Vol. 41 (2021)
No. 4

数 学 杂 志
J. of Math. (PRC)

ONLINE QUANTILE REGRESSION WITH VARYING THRESHOLDS AND NON-IDENTICAL SAMPLING DISTRIBUTIONS

JIANG Ming-qin

(*School of Mathematics and Statistics; Wuhan University, Wuhan 430072, China*)

Abstract: In this paper we study the online quantile regression algorithm with varying thresholds and non-identical sampling distributions, where at each time a sample is drawn independently from different probability distributions and the threshold values decrease with the iteration process. The learning rate of the algorithm is obtained under the assumption that the sequence of marginal distribution converges polynomially fast in the dual of a Hölder space. Several numerical simulations are presented to support our results.

Keywords: sampling with non-identical distributions, online learning, quantile regression, ϵ-insensitive pinball loss, reproducing kernel Hilbert spaces

2010 MR Subject Classification: 62J99

Document code: A **Article ID:** 0255-7797(2021)04-0316-13

1 Introduction

Quantile regression extends the classical least squares regression and provides more information about the distributions of response variables such as stretching or compression tails and multimodality. Since quantile regression can provide a more complete description of the response distribution than a single estimate of the center, such as the mean or median, it has received considerable study in the literature; see [1–3].

An initial form of online learning algorithm was proposed in [4]. It is a type of stochastic gradient descent method, which is applicable to the situations where sample data is presented in a sequential manner and the predictor is updated at each iteration. With linear complexity, online learning provides an important family of efficient and scalable machine learning algorithms for real applications. Thus, a variety of online learning paradigms have been introduced, see [5–10]. Here we aim to study the online quantile regression algorithm generated from a stochastic gradient descent method of regularization schemes in a reproducing Kernel Hilbert space (RKHS) associated with non-identical distributions.

In the literature on learning theory, samples are often drawn independently from an identical distribution (i.i.d.). However, the data in practice are usually not from an identical

* **Received date:** 2020-11-27 **Accepted date:** 2020-12-30

Foundation item: Supported by National Natural Science Foundation of China(11671307).

Biography: Jiang Mingqin(1997–), male, born at Shenzhen, Guangdong, postgraduate, major in machine learning.

distribution. The first case is when the sampling distribution is perturbed by some noise and the noise level decreases as the learning time. The second case is generated by iterative actions of an integral operator associated with a stochastic density kernel. The third case is to induce distributions by dynamical systems. For details, we can refer to papers [11,12].

The rest of this paper is organized as follows. We begin with Section 2 by providing necessary background and notations required for a precise statement of our algorithm. We then present our main theorems on the learning ability of our algorithm. Sections 3 is devoted to the proofs of our results. Lastly, we present simulation results in Section 4 to further explore our theoretical results.

2 Backgrounds and Algorithm

In the standard framework of learning, let a separable metric space $(\mathcal{X}, d)$ be the input space and $\mathcal{Y} \subset \mathbb{R}$ be the output space. Kernel methods provide efficient non-parametric learning algorithms to deal with data of nonlinear structures via feature mapping. Here we shall use a reproducing Kernel Hilbert space (RKHS) as the hypothesis space in the design of learning algorithms. A reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function such that the matrix $(K(u_i, u_j))_{i,j=1}^l$ is positive semidefinite for any finite set of points $\{u_i\}_{i=1}^l \subset \mathcal{X}$. A RKHS $(\mathcal{H}_K, \|\cdot\|_K)$ is the completion of the linear span of the function set $\{K_x = K(x, \cdot) : x \in X\}$ with respect to the inner product given by $\langle K_x, K_u\rangle_K = K(x, u),\ \forall x, u \in \mathcal{X}$. It implies the reproducing property

$$\langle f, K_x\rangle_K = f(x),\ \forall f \in \mathcal{H}_K, x \in \mathcal{X}. \tag{2.1}$$

Throughout the paper, we assume that $\kappa := \sup\limits_{f\in\mathcal{H}_K} \sqrt{K(x, x)}$.

