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 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 assumptions without any further reference. We aim to approximate f_ρ^τ from the sample $\mathbf{z}$ through learning algorithms.

The classical least-squares regression models the relationship between an input $x \in X$ and the conditional mean of a response variable $y \in Y$ given x, which describes the centrality of the conditional distribution. In contrast, quantile regression can provide richer information about the conditional distribution of response variables such as stretching or compressing tails, so it is particularly useful in applications when both lower and upper or all quantiles are of interest. Over the last years, quantile regression has become a popular statistical method in various research fields, such as reference charts in medicine [12], survival analysis [16], economics [15] and so on. For example, in financial risk management, the value at risk (VAR) is an important measure for quantifying daily risks, which is defined directly based on extreme quantiles of risk measures [43]. As our interest here focuses on a particular quantile interval of the response, it is appropriate to adopt quantile regression for VAR modeling. Another example comes from environmental studies where upper quantiles of pollution levels are critical from a public health perspective. In addition, relative to the least-squares regression, quantile regression estimates are more robust against outliers in the response measurements. For more practical applications and attractive features of quantile regression, one may see the book [17] and references therein.

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., [30, 27, 42, 9]). Define the τ-*pinball loss* $L_\tau : \mathbb{R} \to \mathbb{R}^+$ as

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

One can see [17] that the loss function L_τ can be used to model the target function, i.e., the conditional $\tau-$quantile function f_ρ^τ minimizes the *generalization error*

$$\mathcal{E}^\tau(f) = \int_{X\times Y} L_\tau(f(x) - y) d\rho. \tag{1.3}$$

over all measurable functions $f : X \to \mathbb{R}$. Based on this observation, learning algorithms produce estimators of f_ρ^τ by minimizing $\frac{1}{m}\sum_{i=1}^m L_\tau(f(x_i) - y_i)$ or a penalized version when i.i.d. samples $\{(x_i, y_i)\}_{i=1}^m$ are given. In kernel-based 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 width $\sigma > 0$, which is given by

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

The width 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. The adjustable parameter σ plays a major role in the performance of the kernel, and should be carefully tuned to the problem at hand. A small σ will lead to over-fitting and the resulting predictive model will be highly sensitive to noise in sample data. Conversely, a large σ will make the learning algorithms perform unsatisfactorily and thus under-fitting will happen. In machine learning community, choosing the width σ is related to the *model selection* problem, which adjusts the capacity or complexity of the models to the available amount of training data to avoid either under-fitting or over-fitting. It thus motivates the theoretical studies on the convergence behavior of algorithms with Gaussian kernels (e.g. [24, 41]). In particular, [42, 9] consider approximating f_ρ^τ by a solution of the optimization scheme

$$\arg\min_{f\in\mathscr{H}_\sigma} \left\{\frac{1}{m}\sum_{i=1}^m L_\tau(f(x_i) - y_i) + \lambda\|f\|_\sigma^2\right\}, \tag{1.4}$$

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 [36], the solution of algorithm (1.4) belongs to a data-dependent hypothesis space

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

The basis functions $\{K^\sigma(\cdot, x_i)\}_{i=1}^m$ are referred as to features generated by the input data $\{x_i\}_{i=1}^m$ and the feature map $x \mapsto K^\sigma(\cdot, x)$ which is well established from X to $\mathscr{H}^\sigma$ [8].

In this paper, for pursuing sparsity and achieving feature selections in $\mathscr{H}_{\mathbf{z},\sigma}$, 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

$$\hat{f}_{\mathbf{z}}^\tau = \arg\min_{f \in \mathscr{H}_{\mathbf{z},\sigma}} \left\{ \frac{1}{m} \sum_{i=1}^m L_\tau(f(x_i) - y_i) + \lambda \Omega(f) \right\}, \tag{1.5}$$

where the regularization term is given by

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

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 ℓ_1-regularization term not only shrinks the coefficients in the kernel expansion toward zero but also causes some of coefficients to be exactly zero when making a sufficiently large λ. The latter property will bring sparsity in the expression of the output function $\hat{f}_{\mathbf{z}}^\tau$. As is well known, the RKHS-based regularization which is essentially a squared penalty may have the disadvantage that even though some features $K^\sigma(x, x_i)$ may not contribute much to the overall solution, they still appear in the kernel expansion. Therefore, in situations where there are a lot of irrelevant noise features, the ℓ_1-norm regularization may perform superior to the RKHS-based regularization and offer more compact predictive models. As in algorithm (1.4), the parameters λ and σ are both free-determined which provides adaptivity of the algorithm.

The scheme with ℓ_1-regularization is often related to LASSO algorithm [31] 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, 45]). In kernel-based learning, the ℓ_1-regularization was first introduced to design the linear programming support vector machine (e.g. [19, 33, 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 [28, 23]). The ℓ_1-regularization is a very important regularization form as it is robust to irrelevant features and also serves as a methodology for feature selection. 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 (1.5) is essentially a linear programming optimization problem and thus can be efficiently solved by existing codes for large scale problems.

As a linear combination expressed by a Gaussian kernel K^σ and the input data $\{x_i\}_{i=1}^m$, functions from the space $\mathscr{H}_{\mathbf{z},\sigma}$ are often used for scattered data interpolation in computer aided geometric design (CAGD) and approximation theory [35]. Functions of this form also has a wide application in the radial basis function networks [20]. Additionally, for $i = 1, \cdots, m$, by taking $\alpha_i = \frac{1}{m\sigma^n}$ and $\alpha_i = \frac{y_i}{m\sigma^n}$ with a suitable chosen $\sigma = \sigma(m)$, the formula $\sum_{i=1}^m \alpha_i K^\sigma(x, x_i)$ also can be used to estimate the density function of ρ_X and the conditional mean of ρ [44]. In the present scenario, the parameters $\{\alpha_i\}_{i=1}^m$ are obtained by solving a convex optimization problem in $\mathbb{R}^m$, which is induced from the learning algorithms such as (1.4) and (1.5). Recall that the target function f_ρ^τ gives the smallest generalization error over all possible solutions. The performance of the algorithm (1.5) is measured by the *excess generalization error* $\mathcal{E}^\tau(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau)$. For any $\sigma > 0$, when X is a compact subset of $\mathbb{R}^n$, the linear span of the function set $\{K^\sigma(x,t)|t \in X\}$ is dense in the space of continuous functions on X [26, 18]. We thus can expect that the learning scheme (1.5) is *consistent*, i.e., as the sample size m increases, the excess generalization error will tend to zero with high probabilities.

