta}(M+M_\tau)(c+1)}{\gamma\sqrt{m}}, \quad \forall \mathbf{z}\in Z_2.\end{aligned} \tag{4.5}$$

We shall directly use Lemma 5 to bound $\mathscr{S}_1$ by some properly chosen R. When $\sigma=m^{-\alpha}$ with $0<\alpha<\frac{1}{2(n+1)}$, from the inequality

$$\exp\{-cx\} \le \left(\frac{a}{ec}\right)^a x^{-a}, \qquad \forall x,c,a>0,$$

we have $(\log m)^{n+1} \le (\frac{1}{2e\alpha})^{n+1}m^{2(n+1)\alpha}$, then

$$\eta_\delta = \log\frac{1}{\delta} + \sigma^{-\frac{2(n+1)}{2-\theta}} + (\Delta\log m)^{\frac{n+1}{2-\theta}} \le \log\frac{1}{\delta} + \left(1+\left(\frac{\Delta}{2e\alpha}\right)^{\frac{n+1}{2-\theta}}\right)m^{\frac{2(n+1)\alpha}{2-\theta}}.$$

For $R\ge 1$, denote

$$\mathscr{W}(R) = \left\{\mathbf{z}\in Z^m : \|\hat{f}^\tau_{\mathbf{z}}\|_\sigma \le R\right\}. \tag{4.6}$$

By Lemma 5, there exits $Z_3\subset Z^m$ with the measure at least $1-\delta/5$ such that for any $\Delta\ge 1$ and $B\ge M_\tau$, there holds

$$\begin{aligned}\left\{\mathcal{E}^\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})-\mathcal{E}^\tau(f^\tau_\rho)\right\} - \left\{\mathcal{E}^\tau_{\mathbf{z}}(\pi_B(\hat{f}^\tau_{\mathbf{z}}))-\mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho)\right\} \le \frac{1}{2}\left\{\mathcal{E}^\tau(\pi_B(\hat{f}^\tau_{\mathbf{z}})-\mathcal{E}^\tau(f^\tau_\rho)\right\}\\ + C_{X,\rho}\left(2+\left(\frac{\Delta}{2e\alpha}\right)^{\frac{n+1}{2-\theta}}\right)\log\frac{5}{\delta}Bm^{-\frac{1-2(n+1)\alpha}{2-\theta}} + 20Rm^{-\Delta}, \quad \forall \mathbf{z}\in\mathscr{W}(R)\cap Z_3.\end{aligned} \tag{4.7}$$

Let $\gamma = \sigma^{n+s}$ and $\lambda = m^{-\beta}$ with $\beta > \alpha(n+s)$. Combining the bounds (4.4), (4.5), (4.7), (3.7) and (3.9), we obtain that

$$\begin{aligned}\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) \leq\ & 24(C_\theta+1)(\widetilde{B}+M_\tau)\log\frac{5}{\delta}m^{-\frac{1}{2-\theta}} \\ & + 2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k} + 6M2^{k+1}\log\frac{5}{\delta}m^{-1} \\ & + 25\log\frac{5}{\delta}(M+M_\tau)(c+1)m^{-(\beta-(n+s)\alpha)} \\ & + 2C_{X,\rho}\left(2+\left(\frac{\Delta}{2e\alpha}\right)^{\frac{n+1}{2-\theta}}\right)\log\frac{5}{\delta}Bm^{-\frac{1-2(n+1)\alpha}{2-\theta}} + 40Rm^{-\Delta} \\ & + 9\widetilde{B}m^{-\alpha s}, \quad \forall \mathbf{z} \in \mathscr{W}(R) \cap Z_3 \cap Z_2 \cap Z_1.\end{aligned} \tag{4.8}$$

Recall that the output function takes the from $\hat{f}_{\mathbf{z}}^\tau(x) = \sum_{k=1}^m \alpha_k^{\mathbf{z}} K^\sigma(x, x_k)$, from the definition of the RKHS-norm, we have

$$\|\hat{f}_{\mathbf{z}}^\tau\|_\sigma \leq \sqrt{\sum_{i,j=1}^m \alpha_i^{\mathbf{z}}\alpha_j^{\mathbf{z}}K^\sigma(x_i,x_j)} \leq \Omega(\hat{f}_{\mathbf{z}}^\tau).$$

In order to find a $R > 0$ such that $\hat{f}_{\mathbf{z}}^\tau \in \mathscr{B}_R^\sigma$, we turn to give a bound for $\Omega(\hat{f}_{\mathbf{z}}^\tau)$. From the definition of $\hat{f}_{\mathbf{z}}^\tau$ (1.4), we have

$$\lambda\Omega(\hat{f}_{\mathbf{z}}^\tau) \leq \mathcal{E}_{\mathbf{z}}^\tau(\hat{f}_{\mathbf{z}}^\tau) + \lambda\Omega(\hat{f}_{\mathbf{z}}^\tau) \leq \mathcal{E}_{\mathbf{z}}^\tau(0) \leq \frac{1}{m}\sum_{i=1}^m |y_i|.$$

We use Lemma 7 again and find that with confidence $1-\delta/5$, there holds

$$\frac{1}{m}\sum_{i=1}^m |y_i| \leq cM + 4M(1+\sqrt{2c})\frac{\log\frac{5}{\delta}}{\sqrt{m}} \leq (3cM+4M)\log\frac{5}{\delta} := M_\delta. \tag{4.9}$$

This yields the measure of the set $\mathscr{W}(\frac{M_\delta}{\lambda})$ is at least $1-\delta/5$, thus the measure of the set $\mathscr{W}(\frac{M_\delta}{\lambda}) \cap Z_3 \cap Z_2 \cap Z_1$ is at least $1-\delta$. We substitute $R = \frac{M_\delta}{\lambda}$ to (4.8) and let $\Delta = 1+\beta$, then $\frac{R}{m^\Delta} \leq \frac{M_\delta}{m}$ and the conclusion follows. □

Now we are in the position to give the proof of Theorem 1.

