Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 148490, 6 pages
http://dx.doi.org/10.1155/2013/148490

Research Article

ERM Scheme for Quantile Regression

Dao-Hong Xiang

Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang 321004, China

Correspondence should be addressed to Dao-Hong Xiang; daohongxiang@gmail.com

Received 30 November 2012; Accepted 21 February 2013

Academic Editor: Ding-Xuan Zhou

This paper considers the ERM scheme for quantile regression. We conduct error analysis for this learning algorithm by means of a variance-expectation bound when a noise condition is satisfied for the underlying probability measure. The learning rates are derived by applying concentration techniques involving the ℓ^2-empirical covering numbers.

1. Introduction

In this paper, we study empirical risk minimization scheme (ERM) for quantile regression. Let X be a compact metric space (input space) and $Y = \mathbb{R}$. Let ρ be a fixed but unknown probability distribution on $Z := X \times Y$ which describes the noise of sampling. The conditional quantile regression aims at producing functions to estimate quantile regression functions. With a prespecified quantile parameter $\tau \in (0,1)$, a *quantile regression function* $f_{\tau,\rho}$ is defined by its value $f_{\tau,\rho}(x)$ to be a τ-quantile of $\rho(\cdot \mid x)$, that is, a value $t \in Y$ satisfying

$$\rho((-\infty, t] \mid x) \geq \tau, \quad \rho([t, \infty) \mid x) \geq 1 - \tau, \quad x \in X, \quad (1)$$

where $\rho(\cdot \mid x)$ is the conditional distribution of ρ at $x \in X$.

We consider a learning algorithm generated by ERM scheme associated with pinball loss and hypothesis space $\mathscr{H}$. The *pinball loss* $L_\tau : \mathbb{R} \to [0, \infty)$ is defined by

$$L_\tau(r) = \begin{cases} (\tau - 1)\, r, & \text{if } r \leq 0, \\ \tau r, & \text{if } r \geq 0. \end{cases} \quad (2)$$

The hypothesis space $\mathscr{H}$ is a compact subset of $C(X)$. So there exists some $M > 0$ such that $\|f\|_{C(X)} \leq M$ for any $f \in \mathscr{H}$. We assume without loss of generality $\|f\|_{C(X)} \leq 1$ for any $f \in \mathscr{H}$.

The ERM scheme for quantile regression is defined with a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ drawn independently from ρ as follows:

$$f_{\mathbf{z}} = \arg\min_{f \in \mathscr{H}} \frac{1}{m} \sum_{i=1}^m L_\tau(y_i - f(x_i)). \quad (3)$$

A family of kernel based learning algorithms for quantile regression has been widely studied in a large literature [1–4] and references therein. The form of the algorithms is a regularized scheme in a reproducing kernel Hilbert space $\mathscr{H}_K$ (RKHS, see [5] for details) associated with a Mercer kernel K. Given a sample $\mathbf{z}$ the kernel based regularized scheme for quantile regression is defined by

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathscr{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m L_\tau(y_i - f(x_i)) + \lambda \|f\|_K^2 \right\}. \quad (4)$$

In [1, 3, 4], error analysis for general $\mathscr{H}_K$ has been done. Learning with varying Gaussian kernel was studied in [2].

ERM scheme (3) is very different from kernel based regularized scheme (4). The output function $f_{\mathbf{z}}$ produced by the ERM scheme has a uniform bound, under our assumption, $\|f_{\mathbf{z}}\|_{C(X)} \leq 1$. However, we cannot expect it for $f_{\mathbf{z},\lambda}$. It is easy to see that $\lambda \|f_{\mathbf{z},\lambda}\|_K^2 \leq \sum_{i=1}^m |y_i|$ by choosing $f = 0$. It happens often that $\|f_{\mathbf{z},\lambda}\|_K \to \infty$ as $\lambda \to 0$. The lack of a uniform bound for $f_{\mathbf{z},\lambda}$ has a serious negative impact on the learning rates. So in the literature of kernel based regularized schemes for quantile regression, values of the output function $f_{\mathbf{z},\lambda}$ are always projected onto the interval $[-1, 1]$, and error analysis is conducted for the projected function, not $f_{\mathbf{z},\lambda}$ itself.