2.1 Online Quantile Regression Algorithm

Let ρ be a Borel probability measure defined on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Denote by ρ_x the conditional distribution ρ at $x \in X$. The goal of non-parametric quantile regression is to learn a quantile function $f_{\rho,\tau} : \mathcal{X} \to \mathcal{Y}$ from the sample set $\mathbf{z} = \{z_i\}_{i=1}^T := \{(x_i, y_i)\}_{i=1}^T \subset \mathcal{Z}$, whose value $f_{\rho,\tau}(x)$ is defined as the τ-quantile $(0 < \tau < 1)$ of the conditional ρ_x at $x \in \mathcal{X}$. Here a τ-quantile of ρ_x means a value $u \in Y$ satisfying

$$\rho_x(y \in Y : y \le u) \ge \tau \quad \text{and} \quad \rho_x(y \in Y : y \ge u) \ge 1 - \tau.$$

For quantile regression, the pinball loss $\psi_\tau : \mathbb{R} \to \mathbb{R}_+$ is usually taken as the corresponding loss function in learning schemes, which is defined as

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

To produce sparse estimators, the alternative ϵ-insensitive pinball loss $\psi_\tau^\epsilon : \mathbb{R} \to \mathbb{R}_+$ is introduced in [5], that is,

$$\psi_\tau^\epsilon(u) = \begin{cases} (1-\tau)(u-\epsilon) & \text{if } u > \epsilon, \\ -\tau(u+\epsilon) & \text{if } u \le -\epsilon, \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where $\epsilon > 0$ is the insensitive parameter. This loss function has been applied to various online and batch algorithms, see [6,13,14]. In the following, we consider the online learning algorithm for quantile regression with a varying threshold sequence $\{\epsilon_t > 0\}_t$.

Definition 2.1 Given the sample set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^T \subset \mathcal{Z}$, the online algorithm for quantile regression is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t\{(\psi_\tau^{\epsilon_t})_-'(f_t(x_t) - y_t)K_{(x_t)} + \lambda_t f_t\}, \quad t = 1, 2, \ldots, \tag{2.3}$$

where $\lambda_t > 0$ is a regularization parameter, $\eta_t > 0$ is a step size, $(\psi_\tau^{\epsilon_t})_-'$ is the left derivative of $\psi_\tau^{\epsilon_t}$, the insensitive parameters $\epsilon_t > 0$ converge to zero as the learning step t increases.

With (2.2), the learning sequence $\{f_t\}$ can be expressed as $f_1 = 0$ and

$$f_{t+1} = \begin{cases} (1-\lambda_t\eta_t)f_t - (1-\tau)\eta_t K_{x_t} & \text{if } f_t(x_t) - y_t > \epsilon_t, \\ (1-\lambda_t\eta_t)f_t + \tau\eta_t K_{x_t} & \text{if } f_t(x_t) - y_t \le \epsilon_t, \\ (1-\lambda_t\eta_t)f_t & \text{if } -\epsilon_t < f_t(x_t) - y_t \le \epsilon_t. \end{cases} \tag{2.4}$$

The main purpose of this paper is to investigate how the output function f_{T+1} given by (2.3) converges to the quantile function $f_{\rho,\tau}$ with the non-identical sampling process and how explicit learning rates can be obtained with suitable choices of step sizes and threshold values based on a prior conditions on sampling distributions.

2.2 Sampling with Non-Identical Distributions

In this work, the data pairs $\{z_i\}_{i=1}^T := \{(x_i, y_i)\}_{i=1}^T \subset \mathcal{Z}$ are drawn from a probability distribution $\rho^{(t)}$ on $\mathcal{Z}$ at each step $t = 1, 2, \ldots$. The sampling sequence of probability distributions $\{\rho^{(t)}\}$ is independent but not identical. We assume the marginal distributions sequence $\left\{\rho_{\mathcal{X}}^{(t)}\right\}$ converges polynomially on the dual of the Hölder space $C^s(\mathcal{X})$ for some $0 < s \le 1$. Define Hölder space $C^s(\mathcal{X})$ is the span of all continuous functions on $\mathcal{X}$ with the norm $\|f\|_{C^s(X)} = \|f\|_{C(\mathcal{X})} + |f|_{C^s(\mathcal{X})}$ finite, where $|f|_{C^s(\mathcal{X})} := \sup_{x \ne y} \frac{|f(x)-f(y)|}{(d(x,y))^s}$.