Proof of Theorem 1. It follows from Proposition 2 and Proposition 4 that for any $B \geq M_\tau$, with confidence $1-\delta$, there holds

$$\begin{aligned}&\|\pi_{M_\tau}(\hat{f}_{\mathbf{z}}^\tau) - f_\rho^\tau\|_{L^r_{\rho_X}} \leq \|\pi_B(\hat{f}_{\mathbf{z}}^\tau) - f_\rho^\tau\|_{L^r_{\rho_X}} \\ &\leq c_\rho\left(5C_{X,\rho,\alpha,\beta}\log\frac{5}{\delta}\right)^{1/q}Bm^{-\Theta/q} + B\left(2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k}\right)^{1/q} \\ &+ \left(6M2^{k+1}\log\frac{5}{\delta}\right)^{1/q}B^{1-1/q}m^{-1/q},\end{aligned} \tag{4.10}$$

where Θ is given by (2.5) and the first inequality holds since $|f_\rho^\tau| \leq M_\tau$ almost surely. Since $(k+1)! \leq k^k 2^k$, then we have

$$\left(2c\left\{(k+1)! + 2^{k+2}k^k\right\} M^{k+1}B^{-k}\right)^{1/q} \leq 2^{\frac{k+1}{q}}(5cM)^{1/q}\left\{(MkB^{-1})^k\right\}^{1/q} \tag{4.11}$$

and

$$\left(6M2^{k+1}\log\frac{5}{\delta}\right)^{1/q} B^{1-1/q}m^{-1/q} \leq 2^{\frac{k+1}{q}}\left(6M\log\frac{5}{\delta}\right)^{1/q} Bm^{-1/q}. \tag{4.12}$$

For any $\epsilon < \Theta/q$, chose k to be the integer part of $\frac{\Theta}{\epsilon} + 1$ and $B = \max\{M, M_\tau\}\Theta\epsilon^{-1}m^\epsilon$, then $(MkB^{-1})^k \leq 4m^{-\Theta}$, thus we find

$$2^{\frac{k+1}{q}}(5cM)^{1/q}\left\{(MkB^{-1})^k\right\}^{1/q} \leq 2^{\frac{k+1}{q}}(20cM)^{1/q}m^{-\Theta/q} \leq 2^{\frac{\Theta+2\epsilon}{q\epsilon}}(20cM)^{1/q}m^{-\Theta/q}. \tag{4.13}$$

Finally, we complete the proof by substituting the bounds (4.12), (4.11) and (4.13) to (4.10) with

$$C^\epsilon_{X,\rho,\alpha,\beta} = 3(M+M_\tau)\max\left\{c_\rho(5C_{X,\rho,\alpha,\beta})^{1/q}, (20cM)^{1/q}, (6M)^{1/q}\right\} 2^{\frac{\Theta+2\epsilon}{q\epsilon}}\Theta\epsilon^{-1}.$$

□

Next, we prove Theorem 2 mainly based on Lemma 6.

Proof of Theorem 2. We first establish a similar result as Proposition 4 based on Lemma 6. Recall the definition of $\mathscr{W}(R)$ in (4.6). Lemma 6 yields that, when $B \geq M_\tau$, there exists $Z_3' \subset Z^m$ with the measure at least $1 - 5/\delta$, such that

$$\begin{aligned}
&\left\{\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau)\right\} - \left\{\mathcal{E}_{\mathbf{z}}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}_{\mathbf{z}}^\tau(f_\rho^\tau)\right\} \\
&\quad\leq \frac{1}{2}\eta_R^{1-\theta}\left\{\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau)\right\}^\theta + c_\mu\eta_R \\
&\quad+ \left(2C_\theta^{\frac{1}{2-\theta}} + 36\right)\log\frac{5}{\delta}Bm^{-\frac{1}{2-\theta}}, \qquad \forall \mathbf{z} \in \mathscr{W}(R) \cap Z_3',
\end{aligned}$$

where η_R is given by (3.18). From the bound above and error decomposition (3.8), we obtain

$$\begin{aligned}
\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) &\leq \mathscr{S}_1 + \mathscr{S}_2 + \mathscr{S}_3 + \mathscr{D} \\
&\leq \frac{1}{2}\eta_R^{1-\theta}\left\{\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau)\right\}^\theta + \mathscr{S}_1' + \mathscr{S}_2 + \mathscr{S}_3 + \mathscr{D}, \quad \forall \mathbf{z} \in \mathscr{W}(R) \cap Z_3',
\end{aligned} \tag{4.14}$$

where

$$\mathscr{S}_1' = c_\mu\eta_R + \left(2C_\theta^{\frac{1}{2-\theta}} + 36\right)\log\frac{5}{\delta}Bm^{-\frac{1}{2-\theta}}.$$

Since for $x > 0$, the inequality $x \leq ax^\theta + b$ implies $x \leq \max\{(2a)^{\frac{1}{1-\theta}}, 2b\}$, then (4.14) implies

$$\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) \leq \max\{\eta_R, 2(\mathscr{S}_1' + \mathscr{S}_2 + \mathscr{S}_3 + \mathscr{D})\}, \quad \forall \mathbf{z} \in \mathscr{W}(R) \cap Z_3'. \tag{4.15}$$