In this paper, we aim at establishing convergence and learning rates for the error $\|f_{\mathbf{z}} - f_{\tau,\rho}\|_{L^r_{\rho_X}}$ in the space $L^r_{\rho_X}$. Here $r > 0$ depends on the pair (ρ, τ) which will be decided in Section 2 and ρ_X is the marginal distribution of ρ on X. In the rest of this paper, we assume $Y = [-1, 1]$ which in turn leads to values of the target function $f_{\tau,\rho}$ lie in the same interval.

2. Noise Condition and Main Results

There has been a large literature in learning theory (see [6] and references therein) devoted to the least square regression. It aims at learning the regression function $f_\rho(x) = \int_Y y d\rho(y \mid x)$. The identity $\mathcal{E}_{\ls}(f) - \mathcal{E}_{\ls}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}}$ for the generalization error $\mathcal{E}_{\ls}(f) = \int_Z (y - f(x))^2 d\rho$ leads to a variance-expectation bound with the form of $\mathbf{E}\xi^2 \le 4\mathbf{E}\xi$, where $\xi = (y - f(x))^2 - (y - f_\rho(x))^2$ on (Z, ρ). It plays an essential role in error analysis of kernel based regularized schemes.

However, this identity relation and expectation-variance bound fail in the setting of the quantile regression. The reason is that the pinball loss is lack of strong convexity. If we add some noise condition on distribution ρ named τ-quantile of p-average type q (see Definition 1), we can also get a similar identity relation which in turn enables us to have a variance-expectation bound stated in the following which is proved by Steinwart and Christman [1].

Definition 1. Let $p \in (0, \infty]$ and $q \in [1, \infty)$. A distribution ρ on $X \times [-1, 1]$ is said to have a τ-quantile of p-average type q if for ρ_X-almost every $x \in X$, there exist a τ-quantile $t^* \in \mathbb{R}$ and constants $\alpha_{\rho(\cdot|x)} \in (0, 2]$, $b_{\rho(\cdot|x)} > 0$ such that for each $s \in [0, \alpha_{\rho(\cdot|x)}]$,

$$\begin{aligned} \rho\left((t^* - s, t^*) \mid x\right) &\ge b_{\rho(\cdot|x)} s^{q-1}, \\ \rho\left((t^*, t^* + s) \mid x\right) &\ge b_{\rho(\cdot|x)} s^{q-1}, \end{aligned} \tag{5}$$

and that the function γ on X defined by $\gamma(x) = b_{\rho(\cdot|x)} \alpha^{q-1}_{\rho(\cdot|x)}$ satisfies $\gamma^{-1} \in L^p_{\rho_X}$.

We also need capacity of the hypothesis space $\mathscr{H}$ for our learning rates. Here in this paper, we measure the capacity by empirical covering numbers.

Definition 2. Let $(\mathfrak{M}, d)$ be a pseudometric space and S be a subset of $\mathfrak{M}$. For every $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon, d)$ of S with respect to ε and d is defined as the minimal number of balls of radius ε whose union covers S, that is,

$$\begin{aligned} &\mathcal{N}(S, \varepsilon, d) \\ &= \min\left\{\ell \in \mathbb{N} : S \subset \cup_{j=1}^{\ell} B\left(s_j, \varepsilon\right) \text{ for some } \left\{s_j\right\}_{j=1}^{\ell} \subset \mathfrak{M}\right\}, \end{aligned} \tag{6}$$

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

Definition 3. Let $\mathscr{F}$ be a set of functions on X, $\mathbf{x} = (x_i)_{i=1}^k \subset X^k$ and $\mathscr{F}|_{\mathbf{x}} = \{(f(x_i))_{i=1}^k : f \in \mathscr{F}\} \subset \mathbb{R}^k$. Set $\mathcal{N}_{2,\mathbf{x}}(\mathscr{F}, \varepsilon) = \mathcal{N}(\mathscr{F}|_{\mathbf{x}}, \varepsilon, d_2)$. The ℓ^2-empirical covering number of $\mathscr{F}$ is defined by

$$\mathcal{N}_2(\mathscr{F}, \varepsilon) = \sup_{k \in \mathbb{N}} \sup_{\mathbf{x} \in X^k} \mathcal{N}_{2,\mathbf{x}}(\mathscr{F}, \varepsilon), \quad \varepsilon > 0. \tag{7}$$

Here d_2 is the normalized ℓ^2-metric on the Euclidean space $\mathbb{R}^k$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{k}\sum_{i=1}^k |a_i - b_i|^2\right)^{1/2} \quad \text{for } \mathbf{a} = (a_i)_{i=1}^k,\ \mathbf{b} = (b_i)_{i=1}^k \in \mathbb{R}^k. \tag{8}$$

