2015 8th International Conference on BioMedical Engineering and Informatics (BMEI 2015)

Conditional Quantile Regression with ℓ_1-regularization and ϵ-insensitive Pinball Loss

Meng Li, Meijian Zhang, Hongwei Sun
School of Mathematical Sciences
University of Jinan
Jinan, China

***Abstract*—This paper considers the regularized learning schemes based on ℓ_1-regularizer and the ϵ-insensitive pinball loss in a data dependent hypothesis space. The target is the error analysis for the conditional quantile regression learning. Except for continuity and boundedness, the kernel function is not necessary to satisfy any further regularity conditions. The data dependent nature of the algorithm leads to an extra error term called hypothesis error. By concentration inequality with ℓ_2-empirical covering numbers and operator decomposition techniques, satisfied error bounds and convergence rates are explicitly derived.**

Keywords: Learning theory; conditional quantile regression; ℓ_1-regularization; ϵ-insensitive pinball loss.

I. Introduction

In this paper, we study ℓ_1-regularized quantile regression, which is generated by the ϵ-insensitive pinball loss and data dependent hypothesis space.

In our setting, functions are defined on a compact subset X of $\mathbb{R}^n$ and take values in $Y = \mathbb{R}$. We assume that the sampling process is controlled by a Borel probability distribution ρ on $Z = X \times Y$. The target of the conditional quantile regression problem is the conditional $\tau-$quantile function f_ρ^τ. Here $f_\rho^\tau(x) = t$ is defined by

$$\rho(\{y \in (-\infty, t]\}|x) \geq \tau \quad and \quad \rho(\{y \in [t, \infty)\}|x) \geq 1 - \tau, \tag{1}$$

where $\rho(\cdot|x)$ is the conditional distribution of ρ at $x \in X$ and $\tau \in (0, 1)$ is a fixed constant associated with the desired quantile level. Throughout this paper we assume that the conditional distribution $\rho(\cdot|x)$ is supported on $[-1, 1]$ for almost every $x \in X$. This assumption ensures that

$$|f_\rho^\tau(x)| \leq 1, \ a.e., x \in X \ with \ respect \ to \ \rho_X, \tag{2}$$

where ρ_X is the marginal distribution of ρ on X.

In the setting of learning theory, the distribution ρ is unknown. All we have is only a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ which is assumed to be drawn identically and independently distributed according to ρ. The empirical method for estimating the conditional τ-quantile function is based on the so-called *τ-pinball loss* $\psi_\tau(u) : \mathbb{R} \to \mathbb{R}_+$,

$$\psi_\tau(u) = \begin{cases} (1-\tau)u, & \text{if } u > 0, \\ -\tau u, & \text{if } u \leq 0. \end{cases} \tag{3}$$

We use the τ-pinball loss and define the *generalization error* for any measurable function $f : X \to \mathbb{R}$ by

$$\varepsilon^\tau(f) := \int_X \int_Y \psi_\tau(f(x) - y) d\rho(y|x) d\rho_X \tag{4}$$

Obviously, the conditional τ-quantile function f_ρ^τ minimizes the generalization error.

An *ϵ-insensitive loss* $\psi^{(\epsilon)}(u) : \mathbb{R} \to \mathbb{R}_+$ is defined for $\epsilon \geq 0$ by

$$\psi^{(\epsilon)}(u) = \max\{|u| - \epsilon, 0\} = \begin{cases} |u| - \epsilon, & \text{if } |u| \geq \epsilon, \\ 0, & otherwise. \end{cases} \tag{5}$$

Combined the ϵ-insensitive loss $\psi^{(\epsilon)}$ with the τ-pinball loss ψ_τ, we propose the *ϵ-insensitive pinball loss* $\psi_\tau^{(\epsilon)} : \mathbb{R} \to \mathbb{R}_+$ with a parameter $\epsilon \geq 0$ defined as

$$\psi_\tau^{(\epsilon)}(u) = \begin{cases} (1-\tau)(u - \epsilon), & \text{if } u > \epsilon, \\ -\tau(u + \epsilon), & \text{if } u \leq -\epsilon, \\ 0, & otherwise. \end{cases} \tag{6}$$

This loss function has been applied to a regularization scheme in the RKHS [17] as

$$f_{\mathbf{z},\lambda,\tau}^{(\epsilon)} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m \psi_\tau^{(\epsilon)}(f(x_i) - y_i) + \lambda \|f\|_K^2 \right\}. \tag{7}$$

Here, $K : X \times X \to \mathbb{R}$ is a continuous, symmetric and positive semi-definite function called a *Mercer kernel*. It generates a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ as the linear span of the set of function $\{K_x = K(\cdot, x) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$ (see [2]).

Now, we restrict our attention to coefficient-based regularization schemes in a data dependent hypothesis space. Given a continuous and bounded function $K : X \times X \to \mathbb{R}$ called *kernel*[1], the data dependent hypothesis space associated with the kernel K and sample $\mathbf{z}$ is defined as

$$\mathcal{H}_{K,\mathbf{z}} = \Big\{ \sum_{i=1}^m \alpha_i K_{x_i} : (\alpha_1, ..., \alpha_m) \in \mathbb{R}^m \Big\}. \tag{8}$$

[1]In the literature, the term "kernel" is usually used for positive definite functions on $X \times X$. Here we extend it to general functions.

Here $K_t(\cdot) = K(\cdot, t)$. The functions belonging to $\mathcal{H}_{K,\mathbf{z}}$ are entirely determined by the coefficient vector $\alpha = (\alpha_1, ..., \alpha_m) \in \mathbb{R}^m$. Hence, the penalty term $\Omega(f)$ which is a positive functional on the hypothesis space, could be imposed on the corresponding coefficients of f.

In this paper, we focus on investigating the coefficient-based regularized regression scheme with ℓ_1-regularizer

$$f_{\mathbf{z},\lambda}^{(\epsilon)} = arg \min_{f \in \mathcal{H}_{K,\mathbf{z}}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \psi_\tau^{(\epsilon)}(f(x_i) - y_i) + \lambda \Omega(f) \right\}, \tag{9}$$

where

$$\Omega(f) = \sum_{i=1}^{m} |\alpha_i| \qquad for \quad f = \sum_{i=1}^{m} \alpha_i K_{x_i}, \tag{10}$$

and $\lambda > 0$ is the penalty parameter.

In recent years, people pay close attention to the ℓ_1-regularizer. For example, coefficient-based regularization schemes with ℓ_1-penalty are studied in the least-square regression setting, e.g. [3], [6], [18]. In the quantile regression learning, Shi studies a learning scheme with Gaussian kernels in [7]. Compared with [7], we take more general kernels and ϵ-insensitive pinball loss instead of Gaussian kernels and pinball loss. Our paper is devoted to investigate how the output function $f_{\mathbf{z},\lambda}^{(\epsilon)}$ approximates the quantile regression function f_ρ^τ with suitable chosen $\lambda = m^{-\beta}$ and $\epsilon = m^{-\varpi}$ as $m \to \infty$.

In the rest of this paper, we first give the main result in Section 2. After that, we give an error decomposition technique introduced in [5] and bound the approximation error in Section 3. In Section 4, we estimate hypothesis errors, sample errors and total error by classical learning analysis. In Section 5, we deduce the error bound and learning rate by iteration method.

II. Main result

In order to improve the performance of $f_{\mathbf{z},\lambda}^{(\epsilon)}$ approximating the conditional quantile regression f_ρ^τ, we take the *projection operator* technique which has been widely used in algorithm analysis with bounded output samples, see [1].

Definition 1. *The projection operator π on the space of function on X is defined by*

$$\pi(f)(x) = \begin{cases} 1, & if\ f(x) > 1, \\ -1, & if\ f(x) < -1, \\ f(x), & if\ -1 \le f(x) \le 1. \end{cases} \tag{11}$$

We see from (2) that f_ρ^τ takes values in $[-1, 1]$. Hence, it is natural to measure the approximation ability by the distance $\|\pi(f_{\mathbf{z},\lambda}^{(\epsilon)}) - f_\rho^\tau\|_{L_{\rho_X}^{p^*}}$. Here the index $p^* > 0$ depends on the pair (ρ, τ) and takes the value $p^* = \frac{pq}{p+q}$ when the following distribution condition is satisfied.