When $\sigma = m^{-\alpha}$ with $\alpha < \frac{1}{n}$ and $\lambda = m^{-\beta}$ with $\beta > (n+s)\alpha$, let $\gamma = \sigma^{n+s}$ and $R = \frac{M_\delta}{\lambda}$ with M_δ given by (4.9), then combining the bounds (4.4), (4.5), (4.15), (3.7) and (3.9), we obtain

$$\begin{aligned}
&\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) \\
&\leq \max\{1, 2c_\mu\}\eta_{\frac{M_\delta}{\lambda}} + \left(4C_\theta^{\frac{1}{2-\theta}} + 72 + 24(C_\theta+1)(\widetilde{B}+M_\tau)\right)\log\frac{5}{\delta}m^{-\frac{1}{2-\theta}} \\
&+ 2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k} + 6M2^{k+1}\log\frac{5}{\delta}m^{-1} \\
&+ 25\log\frac{5}{\delta}(M+M_\tau)(c+1)m^{-(\beta-(n+s)\alpha)} + 9\widetilde{B}m^{-\alpha s}, \quad \forall \mathbf{z} \in \mathscr{W}(\frac{M_\delta}{\lambda}) \cap Z_3' \cap Z_2 \cap Z_1,
\end{aligned}$$

where

$$\eta_{\frac{M_\delta}{\lambda}} \leq C_{\mu,\nu,\rho}(3cM+4M)\log\frac{5}{\delta}Bm^{-\frac{1-(1-\mu)(1+\nu)n\alpha}{2-\theta+\mu\theta}+\frac{2\mu\beta}{1+\mu}}.$$

For any $\epsilon < \Theta'/q$, where Θ' is given by (2.7), let $\mu = \min\{\frac{\epsilon}{12\beta-\epsilon}, \frac{\epsilon(2-\theta)^2}{6\theta}\}$ and $\nu = \frac{\epsilon(2-\theta)}{6n\alpha}$, we then have $\frac{2\mu\beta}{1+\mu} \leq \frac{\epsilon}{6}$ and

$$\begin{aligned}
&\frac{1-n\alpha}{2-\theta} - \frac{1-(1-\mu)(1+\nu)n\alpha}{2-\theta+\mu\theta} \\
&= \frac{(1-n\alpha)\mu\theta + (2-\theta)n\alpha\left\{(1-\mu)(1+\nu)-1\right\}}{(2-\theta)(2-\theta+\mu\theta)} \\
&\leq \frac{\mu\theta}{(2-\theta)^2} + \frac{n\alpha\nu}{2-\theta} \leq \frac{\epsilon}{3}.
\end{aligned}$$

Hence we get

$$\eta_{\frac{M_\delta}{\lambda}} \leq C_{\mu,\nu,\rho}(3cM+4M)\log\frac{5}{\delta}Bm^{\frac{\epsilon}{2}-\frac{1-n\alpha}{2-\theta}}.$$

From the proof of Proposition 4, the measure of $\mathscr{W}(\frac{M_\delta}{\lambda})$ is at least $1-\delta/5$, thus the measure of the set $\mathscr{W}(\frac{M_\delta}{\lambda}) \cap Z_3' \cap Z_2 \cap Z_1$ is at least $1-\delta$. Finally, with confidence $1-\delta$, we have

$$\begin{aligned}
\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) \leq 4C_\epsilon'\log\frac{5}{\delta}\left\{Bm^{\frac{\epsilon}{2}-\frac{1-n\alpha}{2-\theta}} + m^{-(\beta-(n+s)\alpha)} + m^{-\alpha s}\right\} \\
+ 2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k} + 6M2^{k+1}\log\frac{5}{\delta}m^{-1},
\end{aligned}$$

where

$$\begin{aligned}
C_\epsilon' = \max\Big\{&(1+2c_\mu)C_{\mu,\nu,\rho}(3cM+4M), 25(M+M_\tau)(c+1), \\
&4C_\theta^{\frac{1}{2-\theta}} + 72 + 24(C_\theta+1)(\widetilde{B}+M_\tau), 9\widetilde{B}\Big\}.
\end{aligned}$$

Next, completely following the proof of Theorem 1, we choose k to be the integer part of $\frac{2\Theta}{\epsilon} + 1$ and $B = \max\{M, M_\tau\}2\Theta\epsilon^{-1}m^{\epsilon/2}$. The bound (2.6) achieves with

$$\widetilde{C}_{X,\rho,\alpha,\beta}^\epsilon = 3(M+M_\tau)\max\left\{c_\rho(4C_\epsilon')^{1/q}, (20cM)^{1/q}, (6M)^{1/q}\right\}2^{\frac{2\Theta'+2\epsilon+q\epsilon}{q\epsilon}}\Theta'\epsilon^{-1}.$$

□

Finally, we give the proof of Proposition 1.

Proof of Proposition 1. For any given $x \in X$, the density function of conditional distribution $\rho(y|x)$ is $\frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(y-f_\rho(x))^2}{2\sigma_x^2}}$. Then for any $\ell \in \mathbb{N}$, we have

$$\begin{aligned}\int_Y |y|^\ell d\rho(y|x) &= \int_{\mathbb{R}} |y|^\ell \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(y-f_\rho(x))^2}{2\sigma_x^2}} dy = \frac{1}{\sqrt{2\pi}\sigma_x} \int_{\mathbb{R}} |y + f_\rho(x)|^\ell e^{-\frac{y^2}{2\sigma_x^2}} dy \\ &\leq \frac{2^{\ell+1}}{\sqrt{2\pi}\sigma_x} \int_0^\infty |y|^\ell e^{-\frac{y^2}{2\sigma_x^2}} dy + 2^\ell |f_\rho(x)|^\ell = \frac{(\sqrt{2}\sigma_x)^\ell}{\sqrt{\pi}} \Gamma(\frac{\ell+1}{2}) + 2^\ell |f_\rho(x)|^\ell,\end{aligned}$$

where $\Gamma(t) = \int_0^\infty e^{-s} s^{t-1} ds$. Since $\Gamma(\frac{\ell+1}{2}) \leq \ell! \sqrt{\pi}$, we have

$$\int_Y |y|^\ell d\rho(y|x) \leq \ell! (\sqrt{2}\sigma_x)^\ell + 2^\ell |f_\rho(x)|^\ell \leq \ell! (\sqrt{2}\sigma_x + 2|f_\rho(x)|)^\ell.$$