Assumption. Assume that the empirical covering number of the hypothesis space $\mathscr{H}$ is bounded for some $a > 0$ and $\iota \in (0, 2)$,

$$\log \mathcal{N}_2(\mathscr{H}, \varepsilon) \le a\varepsilon^{-\iota}, \quad \forall \varepsilon > 0. \tag{9}$$

Theorem 4. *Assume that ρ satisfies (5) with some $p \in (0, \infty]$ and $q \in [1, \infty)$. Denote $r = pq/(p+1)$. One further assumes that $f_{\tau,\rho}$ is uniquely defined. If $f_{\tau,\rho} \in \mathscr{H}$ and $\mathscr{H}$ satisfies (9) with $\iota \in (0, 2)$, then for any $0 < \delta < 1$, with confidence $1 - \delta$, one has*

$$\left\|f_{\mathbf{z}} - f_{\tau,\rho}\right\|_{L^r_{\rho_X}} \le \widetilde{C} m^{-\vartheta} \log \frac{2}{\delta}, \tag{10}$$

where

$$\vartheta = \frac{2(p+1)}{4q(p+1) - (2-\iota)\min\{2(p+1), pq\}} \tag{11}$$

and $\widetilde{C}$ is a constant independent of m and δ.

Remark 5. In the ERM scheme, we can choose $f_{\tau,\rho} \in \mathscr{H}$ which in turn makes the approximation error described by (23) equal to zero. However, it is impossible for the kernel based regularized scheme because of the appearance of the penalty term $\lambda\|f\|^2_K$.

If $q \le 2$, all conditional distributions around the quantile behave similar to the uniform distribution. In this case $r = pq/(p+1) \le 2$ and $\theta = \min\{2/q, p/(p+1)\} = p/(p+1)$ for all $p > 0$. Hence, $\vartheta = 2(p+1)/((2+\iota)pq + 4q)$. Furthermore, when p is large enough, the parameter r tends to q and the power index for the above learning rate arbitrarily approaches to $2/(2+\iota)q$ which shows that the learning rate power index for $\|f_{\mathbf{z}} - f_{\tau,\rho}\|^q_{L^q_{\rho_X}}$ is arbitrarily close to $2/(2+\iota)$ independent of q. In particular, ι can be arbitrarily small when $\mathscr{H}$ is smooth enough. In this case, the power index of the learning rates $2/(2+\iota)$ can be arbitrarily close to 1 which is the optimal learning rate for the least square regression.

Let us take some examples to demonstrate the above main result.

Example 6. Let $\mathscr{H}$ be a unit ball of the sobolev space H^s with $s > 0$. Observe that the empirical covering number is bounded above by the uniform covering number defined in Definition 2. Hence we have (see [6, 7])

$$\log \mathcal{N}_2(\mathscr{H}, \varepsilon) \le C_s\left(\frac{1}{\varepsilon}\right)^{n/s}, \tag{12}$$

where n is the dimension of the input space X and $C_s > 0$.

Under the same assumptions as Theorem 4, we get that by replacing ι by n/s, for any $0 < \delta < 1$, with confidence $1-\delta$,

$$\left\|f_{\mathbf{z}} - f_{\tau,\rho}\right\|_{L^r_{\rho_X}} \le \widetilde{C} m^{-\vartheta} \log \frac{2}{\delta}, \tag{13}$$

where

$$\vartheta = \frac{2(p+1)}{4q(p+1) - (2-n/s)\min\{2(p+1), pq\}} \tag{14}$$

and $\widetilde{C}$ is a constant independent of m and δ.

We carry out the same discussions on the case of $q \le 2$ and large enough p as Remark 5. Therefore the power index of the learning rates for $\|f_{\mathbf{z}} - f_{\tau,\rho}\|^q_{L^q_{\rho_X}}$ is arbitrarily close to $2/(2+n/s)$ independent of q. Furthermore, s can be arbitrarily large if the Sobolev space is smooth enough. In this special case, the learning rate power index arbitrarily approaches to 1.

Example 7. Let $\mathcal{H}$ be a unit ball of the reproducing kernel Hilbert space $\mathcal{H}_\sigma$ generated by a Gaussian kernel (see [5]). Reference [7] tells us that

$$\log \mathcal{N}(\mathcal{H}, \varepsilon) \le C_{n,\sigma}\left(\log \frac{1}{\varepsilon}\right)^{n+1}, \quad \forall \varepsilon > 0, \tag{15}$$

where $C_{n,\sigma} > 0$ depends only on n and $\sigma > 0$. Obviously, the right-hand side of (15) is bounded by $C_{n,\sigma}(1/\varepsilon)^{n+1}$.