Definition 2.2 We say that the sequence $\{\rho_{\mathcal{X}}^{(t)}\}_{t=1,2,\ldots}$ converges polynomially to a probability distribution $\rho_{\mathcal{X}}$ in $(C^s(\mathcal{X}))^*(0 \le s \le 1)$ if there exist $C > 0$ and $b > 0$ such that

$$\|\rho_{\mathcal{X}}^{(t)} - \rho_{\mathcal{X}}\|_{(C^s(\mathcal{X}))^*} \le Ct^{-b}, \quad t \in \mathbb{N}. \tag{2.5}$$

The power index b measures the differences from non-identical sampling to i.i.d case and impact on the learning rate of the online algorithm. Specially, when $b = \infty$ the sampling

is the i.i.d case. For example, let $h^{(t)}$ be a sequence of bounded functions on $\mathcal{X}$ such that $\sup_{x\in\mathcal{X}}|h^{(t)}(x)| \le Ct^{-b}$. Then the sequence $\{\rho_{\mathcal{X}}^{(t)}\}_{t=1,2,\dots}$ defined by $d\rho_{\mathcal{X}}^{(t)} = d\rho_{\mathcal{X}} + h^{(t)}(x)d\rho_{\mathcal{X}}$ satisfies the decay condition(2.5) for any $0 \le s \le 1$. In this example, $h^{(t)}$ is the density function of the noise distribution and we assume its noise level to decay polynomially as t increases.

2.3 Learning Errors

Usually we measure the learning performance of algorithms by generalization errors. In this paper, the *generalization error* $\mathcal{E}(f)$ of a function $f : \mathcal{X} \to \mathcal{Y}$ is defined by means of the pinball loss ψ_τ as

$$\mathcal{E}(f) = \int_{\mathcal{Z}} \psi_\tau(f(x) - y)d\rho.$$

Throughout the paper, we assume that $\int |y|d\rho < \infty$ and the value of the quantile regression function $f_{\rho,\tau}$ is uniquely determined at each $x \in \mathcal{X}$. With this assumption, if f is bounded on $\mathcal{X}$ or $f \in L^2_{\rho_{\mathcal{X}}}$, $\mathcal{E}(f)$ is finite since $\psi_\tau(u) \le |u|$. By decomposing the measure ρ into the marginal distribution $\rho_{\mathcal{X}}$ and the conditional distribution ρ_x at $x \in \mathcal{X}$, we see that $f_{\rho,\tau}$ is the only minimizer of $\mathcal{E}(f)$ among all measurable functions on $\mathcal{X}$.

This work will investigate the approximation or learning ability of algorithm (2.3) by the *excess generalization error* $\mathcal{E}(f) - \mathcal{E}(f_{\rho,\tau})$. To this end, we introduce some necessary conditions. The first one is involved with the approximation ability of the hypothesis space $\mathcal{H}_K$, which is characterized by the *approximation error*.

Definition 2.3 The approximation error $\mathcal{D}(\lambda)$ of the triple(K, V, ρ) is defined by

$$\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) - \mathcal{E}(f_{\rho,\tau}) + \frac{\lambda}{2}\|f\|_K^2\right\} \tag{2.6}$$

and f_λ is a minimizer of (2.6), called the regularizing function.

A usual assumption on the regularization error $\mathcal{D}(\lambda)$ which imposes certain smoothness on $\mathcal{H}_K$ is

$$\mathcal{D}(\lambda) \le \mathcal{D}_0\lambda^\gamma, \quad \forall\lambda > 0 \tag{2.7}$$

with some $0 < \gamma < 1$ and $\mathcal{D}_0 > 0$.

The second one is respect to the continuity of the conditional distribution $\{\rho_x\}_{x\in X}$ introduced in [11].