Hence assumption (1.5) is satisfied with $c = 1$ and $M = 2\vartheta_1 + \sqrt{2}\vartheta_2$. Note that the medium of $\rho(\cdot|x)$ is $f_\rho(x)$, we next verify condition (2.2). For $s \in [0, \sigma_x]$, there holds

$$\rho((f_\rho(x), f_\rho(x) + s)|x) = \int_{f_\rho(x)}^{f_\rho(x)+s} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(y-f_\rho(x))^2}{2\sigma_x^2}} dy \geq s(2\pi)^{-1/2} \sigma_x^{-1} e^{-\frac{1}{2}}.$$

Symmetry of $\rho(y|x)$ directly yields that $\rho((f_\rho(x) - s, f_\rho(x))|x) \geq s(2\pi)^{-1/2}\sigma_x^{-1}e^{-\frac{1}{2}}$ also holds. Thus the condition (2.2) holds true with $a_x = \sigma_x$, $b_x = (2\pi)^{-1/2}\sigma_x^{-1}e^{-\frac{1}{2}}$ and $q = 2$. Hence for any $x \in X$, $(b_x a_x^{q-1})^{-1} = \sqrt{2\pi}e^{1/2}$ is a constant, then we can take $p = \infty$. Therefore, we further get $\theta = \min\left\{\frac{2}{q}, \frac{p}{p+1}\right\} = 1$ and $r = \frac{pq}{p+1} = 2$. Since X has a Lipschitz boundary, the extension Theorem [15] guarantees the existence of function $\tilde{f}_\rho \in H^s(\mathbb{R}^n)$ such that $\tilde{f}_\rho|_X = f_\rho$. Because of $s > \frac{n}{2}$, we know that the Sobolev space $H^s(\mathbb{R}^n)$ can be embedded into $\mathscr{C}(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$, then the regularity condition for f_ρ is satisfied. Finally, our desired results follows from Theorem 1 and Theorem 2. □

5 Numerical Examples

In the above sections, we have given the convergence analysis for quantile regression with ℓ_1-regularization and Gaussian Kernels. In this section, we evaluate the theoretical results by numerical experiments. We shall compare the performances of RKHS-based algorithm (1.3) and the concerned ℓ_1-regularized algorithm (1.4) on artificial data sets. In the proof of Lemma 1, we restate the RKHS-based algorithm (3.2) as an optimization problem (3.4). Using that form, the analysis on the optimal solution is easy to understand, but in view

of computation, it can be further simplified. In this section, by setting $C = \frac{1}{2\lambda m}$, we solve (1.3) via the following problem

$$\begin{array}{ll} \underset{\alpha_i\in\mathbb{R},\xi_i\in\mathbb{R}}{\text{minimize}} & \frac{1}{2}\sum_{i,j=1}^m \alpha_i\alpha_j K^\sigma(x_i,x_j) + C\sum_{i=1}^m \xi_i \\ \text{subject to} & \sum_{j=1}^m \alpha_j K^\sigma(x_i,x_j) - y_i \le \frac{1}{1-\tau}\xi_i, \\ & y_i - \sum_{j=1}^m \alpha_j K^\sigma(x_i,x_j) \le \frac{1}{\tau}\xi_i, \quad \text{for all } i=1,\cdots,m, \end{array} \tag{5.1}$$

where the output function is given by $\sum_{i=1}^m \alpha_i^* K^\sigma(x,x_i)$ with $\{\alpha_i^*\}_{i=1}^m$ being the solution of (5.1). One can verify the equivalence between (1.3) and (5.1). To make a comparison, we also consider the ℓ_1-regularization algorithm (1.3) with $\hat{f}_{\mathbf{z}}^\tau = \sum_{i=1}^m \alpha_i^* K^\sigma(x,x_i)$, where $\{\alpha_i^*\}_{i=1}^m$ is given by

$$\begin{array}{ll} \underset{\alpha_i\in\mathbb{R},\xi_i\in\mathbb{R}}{\text{minimize}} & \frac{1}{2}\sum_{i=1}^m |\alpha_i| + C\sum_{i=1}^m \xi_i \\ \text{subject to} & \sum_{j=1}^m \alpha_j K^\sigma(x_i,x_j) - y_i \le \frac{1}{1-\tau}\xi_i, \\ & y_i - \sum_{j=1}^m \alpha_j K^\sigma(x_i,x_j) \le \frac{1}{\tau}\xi_i, \quad \text{for all } i=1,\cdots,m. \end{array} \tag{5.2}$$

We first consider the testing functions provided in [5]. These functions have been used in many papers to examine the regression performance, see e.g. [14], [9] and [31]. Below are the expressions of two of these functions, where the domain $D = [a,b]^n = \{x | x \in \mathbb{R}^n, a \le x(i) \le b, \forall 1 \le i \le n\}$ and $x(i)$ stands for the i-th component of x.

$f_1(x) = \exp\big(x(1)\sin(\pi x(2))\big)$, $D = [-1,1]^2$.

$f_2(x) = \dfrac{1+\sin(2x(1)+3x(2))}{3.5+\sin(x(1)-x(2))}$, $D = [-2,2]^2$.

The examples above are smooth functions and we also concern the approximation abilities of the algorithms when the target functions are non-smooth. For this purpose, we will construct numerical examples for continuous piecewise linear functions. Recall that, a piecewise linear function equals a linear or an affine function in each subregion of the domain. Continuous piecewise linear functions require the continuity in the boundaries of adjacent subregions. It is easy to see that a continuous piecewise linear function is non-smooth since it is non-differentiable at the boundaries. To construct a piecewise linear function with sufficient nonlinearity, we apply the identification algorithm proposed in [11] to provide continuous piecewise linear approximations for $f_1(x)$ and $f_2(x)$. The obtained functions are denoted by f_1^{pw} and f_2^{pw}, respectively and then we use algorithms (5.1) and (5.2) to approximate them.