So from Theorem 4, we can get different learning rates with power index

$$\vartheta = \frac{2(p+1)}{4q(p+1) - (1-n)\min\{2(p+1), pq\}}. \tag{16}$$

If $q \le 2$ and p is large enough, the power index of the learning rates for $\|f_{\mathbf{z}} - f_{\tau,\rho}\|^q_{L^q_{\rho_X}}$ is arbitrarily close to $2/(3+n)$ which is very slow if n is large. However, in most data sets the data are concentrated on a much lower dimensional manifold embedded in the high dimensional space. In this setting an analysis that replaces n by the intrinsic dimension of the manifold would be of great interest (see [8] and references therein).

3. Error Analysis

Define the noise-free error called *generalization error* associated with the pinball loss L_τ as

$$\mathcal{E}_\tau(f) = \int_Z L_\tau(y - f(x))\, d\rho \quad \text{for } f : X \to [-1,1]. \tag{17}$$

Then the measurable function $f_{\tau,\rho}$ is a minimizer of $\mathcal{E}_\tau$. Obviously, $f_{\tau,\rho}(x) \in [-1,1]$.

We need the following results from [1] for our error analysis.

Proposition 8. *Let L_τ be the pinball loss. Assume that ρ satisfies* (5) *with some $p \in (0,\infty]$ and $q \in [1,\infty)$. Then for all $f : X \to [-1,1]$ one has*

$$\left\|f - f_{\tau,\rho}\right\|_{L^r_{\rho_X}} \le 2^{1-1/q} q^{1/q} \left\|\gamma^{-1}\right\|^{1/q}_{L^p_{\rho_X}} \left(\mathcal{E}_\tau(f) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right)\right)^{1/q}. \tag{18}$$

Furthermore, with

$$\theta = \min\left\{\frac{2}{q}, \frac{p}{p+1}\right\} \in (0,1], \qquad C_\theta = 2^{2-\theta} q^\theta \left\|\gamma^{-1}\right\|^\theta_{L^p_{\rho_X}} > 0, \tag{19}$$

one has

$$\mathbf{E}\left\{\left(L_\tau(y - f(x)) - L_\tau\left(y - f_{\tau,\rho}(x)\right)\right)^2\right\} \le C_\theta\left(\mathcal{E}_\tau(f) - \mathcal{E}_\tau(f_{\tau,\rho})\right)^\theta, \quad \forall f : X \to Y. \tag{20}$$

The above result implies that we can get convergence rates of $f_{\mathbf{z}}$ in the space $L^r_{\rho_X}$ by bounding the excess generalization error $\mathcal{E}_\tau(f_{\mathbf{z}}) - \mathcal{E}_\tau(f_{\tau,\rho})$.

To bound $\mathcal{E}_\tau(f_{\mathbf{z}}) - \mathcal{E}_\tau(f_{\tau,\rho})$, we need a standard error decomposition procedure [6] and a concentration inequality.

3.1. Error Decomposition. Define the empirical error associated with the pinball loss L_τ as

$$\mathcal{E}_{\mathbf{z},\tau}(f) = \frac{1}{m}\sum_{i=1}^m L_\tau(y_i - f(x_i)) \quad \text{for } f : X \to [-1,1]. \tag{21}$$

Define

$$f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_\tau(f) = \arg\min_{f\in\mathcal{H}} \left\{\mathcal{E}_\tau(f) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right)\right\}. \tag{22}$$

Lemma 9. *Let L_τ be the pinball loss, $f_{\mathbf{z}}$ be defined by* (3) *and $f_{\mathcal{H}} \in \mathcal{H}$ by* (22). *Then*

$$\begin{aligned}
&\mathcal{E}_\tau(f_{\mathbf{z}}) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right) \\
&\le \left[\mathcal{E}_\tau(f_{\mathbf{z}}) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right)\right] - \left[\mathcal{E}_{\mathbf{z},\tau}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z},\tau}\left(f_{\tau,\rho}\right)\right] \\
&\quad + \left[\mathcal{E}_{\mathbf{z},\tau}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z},\tau}\left(f_{\tau,\rho}\right)\right] - \left[\mathcal{E}_\tau(f_{\mathcal{H}}) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right)\right] \\
&\quad + \mathcal{E}_\tau(f_{\mathcal{H}}) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right).
\end{aligned} \tag{23}$$