Up to now, the kernel-based quantile regression mainly focuses on estimating f_ρ^τ by regularization schemes in RKHS and the consistency of the algorithms is well understood due to the literature [27, 42, 9]. All theoretical results are stated under the boundedness assumption for the output, i.e., for some constant $M > 0$, $|y| \leq M$ almost surely. However, the regularization algorithm (1.5) is essentially different from its counterpart in RKHS, as the minimization procedure directly carries out in $\mathscr{H}_{\mathbf{z},\sigma}$ which varies with samples. The sample dependent nature of the hypothesis space causes technical difficulties in the analysis [39]. The consistency study on such kind of algorithm is still open. Our paper is devoted to solving this problem. Specifically, we investigate how the output function $\hat{f}_{\mathbf{z}}^\tau$ given in (1.5) approximates the quantile regression function f_ρ^τ with suitable chosen $\lambda = \lambda(m)$ and $\sigma = \sigma(m)$ as $m \to \infty$. We show that the learning ability of algorithm (1.5) is almost the same as that of RKHS-based algorithm (1.4). It is also worth noting that the consistency of the algorithm generally implies that the estimator $\hat{f}_{\mathbf{z}}^\tau$ is closed to the target function f_ρ^τ in a very weak sense. To obtain strong convergence result, under some mild conditions, we apply a so-called self-calibration inequality [25] to bound the function approximation in a weighted L^r-space by the excess generalization error (see Proposition 2). Our error bounds are obtained under a weaker assumption: for some constants $M \geq 1$ and $c > 0$,

$$\int_Y |y|^\ell d\rho(y|x) \leq c\ell! M^\ell, \qquad \forall \ell \in \mathbb{N}, \quad x \in X. \tag{1.6}$$

Note that the boundedness assumption excludes the Gaussian noise while assumption (1.6) covers it. This assumption is well known in probability theory and was introduced

in learning theory in [34, 11].

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. We concludes the paper in Section 6 by presenting some future topics related to our work.

2 Main Results

In order to illustrate our convergence analysis, we first state the definition of *projection operator* introduced in [7].

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

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

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

Let ν be a Borel measure on X (or $\mathbb{R}^n$). For $p \in (0, \infty]$, a weighted L^p-space with the norm $\|f\|_{L^p_\nu} = \left(\int_X |f(x)|^p d\nu\right)^{1/p}$ is denoted by L^p_ν. When omitting the subscript, the notion L^p is referred to the L^p-space with respect to the Lebesgue measure. Since the target function f^τ_ρ takes value in $[-M_\tau, M_\tau]$ almost surely, it is natural to measure the approximation ability of $\hat{f}^\tau_{\mathbf{z}}$ by the distance $\|\pi_{M_\tau}(\hat{f}^\tau_{\mathbf{z}}) - f^\tau_\rho\|_{L^r_{\rho_X}}$. 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 exist a $\tau-$quantile $t^* \in \mathbb{R}$ and constants $0 < a_x \le 1$, $b_x > 0$ such that for each $s \in [0, a_x]$,*

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

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^τ_ρ and the singleton assumption on F^τ_ρ. For more details and examples about this definition, one may see [27] and references therein.

Denoted by $H^s(\mathbb{R}^n)$ the Sobolev space [22] 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 (1.6). 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 (1.4) presented in [42] 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 (1.4) obtained in [9] for bounded output y. Based on these observations, we claim that the approximation ability of algorithm (1.5) is comparable with that of the RKHS-based algorithm (1.4). Next, 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 (1.5) 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 $B := B(m) \geq M_\tau$. From the proofs of the main results, one can see that B will tend to infinity as the sample size increases.*

Actually, when the kernel function is pre-given, since the pinball loss L_τ is Lipschitz continuous, one may derive the learning rates of kernel-based quantile regression with ℓ_1-regularization under the framework of our previous work [28]. However, besides the uniformly boundedness assumption, the presented approach also require the marginal distribution ρ_X to satisfy some regularity condition (see Definition 1 in [28]), which guarantees that the sampling data will have a certain density in X. Moreover, the analysis approach in [28] can not lead to satisfactory results for non-smooth kernel functions. Our approach in this paper is applicable to investigate the learning behavior of ℓ_1-regularized quantile regression with a fixed Mercer kernel and will derive fast learning rates even for rough kernels. It also should be pointed out that, as the kernel width σ need to be tuned in the present scheme, the previous analysis methods that are available for the fixed kernel case can not be directly applied to our setting.

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 $L_K(f)(x) = \int_X K(x,y)f(y)d\rho_X(y)$. Specifically, 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 [29] indicate that, the approximation ability of a Gaussian kernel with a fixed width 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 (1.5). Given $f : X \to \mathbb{R}$, recall the generalization error $\mathcal{E}^\tau(f)$ defined by (1.3) and correspondingly the excess generalization error is given by $\mathcal{E}^\tau(f) - \mathcal{E}^\tau(f_\rho^\tau)$. Compared to the consistency of the algorithm, people may be more concerned with the approximation of f_ρ^τ by the obtained estimator in some kind of function spaces. We thus need the following inequality, 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 $f : X \to [-B, B]$ with $B > 0$, 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 [27], and we move the proof to the Appendix just for completeness. For the least-square regression, the excess generalization error is exactly the distance in the space $L^2_{\rho_X}(X)$ due to the strong convexity of the loss function (e.g., see Proposition 1.8 in [8]). However, as the pinball loss is not strictly convex, the established inequality (3.1) is non-trivial and noise condition on the distribution ρ is needed to derive the result.

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 $\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau)$. This will be done by conducting an error decomposition which has been developed in the literature for RKHS-based regularization schemes (e.g. [8, 26]). 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 (1.5).

In this paper, we shall overcome this difficulty by a stepping stone method [37]. We use $\hat{f}^{\tau}_{\mathbf{z},\gamma}$ to denote the solution of algorithm (1.4) 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 (1.5), 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 ξ_i and $\tilde{\xi}_i$, 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 $\frac{\partial\mathcal{L}}{\partial\mathscr{H}_\sigma}(f) = f + \sum_{i=1}^{m}\alpha_i K^\sigma(x,x_i) - \sum_{i=1}^{m}\tilde{\alpha}_i K^\sigma(x,x_i), \forall f\in\mathscr{H}_\sigma$. In order to derive the dual problem of (3.4), we first let

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

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,

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

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

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

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 [41] for regularized classification schemes. Since the pinball loss is Lipschitz continuous, the proof of Lemma 2 is exactly the same as [41]. 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 [42] to analyze the conditional quantile regression algorithm (1.4). Recently, a more general version is presented by [9].