To evaluate the performance of the RKHS-based regularization (5.1) and the ℓ_1-regularization (5.2) in the examples above, we generate 400 points $x_i \in \mathbb{R}^2$, $1 \le i \le 400$, evenly spaced along its domain axes. For $f = f_1, f_2, f_1^{\text{pw}}$ and f_2^{pw}, we compute three groups of noise-polluted function values, $y_i = f\big(x_i\big) + e_i$, $1 \le i \le 400$, each with different

levels of Gaussian noises with zero mean. The levels are selected so as to make the ratio of the variance of the noises e_i to that of y_i, denoted as r_n, equal to $0.05, 0.1$ or 0.2. We take $\tau = \frac{1}{2}$ in the algorithms and adopt the 10-fold cross-validation method to determine the parameters, i.e. C and σ in (5.1) and (5.2). Then using the obtained parameters, (5.1) and (5.2) are solved to get the corresponding regression results. To evaluate the performance, we randomly generate 100 points uniformly distributed in the domain and calculate the relative sum of squared error (RSSE) on this validation data V, defined below,

$$\mathrm{RSSE}_V = \frac{\sum_{x\in V}\big(f(x)) - \hat{f}(x)\big)^2}{\sum_{x\in V}\big(f(x) - E(f)\big)^2},$$

where $\hat{f}$ denotes the output function of the algorithms and $E(f)$ is the average value of f on V. Obviously, if we use the average value to approach the original function, the corresponding RSSE equals to one. Thus, the RSSE for any reasonable regressor is larger than zero and smaller than one. Empirically, when RSSE is smaller than 0.1, the regression precision is satisfactory. Except for RSSE, we also count the number of nonzero α_i ($|\alpha_i| \geq 10^{-4}$). Then the regression performance of the RKHS-based regularization (5.1) and the ℓ_1-regularization (5.2) are reported in Table 1, including RSSEs and the numbers of non-zero coefficients (in brackets). From the results, one can see that for both smooth and non-smooth functions, the ℓ_1-regularization can generally provide almost the same precision as that of RKHS-based regularization, which are supportive of the theoretical prediction. At last, we should point out that, when training ℓ_1-regularized scheme (5.2) with high dimensional samples, it usually leads to less sparse solutions if the sample size m is small.

Table 1: Regression Error and Number of Non-zero Coefficients

$f_1(x)$	$r_n = 0.05$	$r_n = 0.1$	$r_n = 0.2$
RKHS-based regularization	3.101×10^{-2} (400)	4.390×10^{-2} (400)	6.453×10^{-2} (400)
ℓ_1-regularization	3.043×10^{-2} (179)	4.585×10^{-2} (149)	6.721×10^{-2} (145)
$f_2(x)$	$r_n = 0.05$	$r_n = 0.1$	$r_n = 0.2$
RKHS-based regularization	1.614×10^{-2} (400)	1.921×10^{-2} (400)	3.097×10^{-2} (400)
ℓ_1-regularization	1.586×10^{-2} (164)	1.908×10^{-2} (144)	2.793×10^{-2} (109)
$f_1^{\mathrm{pw}}(x)$	$r_n = 0.05$	$r_n = 0.1$	$r_n = 0.2$
RKHS-based regularization	3.253×10^{-2} (400)	4.443×10^{-2} (400)	5.581×10^{-2} (400)
ℓ_1-regularization	3.211×10^{-2} (171)	4.417×10^{-2} (150)	5.396×10^{-2} (129)
$f_2^{\mathrm{pw}}(x)$	$r_n = 0.05$	$r_n = 0.1$	$r_n = 0.2$
RKHS-based regularization	1.731×10^{-2} (400)	2.387×10^{-2} (400)	3.057×10^{-2} (400)
ℓ_1-regularization	1.456×10^{-2} (124)	2.059×10^{-2} (123)	3.019×10^{-2} (111)

6 Appendix

In this appendix, we give the proofs of Proposition 2, Lemma 5 and Lemma 6.

Proof of Proposition 2. By the definition of $\mathcal{E}^\tau(f)$, we know that

$$\mathcal{E}^\tau(f) - \mathcal{E}^\tau(f_\rho^\tau) = \int_X \int_Y \left\{ L_\tau(f(x) - y) - L_\tau(f_\rho^\tau(x) - y) \right\} d\rho(y|x)\rho_X(x).$$

For a fixed $x \in X$, let $g = g_x$ be a convex function in $\mathbb{R}$ given by $g(t) = \int_Y L_\tau(t-y)d\rho(y|x)$ and $a = f_\rho^\tau(x), b = f(x)$. Then

$$\mathcal{E}^\tau(f) - \mathcal{E}^\tau(f_\rho^\tau) = \int_X (g(b) - g(a))d\rho_X(x) = \int_X \int_a^b g'_-(t)dtd\rho_X(x).$$

The left derivative of the function g equals

$$g'_-(t) = \int_{-\infty}^{t-0} (1-\tau)d\rho(y|x) + \int_t^{+\infty} \tau d\rho(y|x) = (1-\tau)\rho((-\infty, t)|x) - \tau\rho([t, +\infty)|x).$$

Therefore

$$\begin{aligned}\int_a^b g'_-(t)dt &= \int_a^b ((1-\tau)\rho((-\infty, t)|x) - \tau\rho([t, +\infty)|x))dt = \int_a^b (\rho((-\infty, t)|x) - \tau)dt \\ &= (\rho((-\infty, a]|x) - \tau)(b-a) + \int_a^b \rho((a, t)|x)dt.\end{aligned}$$

If $0 < b - a \leq a_x$, then (2.2) tells us that

$$\int_a^b \rho((a, t)|x)dt \geq b_x \int_a^b (t-a)^{q-1}dt = q^{-1}b_x(b-a)^q.$$