Definition 2. *Let $p \in (0, +\infty]$ and $q \in (1, +\infty)$. We say that ρ has a τ-quantile of p-average type q if for almost all $x \in X$ with respect to ρ_X, there exist a τ-quantile $t^* \in \mathbb{R}$ and constants $0 < a_x \le 2, b_x > 0$ such that for each $s \in [0, a_x]$,*

$$\begin{aligned} \rho(\{y \in (t^* - s, t^*)\}|x) &\ge b_x s^{q-1}, \\ \rho(\{y \in (t^*, t^* + s)\}|x) &\ge b_x s^{q-1}, \end{aligned} \tag{12}$$

and that the function $\phi : X \to [0, \infty]$, $\phi(x) = b_x a_x^{q-1}$, satisfies $\phi^{-1} \in L_{\rho_X}^p$.

Note that condition (12) ensures the uniqueness of the conditional τ-quantile of $\rho(\cdot|x)$ at almost every $x \in X$, thus f_ρ^τ is well defined. For more details about this definition, see [10].

For $p \in (0, +\infty]$ and $q \in (1, +\infty)$, denote

$$\theta = \min \left\{ \frac{2}{q}, \frac{p}{p+1} \right\} \in (0, 1]. \tag{13}$$

The following variance-expectation bound can be found in [9].

Lemma 1. *If ρ has a τ-quantile of p-average type q for some $p \in (0, \infty]$ and $q \in (1, \infty)$, then for any measurable function $f : X \to Y$, there holds*

$$\begin{aligned} & E\left\{ \left(\psi_\tau(f(x) - y) - \psi_\tau(f_\rho^\tau(x) - y) \right)^2 \right\} \\ & \le C_\theta \left(\varepsilon^\tau(f) - \varepsilon^\tau(f_\rho^\tau) \right)^\theta, \end{aligned} \tag{14}$$

where the power index θ is given by (13) and the constant $C_\theta = 2^{2-\theta} q^\theta \|\phi^{-1}\|_{L_{\rho_X}^p}^\theta$.

Our Approximation condition is given as

$$f_\rho^\tau = L_{\widetilde{K}}^r g_\rho^\tau, \quad \text{for some } \ 0 < r \le 1, \quad g_\rho^\tau \in L_{\rho_X}^2(X). \tag{15}$$

Here, the kernel $\widetilde{K}$ is defined by

$$\widetilde{K}(x, y) = \int_X K(x, t) K(y, t) d\rho_X(t). \tag{16}$$

Hence, although kernel K is not positive semi-definite, $\widetilde{K}$ is a Mercer kernel, $\mathcal{H}_{\widetilde{K}}$ denotes the associated reproducing kernel Hilbert space. The integral operator $L_{\widetilde{K}} : L_{\rho_X}^2 \to L_{\rho_X}^2$ is defined as

$$L_{\widetilde{K}} f(x) = \int_X \widetilde{K}(x, t) f(t) d\rho_X(t), \quad x \in X. \tag{17}$$

Note that $L_{\widetilde{K}} = L_K L_K^*$ is a self-adjoint positive operator on $L_{\rho_X}^2$. Hence its r-th power $L_{\widetilde{K}}^r$ is well defined for any $r > 0$.

Regularization scheme (9) is different from the standard one in a RKHS. The ℓ_1-regularizer leading to a nonlinear optimization problem and the hypothesis spaces varying with samples will both cause technical difficulties in the error analysis. In order to overcome these difficulties, Zhou uses the ℓ_1-sequence to define a Banach space $\mathcal{H}_1$ in [18] as follows.

Definition 3. *Define a Banach space $\mathcal{H}_1 = \left\{ f : f = \sum_{j=1}^\infty \alpha_j K_{u_j}, \{\alpha_j\} \in \ell_1, \{u_j\} \subset X \right\}$ with the norm*

$$\|f\| = \inf \left\{ \sum_{j=1}^{\infty} |\alpha_j| : f = \sum_{j=1}^{\infty} \alpha_j K_{u_j}, \{\alpha_j\} \in \ell_1, \{u_j\} \subset X \right\} \tag{18}$$

Assume that $\kappa := \sup_{t,x\in X} |K(x,t)| < \infty$, since the kernel K is continuous, the space $\mathcal{H}_1$ can be regarded as a subspace of $C(X)$ with the inclusion map $I : \mathcal{H}_1 \to C(X)$ bounded as

$$\|f\|_\infty \le \kappa \|f\|, \qquad \forall f \in \mathcal{H}_1. \tag{19}$$

This function space contains all possible data dependent hypothesis spaces $\mathcal{H}_{K,\mathbf{z}}$. We shall use the ℓ_2-empirical covering number (see [13]) to describe the capacity property of $\mathcal{H}_1$.

Definition 4. *Let $\mathcal{F}$ be a set of functions on X, $\mathbf{x} = (x_i)_{i=1}^k \in X^k$. The metric $d_{2,\mathbf{x}}$ between functions on X is*

$$d_{2,\mathbf{x}}(f,g) = \left\{ \frac{1}{k}\sum_{i=1}^{k} \big(f(x_i) - g(x_i)\big)^2 \right\}^{\frac{1}{2}}, \quad \forall f, g \in \mathcal{F}. \tag{20}$$

For every $\zeta > 0$, the ℓ_2-empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \zeta) = \sup_{k\in\mathbb{N}} \sup_{\mathbf{x}\in X^k} \inf \Big\{ l \in \mathbb{N} : \exists \{f_i\}_{i=1}^l \subset \mathcal{F} \; such\ that \; for\ all\ f \in \mathcal{F}, there\ is \min_{1\le i\le l} d_{2,\mathbf{x}}(f, f_i) \le \zeta \Big\}. \tag{21}$$

Denote the ball of radius $R \ge 1$ as $B_R = \Big\{ f \in \mathcal{H}_1 : \|f\| \le R \Big\}$. We assume $\mathcal{H}_1$ satisfies the following capacity assumption (see [3] for more details).

Capacity assumption There exist an exponent μ with $0 < \mu < 2$ and a constant $c_{\mu,K} > 0$ such that

$$\log \mathcal{N}_2(B_1, \zeta) \le c_{\mu,K}\zeta^{-\mu}, \quad \forall\ \zeta > 0. \tag{22}$$

Suppose that $K \in C^s(X \times X)$, then the capacity assumption (22) is satisfied with

$$\mu = \begin{cases} 2n/(n+2s), & when\ 0 < s \le 1, \\ 2n/(n+2), & when\ 1 < s \le 1 + n/2, \\ n/s, & when\ s > 1 + n/2. \end{cases} \tag{23}$$

Now we can give our main result which will be proved in Section 5.

Theorem 1. *Assume Approximation condition* (15) *and Capacity condition* (22) *hold. Taking $\lambda = m^{-\frac{1}{2}}$ and $\epsilon = m^{-\varpi}$ with $\frac{r}{2} \le \varpi \le \frac{1}{2}$. Suppose that ρ has a τ-quantile of p-average type q for some $p \in (0, +\infty]$ and $q \in (1, +\infty)$, $p^* = \frac{pq}{p+1} > 0$. Then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have*

$$\left\|\pi\big(f_{\mathbf{z},\lambda}^{(\epsilon)}\big) - f_\rho^\tau\right\|_{L_{\rho_X}^{p^*}}^q \le \tilde{a} \max\left\{ b\big(\theta, \mu, \frac{\delta}{2}\big)^{\frac{2\mu}{2+\mu}}, \Big(1 + \frac{1}{m}\log\frac{20}{\delta}\Big)\log\frac{20}{\delta} \right\} m^{-\frac{r}{2}}. \tag{24}$$

Here $\tilde{a}$ is a constant independent of m or δ, $b\big(\theta, \mu, \frac{\delta}{2}\big)$ is given by (79).

In Theorem 1, the convergence rate is deduced to $O\big(m^{-\frac{r}{2}}\big)$ by taking $\beta = \frac{1}{2}$. This rate mainly depends on the smoothness of quantile function f_ρ^τ because of approximation error making the main part of total error. Under the assumption that $f_\rho^\tau \in H^s(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ and taking the Gaussian kernel $K^\sigma(x\, y)$ with $\sigma = m^{-\alpha}$, the convergence rate is $O\big(m^{\varepsilon - \frac{s}{q(n+(2-\theta)s)}}\big)$ with an arbitrarily small (but fixed) $\varepsilon > 0$ (see Theorem 2 in [7]). Compared with [7], we take more general kernels and ϵ-insensitive pinball loss instead of Gaussian kernels and pinball loss.

III. Error decomposition and approximation error

The following inequality for quantile regression in [9] plays an important role in our mathematical analysis.