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. *Under the assumption of Lemma 2, let $(\lambda,\sigma,\gamma) \in (0,1]^3$, $\hat{f}^\tau_{\mathbf{z}}$ be defined by (1.5) and $f^\tau_{\sigma,\gamma} \in \mathscr{H}_\sigma$ satisfying (3.6) and (3.7). Then for any $B > 0$, there holds*

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

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 $\hat{f}^\tau_{\mathbf{z}}$ (1.5), 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 (1.6) 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 technical details will be explained in the next section.

4 Concentration Estimates

This section is devoted to estimating $\mathscr{S}_i$ $(i = 1, 2, 3)$ and deriving convergence rates. 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). For example, in order to estimate $\mathscr{S}_1$ and $\mathscr{S}_2$, we define random variables as

$$\xi_i := \xi(z_i) = L_\tau(f(x_i) - y_i) - L_\tau(f_\rho^\tau(x_i) - y_i), \tag{4.1}$$

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 (4.1) are bounded random variables even if y_i is unbounded. Similarly, by considering other suitable random variables on (Z, ρ), $\mathscr{S}_3$ can also be estimated by bounding the difference between the empirical mean and its expectation.

4.1 Bounding $\mathscr{S}_2$

When f is fixed in (4.1), which is exactly the case as we estimate $\mathscr{S}_2$, the convergence is guaranteed by the following probability inequality [8].

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

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi}{\sqrt{(\mathbb{E}\xi)^\theta + \epsilon^\theta}} > \epsilon^{1-\frac{\theta}{2}}\right\} \leq \exp\left\{-\frac{m\epsilon^{2-\theta}}{2c_1 + \frac{2}{3}Q\epsilon^{1-\theta}}\right\}. \tag{4.2}$$

When the random variables are given by (4.1), for a general distribution ρ, the variance-expectation condition $\mathbb{E}\xi^2 \leq c_1(\mathbb{E}\xi)^\theta$ 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*

$$\mathbb{E}\left\{(L_\tau(f(x) - y) - L_\tau(f_\rho^\tau(x) - y))^2\right\} \leq C_\theta \max\{B, M_\tau\}^{2-\theta}\left(\mathcal{E}^\tau(f) - \mathcal{E}^\tau(f_\rho^\tau)\right)^\theta, \tag{4.3}$$

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 [27, 42] 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.

Now we can bound $\mathscr{S}_2$ by the following proposition.

Proposition 4. *Under the assumptions of Proposition 2 and Lemma 2, for any $0 < \delta < 1$, there exists a subset of Z_1 of Z^m with measure at least $1 - \delta/5$, such that*

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

where the constant C_θ is given in Lemma 4.

Proof. We bound $\mathscr{S}_2$ by considering the random variable

$$\xi(z) = L_\tau(f^\tau_{\sigma,\gamma}(x) - y) - L_\tau(f^\tau_\rho(x) - y)$$

on (Z, ρ). From Lemma 2, $|\xi(z)| \leq |f^\tau_{\sigma,\gamma}(x) - f^\tau_\rho(x)| \leq \widetilde{B} + M_\tau$ 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

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi \leq \sqrt{(\mathbb{E}\xi)^\theta + \epsilon^\theta}\,\epsilon^{1-\frac{\theta}{2}} \leq (\mathbb{E}\xi)^{\frac{\theta}{2}}\epsilon^{1-\frac{\theta}{2}} + \epsilon \leq \frac{\theta}{2}\mathbb{E}\xi + \left(2 - \frac{\theta}{2}\right)\epsilon. \tag{4.5}$$

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 [8], 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 (4.5), 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, we prove the existence of the subset Z_1 and the inequality (4.4). □

4.2 Bounding $\mathscr{S}_1$

The only difference between $\mathscr{S}_1$ and $\mathscr{S}_2$ is that the first term involves $\hat{f}^\tau_{\mathbf{z}}$ which varies with samples. Thus a uniform concentration inequality for a family of functions containing $\hat{f}^\tau_{\mathbf{z}}$ is needed to estimate $\mathscr{S}_1$. Since $\hat{f}^\tau_{\mathbf{z}} \in \mathscr{H}_\sigma$, let $\mathscr{B}^\sigma_R = \{f \in \mathscr{H}_\sigma : \|f\|_\sigma \le R\}$, 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 $R \ge 1$, $\Delta \ge 1$, $B > 0$ and $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned}
\left\{\mathcal{E}^\tau(\pi_B(f)) - \mathcal{E}^\tau(f^\tau_\rho)\right\} - \left\{\mathcal{E}^\tau_{\mathbf{z}}(\pi_B(f)) - \mathcal{E}^\tau_{\mathbf{z}}(f^\tau_\rho)\right\} &\le \frac{1}{2}\left\{\mathcal{E}^\tau(\pi_B(f)) - \mathcal{E}^\tau(f^\tau_\rho)\right\} \\
&+ C_{X,\rho}\max\{B,M_\tau\}\eta_\delta m^{-\frac{1}{2-\theta}} + 20Rm^{-\Delta}, \quad \forall f \in \mathscr{B}^\sigma_R,
\end{aligned} \tag{4.6}$$

where

$$\eta_\delta = \log\frac{1}{\delta} + \sigma^{-\frac{2(n+1)}{2-\theta}} + (\Delta\log m)^{\frac{n+1}{2-\theta}} \tag{4.7}$$

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 [32] 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 $B(s_j, \epsilon) = \{s \in \mathscr{M} : d(s, s_j) \leq \epsilon\}$ 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 $\mathscr{N}(\mathscr{B}_1^\sigma, \epsilon, \|\cdot\|_\infty)$ [8]: there exists a constant C_X depending only on X and n such that

$$\log \mathscr{N}(\mathscr{B}_1^\sigma, \epsilon, \|\cdot\|_\infty) \leq C_X \left(\left(\log \frac{1}{\epsilon} \right)^{n+1} + \frac{1}{\sigma^{2(n+1)}} \right), \quad \forall \epsilon > 0, \ \sigma > 0. \tag{4.8}$$

The upper bound appearing in the right hand side of (4.8) 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 $\omega = \{\omega_1, \cdots, \omega_k\} \subset X^k$ for $k \in \mathbb{N}$. 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 [26] 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

$$\log \mathscr{N}_2(\mathscr{B}_1^\sigma, \epsilon) \leq c_{\mu,\nu} \sigma^{-(1-\mu)(1+\nu)n} \left(\frac{1}{\epsilon} \right)^{2\mu}. \tag{4.9}$$