If $b - a > a_x$, then we have

$$\begin{aligned}\int_a^b \rho((a, t)|x)dt &= \int_a^{a_x+a} \rho((a, t)|x)dt + \int_{a_x+a}^b \rho((a, t)|x)dt \\ &\geq b_x \int_a^{a_x+a} (t-a)^{q-1}dt + (b-a-a_x)\rho((a, a+a_x)|x) \\ &\geq b_x q^{-1} a_x^q + (b-a-a_x)b_x a_x^{q-1} \\ &= q^{-1}b_x(qa_x^{q-1}(b-a) - (q-1)a_x^q).\end{aligned}$$

Since $a = f_\rho^\tau(x) \leq M_\tau$ with $M_\tau \geq 1$, $b = f(x) \leq B$ and $0 < a_x \leq 1$ holds true for almost every $x \in X$, we have $qa_x^{q-1}(b-a) - (q-1)a_x^q \geq \left(\frac{a_x}{B+M_\tau}\right)^{q-1}(b-a)^q$ and $(b-a)^q \geq \left(\frac{a_x}{B+M_\tau}\right)^{q-1}(b-a)^q$ for almost every $x \in X$. Note that $\rho((-\infty, a]|x) - \tau \geq 0$, then when $b - a > 0$, for almost every $x \in X$, there holds

$$\int_a^b g'_-(t)dt \geq q^{-1}b_x a_x^{q-1}(b-a)^q.$$

The same inequality can be proved when $b-a\leq 0$. Therefore,

$$\int_a^b g'_-(t)dt \geq q^{-1}b_x a_x^{q-1}|b-a|^q,$$

for almost every $x\in X$. Hence

$$\begin{aligned}\mathcal{E}^\tau(f)-\mathcal{E}^\tau(f_\rho^\tau) &= \int_X\int_a^b g'_-(t)dtd\rho_X(x)\\ &\geq q^{-1}(B+M_\tau)^{1-q}\int_X b_x a_x^{q-1}|f(x)-f_\rho^\tau(x)|^q d\rho_X(x)\\ &\geq q^{-1}2^{1-q}\max\{B,M_\tau\}^{1-q}\int_X b_x a_x^{q-1}|f(x)-f_\rho^\tau(x)|^q d\rho_X(x).\end{aligned}$$

Then for $r=\frac{pq}{p+1}$, applying Hölder inequality, we have

$$\begin{aligned}&\int_X |f(x)-f_\rho^\tau(x)|^r d\rho_X(x)\\ &=\int_X (b_x a_x^{q-1})^{-\frac{p}{p+1}}(b_x a_x^{q-1})^{\frac{p}{p+1}}|(f(x)-f_\rho^\tau(x))|^r d\rho_X(x)\\ &\leq \left\{\int_X (b_x a_x^{q-1})^{-p}d\rho_X(x)\right\}^{\frac{1}{p+1}}\left\{\int_X b_x a_x^{q-1}|f(x)-f_\rho^\tau(x)|^q d\rho_X(x)\right\}^{\frac{p}{p+1}}\\ &\leq \left\{\int_X (b_x a_x^{q-1})^{-p}d\rho_X(x)\right\}^{\frac{1}{p+1}}\left\{q2^{q-1}\max\{B,M_\tau\}^{q-1}(\mathcal{E}^\tau(f)-\mathcal{E}^\tau(f_\rho^\tau))\right\}^{\frac{p}{p+1}}.\end{aligned}$$

Finally we complete the proof of Proposition 2 by taking $r-$th root of the both sides in the last inequality. □

Next we prove Lemma 5.

Proof of Lemma 5. Define the function set

$$\mathscr{G}=\left\{g(z)=L_\tau(\pi_B(f)(x)-y)-L_\tau(f_\rho^\tau(x)-y): f\in\mathscr{B}_R^\sigma\right\}. \tag{6.1}$$

For $\forall g\in\mathscr{G}$, we have $\mathbb{E}(g)\geq 0$, $|g(z)|\leq B+M_\tau\leq 2\max\{B,M_\tau\}$. Additionally, Lemma 4 tells us that $\mathbb{E}(g^2)\leq C_\theta\max\{B,M_\tau\}^{2-\theta}(\mathbb{E}(g))^\theta$.

We consider $\mathscr{G}$ as a subset of continuous functions on $X\times Y$, then for any $\epsilon>0$, the Lipschitz property of pinball loss yields $\mathscr{N}(\mathscr{G},\epsilon,\|\cdot\|_\infty)\leq\mathscr{N}(\mathscr{B}_R,\epsilon,\|\cdot\|_\infty)=\mathscr{N}(\mathscr{B}_1,\epsilon R^{-1},\|\cdot\|_\infty)$. Then we apply a standard covering number argument (see [7]) with Lemma 3 to $\mathcal{G}$ and find

$$\begin{aligned}&\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{\|f\|_\sigma\leq R}\frac{\left[\mathcal{E}^\tau(\pi_B(\hat{f}_\mathbf{z}^\tau))-\mathcal{E}^\tau(f_\rho^\tau)\right]-\left[\mathcal{E}_\mathbf{z}^\tau(\pi_B(\hat{f}_\mathbf{z}^\tau))-\mathcal{E}_\mathbf{z}^\tau(f_\rho^\tau)\right]}{\sqrt{\left[\mathcal{E}^\tau(\pi_B(f_\mathbf{z}^\tau))-\mathcal{E}^\tau(f_\rho^\tau)\right]^\theta+\epsilon^\theta}}>4\epsilon^{1-\frac{\theta}{2}}\right\}\\ &\qquad\leq\mathscr{N}(\mathscr{B}_1^\sigma,\epsilon R^{-1},\|\cdot\|_\infty)\exp\left\{-\frac{m\epsilon^{2-\theta}}{2C_\theta\max\{B,M_\tau\}^{2-\theta}+\frac{4}{3}\max\{B,M_\tau\}\epsilon^{1-\theta}}\right\}.\end{aligned}$$