Lemma 2. *Let $p \in (0, +\infty]$ and $q \in (1, +\infty)$. Denote $p^* = \frac{pq}{p+1} > 0$. If ρ has a τ-quantile of p-average type q, then for any measurable function f on X, we have*

$$\|f - f_\rho^\tau\|_{L_{\rho_X}^{p^*}} \le C_{q,\rho}\{\varepsilon^\tau(f) - \varepsilon^\tau(f_\rho^\tau)\}^{1/q}, \tag{25}$$

where $C_{q,\rho} = 2^{1-1/q}q^{1/q}\|\{(b_x a_x^{q-1})^{-1}\}_{x\in X}\|_{L_{\rho_X}^p}^{1/q}$.

Hence, to estimate $\|\pi(f_{\mathbf{z},\lambda}^{(\epsilon)}) - f_\rho^\tau\|_{L_{\rho_X}^{p^*}}$, we only need to bound the excess generalization error $\varepsilon^\tau(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})) - \varepsilon^\tau(f_\rho^\tau)$. According to the error estimation scheme in [16], a useful approach to do the error analysis for regularization schemes with sample dependent hypothesis spaces is to introduce a hypothesis error, and utilizes a suitable *error decomposition*. In this paper, we take the RKHS $\mathcal{H}_{\widehat{K}}$ with

$$\widehat{K}(x,y) = \int_X \widetilde{K}(x,t)\widetilde{K}(y,t)d\rho_X(t). \tag{26}$$

It is easy to see $L_{\widehat{K}} = L_{\widetilde{K}}^2$, so that any function $f \in \mathcal{H}_{\widehat{K}}$ can be expressed as $L_{\widetilde{K}}g$ for some $g \in L_{\rho_X}^2$. The performance of $\mathcal{H}_{\widehat{K}}$ approaching f_ρ^τ can be described through the regularizing function f_λ defined as

$$f_\lambda = arg \min_{f\in\mathcal{H}_{\widehat{K}}} \left\{ \varepsilon^\tau(f) - \varepsilon^\tau(f_\rho^\tau) + \lambda\|f\|_{\widehat{K}} \right\}. \tag{27}$$

Hence, we define

$$\mathscr{D}(\lambda) = \varepsilon^\tau(f_\lambda) - \varepsilon^\tau(f_\rho^\tau) + \lambda\|f_\lambda\|_{\widehat{K}}. \tag{28}$$

Lemma 3. *The function f_λ given by* (27) *can be expressed as*

$$f_\lambda = L_{\widetilde{K}}h_\lambda = L_K g_\lambda, \tag{29}$$

where $g_\lambda = L_K^* h_\lambda$. Moreover, g_λ is a continuous function on X and

$$\|g_\lambda\|_{L_{\rho_X}^2} = \|f_\lambda\|_{\widetilde{K}}, \quad \|f_\lambda\|_{\widetilde{K}} \le \kappa\|f_\lambda\|_{\widehat{K}} = \kappa\|h_\lambda\|_{L_{\rho_X}^2}. \tag{30}$$

The proof of Lemma 3 can be founded in [4]. Since hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ changes with samples $\mathbf{z}$, and f_λ may not belong to $\mathcal{H}_{K,\mathbf{z}}$, so we need other function as a bridge between f_λ and $f_{\mathbf{z},\lambda}^{(\epsilon)}$. As $f_\lambda = L_K g_\lambda$, we take its empirical expression, and define

$$\hat{f}_{\mathbf{z},\lambda} = \frac{1}{m}\sum_{i=1}^{m} g_\lambda(x_i)K_{x_i}. \tag{31}$$

The caused error between f_λ and $\hat{f}_{\mathbf{z},\lambda}$ is called the hypothesis error.

The other technical difficulty caused by the insensitive parameter ϵ which changes with m can be overcome by the following inequality

$$\psi_\tau(u) - \epsilon \leq \psi_\tau^{(\epsilon)}(u) \leq \psi_\tau(u). \tag{32}$$

The *τ-quantile empirical error* of a function $f: X \to \mathbb{R}$ is defined as

$$\varepsilon_{\mathbf{z}}^\tau(f) := \frac{1}{m}\sum_{i=1}^m \psi_\tau(f(x_i) - y_i). \tag{33}$$

Proposition 1. *Let $f_{\mathbf{z},\lambda}^{(\epsilon)} = \sum_{i=1}^m \alpha_{\mathbf{z},i} K_{x_i}$ be given by* (9) *with $\lambda > 0$. Then*

$$\begin{aligned}&\varepsilon^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) - \varepsilon^\tau\big(f_\rho^\tau\big) + \lambda\Omega\big(f_{\mathbf{z},\lambda}^{(\epsilon)}\big)\\ &\leq \mathscr{F}_1 + \mathscr{F}_2 + \mathscr{H}_1 + \mathscr{H}_2 + (1+\kappa)\mathscr{D}(\lambda) + \epsilon,\end{aligned} \tag{34}$$

where

$$\begin{aligned}\mathscr{F}_1 &= \Big\{\varepsilon^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) - \varepsilon^\tau(f_\rho^\tau)\Big\} - \Big\{\varepsilon_{\mathbf{z}}^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) - \varepsilon_{\mathbf{z}}^\tau(f_\rho^\tau)\Big\},\\ \mathscr{F}_2 &= \Big\{\varepsilon_{\mathbf{z}}^\tau(\hat{f}_{\mathbf{z},\lambda}) - \varepsilon_{\mathbf{z}}^\tau(f_\rho^\tau)\Big\} - \Big\{\varepsilon^\tau(\hat{f}_{\mathbf{z},\lambda}) - \varepsilon^\tau(f_\rho^\tau)\Big\},\\ \mathscr{H}_1 &= \lambda\Omega(\hat{f}_{\mathbf{z},\lambda}) - \lambda\|g_\lambda\|_{L^1_{\rho_X}},\\ \mathscr{H}_2 &= \varepsilon^\tau(\hat{f}_{\mathbf{z},\lambda}) - \varepsilon^\tau(f_\lambda).\end{aligned} \tag{35}$$

Proof: A direct decomposition shows that

$$\begin{aligned}&\varepsilon^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) - \varepsilon^\tau\big(f_\rho^\tau\big) + \lambda\Omega\big(f_{\mathbf{z},\lambda}^{(\epsilon)}\big)\\ &= \Big\{\varepsilon^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) - \varepsilon^\tau(f_\rho^\tau)\Big\} - \Big\{\varepsilon_{\mathbf{z}}^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) - \varepsilon_{\mathbf{z}}^\tau(f_\rho^\tau)\Big\}\\ &\quad + \Big\{\varepsilon_{\mathbf{z}}^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) + \lambda\Omega(f_{\mathbf{z},\lambda}^{(\epsilon)})\Big\} - \Big\{\varepsilon_{\mathbf{z}}^\tau(\hat{f}_{\mathbf{z},\lambda}) + \lambda\Omega(\hat{f}_{\mathbf{z},\lambda})\Big\}\\ &\quad + \Big\{\varepsilon_{\mathbf{z}}^\tau(\hat{f}_{\mathbf{z},\lambda}) - \varepsilon_{\mathbf{z}}^\tau(f_\rho^\tau)\Big\} - \Big\{\varepsilon^\tau(\hat{f}_{\mathbf{z},\lambda}) - \varepsilon^\tau(f_\rho^\tau)\Big\}\\ &\quad + \varepsilon^\tau(\hat{f}_{\mathbf{z},\lambda}) - \varepsilon^\tau(f_\lambda)\\ &\quad + \Big\{\varepsilon^\tau(f_\lambda) - \varepsilon^\tau(f_\rho^\tau) + \lambda\|g_\lambda\|_{L^2_{\rho_X}}\Big\}\\ &\quad + \lambda\Omega(\hat{f}_{\mathbf{z},\lambda}) - \lambda\|g_\lambda\|_{L^1_{\rho_X}}\\ &\quad + \lambda\|g_\lambda\|_{L^1_{\rho_X}} - \lambda\|g_\lambda\|_{L^2_{\rho_X}}.\end{aligned} \tag{36}$$

The fact $|y| \leq 1$ implies that $\varepsilon_{\mathbf{z}}^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) \leq \varepsilon_{\mathbf{z}}^\tau(f_{\mathbf{z},\lambda}^{(\epsilon)})$. The definition of $f_{\mathbf{z},\lambda}^{(\epsilon)}$ and the inequality (32) tell us that

$$\begin{aligned}&\varepsilon_{\mathbf{z}}^\tau(f_{\mathbf{z},\lambda}^{(\epsilon)}) + \lambda\Omega(f_{\mathbf{z},\lambda}^{(\epsilon)})\\ &\leq \frac{1}{m}\sum_{i=1}^m \psi_\tau^{(\epsilon)}\big(f_{\mathbf{z},\lambda}^{(\epsilon)}(x_i) - y_i\big) + \lambda\Omega(f_{\mathbf{z},\lambda}^{(\epsilon)}) + \epsilon\\ &\leq \frac{1}{m}\sum_{i=1}^m \psi_\tau^{(\epsilon)}\big(\hat{f}_{\mathbf{z},\lambda}(x_i) - y_i\big) + \lambda\Omega(\hat{f}_{\mathbf{z},\lambda}) + \epsilon\\ &\leq \varepsilon_{\mathbf{z}}^\tau(\hat{f}_{\mathbf{z},\lambda}) + \lambda\Omega(\hat{f}_{\mathbf{z},\lambda}) + \epsilon.\end{aligned} \tag{37}$$