Definition 2.4 We say that the set of conditional distributions $\{\rho_x : x \in \mathcal{X}\}$ is Lipschitz-s if there exists a constant $C_\rho > 0$ such that

$$\rho_x(\{y \in \mathcal{Y} : u < y \le v\}) \le C_\rho|u - v|^s, \quad u < v \in \mathcal{Y}. \tag{2.8}$$

Notice that if each density function $\frac{d\rho_x(y)}{dy}$ exists and is uniformly bounded on $\mathcal{Y}$ by a constant C_ρ for each ρ_x, then $s = 1$ is valid.

The third one is about the kernel condition of K, which is stated as follows.

Definition 2.5 We say a Mercer kernel K satisfies the kernel condition of order s if $K \in C^s(\mathcal{X} \times \mathcal{X})$ and for some $\kappa_{2s} > 0$,

$$|K(x,x) - 2K(x,x) + K(u,u)| \leq \kappa_{2s}^2 (d(x,u))^{2s}, \quad \forall x, u \in \mathcal{X}. \tag{2.9}$$

When $0 < s \leq \frac{1}{2}$ and $K \in C^{2s}(\mathcal{X} \times \mathcal{X})$, (2.9) holds true.

With these assumptions in place, we are now ready for the statements of our main results.

Theorem 2.6 Suppose assumptions (2.5), (2.7), (2.8) and (2.9) hold. Take the parameters $\eta_t, \lambda_t, \epsilon_t$ as the form $\eta_t = \eta_1 t^{-\alpha}, \lambda_t = \lambda_1 t^{-p}, \epsilon_t = \epsilon_1 t^{-\beta}$ with $\eta_1, \lambda_1, \epsilon_1, \alpha, p, \beta > 0$. If

$$0 < p < \min\left\{\frac{2+\beta}{5}, \frac{2}{5-\gamma}, \frac{\beta+1}{3}s\right\} \tag{2.10}$$

and

$$p < \alpha < \min\left\{\frac{2+p\gamma-3p}{2}, \frac{2+\beta-3p}{2}, (\beta+1)s-2p\right\}, \tag{2.11}$$

then we have

$$\mathbb{E}_{z_1,\ldots,z_T}\left[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_{\lambda_T}^{\epsilon_t})\right] \leq C' T^{-\min\left\{\frac{\theta^*}{2}, \beta s - p, p\gamma\right\}} \tag{2.12}$$

where C' is a constant independent of T and

$$\theta^* := \min\{2+p\gamma-3p-2\alpha, 2+\beta-3p-2\alpha, 2(\beta+1)s-4p-2\alpha, \alpha-p, b-2p\}. \tag{2.13}$$

Furthermore, we shall bound the difference between f_{T+1} and $f_{\rho,\tau}$ in some Banach space by means of the noise condition.

Definition 2.7 Let $0 < \varphi \leq \infty$ and $\xi > 1$. Denote $r = \varphi\xi/(\varphi+1) > 0$. We say that ρ has a τ-quantile of φ-average type ξ if there exist two positive functions w_τ and b_τ on X such that $\{b_\tau w_\tau^{\xi-1}\}^{-1} \in L_{\rho_X}^{\varphi}$ and for any $x \in \mathcal{X}$ and $w \in (0, w_\tau(x)]$, there hold

$$\rho_x(\{y : f_{\rho,\tau}(x) < y < f_{\rho,\tau}(x) + w\}) \geq b_\tau(x) w^{\xi-1}$$

and

$$\rho_x(\{y : f_{\rho,\tau}(x) - w < y < f_{\rho,\tau}(x)\}) \geq b_\tau(x) w^{\xi-1}.$$

Theorem 2.8 Let $0 < \varphi \leq \infty$ and $\xi > 1$. Denote $r = \varphi\xi/(\varphi+1) > 0$. Assume the measure ρ has a τ- quantile of φ-average type ξ. Under the same conditions of Theorem 2.6 , we have

$$\mathbb{E}_{z_1,\ldots,z_T}\|f_{T+1} - f_{\rho,\tau}\|_{L_{\rho_X}^r} \leq C^* \|b_\tau w_\tau^{\xi-1^{-1}}\|_{L_{\rho_X}^{\varphi}}^{1/\xi} T^{-\theta}$$

where C^* is a constant independent of T and

$$\theta = \min\left\{\frac{\theta^*}{2\xi}, \frac{\beta s - p}{\xi}, \frac{p\gamma}{\xi}\right\}$$

with θ^* in (2.13).