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

$$\begin{aligned}\left\{\mathcal{E}^\tau(\pi_B(f)-\mathcal{E}^\tau(f_\rho^\tau)\right\}-\left\{\mathcal{E}_{\mathbf{z}}^\tau(\pi_B(f))-\mathcal{E}_{\mathbf{z}}^\tau(f_\rho^\tau)\right\}&\leq\left(1-\frac{\theta}{2}\right)4^{\frac{2}{2-\theta}}\epsilon^*(m,R,\sigma,B,\delta)\\&+\frac{\theta}{2}\left\{\mathcal{E}^\tau(\pi_B(f)-\mathcal{E}^\tau(f_\rho^\tau)\right\}+4\epsilon^*(m,R,\sigma,B,\delta),\quad\forall f\in\mathscr{B}_R^\sigma,\end{aligned}\tag{6.2}$$

where $\epsilon^*(m,R,\sigma,B,\delta)$ is the smallest positive number ϵ satisfying

$$\log\mathscr{N}(\mathscr{B}_1^\sigma,\epsilon R^{-1},\|\cdot\|_\infty)-\frac{m\epsilon^{2-\theta}}{2C_\theta\max\{B,M_\tau\}^{2-\theta}+\frac{4}{3}\max\{B,M_\tau\}\epsilon^{1-\theta}}\leq\log\delta.$$

Based on bound (3.15), following the same idea in the proof of Proposition 4 in [32], we find

$$\begin{aligned}\epsilon^*(m,R,\sigma,B,\delta)\leq\frac{R}{m^\Delta}+\max\{B,M_\tau\}&\left(\frac{4C_\theta(\log\frac{1}{\delta}+C_X\sigma^{-2(n+1)})+4C_\theta C_X(\Delta\log m)^{n+1}}{m}\right)^{\frac{1}{2-\theta}}\\&+\frac{8\max\{B,M_\tau\}}{3m}\left\{\log\frac{1}{\delta}+C_X\sigma^{-2(n+1)}+C_X(\Delta\log m)^{n+1}\right\}.\end{aligned}\tag{6.3}$$

Finally substituting (6.3) to (6.2), we derive our desired result with $C_{X,\rho}=(140+80C_\theta)(C_X+1)$. □

Finally, we prove Lemma 6 by applying the following result in [30].

Lemma 8. *Let $\mathscr{F}$ be a class of bounded measurable functions. Assume that there are constants $Q,c_1>0$ and $\theta\in[0,1]$ such that $\|f\|_\infty\leq Q$ and $\mathbb{E}f^2\leq c_1(\mathbb{E}f)^\theta$ for every $f\in\mathscr{F}$. If for some $c_2>0$ and $0<\varsigma<2$,*

$$\log\mathscr{N}_2(\mathscr{F},\epsilon)\leq c_2\epsilon^{-\varsigma},\quad\forall\epsilon>0,$$

then there exists a constant c'_ς depending only on ς such that for any $t>0$, with probability at least $1-e^{-t}$, there holds

$$\mathbb{E}f-\frac{1}{m}\sum_{i=1}^m f(z_i)\leq\frac{1}{2}\eta^{1-\theta}(\mathbb{E}f)^\theta+c'_\varsigma\eta+2(\frac{c_1t}{m})^{\frac{1}{2-\theta}}+\frac{18Qt}{m},\qquad\forall f\in\mathscr{F},$$

where

$$\eta:=\max\left\{c_1^{\frac{2-p}{4-2\alpha+p\alpha}}\left(\frac{c_2}{m}\right)^{\frac{2}{4-2\alpha+p\alpha}},\quad Q^{\frac{2-p}{2+p}}\left(\frac{c_2}{m}\right)^{\frac{2}{2+p}}\right\}.$$

Proof of Lemma 5. Recall the function set $\mathscr{G}$ defined by (6.1), for any $\epsilon>0$, the Lipschitz property of pinball loss also yields $\mathscr{N}_2(\mathscr{G},\epsilon)\leq\mathscr{N}_2(\mathscr{B}_R,\epsilon)=\mathscr{N}_2(\mathscr{B}_1,\epsilon R^{-1})$. Under the assumption that X is contained in the unit ball of $\mathbb{R}^n$, the covering number bound (3.16) come to existence, hence

$$\log\mathscr{N}_2(\mathscr{G},\epsilon)\leq c_{\mu,\nu}R^{2\mu}\sigma^{-(1-\mu)(1+\nu)n}\epsilon^{-2\mu}.$$

We apply Lemma 8 to $\mathscr{G}$ with $c_1 = C_\theta \max\{B, M_\tau\}^{2-\theta}$, $Q = 2\max\{B, M_\tau\}$, $\varsigma = 2\mu$ and $c_2 = c_{\mu,\nu} R^{2\mu} \sigma^{-(1-\mu)(1+\nu)n}$. Next we need to bound η, since

$$\begin{aligned}
&\left(C_\theta \max\{B, M_\tau\}^{2-\theta}\right)^{\frac{1-\mu}{2-\theta+\mu\theta}} \left(\frac{c_{\mu,\nu} R^{2\mu} \sigma^{-(1-\mu)(1+\nu)n}}{m}\right)^{\frac{1}{2-\theta+\mu\theta}} \\
&= C_\theta^{\frac{1-\mu}{2-\theta+\mu\theta}} c_{\mu,\nu}^{\frac{1}{2-\theta+\mu\theta}} \max\{B, M_\tau\}^{\frac{(2-\theta)(1-\mu)}{2-\theta+\mu\theta}} R^{\frac{2\mu}{2-\theta+\mu\theta}} \left(\frac{\sigma^{-(1-\mu)(1+\nu)n}}{m}\right)^{\frac{1}{2-\theta+\mu\theta}}
\end{aligned}$$

and

$$\begin{aligned}
&\left(2\max\{B, M_\tau\}\right)^{\frac{1-\mu}{1+\mu}} \left(\frac{c_{\mu,\nu} R^{2\mu} \sigma^{-(1-\mu)(1+\nu)n}}{m}\right)^{\frac{1}{1+\mu}} \\
&= 2^{\frac{1-\mu}{1+\mu}} c_{\mu,\nu}^{\frac{1}{1+\mu}} \max\{B, M_\tau\}^{\frac{1-\mu}{1+\mu}} R^{\frac{2\mu}{1+\mu}} \left(\frac{\sigma^{-(1-\mu)(1+\nu)n}}{m}\right)^{\frac{1}{1+\mu}},
\end{aligned}$$

we get our conclusion by the fact that when $\mu, \theta \in [0, 1]^2$, there hold $1 + \mu \leq 2 - \theta + \mu\theta$ and $\frac{(2-\theta)(1-\mu)}{2-\theta+\mu\theta} \geq \frac{1-\mu}{1+\mu}$. □