Although the term $\epsilon^{-2\mu}$ increases polynomially, as μ and ν can be chosen arbitrarily small, the bound (4.9) is tighter than the bound (4.8) and thus can lead to sharper estimates. We can bound $\mathscr{S}_1$ by the following lemma when bound (4.9) 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 $R \geq 1$, $B > 0$ and $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\begin{aligned} &\left\{\mathcal{E}^\tau(\pi_B(f)) - \mathcal{E}^\tau(f_\rho^\tau)\right\} - \left\{\mathcal{E}_{\mathbf{z}}^\tau(\pi_B(f)) - \mathcal{E}_{\mathbf{z}}^\tau(f_\rho^\tau)\right\} \\ &\quad \leq \frac{1}{2}\eta_R^{1-\theta}\left\{\mathcal{E}^\tau(\pi_B(f)) - \mathcal{E}^\tau(f_\rho^\tau)\right\}^\theta + c_\mu \eta_R \\ &\quad + \left(2C_\theta^{\frac{1}{2-\theta}} + 36\right)\log\frac{1}{\delta}\max\{B, M_\tau\}m^{-\frac{1}{2-\theta}}, \qquad \forall f \in \mathscr{B}_R^\sigma, \end{aligned} \tag{4.10}$$

where

$$\eta_R = c_{\mu,\nu,\rho}\max\{B, M_\tau\}^{\frac{(2-\theta)(1-\mu)}{2-\theta+\mu\theta}} R^{\frac{2\mu}{1+\mu}}\left(\frac{\sigma^{-(1-\mu)(1+\nu)n}}{m}\right)^{\frac{1}{2-\theta+\mu\theta}} \tag{4.11}$$

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. To bound $\mathscr{S}_1$, one just need to determine a suitable R and directly apply Lemma 5 or Lemma 6.

4.3 Bounding $\mathscr{S}_3$

The estimation of $\mathscr{S}_3$ involves the unboundedness of the output y. 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 $\mathbb{E}|X_i|^\ell \leq \frac{1}{2}\ell! M_1^{\ell-2} v_1$ holds for every $2 \leq \ell \in \mathbb{N}$, then*

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

The term $\mathscr{S}_3$ can be bounded as follows.

Proposition 5. *Under the assumptions (1.2) and (1.6), for any $k \in \mathbb{N}$ and $0 < \delta < 1$, there exists a subset of Z_2 of Z^m with measure at least $1 - 2\delta/5$, such that*

$$\begin{aligned} \mathscr{S}_3 \leq c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k} + \frac{3M2^{k+1}\log\frac{5}{\delta}}{m} \\ + \frac{12\lambda\log\frac{5}{\delta}(M + M_\tau)(c+1)}{\gamma\sqrt{m}}, \quad \forall \mathbf{z} \in Z_2. \end{aligned} \tag{4.12}$$

Proof. We directly 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 $\{|y| \geq B\}$ as $I_{|y|\geq B}$. It follows

from assumption (1.6) and the inequalities $I_{|y|\geq B} \leq B^{-k}|y|^k$, $|\zeta - \mathbb{E}\zeta|^\ell \leq 2^\ell(|\zeta|^\ell + \mathbb{E}|\zeta|^\ell)$ and $(k+\ell)! \leq \ell! k^k 2^{\ell k}$ that

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

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

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

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

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

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

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

Similarly, we can estimate $\mathcal{E}_{\mathbf{z}}^\tau(f_\rho^\tau) - \mathcal{E}^\tau(f_\rho^\tau)$ by considering a random variable $\zeta(z) = L_\tau(f_\rho^\tau(x) - y)$ defined on (Z, ρ). It follows from the assumptions (1.6) and (1.2) that

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

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

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

Finally, combining the above two bounds, we obtain the desired result. □

4.4 Proofs of main results

In this subsection, based on the estimation of $\mathscr{S}_i$, we give the proofs of main results. We first present a proposition, which bounds the excess generalization error.

Proposition 6. *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 $B \geq M_\tau$, $k \in \mathbb{N}$ and $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned}\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau) \leq 5C_{X,\rho,\alpha,\beta} \log\frac{5}{\delta}\left\{Bm^{-\frac{1-2(n+1)\alpha}{2-\theta}} + m^{-(\beta-(n+s)\alpha)} + m^{-\alpha s}\right\} \\ + 2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k} + 6M2^{k+1}\log\frac{5}{\delta}m^{-1},\end{aligned} \tag{4.13}$$

where $C_{X,\rho,\alpha,\beta}$ is a constant independent of m, δ, k or B.

Proof. We just directly use Lemma 5 to bound $\mathscr{S}_1$ by some properly chosen R. And the desired result follows by combining the estimates in the previous subsections. When $\sigma = m^{-\alpha}$ with $0 < \alpha < \frac{1}{2(n+1)}$, from the inequality

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

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

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

For $R \geq 1$, denote

$$\mathscr{W}(R) = \left\{\mathbf{z} \in Z^m : \|\hat{f}_{\mathbf{z}}^\tau\|_\sigma \leq R\right\}. \tag{4.14}$$

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

$$\begin{aligned}\left\{\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau)\right\} - \left\{\mathcal{E}_{\mathbf{z}}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}_{\mathbf{z}}^\tau(f_\rho^\tau)\right\} \leq \frac{1}{2}\left\{\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau)\right\} \\ + C_{X,\rho}\left(2 + \left(\frac{\Delta}{2e\alpha}\right)^{\frac{n+1}{2-\theta}}\right)\log\frac{5}{\delta}Bm^{-\frac{1-2(n+1)\alpha}{2-\theta}} + 20Rm^{-\Delta}, \quad \forall \mathbf{z} \in \mathscr{W}(R) \cap Z_3.\end{aligned} \tag{4.15}$$

Let $\gamma = \sigma^{n+s}$ and $\lambda = m^{-\beta}$ with $\beta > \alpha(n+s)$. Combining the bounds (4.4), (4.12),