Proof. The excess generalization error can be written as

$$\begin{aligned}
\mathcal{E}_\tau(f_{\mathbf{z}}) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right) = {} & \left[\mathcal{E}_\tau(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z},\tau}(f_{\mathbf{z}})\right] \\
& + \left[\mathcal{E}_{\mathbf{z},\tau}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z},\tau}(f_{\mathcal{H}})\right] \\
& + \left[\mathcal{E}_{\mathbf{z},\tau}(f_{\mathcal{H}}) - \mathcal{E}_\tau(f_{\mathcal{H}})\right] \\
& + \left[\mathcal{E}_\tau(f_{\mathcal{H}}) - \mathcal{E}_\tau\left(f_{\tau,\rho}\right)\right].
\end{aligned} \tag{24}$$

The definition of $f_{\mathbf{z}}$ implies that $\mathcal{E}_{\mathbf{z},\tau}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z},\tau}(f_{\mathcal{H}}) \le 0$. Furthermore, by subtracting and adding $\mathcal{E}_\tau(f_{\tau,\rho})$ and $\mathcal{E}_{\mathbf{z},\tau}(f_{\tau,\rho})$ in the first term and third term, we see that Lemma 9 holds true. □

We call the term (23) *approximation error*. It has been studied in [9].

3.2. Concentration Inequality and Sample Error. Let us recall the one-sided Bernstein inequality as follows.

Lemma 10. *Let ξ be a random variable on a probability space Z with variance σ^2 satisfying $|\xi - \mathbb{E}(\xi)| \le M_\xi$ for some constant M_ξ. Then for any $0 < \delta < 1$, with confidence $1 - \delta$, one has*

$$\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}(\xi) \le \frac{2M_\xi \log(1/\delta)}{3m} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{m}}. \tag{25}$$

Proposition 11. *Let $f_{\mathscr{H}} \in \mathscr{H}$. Assume that ρ on $X \times [-1,1]$ satisfies the variance bound (20) with index θ indicated in (19). For any $0 < \delta < 1$, with confidence $1 - \delta/2$, (3.6) can be bounded as*

$$\begin{aligned}&\left[\mathscr{E}_{\mathbf{z},\tau}(f_{\mathscr{H}}) - \mathscr{E}_{\mathbf{z},\tau}\left(f_{\tau,\rho}\right)\right] - \left[\mathscr{E}_\tau(f_{\mathscr{H}}) - \mathscr{E}_\tau\left(f_{\tau,\rho}\right)\right]\\ &\quad\le \frac{4\log(2/\delta)}{3m} + \left(\frac{2C_\theta \log(2/\delta)}{m}\right)^{1/(2-\theta)}\\ &\qquad + \mathscr{E}_\tau(f_{\mathscr{H}}) - \mathscr{E}_\tau\left(f_{\tau,\rho}\right).\end{aligned} \tag{26}$$

Proof. Let $\xi(z) = L_\tau(y - f_{\mathscr{H}}(x)) - L_\tau(y - f_{\tau,\rho}(x))$ which satisfies $|\xi| \le 2$ and in turn $|\xi - \mathbf{E}(\xi)| \le 2$. The variance bound (20) implies that

$$\mathbf{E}(\xi - \mathbf{E}(\xi))^2 \le \mathbf{E}\xi^2 \le C_\theta\left(\mathscr{E}_\tau(f_{\mathscr{H}}) - \mathscr{E}_\tau\left(f_{\tau,\rho}\right)\right)^\theta. \tag{27}$$

Using (25) on the random variable $\xi(z) = L_\tau(y - f_{\mathscr{H}}(x)) - L_\tau(y - f_{\tau,\rho}(x))$, we can get the desired bound (28) with the help of Young's inequality. □

Let us turn to estimate the sample error (3.5) involving the function $f_{\mathbf{z}}$ which runs over a set of functions since $\mathbf{z}$ is a random sample itself. To estimate it, we use a concentration inequality below involving empirical covering numbers [10–12].