Thus the second term on the right hand in (36) is less than ϵ. Due to $\|g_\lambda\|_{L^1_{\rho_X}} \leq \|g_\lambda\|_{L^2_{\rho_X}}$, we see that the last term is at most zero. The fifth item is less than $(1+\kappa)\mathscr{D}(\lambda)$ by the fact that $\|g_\lambda\|_{L^2_{\rho_X}} = \|f_\lambda\|_{\widetilde{K}} \leq \kappa\|f_\lambda\|_{\widehat{K}}$. Thus we complete the proof. ∎

$\mathscr{F}_i$ $(i = 1, 2)$ in (34) are called the *sample error*. $\mathscr{H}_i$ $(i = 1, 2)$ are called the *hypothesis error* which characterizes the approximation ability of $\hat{f}_{\mathbf{z},\lambda}$ approaching f_λ. $\mathscr{D}(\lambda)$ is called the *approximation error*.

Proposition 2. *Under the Approximation condition* (15)*, let $0 < \lambda \leq 1$. Then we have*

$$\mathscr{D}(\lambda) \leq C_0\lambda^r, \tag{38}$$

where the constant $C_0 = 2\|g_\rho^\tau\|_{L^2_{\rho_X}}$.

Proof: Let

$$\widetilde{f}_\nu = arg\min_{f\in\mathcal{H}_{\widetilde{K}}}\Big\{\|f - f_\rho^\tau\|^2_{L^2_{\rho_X}} + \nu\|f\|^2_{\widetilde{K}}\Big\}. \tag{39}$$

From the definition of f_λ, since ψ_τ is Lipschitz, we have

$$\begin{aligned}\mathscr{D}(\lambda) &\leq \varepsilon^\tau(\widetilde{f}_\nu) - \varepsilon^\tau(f_\rho^\tau) + \lambda\|\widetilde{f}_\nu\|_{\widehat{K}}\\ &\leq \|\widetilde{f}_\nu - f_\rho^\tau\|_{L^2_{\rho_X}} + \lambda\|\widetilde{f}_\nu\|_{\widehat{K}}\\ &\leq \Big\{2\big(\|\widetilde{f}_\nu - f_\rho^\tau\|^2_{L^2_{\rho_X}} + \lambda^2\|\widetilde{f}_\nu\|^2_{\widehat{K}}\big)\Big\}^{\frac{1}{2}}.\end{aligned} \tag{40}$$

$\widetilde{f}_\nu$ has the following explicit expression (see [12] for more details),

$$\widetilde{f}_\nu = L^2_{\widetilde{K}}(\nu I + L^2_{\widetilde{K}})^{-1} f_\rho^\tau. \tag{41}$$

Then the following inequalities follow from Approximation condition (15) (see [4]),

$$\|\widetilde{f}_\nu - f_\rho^\tau\|_{L^2_{\rho_X}} \leq \nu^{\frac{r}{2}}\|g_\rho^\tau\|_{L^2_{\rho_X}}, \quad \|\widetilde{f}_\nu\|^2_{\widehat{K}} \leq \nu^{r-1}\|g_\rho^\tau\|^2_{L^2_{\rho_X}}. \tag{42}$$

Taking $\nu = \lambda^2$, then the inequality (38) holds. ∎

IV. Error Analysis

In this section, we estimate the hypothesis error $\mathscr{H}_i$ $(i = 1, 2)$ and the sample error $\mathscr{F}_i$ $(i = 1, 2)$ respectively, furthermore, give the bound of excess generalization error $\varepsilon^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big) - \varepsilon^\tau\big(f_\rho^\tau\big)$.

A. Hypothesis error estimates

This subsection is devoted to estimate the hypothesis errors. Under the assumption that the sample is identically and independently drawn from ρ and $|y| \leq 1$ almost surely, we estimate $\mathscr{H}_i$ $(i = 1, 2)$ by the following lemma (see [8]).

Lemma 4. *Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable on a probability space (Z, ρ) with values in $\mathcal{H}$. Assume $\|\xi\| \leq \widetilde{M} < \infty$ almost surely. Denote $\sigma^2(\xi) = \boldsymbol{E}(\|\xi\|^2)$.*

Let $\{\xi_i\}_{i=1}^m$ be independent random drawers of ξ. For any $0<\delta<1$, with confidence $1-\delta$,

$$\left\|\frac{1}{m}\sum_{i=1}^m[\xi_i-\boldsymbol{E}(\xi_i)]\right\| \le \frac{2\widetilde{M}\log(2/\delta)}{m}+\sqrt{\frac{2\sigma^2(\xi)\log(2/\delta)}{m}}. \tag{43}$$

Now, we apply Lemma 4 to bound $\mathscr{H}_i$ $(i=1,2)$.

Proposition 3. *For any $0<\delta<1$, with confidence $1-\delta$, we have*

$$\mathscr{H}_1 \le \frac{3\kappa\mathscr{D}(\lambda)\log(4/\delta)}{m}+\frac{\kappa}{2}\mathscr{D}(\lambda) \tag{44}$$

and

$$\mathscr{H}_2 \le \kappa^2\frac{\mathscr{D}(\lambda)}{\lambda\sqrt{m}}\left\{\frac{2\log(4/\delta)}{\sqrt{m}}+\sqrt{2\log(4/\delta)}\right\}. \tag{45}$$

Proof: Firstly, we estimate $\mathscr{H}_1$. Recall $\hat{f}_{\mathbf{z},\lambda}=\frac{1}{m}\sum_{i=1}^m g_\lambda(x_i)K_{x_i}$, then $\Omega(\hat{f}_{\mathbf{z},\lambda})=\frac{1}{m}\sum_{i=1}^m|g_\lambda(x_i)|$. Applying Lemma 4 to the random variable $\xi=|g_\lambda(x)|$ on (X,ρ_X) with values in $\mathbb{R}$, then $|\xi|\le\|g_\lambda\|_\infty$. From (30) and the definition of $\mathscr{D}(\lambda)$, there is

$$\begin{aligned}|g_\lambda(x)|=|L_K^*h_\lambda(x)|&=\Big|\int_X K(t,x)h_\lambda(t)d\rho_X(t)\Big|\\ &\le\kappa\|h_\lambda\|_{L^2_{\rho_X}}\le\kappa\frac{\mathscr{D}(\lambda)}{\lambda}.\end{aligned} \tag{46}$$

The expectation $\mathbf{E}(\xi_i)$ and $\sigma^2(\xi)$ satisfy that

$$\begin{aligned}&\mathbf{E}\xi=\int_X|g_\lambda(x)|d\rho_X=\|g_\lambda\|_{L^1_{\rho_X}},\\ &\sigma^2(\xi)=\mathbf{E}(\xi^2)=\int_X g_\lambda^2 d\rho_X=\|g_\lambda\|^2_{L^2_{\rho_X}}\le\kappa^2\big(\frac{\mathscr{D}(\lambda)}{\lambda}\big)^2.\end{aligned} \tag{47}$$

Thus, with confidence $1-\delta/2$, there holds

$$\begin{aligned}&\Omega(\hat{f}_{\mathbf{z},\lambda})-\|g_\lambda\|_{L^1_{\rho_X}}\\ &\le\frac{2\kappa\mathscr{D}(\lambda)\log(4/\delta)}{\lambda m}+\sqrt{\frac{2\kappa^2\mathscr{D}^2(\lambda)\log(4/\delta)}{\lambda^2 m}}.\end{aligned} \tag{48}$$

Finally, we have

$$\begin{aligned}\mathscr{H}_1&=\lambda\left\{\Omega(\hat{f}_{\mathbf{z},\lambda})-\|g_\lambda\|_{L^1_{\rho_X}}\right\}\\ &\le\frac{2\kappa\mathscr{D}(\lambda)\log(4/\delta)}{m}+\sqrt{\frac{2\kappa^2\mathscr{D}^2(\lambda)\log(4/\delta)}{m}}\\ &\le\frac{3\kappa\mathscr{D}(\lambda)\log(4/\delta)}{m}+\frac{\kappa}{2}\mathscr{D}(\lambda).\end{aligned} \tag{49}$$