3 Error Decomposition and Technical Estimates

In this section, we shall prove our main results in the previous section. By the standard decomposition, we have that

$$\mathcal{E}(f_{T+1}) - \mathcal{E}(f_{\rho,\tau}) \le \kappa\|f_{T+1} - f_{\lambda_T}^{\epsilon_T}\|_K + \kappa\|f_{\lambda_T}^{\epsilon_T} - f_{\lambda_T}\|_K + \mathcal{D}(\lambda_T). \tag{3.1}$$

For the second term $\|f_{\lambda_T}^{\epsilon_T} - f_{\lambda_T}\|_K$, we can estimate it by the following proposition, whose proof can be found in [5].

Proposition 3.1 If the family of conditional distributions ρ_x at $x \in X$ is Lipschitz-s for some $s > 0$, then for any $0 \le \nu < \mu$. we have

$$\|f_\lambda^\mu - f_\lambda^\nu\|_K \le \frac{C_\rho\kappa|\mu - \nu|^s}{\lambda}. \tag{3.2}$$

In particular, when $\lambda > 0$ and $\epsilon_t = \epsilon_1 t^{-\beta}$with $\beta > 0, \epsilon \ge 0$, there holds

$$\|f_\lambda^{\epsilon_{t-1}} - f_\lambda^{\epsilon_t}\|_K \le \frac{C_\rho\kappa\epsilon_1^s\beta^s 2^{(\beta+1)s}}{\lambda} t^{-(\beta+1)s}, \quad \forall t \ge 2.$$

Thus, our key error analysis is about the *sample error* $\|f_{T+1} - f_{\rho,\tau}\|_K$. To this end, we first estimate the error caused by the non-sampling process.

3.1 Error Caused by Sampling with Non-Identical Distribution

When we take the expectation with respect to $z_t = (x_t, y_t)$ drawn from the non-identical distribution, we get $\int_{\mathcal{Z}} \psi_\tau^{\epsilon_t}(u)d\rho^{(t)}$ instead of $\int_{\mathcal{Z}} \psi_\tau^{\epsilon_t}(u)d\rho$, in this case, an extra error term Δ_t in (3.3) involving the different measure $\rho^{(t)} - \rho$ shows up

$$\Delta_t = \int_{\mathcal{Z}} \{\psi_\tau^{\epsilon_t}(f_{\lambda_t}^{\epsilon_t} - y_t) - \psi_\tau^{\epsilon_t}(f_t - y_t)\}d[\rho^{(t)} - \rho]. \tag{3.3}$$

Lemma 3.2 Let $h, g \in C^s(\mathcal{X})$. If the family of conditional distributions $\{\rho_x\}_{x\in X}$ is Lipschitz-s, then we have

$$\left|\int_{\mathcal{Z}} \psi_\tau^{\epsilon_t}(y - h(x)) - \psi_\tau^{\epsilon_t}(y - g(x))d\left[\rho^{(t)} - \rho\right]\right| \le M_\rho\|\rho_{\mathcal{X}}^{(t)} - \rho_{\mathcal{X}}\|_{C^s(\mathcal{X}))^*}$$

where M_ρ, $B_{h,g}$ and $N_{h,g}$ are given by

$$M_\rho = \left\{B_{h,g}(\|h\|_{C^s(\mathcal{X})} + \|g\|_{C^s(\mathcal{X})}) + 2C_\rho N_{h,g}\right\},$$

and

$$B_{h,g} = \sup\left\{(|\psi_\tau^{\epsilon_t})'(y-f)| : y \in \mathcal{Y}, |f| \le \max\{\|h\|_{C(\mathcal{X})}, \|g\|_{C(\mathcal{X})}\}\right\},$$

and

$$N_{h,g} = \sup\left\{\|\psi_\tau^{\epsilon_t})'(u-f)\|_{C^s(\mathcal{Y})} : |f| \le \max\{\|h\|_{C(\mathcal{X})}, \|g\|_{C(\mathcal{X})}\}\right\}.$$

The proof of Lemma 3.2 can be found in [5].