(4.15), (3.7) and (3.9), we obtain that

$$\begin{aligned}\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau)) - \mathcal{E}^\tau(f_\rho^\tau) \leq\ & 24(C_\theta+1)(\widetilde{B}+M_\tau)\log\frac{5}{\delta}m^{-\frac{1}{2-\theta}} \\ & + 2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k} + 6M2^{k+1}\log\frac{5}{\delta}m^{-1} \\ & + 25\log\frac{5}{\delta}(M+M_\tau)(c+1)m^{-(\beta-(n+s)\alpha)} \\ & + 2C_{X,\rho}\left(2+\left(\frac{\Delta}{2e\alpha}\right)^{\frac{n+1}{2-\theta}}\right)\log\frac{5}{\delta}Bm^{-\frac{1-2(n+1)\alpha}{2-\theta}} + 40Rm^{-\Delta} \\ & + 9\widetilde{B}m^{-\alpha s}, \quad \forall \mathbf{z} \in \mathscr{W}(R)\cap Z_3\cap Z_2\cap Z_1.\end{aligned} \tag{4.16}$$

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 $\hat{f}_{\mathbf{z}}^\tau$ (1.5), 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

$$\frac{1}{m}\sum_{i=1}^m |y_i| \leq cM + 4M(1+\sqrt{2c})\frac{\log\frac{5}{\delta}}{\sqrt{m}} \leq (3cM+4M)\log\frac{5}{\delta} := M_\delta. \tag{4.17}$$

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 (4.16) and let $\Delta = 1+\beta$, then $\frac{R}{m^\Delta} \leq \frac{M_\delta}{m}$ and the conclusion follows with

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

□

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

Proof of Theorem 1. It follows from Proposition 2 and Proposition 6 that for any $B \geq M_\tau$,

with confidence $1-\delta$, there holds

$$\begin{aligned}
&\|\pi_{M_\tau}(\hat{f}_{\mathbf{z}}^\tau) - f_\rho^\tau\|_{L^r_{\rho_X}} \le \|\pi_B(\hat{f}_{\mathbf{z}}^\tau) - f_\rho^\tau\|_{L^r_{\rho_X}} \\
&\le c_\rho \left(5C_{X,\rho,\alpha,\beta}\log\frac{5}{\delta}\right)^{1/q} Bm^{-\Theta/q} + B\left(2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k}\right)^{1/q} \\
&+ \left(6M2^{k+1}\log\frac{5}{\delta}\right)^{1/q} B^{1-1/q}m^{-1/q},
\end{aligned} \tag{4.18}$$

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

$$\left(2c\left\{(k+1)! + 2^{k+2}k^k\right\}M^{k+1}B^{-k}\right)^{1/q} \le 2^{\frac{k+1}{q}}(5cM)^{1/q}\left\{(MkB^{-1})^k\right\}^{1/q} \tag{4.19}$$

and

$$\left(6M2^{k+1}\log\frac{5}{\delta}\right)^{1/q} B^{1-1/q}m^{-1/q} \le 2^{\frac{k+1}{q}}\left(6M\log\frac{5}{\delta}\right)^{1/q} Bm^{-1/q}. \tag{4.20}$$

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 $(MkB^{-1})^k \le 4m^{-\Theta}$, thus we find

$$2^{\frac{k+1}{q}}(5cM)^{1/q}\left\{(MkB^{-1})^k\right\}^{1/q} \le 2^{\frac{k+1}{q}}(20cM)^{1/q}m^{-\Theta/q} \le 2^{\frac{\Theta+2\epsilon}{q\epsilon}}(20cM)^{1/q}m^{-\Theta/q}. \tag{4.21}$$

Finally, we complete the proof by substituting the bounds (4.20), (4.19) and (4.21) to (4.18) 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 6 based on Lemma 6. Recall the definition of $\mathscr{W}(R)$ in (4.14). Lemma 6 yields that, when $B \ge M_\tau$, there exists $Z_3' \subset Z^m$ with the measure at least $1-5/\delta$, such that

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

where η_R is given by (4.11). From the bound above and error decomposition (3.8), we obtain

$$\begin{aligned}
&\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) \le \mathscr{S}_1 + \mathscr{S}_2 + \mathscr{S}_3 + \mathscr{D} \\
&\quad\le \frac{1}{2}\eta_R^{1-\theta}\left\{\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau)\right\}^\theta + \mathscr{S}_1' + \mathscr{S}_2 + \mathscr{S}_3 + \mathscr{D}, \quad \forall \mathbf{z} \in \mathscr{W}(R)\cap Z_3',
\end{aligned} \tag{4.22}$$

where

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

Since for $x > 0$, the inequality $x \le ax^\theta + b$ implies $x \le \max\{(2a)^{\frac{1}{1-\theta}}, 2b\}$, then (4.22) implies

$$\mathcal{E}^\tau(\pi_B(\hat{f}_{\mathbf{z}}^\tau) - \mathcal{E}^\tau(f_\rho^\tau) \le \max\{\eta_R, 2(\mathscr{S}_1' + \mathscr{S}_2 + \mathscr{S}_3 + \mathscr{D})\}, \quad \forall \mathbf{z} \in \mathscr{W}(R) \cap Z_3'. \tag{4.23}$$

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 (4.17), then combining the bounds (4.4), (4.12), (4.23), (3.7) and (3.9), we obtain

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

where

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

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 $\frac{2\mu\beta}{1+\mu} \le \frac{\epsilon}{6}$ and

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

Hence we get

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

From the proof of Proposition 6, 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

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

where

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

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

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

□

Then, 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

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

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

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

Hence assumption (1.6) 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

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

Symmetry of $\rho(y|x)$ directly yields that $\rho((f_\rho(x)-s, f_\rho(x))|x) \ge s(2\pi)^{-1/2}\sigma_x^{-1}e^{-\frac{1}{2}}$ 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 [22] 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 performance of RKHS-based algorithm (1.4) and the concerned ℓ_1-regularized algorithm (1.5) on artificial data sets. Due to the proof of Lemma 1, we can compute the dual optimization problem to the RKHS-based algorithm (1.4) for efficient numerical implementation. Based on its dual form (3.5), the output function is given by $\sum_{i=1}^m \alpha_i^* K^\sigma(x, x_i)$ with $\vec{\alpha}^* = (\alpha_1^*, \cdots, \alpha_m^*)^T$ being the solution of the following quadratic programming problem:

$$\begin{array}{ll} \underset{\vec{\alpha}\in\mathbb{R}^m}{\text{minimize}} & \frac{1}{2}\vec{\alpha}^T\mathbb{K}^\sigma\vec{\alpha} - \vec{\alpha}^T\vec{y} \\ \text{subject to} & C(\tau-1) \le \alpha_i \le C\tau, \quad \text{for all } i = 1, \cdots, m, \end{array} \tag{5.1}$$

where $C = \frac{1}{2\lambda m}$, $\vec{y} = (y_1, \cdots, y_m)^T$ and $\mathbb{K}^\sigma \in \mathbb{R}^{m\times m}$ is the kernel matrix obtained by $\mathbb{K}^\sigma_{i,j} = K^\sigma(x_i, x_j)$.