For $\mathscr{H}_2$, since ψ_τ is a Lipschitz function, the following inequality

$$\begin{aligned}\varepsilon^\tau(\hat{f}_{\mathbf{z},\lambda})-\varepsilon^\tau(f_\lambda)&\le\int_X\Big|\hat{f}_{\mathbf{z},\lambda}(x)-f_\lambda(x)\Big|d\rho_X(x)\\ &\le\|\hat{f}_{\mathbf{z},\lambda}-f_\lambda\|_{L^2_{\rho_X}}\end{aligned} \tag{50}$$

holds true. Now we also apply Lemma 4 to the random variable $\varsigma(x)=g_\lambda(x)K_x$ on (X,ρ_X) with values in the Hilbert space $L^2_{\rho_X}(X)$. It satisfies

$$\begin{aligned}&\mathbf{E}(\varsigma)=L_Kg_\lambda=f_\lambda,\\ &\|\varsigma\|_{L^2_{\rho_X}}\le\kappa\|g_\lambda\|_\infty\le\kappa^2\frac{\mathscr{D}(\lambda)}{\lambda},\\ &\sigma^2(\varsigma)=\mathbf{E}\|\varsigma\|^2_{L^2_{\rho_X}}\le\kappa^2\|g_\lambda\|^2_{L^2_{\rho_X}}\le\kappa^4\big(\frac{\mathscr{D}(\lambda)}{\lambda}\big)^2.\end{aligned} \tag{51}$$

Then with confidence $1-\delta/2$, we have

$$\begin{aligned}\|\hat{f}_{z,\lambda}-f_\lambda\|_{L^2_{\rho_X}}&\le\frac{2\kappa^2\mathscr{D}(\lambda)\log(4/\delta)}{\lambda m}+\sqrt{\frac{2\kappa^4\mathscr{D}^2(\lambda)\log(4/\delta)}{\lambda^2 m}}\\ &=\kappa^2\frac{\mathscr{D}(\lambda)}{\lambda\sqrt{m}}\left\{\frac{2\log(4/\delta)}{\sqrt{m}}+\sqrt{2\log(4/\delta)}\right\}.\end{aligned} \tag{52}$$

Hence we complete the proof. ∎

B. Sample error estimates

Since either $f^{(\epsilon)}_{\mathbf{z},\lambda}$ or $\hat{f}_{\mathbf{z},\lambda}$ is a function-valued random variable which depends on the sample $\mathbf{z}$, we need to estimate the sample error in the data independent space $\mathcal{H}_1$ which contains all possible hypothesis spaces $\mathcal{H}_{K,\mathbf{z}}$. Our estimations for $\mathscr{F}_1$ and $\mathscr{F}_2$ are based on the following concentration inequality, see [6], [14].

Lemma 5. *Let $\mathcal{F}$ be a class of measurable functions on Z. Assume that there are constants B, $c>0$ and $\beta\in[0,1]$ such that $\|f\|_\infty\le B$ and $\boldsymbol{E}f^2\le c(\boldsymbol{E}f)^\beta$ for every $f\in\mathcal{F}$. If for some $a>0$ and $\mu\in(0,2)$,*

$$\log\mathcal{N}_2(\mathcal{F},\zeta)\le a\zeta^{-\mu},\qquad\forall\zeta>0, \tag{53}$$

then there exists a constant c_μ depending only on μ such that for any $0<\delta<1$, with confidence $1-\delta$, there holds

$$\begin{aligned}\boldsymbol{E}f-\frac{1}{m}\sum_{i=1}^m f(z_i)&\le\frac{1}{2}\omega^{1-\beta}(\boldsymbol{E}f)^\beta+c_\mu\omega\\ &+2\big(\frac{c\log\frac{1}{\delta}}{m}\big)^{\frac{1}{2-\beta}}+\frac{18B\log\frac{1}{\delta}}{m},\quad\forall f\in\mathcal{F},\end{aligned} \tag{54}$$

where $\omega=\max\{c^{\frac{2-\mu}{4-2\beta+\mu\beta}}\big(\frac{a}{m}\big)^{\frac{2}{4-2\beta+\mu\beta}},B^{\frac{2-\mu}{2+\mu}}\big(\frac{a}{m}\big)^{\frac{2}{2+\mu}}\}$. The same bound also holds true for $\frac{1}{m}\sum_{i=1}^m f(z_i)-\boldsymbol{E}f$.

The following proposition which has been proved in [4] will be utilized to bound $\mathscr{F}_1$.

Proposition 4. *Suppose that ρ has a τ-quantile of p-average type q for some $p\in(0,+\infty]$ and $q\in(1,+\infty)$. Let $R\ge1$ and $0<\lambda\le1$. Assume B_1 satisfies the Capacity assumption (22) with some $0<\mu<2$. Then, for any $0<\delta<1$, with* *confidence* $1-\delta$*, there holds, for all* $f \in B_R$*,*

$$\begin{aligned}&\left\{\varepsilon^{\tau}\left(\pi(f)\right)-\varepsilon^{\tau}(f_{\rho}^{\tau})\right\}-\left\{\varepsilon_{\mathbf{z}}^{\tau}\left(\pi(f)\right)-\varepsilon_{\mathbf{z}}^{\tau}(f_{\rho}^{\tau})\right\}\\&\le \frac{1}{2}C_1^{1-\theta}R^{\frac{2\mu(1-\theta)}{2+\mu}}m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}\left\{\varepsilon^{\tau}(\pi(f))-\varepsilon^{\tau}(f_{\rho}^{\tau})\right\}^{\theta}\\&+\left(36+2C_{\theta}^{\frac{1}{2-\theta}}\right)\log(1/\delta)m^{-\frac{1}{2-\theta}}+C_2R^{\frac{2\mu}{2+\mu}}m^{-\frac{2}{4-2\theta+\mu\theta}}.\end{aligned} \tag{55}$$

Here C_1 *and* C_2 *are the constants depending on* μ*,* θ*,* $c_{\mu,K}$ *and* C_θ*.*

Proposition 5. *Under the assumptions of Proposition 4. Then, for any* $0<\delta<1$*, with confidence* $1-\delta$*, there holds*

$$\begin{aligned}\mathscr{F}_2 \le& C_3\Big(1+\frac{1}{m}\log\frac{5}{\delta}\Big)\log\frac{5}{\delta}\times m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}\\&\times\left(\lambda^{r-1}m^{-\frac{\theta}{2}}+\lambda^{r-1+\theta}\right).\end{aligned} \tag{56}$$

Here C_3 *is a constant independent of* m*,* λ*, and* δ*.*

Proof: We bound $\mathscr{F}_2$ by considering the function set for $R \ge 1$,

$$\mathcal{G}_R=\left\{\psi_{\tau}(f(x)-y)-\psi_{\tau}(f_{\rho}^{\tau}(x)-y)\ :\ f\in B_R\right\}. \tag{57}$$

Since $|f_{\rho}^{\tau}(x)|\le 1$ and $\|f\|_{\infty}\le \kappa R$, for any $g\in\mathcal{G}_R$, we have

$$|g(\mathbf{z})|\le |f(x)-f_{\rho}^{\tau}(x)|\le \|f\|_{\infty}+1\le \kappa R+1. \tag{58}$$

By Lemma 1, the variance-expectation condition of $g(\mathbf{z})$ is satisfied with θ given by (13) and $c=C_\theta$, $\beta=\theta$. We see from the Lipschitz property of pinball loss that

$$\log\mathcal{N}_2(\mathcal{G}_R,\zeta)\le c_{\mu,K}R^{\mu}\zeta^{-\mu}. \tag{59}$$

Applying Lemma 5 to $\mathcal{G}_R$, then for any $\delta\in(0,1)$, with confidence $1-\delta$, there holds that, for any $f\in B_R$,

$$\begin{aligned}&\left\{\varepsilon_{\mathbf{z}}^{\tau}(f)-\varepsilon_{\mathbf{z}}^{\tau}(f_{\rho}^{\tau})\right\}-\left\{\varepsilon^{\tau}(f)-\varepsilon^{\tau}(f_{\rho}^{\tau})\right\}\\&\le \frac{1}{2}\omega^{1-\theta}\left\{\varepsilon^{\tau}(f)-\varepsilon^{\tau}(f_{\rho}^{\tau})\right\}^{\theta}\\&+2\left(\frac{C_\theta\log\frac{1}{\delta}}{m}\right)^{\frac{1}{2-\theta}}+\frac{18(\kappa R+1)\log\frac{1}{\delta}}{m}+c_{\mu}\omega.\end{aligned} \tag{60}$$

Here

$$\omega=\widetilde{C}Rm^{-\frac{2}{4-2\theta+\mu\theta}}, \tag{61}$$

with $\widetilde{C}=C_{\theta}^{\frac{2-\mu}{4-2\theta+\mu\theta}}c_{\mu,K}^{\frac{2}{4-2\theta+\mu\theta}}+(\kappa+1)^{\frac{2-\mu}{2+\mu}}c_{\mu,K}^{\frac{2}{2+\mu}}$.