Acknowledgement

The work described in this paper is supported in part by the National Science Foundation of China under Grant 11201079; Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), G.0377.12 (Structured models), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); IBBT; EU: ERNSI; ERC AdG A-DATADRIVE-B, FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other: Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger.

References

[1] N. Aronszajn, Theory of reproducing kernels. Trans. Amer. Math. Soc. **68** (1950), 337–404.

[2] A. Belloni and V. Chernozhukov, ℓ_1-penalized quantile regression in high dimensional sparse models. Ann. Statist. **39** (2011), 82–130.

[3] G. Bennett, Probability inequalities for the sum of independent random variables. J. Amer. Stat. Assoc. **57** (1962), 33–45.

[4] P. Bradley and O. Mangasarian, Massive data discrimination via linear support vector machines. Optim. Methods Softw. **13** (2000), 1–10.

[5] V. Cherkassky, D. Gehring, and F. Mulier, Comparison of adaptive methods for function estimation from samples. IEEE Trans. on Neur. Netw. **7** (1996), 969–984.

[6] D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou, Support vector machine soft margin classifiers: error analysis. J. Mach. Learn. Res. **5** (2004), 1143–1175.

[7] F. Cucker and D. X. Zhou, Learing Theory: An Approxiamtion Theory Viewpoint. Cambridge University Press, 2007.

[8] M. Eberts and I. Steinwart, Optimal regression rates for SVMs using Gaussian kernels, preprint.

[9] J. González, I. Rojas, J. Ortega, H. Pomares, F.J. Fernández, and A.F. Díaz, Multiobjective evolutionary optimization of the size, shape, and position parameters of radial basis function networks for function approximation. IEEE Trans. on Neur. Net. **14** (2003), 1478–1495.

[10] Z. C. Guo and D. X. Zhou, Concentration estimates for learning with unbounded sampling. Adv. Comput. Math, doi: 10.1007/s10444-011-9238-8.

[11] X. Huang, X. Jun, and S. Wang, Nonlinear system identification with continuous piecewise linear neural network. Neurocomput. **77** (2012), 167–177.

[12] R. Koenker, Quantile Regression. Cambridge Univeristy Press, 2005.

[13] P. Niyogi and F. Girosi, On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. Neural Comput. **8** (1996), 819–842.

[14] A. Suárez and J.F. Lutsko. Globally optimal fuzzy decision trees for classification and regression. IEEE Trans. on Pat. Analy. and Mach. Intel. **21** (1999), 1297–1311.

[15] E. M. Stein, Singular Integrals and Differentiability Properties of Functions. Princeton University Press, Princeton, New Jersey, 1970.

[16] G. Song, H. Zhang and F. J. Hickernell, Reproducing kernel Banach spaces with the l1 norm, Applied and Computational Harmonic Analysis, accepted.

[17] I. Steinwart, On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res. **2** (2002), 67–93.

[18] I. Steinwart and C. Scovel, Fast rates for support vector machines using Gaussian kernels. Ann. Statist. **35** (2007), 575–607.

[19] I. Steinwart and A. Christmann, Support Vector Machines, Springer-Verlag, 2008.

[20] I. Steinwart and A. Christmann, How SVMs can estimate quantiles and the median. Advance in Neural Information Processing Systems **20** (2008), 305–312.

[21] I. Steinwart and A. Christmann, Estimate conditional quantiles with the help of the pinball loss. Bernoulli **17** (2011), 211–225.

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

[23] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory. Anal. Appl. **1** (2003), 17–41.

[24] R. Tibshirani, Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. Series B. **58** (1996), 267–288.

[25] A. W. Van Der Vaart and J. A. Wellner, Weak Convergence and Empirical Processes. Springer-Verlag, New York, 1996.

[26] V. Vapnik, Statistical Learning Theory. John Wiley & Sons, 1998.

[27] C. Wang and D. X. Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling. J. Complexity **27** (2011), 55–67.

[28] G. Wahba, Spline Models for Observational Data. Society for Industrial Mathematics, 1990.

[29] Q. Wu and D. X. Zhou, SVM soft margin classifiers: linear programming versus quadratic programming. Neural Comput. **17** (2005), 1160–1187.

[30] Q. Wu, Y. Ying and D. X. Zhou, Multi-kernel regularized classifiers. J. Complexity **23** (2007), 108–134.

[31] S. Wang, X. Huang, and Y. Yam. A neural network of smooth hinge functions. IEEE Trans. on Neur. Netw. **21**, (2010), 1381–1395.

[32] D. H. Xiang and D. X. Zhou, Classification with Gaussians and convex loss. J. Mach. Learn. Res. **10** (2009), 1447–1468.

[33] D. H. Xiang, Conditional quantiles with varying Gaussians. Adv. Comput. Math, doi: 10.1007/s10444-011-9257-5.

[34] M. Yuan and H. Zou, Efficient global approximation of generalized nonlinear ℓ_1-regularized solution paths and its applications. J. Amer. Statist. Assoc. **104** (2009), 1562–1574.

[35] D. X. Zhou, The covering number in learning theory. J. Complexity. **18** (2002), 739–767.

[36] P. Zhao and B. Yu, On model selection consistency of Lasso. J. Mach. Learn. Res. **7** (2007), 2541–2567.