To make a comparison, we also consider estimating f_ρ^τ by the ℓ_1-regularization algorithm (1.5). For $a \in \mathbb{R}$, let $a^+ = \max\{0, a\}$ and $a^- = \max\{0, -a\}$. Then $a = a^+ - a^-$ and $|a| = a^+ + a^-$. Hence the optimization problem (1.5) can be written as

$$\begin{array}{ll} \underset{\vec{\alpha}\in\mathbb{R}^m}{\text{minimize}} & \frac{1-\tau}{m}\sum_{i=1}^m \xi_i + \frac{\tau}{m}\sum_{i=1}^m \tilde{\xi}_i + \lambda\sum_{i=1}^m(\alpha_i^+ + \alpha_i^-) \\ \text{subject to} & \xi_i = \left(\sum_{j=1}^m \left(\alpha_j^+ - \alpha_j^-\right)K^\sigma(x_i, x_j) - y_i\right)^+, \\ & \tilde{\xi}_i = \left(\sum_{j=1}^m \left(\alpha_j^+ - \alpha_j^-\right)K^\sigma(x_i, x_j) - y_i\right)^-, \quad \text{for all } i = 1, \cdots, m. \end{array}$$

Let

$$\vec{p} = \left(\underbrace{0, \cdots, 0}_{2m}, \underbrace{\frac{1-\tau}{m}, \cdots, \frac{1-\tau}{m}}_{m}, \underbrace{\frac{\tau}{m}, \cdots, \frac{\tau}{m}}_{m}\right)^T,$$

$$\vec{q} = \left(\underbrace{1, \cdots, 1}_{2m}, \underbrace{0, \cdots, 0}_{2m}\right)^T.$$

Note that the unknown α_j is totally determined by α_j^+ and α_j^-. We thus further define a vector of variables as

$$\vec{v} = \left(\alpha_1^+, \cdots, \alpha_m^+, \alpha_1^-, \cdots, \alpha_m^-, \xi_1, \cdots, \xi_m, \tilde{\xi}_1, \cdots, \tilde{\xi}_m\right)^T.$$

Then we can reformulate algorithm (1.5) as the following linear programming problem:

$$\begin{array}{ll} \underset{\vec{v}\in\mathbb{R}^{4m}}{\text{minimize}} & \vec{p}^T\vec{v} + \lambda\vec{q}^T\vec{v} \\ \text{subject to} & A\vec{v} = \vec{y}, \\ & v_i \ge 0, \quad \text{for all } i = 1, \cdots, 4m, \end{array} \tag{5.2}$$

where $A = (\mathbb{K}^\sigma, -\mathbb{K}^\sigma, I_m, -I_m)_{m\times 4m}$ is a block matrix with I_m being the m by m identity matrix. Denote the solution of (5.2) by $\vec{v}^* = (v_1^*, \cdots, v_{4m}^*)$. The output function $\hat{f}_{\mathbf{z}}^\tau$ can be recovered by the kernel expansion $\sum_{i=1}^m \alpha_i^* K^\sigma(x, x_i)$ with $\alpha_i^* = v_i^* - v_{i+m}^*$.

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. [21], [10] and [40]. Below are the expressions of these functions, where the domain $D = [a, b]^n = \{x | x \in \mathbb{R}^n, a \leq x(i) \leq b, \forall 1 \leq i \leq n\}$ and $x(i)$ stands for the i-th component of x.

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

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

$f_3(x) = \dfrac{40 f_{31}(x)}{f_{32}(x) + f_{33}(x)}$, $D = [0, 1]^2$, where

$$f_{31}(x) = \exp\left(8\left((x(1) - 0.5)^2 + (x(2) - 0.5)^2\right)\right)$$
$$f_{32}(x) = \exp\left(8\left((x(1) - 0.2)^2 + (x(2) - 0.7)^2\right)\right)$$
$$f_{33}(x) = \exp\left(8\left((x(1) - 0.7)^2 + (x(2) - 0.2)^2\right)\right).$$

$f_4(x) = 10\sin(\pi x(1)x(2)) + 20(x(3) - 0.5)^2 + 5x(4) + 10x(5) + 0x(6)$, $D = [-1, 1]^6$.

$f_5(x) = \exp\left(2\pi x(1)(\sin(x(4))) + \sin(x(2)x(3))\right)$, $D = [-0.25, 0.25]^4$.

The examples above are smooth functions and we are also concerned with 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 [13] 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.

Next we use these testing functions to model the target function and evaluate the performance of algorithms (5.1) and (5.2) on artificial samples. The sample pair (x_i, y_i) are generated as follows: x_i is independently and uniformly distributed on the domain D and $y_i = f(x_i) + e_i$, where f is taken to be the testing function and e_i is the Gaussian noises with zero mean. Three different noise levels are considered, i.e., $r_n = 0.05, 01, 0.2$, where r_n denotes the ratio of the variance of the noises e_i to that of y_i. From the data generation process, we see that the testing function models the median of the conditional distribution, which is also the regression function in our case. We thus take a prior selection of $\tau = \frac{1}{2}$

in the algorithms and adopt the cross-validation method to determine the parameters, i.e., λ and σ in (5.1) and (5.2). Concretely, we first generate 200 data points to determine the optimal parameters by 10-fold cross-validation from all combinations $(\lambda, \sigma) \in \Lambda \times \Gamma$, where Λ and Γ are finite sets of candidates given by $\Lambda = \left\{ \left(\frac{6}{5}\right)^{i-1} \times \left(\frac{1}{400}\right)^2, i = 1, \cdots, 40 \right\}$ and $\Gamma = \left\{ \frac{j}{10}, j = 1, \cdots, 10 \right\}$. Then using the obtained parameters, optimization problems (5.1) and (5.2) are solved to get the corresponding estimators on another 200 data points which are generated under the same noise level. To evaluate the performance, we calculate the relative sum of squared error (RSSE) on an independent unlabeled data set $V = \{\tilde{x}_i\}_{i=1}^{1000}$ which is uniformly distributed on the domain D. The RSSE is defined as

$$\text{RSSE} = \frac{\sum_{\tilde{x} \in V} \left(f(\tilde{x}) - \hat{f}(\tilde{x})\right)^2}{\sum_{\tilde{x} \in V} \left(f(\tilde{x}) - E(f)\right)^2},$$

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 approximate 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 observe $\|\vec{\alpha}^*\|_0$, i.e., the number of nonzero coefficients (the absolute value $\geq 10^{-5}$) in the solution vector $\vec{\alpha}^* = (\alpha_1^*, \cdots, \alpha_m^*)^T$. In our experiment, we use the standard optimization package in Matlab to solve (5.1) and (5.2). To see the efficiency, the computational times are recorded as well. For each testing function and each noise level, after choosing the optimal parameters λ and σ by cross-validation, the testing experiment is repeated 100 times and the average results are reported. All the simulations are conducted on a personal computer (Core 2-2.83 GHz, 3.23G RAM) with Matlab R2011a.