From the estimation for $\mathscr{H}_1$, for any $\delta\in(0,1)$, with confidence $1-\frac{2\delta}{5}$, we have

$$\frac{1}{m}\sum_{i=1}^{m}|g_{\lambda}(x_i)|-\|g_{\lambda}\|_{L^1_{\rho_X}}\le\frac{3\kappa\mathscr{D}(\lambda)\log(5/\delta)}{\lambda m}+\frac{\kappa}{2}\frac{\mathscr{D}(\lambda)}{\lambda}, \tag{62}$$

which implies the existence of a subset V_1 of Z^m with measure at most $\frac{2\delta}{5}$ such that

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}g_{\lambda}(x_i)\le&\max\left\{\frac{3\kappa\mathscr{D}(\lambda)\log(5/\delta)}{\lambda m}+\frac{3\kappa}{2}\frac{\mathscr{D}(\lambda)}{\lambda},1\right\}\\&\triangleq R_{\lambda},\quad\forall\mathbf{z}\in Z^m\backslash V_1.\end{aligned} \tag{63}$$

This inequality guarantees that for every $\mathbf{z}\in Z^m\backslash V_1$, we have $\hat{f}_{\mathbf{z},\lambda}\in B_{R_\lambda}$. By Lemma 5 and (60), there exists V_{R_λ} with measure at most $\frac{\delta}{5}$ such that for every $\mathbf{z}\in Z^m\backslash(V_1\cup V_{R_\lambda})$, we have $\hat{f}_{\mathbf{z},\lambda}\in B_{R_\lambda}$, and

$$\begin{aligned}&\left\{\varepsilon_{\mathbf{z}}^{\tau}(\hat{f}_{\mathbf{z},\lambda})-\varepsilon_{\mathbf{z}}^{\tau}(f_{\rho}^{\tau})\right\}-\left\{\varepsilon^{\tau}(\hat{f}_{\mathbf{z},\lambda})-\varepsilon^{\tau}(f_{\rho}^{\tau})\right\}\\\le&\frac{1}{2}\widetilde{C}^{1-\theta}R_{\lambda}^{1-\theta}m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}\left\{\varepsilon^{\tau}(\hat{f}_{\mathbf{z},\lambda})-\varepsilon^{\tau}(f_{\rho}^{\tau})\right\}^{\theta}\\&+18(\kappa+1)R_{\lambda}m^{-1}\log\frac{5}{\delta}+c_{\mu}\widetilde{C}R_{\lambda}m^{-\frac{2}{4-2\theta+\mu\theta}}\\&+2\Big(\frac{C_\theta\log\frac{5}{\delta}}{m}\Big)^{\frac{1}{2-\theta}}\\\le&\frac{1}{2}\widetilde{C}^{1-\theta}R_{\lambda}^{1-\theta}m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}\left|\varepsilon^{\tau}(\hat{f}_{\mathbf{z},\lambda})-\varepsilon^{\tau}(f_{\lambda})\right|^{\theta}+\\&\frac{1}{2}\widetilde{C}^{1-\theta}R_{\lambda}^{1-\theta}m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}\left\{\varepsilon^{\tau}(f_{\lambda})-\varepsilon^{\tau}(f_{\rho}^{\tau})\right\}^{\theta}\\&+18(\kappa+1)R_{\lambda}m^{-1}\log\frac{5}{\delta}+c_{\mu}\widetilde{C}R_{\lambda}m^{-\frac{2}{4-2\theta+\mu\theta}}\\&+2\Big(\frac{C_\theta\log\frac{5}{\delta}}{m}\Big)^{\frac{1}{2-\theta}}.\end{aligned} \tag{64}$$

It follows from Proposition 2 that $\varepsilon^{\tau}(f_\lambda)-\varepsilon^{\tau}(f_{\rho}^{\tau})\le\mathscr{D}(\lambda)\le C_0\lambda^r$, and

$$R_{\lambda}\le(3\kappa C_0+1)\lambda^{r-1}\Big(\frac{1}{m}\log\frac{5}{\delta}+1\Big). \tag{65}$$

According to Proposition 3, we see that there exists a subset V_2 of Z^m with measure at most $\frac{2\delta}{5}$ such that for every $\mathbf{z}\in Z^m\backslash V_2$,

$$\varepsilon^{\tau}(\hat{f}_{\mathbf{z},\lambda})-\varepsilon^{\tau}(f_{\lambda})\le\kappa^2\frac{\mathscr{D}(\lambda)}{\lambda\sqrt{m}}\left\{\frac{2\log(5/\delta)}{\sqrt{m}}+\sqrt{2\log(5/\delta)}\right\}. \tag{66}$$

Let $V=V_1\cup V_2\cup V_{R_\lambda}$. Obviously, the measure of V is at most δ and for every $\mathbf{z}\in Z^m\backslash V$, the above formulas hold. Finally, we plug the above estimates into (64). This completes the proof of Proposition 5. ∎

C. Total error bound

For $R\ge 1$, denote

$$\mathcal{W}(R)=\left\{\mathbf{z}\in Z^m:\left\|f_{\mathbf{z},\lambda}^{(\epsilon)}\right\|\le R\right\}. \tag{67}$$

Proposition 6. *Suppose that* ρ *has a* τ*-quantile of* p*-average type* q *for some* $p\in(0,\infty]$ *and* $q\in(1,\infty)$*, and that Approximation condition* (15) *and Capacity condition* (22) *hold. Let* $0<\lambda\le 1$*,* $R\ge 1$ *and* $0<\delta<1$*. Then, there exists a subset* U_R *of* Z^m *with measure at most* δ *such that for any* $\mathbf{z}\in\mathcal{W}(R)\backslash U_R$*, we have*

$$\begin{aligned}&\varepsilon^{\tau}\left(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\right)-\varepsilon^{\tau}\left(f_{\rho}^{\tau}\right)+\lambda\Omega\left(f_{\mathbf{z},\lambda}^{(\epsilon)}\right)\\&\le\hat{C}m^{-\frac{2}{4-2\theta+\mu\theta}}R^{\frac{2\mu}{2+\mu}}+2\epsilon\\&+C_4\Big(1+\frac{1}{m}\log\frac{10}{\delta}\Big)\Big(\log\frac{10}{\delta}\Big)\Psi(m,\lambda).\end{aligned} \tag{68}$$

Here $\hat{C}$ and C_4 are constants independent of m, λ, δ, and

$$\Psi(m,\lambda)=\lambda^r+\lambda^{r-1}m^{-\frac{1}{2}}+\lambda^{r-1+\theta}m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}. \quad (69)$$

Proof: By Proposition 3, we see that there exists a subset U_1 of Z^m with measure at most $2\delta/5$ such that for any $\mathbf{z}\in Z^m\backslash U_1$, there holds

$$\mathscr{H}_1\le\frac{3\kappa\mathscr{D}(\lambda)\log(10/\delta)}{m}+\frac{\kappa}{2}\mathscr{D}(\lambda) \quad (70)$$

and

$$\mathscr{H}_2\le\kappa^2\frac{\mathscr{D}(\lambda)}{\lambda\sqrt{m}}\left\{\frac{2\log(10/\delta)}{\sqrt{m}}+\sqrt{2\log(10/\delta)}\right\}. \quad (71)$$

Proposition 4 ensures the existence of a subset V_R of Z^m with measure at most $\delta/10$, such that for any $\mathbf{z}\in\mathcal{W}(R)\backslash V_R$,

$$\begin{aligned}\mathscr{F}_1\le\ &\frac{1}{2}C_1^{1-\theta}R^{\frac{2\mu(1-\theta)}{2+\mu}}m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}\left\{\varepsilon^\tau(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)}))-\varepsilon^\tau(f_\rho^\tau)\right\}^\theta\\&+\left(36+2C_\theta^{\frac{1}{2-\theta}}\right)\log\frac{10}{\delta}m^{-\frac{1}{2-\theta}}+C_2R^{\frac{2\mu}{2+\mu}}m^{-\frac{2}{4-2\theta+\mu\theta}}.\end{aligned} \quad (72)$$