Lemma 12. *Let $\mathscr{F}$ be a class of measurable functions on Z. Assume that there are constants $B, c > 0$ and $\alpha \in [0,1]$ and $\|f\|_\infty \le B$ and $\mathbb{E}f^2 \le c(\mathbb{E}f)^\alpha$ for every $f \in \mathscr{F}$. If (7) holds, then there exists a constant c_ι' depending only on ι such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds*

$$\begin{aligned}\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m} f(z_i) &\le \frac{1}{2}\eta^{1-\alpha}(\mathbb{E}f)^\alpha\\ &\quad + c_\iota'\eta + 2\left(\frac{ct}{m}\right)^{1/(2-\alpha)} + \frac{18Bt}{m}, \quad \forall f \in \mathscr{F},\end{aligned} \tag{28}$$

where

$$\begin{aligned}\eta := \max\Bigg\{&c^{(2-\iota)/(4-2\alpha+\iota\alpha)}\left(\frac{a}{m}\right)^{2/(4-2\alpha+\iota\alpha)},\\ &B^{(2-\iota)/(2+\iota)}\left(\frac{a}{m}\right)^{2/(2+\iota)}\Bigg\}.\end{aligned} \tag{29}$$

We apply Lemma 12 to a function set $\mathscr{F}$, where

$$\mathscr{F} = \left\{L_\tau(y - f(x)) - L_\tau\left(y - f_{\tau,\rho}(x)\right) : f \in \mathscr{H}\right\}. \tag{30}$$

Proposition 13. *Assume ρ on $X \times [-1,1]$ satisfies the variance bound (20) with index θ indicated in (19). If $\mathscr{H}$ satisfies (9) with $\iota \in (0,2)$, then for any $0 < \delta < 1$, with confidence $1 - \delta/2$, one has*

$$\begin{aligned}&\left[\mathscr{E}_\tau(f) - \mathscr{E}_\tau\left(f_{\tau,\rho}\right)\right]\\ &\qquad - \left[\mathscr{E}_{\mathbf{z},\tau}(f) - \mathscr{E}_{\mathbf{z},\tau}\left(f_{\tau,\rho}\right)\right]\\ &\quad\le \frac{1}{2}\left[\mathscr{E}_\tau(f) - \mathscr{E}_\tau\left(f_{\tau,\rho}\right)\right]\\ &\qquad + C_{\iota,\theta}\log\frac{2}{\delta}\left(\frac{1}{m}\right)^{2/(4-2\theta+\iota\theta)}, \quad \forall f \in \mathscr{H},\end{aligned} \tag{31}$$

where

$$\begin{aligned}C_{\iota,\theta} = \left(\frac{1}{2} + c_\iota'\right)\max\Big\{&C_\theta^{(2-\iota)/(4-2\theta+\iota\theta)}a^{2/(4-2\theta+\iota\theta)},\\ &2^{(2-\iota)/(2+\iota)}a^{2/(2+\iota)}\Big\} + 2C_\theta^{1/(2-\theta)} + 36.\end{aligned} \tag{32}$$

Proof. Take $g \in \mathscr{F}$ with the form $g(z) = L_\tau(y - f(x)) - L_\tau(y - f_{\tau,\rho}(x))$ where $f \in \mathscr{H}$. Hence $\mathbb{E}g = \mathscr{E}_\tau(f) - \mathscr{E}_\tau(f_{\tau,\rho})$ and $(1/m)\sum_{i=1}^m g(z_i) = \mathscr{E}_{\mathbf{z},\tau}(f) - \mathscr{E}_{\mathbf{z},\tau}(f_{\tau,\rho})$.

The Lipschitz property of the pinball loss L_τ implies that

$$|g(z)| \le \left|f(x) - f_{\tau,\rho}(x)\right| \le 2. \tag{33}$$

For $g_1, g_2 \in \mathscr{F}$, we have

$$\begin{aligned}&|g_1(z) - g_2(z)|\\ &\quad= \left|L_\tau(y - f_1(x)) - L_\tau(y - f_2(x))\right|\\ &\quad\le |f_1(x) - f_2(x)|,\end{aligned} \tag{34}$$

where $f_1, f_2 \in \mathscr{H}$. It follows that

$$\mathscr{N}_{2,\mathbf{z}}(\mathscr{F}, \varepsilon) \le \mathscr{N}_{2,\mathbf{x}}(\mathscr{H}, \varepsilon). \tag{35}$$

Hence

$$\log \mathscr{N}_2(\mathscr{F}, \varepsilon) \le a\varepsilon^{-\iota}. \tag{36}$$