Finally, the performance of the RKHS-based regularization (5.1) and the ℓ_1-regularization (5.2) are reported in Table 1, including the RSSEs, the numbers of non-zero coefficients $\|\vec{\alpha}^*\|_0$ and the computational times. From the results, one can see that for all testing functions, the ℓ_1-regularization can generally provide almost the same accuracy as that of RKHS-based regularization, which are supportive of the theoretical prediction.

6 Conclusion

In this paper, we have studied the ℓ_1-regularized quantile regression with Gaussian kernels under the assumption that the samples are drawn from unbounded sampling processes. By using the solution of previous RKHS-based quantile regression as a stepping stone, we deduced the error bound, proved the asymptotic convergence, and obtained satisfactory

Table 1: Performance of Algorithms

		$r_n = 0.05$			$r_n = 0.1$			$r_n = 0.2$		
Function	algorithm	RSSE	$\|\vec{\alpha}^*\|_0$	Time(s)	RSSE	$\|\vec{\alpha}^*\|_0$	Time(s)	RSSE	$\|\vec{\alpha}^*\|_0$	Time(s)
f_1	RKHS	0.0252	200	1.9407	0.0466	200	1.3841	0.0698	200	1.4544
	ℓ_1	0.0223	27.77	1.2942	0.0390	31.53	1.0641	0.0639	21.12	0.9357
f_2	RKHS	0.0460	200	2.2858	0.0597	200	2.1018	0.0841	200	1.8793
	ℓ_1	0.0343	39.18	1.9636	0.0532	37.47	1.7427	0.0793	38.65	1.7248
f_3	RKHS	0.0174	200	1.1396	0.0324	200	1.0383	0.0455	200	1.0206
	ℓ_1	0.0141	16.90	0.7904	0.0214	15.34	0.7107	0.0448	11.75	0.6965
f_4	RKHS	0.0570	200	0.4950	0.0819	200	0.4421	0.0949	200	0.5084
	ℓ_1	0.0543	36.91	2.0072	0.0722	35.11	1.8978	0.0927	34.19	1.7852
f_5	RKHS	0.0237	200	0.4568	0.0418	200	0.4141	0.0701	200	0.4269
	ℓ_1	0.0193	20.00	1.3999	0.0291	19.92	1.2474	0.0477	18.01	1.1774
f_1^{pw}	RKHS	0.0253	200	1.8558	0.0481	200	1.5867	0.0784	200	1.5178
	ℓ_1	0.0258	26.42	1.2810	0.0419	21.64	1.2123	0.0728	22.93	0.8335
f_2^{pw}	RKHS	0.0463	200	2.1556	0.0672	200	2.2010	0.0828	200	2.2702
	ℓ_1	0.0413	45.17	2.3002	0.0566	42.15	2.0424	0.0836	37.29	1.8081

learning rates of the concerned algorithm. Based on the approach developed in our paper, the derived learning rates for the ℓ_1-regularized quantile regression are essentially no worse than those of its RKHS-regularized version. Considering that ℓ_1-regularized quantile regression can be formulated as a linear programming optimization problem and often leads to sparse solutions, the proposed algorithm (1.5) may perform even better than RKHS-based algorithm in most instances. There are several questions remaining to be further studied.

- Firstly, it would be interesting to investigate the robustness property of ℓ_1-regularized quantile regression. The theoretical analysis for quantile regression in a RKHS has been conducted in [6] by means of the influence function based on Bouligand derivatives. One may consider establishing a similar analysis in the following Banach space

$$\mathscr{B} = \Big\{ f : f = \sum_{j=1}^{\infty} \alpha_j K(x, x_j), \{\alpha_j\} \in \ell^1, \{x_j\} \subset X \Big\}$$

  with norm

$$\|f\| = \inf \Big\{ \sum_{j=1}^{\infty} |\alpha_j| : \quad f = \sum_{j=1}^{\infty} \alpha_j K(x, x_j) \Big\}.$$

  In this topic, one important issue one should consider is that both the loss function and the regularization term are non-smooth. Additionally, the sample dependent nature of the hypothesis space is still an obstacle when generalizing the previous robustness analysis to the case of ℓ_1-regularized algorithm. We need to develop the corresponding mathematical analysis to handle these difficulties.

- Secondly, it would be interesting to apply the τ-pinball loss to the binary classification setting. According to our recent research [14], the following modified pinball

loss

$$L'_\tau(u) = \begin{cases} u, & \text{if } u \geq 0, \\ -\tau u, & \text{if } u < 0, \end{cases}$$

enjoys most good properties possessed by the commonly used *hinge loss* $L_{hinge}(u) = \max\{u, 0\}$. And what is more, it is shown in [14] that the empirical algorithm based on loss $L'_\tau(u)$ can separate the clusters with respect to the quantile distance and thus brings noise insensitivity to the resulting classifiers. Motivated by the learning scheme (1.5), given sample data $\{(x_i, y_i)\}_{i=1}^m$ with $y_i \in \{-1, +1\}$ representing two classes, we may consider the following classification scheme

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

where

$$\mathscr{H}^b_{\mathbf{z},\sigma} = \left\{ \sum_{i=1}^m \alpha_i y_i K^\sigma(x, x_i) + b : \alpha_i \in \mathbb{R}, b \in \mathbb{R} \right\}$$

is the hypothesis space and the ℓ_1-regularization term is defined as $\Omega(f) = \sum_{i=1}^m |\alpha_i|$ for $f \in \mathscr{H}^b_{\mathbf{z},\sigma}$. One may use a similar approach in our paper to study the learning performance of this algorithm.