Proposition 5 tells us that there exists a subset U_2 of Z^m with measure at most $\delta/2$ such that

$$\begin{aligned}\mathscr{F}_2\le&C_3\Big(1+\frac{1}{m}\log\frac{10}{\delta}\Big)\log\frac{10}{\delta}\times m^{-\frac{2(1-\theta)}{4-2\theta+\mu\theta}}\times\\&\Big(\lambda^{r-1}m^{-\frac{\theta}{2}}+\lambda^{r-1+\theta}\Big),\qquad\forall\mathbf{z}\in Z^m\backslash U_2.\end{aligned} \quad (73)$$

Taking $U_R=U_1\cup U_2\cup V_R$, the measure of U_R is at most δ, plugging the above estimates into (34), then for every $z\in\mathcal{W}(R)\backslash U_R$, we get

$$\begin{aligned}&\varepsilon^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big)-\varepsilon^\tau\big(f_\rho^\tau\big)+\lambda\Omega\big(f_{\mathbf{z},\lambda}^{(\epsilon)}\big)\\&\le\frac{1}{2}C_4\Big(1+\frac{1}{m}\log\frac{10}{\delta}\Big)\big(\log\frac{10}{\delta}\big)\Psi(m,\lambda)+\epsilon\\&+\frac{1}{2}C_1^{1-\theta}R^{\frac{2\mu(1-\theta)}{2+\mu}}m^{-\frac{2(1-\theta)}{4+\mu\theta-2\theta}}\left\{\varepsilon_\tau(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)}))-\varepsilon_\tau(f_\rho^\tau)\right\}^\theta\\&+C_2R^{\frac{2\mu}{2+\mu}}m^{-\frac{2}{4+\mu\theta-2\theta}}.\end{aligned} \quad (74)$$

Here C_4 is a constant independent of m, λ, δ, and $\Psi(m,\lambda)$ is defined by (69).

Next, let $t=\varepsilon^\tau\big(\pi(f_{\mathbf{z},\lambda}^{(\epsilon)})\big)-\varepsilon^\tau\big(f_\rho^\tau\big)+\lambda\Omega\big(f_{\mathbf{z},\lambda}^{(\epsilon)}\big)$. Hence, the inequality (74) can be expressed as

$$t-\frac{1}{2}C_1^{1-\theta}R^{\frac{2\mu(1-\theta)}{2+\mu}}m^{-\frac{2(1-\theta)}{4+\mu\theta-2\theta}}t^\theta-\Pi\le 0, \quad (75)$$

where Π is the rest terms. From Lemma 7.2 in [2], the (75) has a unique positive solution t^* which can be bounded as

$$\begin{aligned}t^*&\le\max\left\{C_1R^{\frac{2\mu}{2+\mu}}m^{-\frac{2}{4+\mu\theta-2\theta}},2\Pi\right\}\\&\le C_1R^{\frac{2\mu}{2+\mu}}m^{-\frac{2}{4+\mu\theta-2\theta}}+2\Pi.\end{aligned} \quad (76)$$

Our proof is complete. ■

V. Deriving convergence rates by iteration

The definition of $f_{\mathbf{z},\lambda}^{(\epsilon)}$ and $|y|\le 1$ tell us that

$$\begin{aligned}\lambda\big\|f_{\mathbf{z},\lambda}^{(\epsilon)}\big\|\le\lambda\Omega(f_{\mathbf{z},\lambda}^{(\epsilon)})&\le\frac{1}{m}\sum_{i=1}^m\psi_\tau^{(\epsilon)}(f_{\mathbf{z},\lambda}^{(\epsilon)}(x_i)-y_i)+\lambda\Omega(f_{\mathbf{z},\lambda}^{(\epsilon)})\\&\le\frac{1}{m}\sum_{i=1}^m\psi_\tau^{(\epsilon)}(-y_i)+\lambda\Omega(0)\le 1.\end{aligned} \quad (77)$$

So $\big\|f_{\mathbf{z},\lambda}^{(\epsilon)}\big\|\le\frac{1}{\lambda}$ holds almost surely. Hence one may choose $R=\frac{1}{\lambda}$, but this choice is too rough. This motivates us to deduce a tight bound for $\big\|f_{\mathbf{z},\lambda}^{(\epsilon)}\big\|$ by the iteration technique, which has been widely utilized in learning error estimate, see [11], [15]. In this section, we denote $\omega_0=\frac{2}{4-2\theta+\mu\theta}$ for simplicity, it follows that $\frac{1}{2}<\omega_0<1$ from $0<\mu<2$ and $0<\theta\le 1$.

Lemma 6. *Under the assumptions in Proposition 6. Taking $\lambda=m^{-\beta}$ and $\epsilon=m^{-\varpi}$ with $0<\beta,\varpi\le\frac{1}{2}$. Then, for any $0<\delta<1$, with confidence $1-\delta$, there holds*

$$\big\|f_{\mathbf{z},\lambda}^{(\epsilon)}\big\|\le\big((\hat{C}+1)^{\frac{2+\mu}{2-\mu}}+\bar{C}\big)b(\theta,\mu,\delta)m^\gamma, \quad (78)$$

where

$$\begin{aligned}b(\theta,\mu,\delta)=(1+L_{\theta,\mu})\times\Big(\log\frac{10}{\delta}+\log L_{\theta,\mu}\Big)\\\times\Big(1+\frac{1}{m}\big(\log\frac{10}{\delta}+\log L_{\theta,\mu}\big)\Big),\end{aligned} \quad (79)$$

with γ is given by (82) *and $L_{\theta,\mu}$ is given by* (91).

Proof: Applying $\lambda=m^{-\beta}$ and $\epsilon=m^{-\varpi}$ with $0<\beta,\varpi\le\frac{1}{2}<\omega_0$ to Proposition 6, then for any $R\ge 1$, there exists a subset V_R of Z^m with measure at most δ such that

$$\big\|f_{\mathbf{z},\lambda}^{(\epsilon)}\big\|\le a_mR^{\frac{2\mu}{2+\mu}}+b_m,\qquad\forall\,\mathbf{z}\in\mathcal{W}(R)\backslash V_R. \quad (80)$$

The constants are given by

$$\begin{aligned}a_m&=\hat{C}m^{\beta-\omega_0},\\b_m&=\bar{C}\Big(1+\frac{1}{m}\log\frac{10}{\delta}\Big)\big(\log\frac{10}{\delta}\big)m^\gamma\triangleq b_\delta m^\gamma,\end{aligned} \quad (81)$$

where $\bar{C}$ is a constant independent of m, λ, δ, and

$$\gamma=\max\Big\{\beta(2-r)-\frac{1}{2},\,\beta(1-r),\,\beta-\varpi\Big\}\ge 0. \quad (82)$$

It follows that

$$\mathcal{W}(R)\subset\mathcal{W}\left(a_mR^{\frac{2\mu}{2+\mu}}+b_m\right)\cup V_R. \quad (83)$$

Define a sequence $\{R^{(l)}\}_{l=0}^L$ by $R^{(0)}=\lambda^{-1}$ and, for $l\ge 1$,

$$R^{(l)}=a_m\left(R^{(l-1)}\right)^{\frac{2\mu}{2+\mu}}+b_m,\quad l\in\mathbb{N}. \quad (84)$$

Inequality (77) ensures that $\mathcal{W}(R^{(0)})=Z^m$. Thus we have

$$\begin{aligned}Z^m=&\mathcal{W}(R^{(0)})\subseteq\mathcal{W}(R^{(1)})\cup V_{R^{(0)}}\\&\subseteq\ldots\subseteq\mathcal{W}(R^{(L)})\cup\left(\bigcup_{l=0}^{L-1}V_{R^{(L)}}\right).\end{aligned} \quad (85)$$

But $\rho\Big(\bigcup_{l=0}^{L-1} V_{R^{(L)}}\Big) \leq L\delta$. Hence the measure of $\mathcal{W}(R^{(L)})$ is at least $1 - L\delta$.