Applying Lemma 12 with $B = 2, \alpha = \theta$, and $c = C_\theta$, we know that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\begin{aligned}&\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\\ &\quad\le \frac{1}{2}\eta^{1-\theta}(\mathbb{E}g)^\theta + c_\iota'\eta + 2\left(\frac{C_\theta \log(2/\delta)}{m}\right)^{1/(2-\theta)}\\ &\qquad + \frac{36\log(2/\delta)}{m}\\ &\quad\le \frac{1}{2}\mathbb{E}g + \left(\frac{1}{2} + c_\iota'\right)\eta + 2\left(\frac{C_\theta \log(2/\delta)}{m}\right)^{1/(2-\theta)}\\ &\qquad + \frac{36\log(2/\delta)}{m}.\end{aligned} \tag{37}$$

Here

$$\eta \le \max\left\{C_\theta^{(2-\iota)/(4-2\theta+\iota\theta)} a^{2/(4-2\theta+\iota\theta)}, 2^{(2-\iota)/(2+\iota)} a^{2/(2+\iota)}\right\} \times \left(\frac{1}{m}\right)^{2/(4-2\theta+\iota\theta)}. \tag{38}$$

Note that

$$\left(\frac{1}{2}+c_\iota'\right)\eta + 2\left(\frac{C_\theta \log(2/\delta)}{m}\right)^{1/(2-\theta)} + \frac{36\log(2/\delta)}{m} \le C_{\iota,\theta} \log\frac{2}{\delta}\left(\frac{1}{m}\right)^{2/(4-2\theta+\iota\theta)}, \tag{39}$$

where $C_{\iota,\theta}$ is indicated in (32). Then our desired bound holds true. □

Proposition 14. *Assume ρ on $X\times[-1,1]$ satisfies the variance bound* (20) *with index θ indicated in* (19). *If $\mathcal{H}$ satisfies* (9) *with $\iota \in (0,2)$ Then for any $0<\delta<1$, with confidence $1-\delta$, there holds*

$$\begin{aligned}\mathcal{E}_\tau(f_{\mathbf{z}}) - \mathcal{E}_\tau(f_{\tau,\rho}) &\le 2\left(\mathcal{E}_\tau(f_{\mathcal{H}}) - \mathcal{E}_\tau(f_{\tau,\rho})\right) + \frac{8\log(2/\delta)}{3m} + 2\left(\frac{2C_\theta\log(2/\delta)}{m}\right)^{1/(2-\theta)} \\ &\quad + 2C_{\iota,\theta}\log\frac{2}{\delta}\left(\frac{1}{m}\right)^{2/(4-2\theta+\iota\theta)}.\end{aligned} \tag{40}$$

The above bound follows directly from Propositions 11 and 13 with the fact that $f_{\mathbf{z}} \in \mathcal{H}$.

3.3. Bounding the Total Error. Now we are in a position to present our general result on error analysis for algorithm (3).

Theorem 15. *Assume that ρ satisfies* (5) *with some $p\in(0,\infty]$ and $q\in[1,\infty)$. Denote $\gamma = pq/(p+1)$. Further assume that $\mathcal{H}$ satisfies* (9) *with $\iota\in(0,2)$ and $f_{\tau,\rho}$ is uniquely defined. Then for any $0<\delta<1$, with confidence $1-\delta$, one has*

$$\left\|f_{\mathbf{z}} - f_{\tau,\rho}\right\|_{L^r_{\rho_X}} \le \widetilde{C}\inf_{f\in\mathcal{H}}\left\{\mathcal{E}_\tau(f) - \mathcal{E}_\tau(f_{\tau,\rho})\right\}^{1/q} + \widetilde{C}m^{-\vartheta}\log\frac{2}{\delta}, \tag{41}$$

where

$$\vartheta = \frac{2(p+1)}{4q(p+1) - (2-\iota)\min\{2(p+1), pq\}} \tag{42}$$

and $\widetilde{C}$ are constant independent of m and δ.