7 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

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

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 $a = f^\tau_\rho(x), b = f(x)$. Then

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

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

$$\mathscr{G} = \left\{ g(z) = L_\tau(\pi_B(f)(x) - y) - L_\tau(f_\rho^\tau(x) - y) : f \in \mathscr{B}_R^\sigma \right\}. \tag{7.1}$$

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 [8]) 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

$$\begin{aligned}
\left\{\mathcal{E}^\tau(\pi_B(f) - \mathcal{E}^\tau(f_\rho^\tau)\right\} - \left\{\mathcal{E}_\mathbf{z}^\tau(\pi_B(f)) - \mathcal{E}_\mathbf{z}^\tau(f_\rho^\tau)\right\} &\leq \left(1 - \frac{\theta}{2}\right) 4^{\frac{2}{2-\theta}} \epsilon^*(m, R, \sigma, B, \delta) \\
+ \frac{\theta}{2}\left\{\mathcal{E}^\tau(\pi_B(f) - \mathcal{E}^\tau(f_\rho^\tau)\right\} + 4\epsilon^*(m, R, \sigma, B, \delta), &\quad \forall f \in \mathscr{B}_R^\sigma,
\end{aligned} \tag{7.2}$$

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 (4.8), following the same idea in the proof of Proposition 4 in [41], we find

$$\begin{aligned}
\epsilon^*(m, R, \sigma, B, \delta) \leq \frac{R}{m^\Delta} + \max\{B, M_\tau\} &\left(\frac{4C_\theta(\log\frac{1}{\delta} + C_X\sigma^{-2(n+1)}) + 4C_\theta C_X(\Delta\log m)^{n+1}}{m} \right)^{\frac{1}{2-\theta}} \\
&+ \frac{8\max\{B, M_\tau\}}{3m}\left\{ \log\frac{1}{\delta} + C_X\sigma^{-2(n+1)} + C_X(\Delta\log m)^{n+1} \right\}.
\end{aligned} \tag{7.3}$$

Finally substituting (7.3) to (7.2), 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 [38].

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 $\|f\|_\infty \le Q$ and $\mathbb{E}f^2 \le c_1(\mathbb{E}f)^\theta$ for every $f \in \mathscr{F}$. If for some $c_2 > 0$ and $0 < \varsigma < 2$,*

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

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

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i) \le \frac{1}{2}\eta^{1-\theta}(\mathbb{E}f)^\theta + c'_\varsigma \eta + 2(\frac{c_1 t}{m})^{\frac{1}{2-\theta}} + \frac{18Qt}{m}, \qquad \forall f \in \mathscr{F},$$

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 (7.1), for any $\epsilon > 0$, the Lipschitz property of pinball loss also yields $\mathscr{N}_2(\mathscr{G}, \epsilon) \le \mathscr{N}_2(\mathscr{B}_R, \epsilon) = \mathscr{N}_2(\mathscr{B}_1, \epsilon R^{-1})$. Under the assumption that X is contained in the unit ball of $\mathbb{R}^n$, the covering number bound (4.9) come to existence, hence

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

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

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

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

Acknowledgement

The authors are grateful to the anonymous reviewers for their invaluable suggestions and comments which greatly improve the paper. 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. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] A. Belloni and V. Chernozhukov, ℓ_1-penalized quantile regression in high dimensional sparse models. *Annals of Statistics*, 39:82–130, 2011.

[3] G. Bennett, Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[4] P. Bradley and O. Mangasarian, Massive data discrimination via linear support vector machines. *Optimization Methods & software*, 13:1–10, 2000.

[5] V. Cherkassky, D. Gehring and F. Mulier, Comparison of adaptive methods for function estimation from samples. *IEEE Transactions on Neural Networks*, 7:969–984, 1996.

[6] A. Christmann and A. V. Messem, Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, 9:915-936, 2008.

[7] D. R. 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.

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

[9] M. Eberts and I. Steinwart, Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, 7:1–42, 2013.

[10] 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 Transactions on Neural Networks*, 14:1478–1495, 2003.

[11] Z. C. Guo and D. X. Zhou, Concentration estimates for learning with unbounded sampling. *Advances in Computational Mathematics*, 38:207–223, 2013.

[12] P. Heagerty and M. Pepe, Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Journal of the Royal Statistical Society: Series C*, 48:533–551, 1999.

[13] X. Huang, X. Jun, and S. Wang, Nonlinear system identification with continuous piecewise linear neural network. *Neurocomputing*, 77:167–177, 2012.

[14] X. Huang, L. Shi and J. A. K. Suykens, Support vector machine classifier with pinball loss. Internal Report 13-31, ESAT-SISTA, KU Leuven, Leuven, Belgium.

[15] R. Koenker and K. Hallock, Quantile regression: An introduction. *Journal of Economic Perspectives*, 15:43–56, 2001.

[16] R. Koenker and O. Geling, Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*, 96:458–468, 2001.

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

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

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

[20] T. Poggio and F. Girosi, Networks for approximation and learning. *Proceedings of the IEEE*, 9:1481–1497, 1990.

[21] A. Suárez and J.F. Lutsko. Globally optimal fuzzy decision trees for classification and regression. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 21:1297–1311, 1999.

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

[23] G. Song, H. Zhang and F. J. Hickernell, Reproducing kernel Banach spaces with the l1 norm. *Applied and Computational Harmonic Analysis*, 34:96–116, 2013.

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

[25] I. Steinwart, How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007.

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

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

[28] 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.

[29] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory. *Applicable Analysis*, 1:17–41, 2003.

[30] I. Takeuchi, Q. V. Le, T. D. Sears and A. J. Smola, Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

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

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

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

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

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

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

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

[38] Q. Wu, Y. Ying and D. X. Zhou, Multi-kernel regularized classifiers. *Journal of Complexity*, 23:108–134, 2007.

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

[40] S. Wang, X. Huang, and Y. Yam. A neural network of smooth hinge functions. *IEEE Transactions on Neural Networks*, 21:1381–1395, 2010.

[41] D. H. Xiang and D. X. Zhou, Classification with Gaussians and convex loss. *Journal of Machine Learning Research*, 10:1447–1468, 2009.

[42] D. H. Xiang, Conditional quantiles with varying Gaussians. *Advances in Computational Mathematics*, 38:723-735, 2013.

[43] K. Yu, Z. Lu and J. Stander, Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D*, 52:331–350, 2003.

[44] X. J. Zhou and D. X. Zhou, High order Parzen windows and randomized sampling. *Advances in Computational Mathematics*, 31:349–368, 2009.

[45] P. Zhao and B. Yu, On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2007.