Denote $\Delta = \frac{2\mu}{2+\mu} < 1$. By the iteration formula (84) , we have

$$\begin{aligned} R^{(L)} &\leq a_m^{1+\Delta+\Delta^2+\ldots+\Delta^{L-1}} (R^{(0)})^{\Delta^L} \\ &+ \sum_{l=1}^{L-1} a_m^{1+\Delta+\Delta^2+\ldots+\Delta^{l-1}} b_m^{\Delta^l} + b_m \\ &= a_m^{\frac{1-\Delta^L}{1-\Delta}} m^{\beta\Delta^L} + \sum_{l=1}^{L-1} a_m^{\frac{1-\Delta^l}{1-\Delta}} b_m^{\Delta^l} + b_m \\ &\leq (1+\hat{C})^{\frac{1}{1-\Delta}} m^{(\beta-\omega_0)\frac{1}{1-\Delta} + \Delta^L[\beta - \frac{1}{1-\Delta}(\beta-\omega_0)]} \\ &+ a_m^{\frac{1}{1-\Delta}} \sum_{l=1}^{L-1} \Big(a_m^{\frac{-1}{1-\Delta}} b_m\Big)^{\Delta^l} + b_m \\ &\leq (1+\hat{C})^{\frac{2+\mu}{2-\mu}} m^{(\beta-\omega_0)\frac{2+\mu}{2-\mu} + \Delta^L(\frac{2+\mu}{2-\mu}\omega_0 - \frac{2\beta\mu}{2-\mu})} \\ &+ L a_m^{\frac{1}{1-\Delta}} \max\{a_m^{\frac{-1}{1-\Delta}} b_m, 1\} + b_m \\ &\leq (1+\hat{C})^{\frac{2+\mu}{2-\mu}} m^{(\beta-\omega_0)\frac{2+\mu}{2-\mu} + \Delta^L(\frac{2+\mu}{2-\mu}\omega_0 - \frac{2\beta\mu}{2-\mu})} \\ &+ (L+1) b_\delta m^\gamma + L\hat{C}^{\frac{2+\mu}{2-\mu}} m^{(\beta-\omega_0)\frac{2+\mu}{2-\mu}}. \end{aligned} \tag{86}$$

Note that $0 < \beta \leq \frac{1}{2} < \omega_0$, to ensure that

$$(\beta-\omega_0)\frac{2+\mu}{2-\mu} + \Delta^L(\frac{2+\mu}{2-\mu}\omega_0 - \frac{2\beta\mu}{2-\mu}) \leq \beta(1-r) \leq \gamma, \tag{87}$$

we only need

$$\Delta^{-L} \geq \frac{\omega_0 - \Delta\beta}{\omega_0 - \Delta\beta - (1-\Delta)r\beta}. \tag{88}$$

Thus, denote

$$L_0 = \max\left\{\left[\log_{\frac{2+\mu}{2\mu}} \frac{\omega_0 - \Delta\beta}{\omega_0 - \Delta\beta - (1-\Delta)r\beta}\right] + 1, 1\right\}. \tag{89}$$

Then, with confidence $1 - L_0\delta$, we have

$$\big\|f_{\mathbf{z},\lambda}^{(\epsilon)}\big\| \leq \Big(b_\delta + (\hat{C}+1)^{\frac{2+\mu}{2-\mu}}\Big)(1+L_0)m^\gamma. \tag{90}$$

A simple computation shows that

$$L_0 \leq \log_{\frac{2+\mu}{2\mu}} \frac{\omega_0 - \frac{\Delta}{2}}{\omega_0 - \frac{1}{2}} + 1 \triangleq L_{\theta,\mu}. \tag{91}$$

Then our result follows by replacing δ by $\frac{\delta}{L_{\theta,\mu}}$. ■

Next, we will give the proof of our main result.

Proof of Theorem 1. Applying Lemma 2, Lemma 6 and Proposition 6, and replacing δ by $\frac{\delta}{2}$ in both results, with confidence $1-\delta$, we have

$$\begin{aligned} &\Big\|\pi\big(f_{\mathbf{z},\lambda}^{(\epsilon)}\big) - f_\rho^\tau\Big\|_{L_{\rho_X}^{p^*}}^q \\ &\leq \tilde{a}\max\Big\{b\big(\theta,\mu,\frac{\delta}{2}\big)^{\frac{2\mu}{2+\mu}}, \big(1+\frac{1}{m}\log\frac{20}{\delta}\big)\log\frac{20}{\delta}\Big\}\Gamma(m,\lambda). \end{aligned} \tag{92}$$

Here, $\tilde{a}$ is a constant independent of m, δ, and

$$\Gamma(m,\lambda) = \Psi(m,\lambda) + m^{\frac{2\mu}{2+\mu}\gamma - \omega_0} + m^{-\varpi} = O\big(m^{-\vartheta(\beta,\varpi)}\big), \tag{93}$$

where

$$\begin{aligned} \vartheta(\beta,\varpi) = \min\Big\{&\varpi, \beta r, \frac{1}{2} - \beta(1-r), \\ &\omega_0(1-\theta) - \beta(1-r-\theta), \omega_0 - \frac{2\mu}{2+\mu}\gamma\Big\}. \end{aligned} \tag{94}$$

Note that $0 < \beta, \varpi \leq \frac{1}{2} < \omega_0$, we find that

$$\gamma = \max\Big\{\beta(1-r), \beta - \varpi\Big\}. \tag{95}$$

Thus

$$\vartheta(\beta,\varpi) = \min\Big\{\varpi, \beta r, \omega_0 - \frac{2\mu\beta(1-r)}{2+\mu}, \omega_0 - \frac{2\mu(\beta-\varpi)}{2+\mu}\Big\}. \tag{96}$$

Our main conclusion follows by taking $\beta = \frac{1}{2}$ and $\frac{r}{2} \leq \varpi < \omega_0$. The proof of Theorem 1 is complete. □

References

[1] D.R. Chen, Q.Y. Ying and D.X. Zhou, "Support vector machine soft margin classifiers: error analysis," J. Machine Learning Research, vol. 5, pp. 1143-1175, 2004.

[2] F. Cucker and D.X. Zhou, *Learning Theory: an Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[3] Z.C. Guo, L. Shi, "Learning with coefficient-based regularization and ℓ^1-penalty," Adv. Comput. Math., vol. 39, no. 3-4, pp. 493-510, 2013.

[4] M. Li, H.W. Sun, "Asymptotic analysis of quantile regression learning based on coefficient dependent regularization," International Journal of Wavelets, Multiresolution and Information Processing, vol. 13, no. 4, 2015.

[5] L. Shi, "Learning theory estimates for coefficient-based regularized regression," Appl. Comput. Harmon. Anal., vol. 34, no. 2, pp. 252-265, 2013.

[6] L. Shi, Y.L. Feng and D.X. Zhou, "Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces," Appl. Comput. Harmon. Anal., vol. 31, no. 2, pp. 286-302, 2011.

[7] L. Shi, X. Huang, Z. Tian and J.A.K. Suykens, "Quantile regression with ℓ_1-regularization and Gaussian kernels," Adv. Comput. Math., vol. 40, no. 2, pp. 517-551, 2014.

[8] S. Smale and D.X. Zhou, "Learning theory estimates via integral operators and their approximations," Constr. Approx., vol. 26, no. 2, pp. 153-172, 2007.

[9] I. Steinwart and A. Christman, "How SVMs can estimate quantiles and the median," Advances in Neural Information Processing Systems, vol. 20, pp. 305-312, 2008.

[10] I. Steinwart and A. Christman, "Estimating conditional quantiles with the help of the pinball loss," Bernoulli, vol. 17, no. 1, pp. 211-225, 2011.

[11] I. Steinwart and C. Scovel, "Fast rates for support vector machines using Gaussian kernels," Ann. Stat., vol. 35, no. 2, pp. 575-607, 2007.

[12] H.W. Sun and Q. Wu, "Least square regression with indefinite kernel and coefficient regularization," Applied and Computational Harmonic Analysis, vol. 3, no. 1, pp. 96-109, 2011.

[13] A.W. van der Vaart and J.A. Wellner, *Weak Convergence and Empirical Processes*, Springer Verlag, New York, 1996.

[14] Q. Wu, Y. Ying and D.X. Zhou, "Multi-kernel regularized classifiers," J. Complexity, vol. 23, no. 1, pp. 108-134, 2007.

[15] Q. Wu, Y. Ying and D.X. Zhou, "Learning rates of least-square regularized regression," Found. Comput. Math., vol. 6, no. 2, pp. 171-192, 2006.

[16] Q. Wu and D.X. Zhou, "Learning with sample dependent hypothesis spaces," Computers and Mathematics with Applications, vol. 56, no. 11, pp. 2896-2907, 2008.

[17] D.H. Xiang, T. Hu and D.X. Zhou, "Approximation analysis of learning algorithms for support vector regression and quantile regression," Journal of Applied Mathematics, vol. 2012, 2012.
[18] Q.W. Xiao and D.X. Zhou, "Learning by nonsymmetric kernels with data dependent spaces and ℓ^1-regularizer," Taiwanese Journal of Mathematics, vol. 14, no. 5, pp. 1821-1836, 2010.