Proof. Combining (18), (19), and (40), with confidence $1-\delta$, we have

$$\begin{aligned}\left\|f_{\mathbf{z}} - f_{\tau,\rho}\right\|_{L^r_{\rho_X}} &\le \widetilde{C}\left(\mathcal{E}_\tau(f_{\mathcal{H}}) - \mathcal{E}_\tau(f_{\tau,\rho})\right)^{1/q} + \widetilde{C}\log\frac{2}{\delta}\left(\frac{1}{m}\right)^{2/q(4-2\theta+\iota\theta)} \\ &\le \widetilde{C}\inf_{f\in\mathcal{H}}\left\{\mathcal{E}_\tau(f) - \mathcal{E}_\tau(f_{\tau,\rho})\right\}^{1/q} + \widetilde{C}m^{-\vartheta}\log\frac{2}{\delta},\end{aligned} \tag{43}$$

where

$$\vartheta = \frac{2(p+1)}{4q(p+1) - (2-\iota)\min\{2(p+1), pq\}}, \qquad \widetilde{C} = 2q^{1/q}\left\|\gamma^{-1}\right\|^{1/q}_{L^p_{\rho_X}}\left(\frac{4}{3} + (2C_\theta)^{1/(2-\theta)} + C_{\iota,\theta}\right)^{1/q}. \tag{44}$$

□

Proof of Theorem 1. The assumption $f_{\tau,\rho}\in\mathcal{H}$ implies that

$$\inf_{f\in\mathcal{H}}\left\{\mathcal{E}_\tau(f) - \mathcal{E}_\tau(f_{\tau,\rho})\right\} = 0. \tag{45}$$

Therefore, our desired result comes directly from Theorem 15.

4. Further Discussions

In this paper, we studied ERM algorithm (3) for quantile regression and provide convergence and learning rates. We showed some essential differences between ERM scheme and kernel based regularized scheme for quantile regression. We also point out the difficulty to deal with quantile regression: the lack of strong convexity of the pinball loss. To overcome this difficulty, some noise condition on ρ is proposed to enable us to get a variance-expectation bound similar to the one for the least square regression.

In our analysis we just consider $f\in\mathcal{H}$ and $\|f\|_{C(X)}\le 1$. The case $\|f\|_{C(X)}\le R$ for $R\ge 1$ would be interesting in the future work. The approximation error involving R can be estimated by the knowledge of interpolation space.

In our setting, the sample is drawn independently from the distribution ρ. However, in many practical problems, the i.i.d condition is a little demanding, so it would be interesting to investigate the ERM scheme for quantile regression with nonidentical distributions [13, 14] or dependent sampling [15].

Acknowledgment

This work described in this paper is supported by NSF of China under Grant 11001247 and 61170109.

References

[1] I. Steinwart and A. Christman, *How SVMs Can Estimate Quantile and the Median*, vol. 20 of *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, Mass, USA, 2008.

[2] D. H. Xiang, "Conditional quantiles with varying Gaussians," *Advances in Computational Mathematics*, 2011.

[3] D.-H. Xiang, T. Hu, and D.-X. Zhou, "Learning with varying insensitive loss," *Applied Mathematics Letters*, vol. 24, no. 12, pp. 2107–2109, 2011.

[4] 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, Article ID 902139, 17 pages, 2012.

[5] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[6] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, vol. 24, Cambridge University Press, Cambridge, UK, 2007.

[7] D.-X. Zhou, "The covering number in learning theory," *Journal of Complexity*, vol. 18, no. 3, pp. 739–767, 2002.

[8] S. Mukherjee, Q. Wu, and D.-X. Zhou, "Learning gradients on manifolds," *Bernoulli*, vol. 16, no. 1, pp. 181–207, 2010.

[9] S. Smale and D.-X. Zhou, "Estimating the approximation error in learning theory," *Analysis and Applications*, vol. 1, no. 1, pp. 17–41, 2003.

[10] S. Mukherjee and Q. Wu, "Estimation of gradients and coordinate covariation in classification," *Journal of Machine Learning Research*, vol. 7, pp. 2481–2514, 2006.

[11] Y. Yao, "On complexity issues of online learning algorithms," *Institute of Electrical and Electronics Engineers*, vol. 56, no. 12, pp. 6470–6481, 2010.

[12] Y. Ying, "Convergence analysis of online algorithms," *Advances in Computational Mathematics*, vol. 27, no. 3, pp. 273–291, 2007.

[13] T. Hu and D.-X. Zhou, "Online learning with samples drawn from non-identical distributions," *Journal of Machine Learning Research*, vol. 10, pp. 2873–2898, 2009.

[14] S. Smale and D.-X. Zhou, "Online learning with Markov sampling," *Analysis and Applications*, vol. 7, no. 1, pp. 87–113, 2009.

[15] Z.-C. Guo and L. Shi, "Classification with non-i.i.d. sampling," *Mathematical and Computer Modelling*, vol. 54, no. 5-6, pp. 1347–1364, 2011.