3.2 One-Step Analysis

Now we turn to bound the sample error $\|f_{T+1} - f_{\lambda_T}^{\epsilon_T}\|_K$. This will be conducted by one-step iteration analysis which aims at bounding $\|f_{t+1} - f_{\lambda_t}^{\epsilon_t}\|_K$ in terms of $\|f_t - f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K$. We define the errors caused by the changing parameters ϵ_t and λ_t.

Definition 3.3 The *insensitive error* is defined as

$$h_t = \|f_{\lambda_{t-1}}^{\epsilon_{t-1}} - f_{\lambda_{t-1}}^{\epsilon_t}\|_K, \quad t \in \mathbb{N}. \tag{3.4}$$

The drift error is defined as

$$d_t = \|f_{\lambda_{t-1}}^{\epsilon_t} - f_{\lambda_t}^{\epsilon_t}\|_K, \quad t \in \mathbb{N}. \tag{3.5}$$

Now we bound the sample error $\|f_{T+1} - f_{\lambda_T}^{\epsilon_T}\|_K$ through $\|f_T - f_{\lambda_T}^{\epsilon_T}\|_K$, h_t, d_t and Δ_t.

Lemma 3.4 Define $\{f_t\}$ by (2.4). Then we have

$$\begin{aligned}\mathbb{E}_{z_t}(\|f_{t+1} - f_{\lambda_t}^{\epsilon_t}\|_K^2) \le &(1-\eta_t\lambda_t)(1+A_1 d_t^{q_1})(1+A_2 h_t^{q_2})\|f_t - f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2 + 2\eta_t\Delta_t + \eta_t^2\mathbb{E}_{z_t}\|G_t\|_K^2 \\ &+ (1+A_1 d_1^{q_1})(h^{2-q_2}/A_2 + h_t^2) + d_t^{2-q_1}/A_1 + d_t^2, \end{aligned} \tag{3.6}$$

where G_t is defined as

$$G_t = (\psi_\tau^{\epsilon_t})'_-(f_t(x_t) - y_t)K_{x_t} + \lambda_t f_t.$$

Proof First, we claim that $\|f_t\| \le \frac{\kappa}{\lambda_t}, \forall t \in \mathbb{N}$. It can be easily seen by induction from $f_1 = 0$ and the following estimate is derived from (2.2)

$$\|f_{t+1}\|_K \le (1-\lambda_t\eta_t)\|f_t\|_K + \eta_t\kappa \le (1-\lambda_t\eta_t)\frac{\kappa}{\lambda_t} + \eta_t\kappa = \frac{\kappa}{\lambda_t} \le \frac{\kappa}{\lambda_{t+1}}.$$

From (2.3), we see by inner products that

$$\|f_{t+1} - f_{\lambda_t}^{\epsilon_t}\|_K^2 = \|f_t - f_{\lambda_t}^{\epsilon_t}\|_K^2 + 2\eta_t\langle f_{\lambda_t}^{\epsilon_t} - f_t, G_t\rangle_K + \eta_t^2\|G_t\|_K^2. \tag{3.7}$$

By the reproducing property (2.1),

$$\langle f_{\lambda_t}^{\epsilon_t} - f_t, G_t\rangle_K = (\psi_\tau^{\epsilon_t})'_-(f_t(x_t) - y_t)\{f_{\lambda_t}^{\epsilon_t}(x_t) - f_t(x_t)\} + \lambda_t\langle f_{\lambda_t}^{\epsilon_t} - f_t, f_t\rangle_K.$$

The convexity of the loss function $\psi_\tau^{\epsilon_t}$ tells us that

$$\begin{aligned}(\psi_\tau^{\epsilon_t})'_-(f_t(x_t)-y_t)\{f_{\lambda_t}^{\epsilon_t}(x_t)-f_t(x_t)\} =&(\psi_\tau^{\epsilon_t})'_-(f_t(x_t)-y_t)\{[f_{\lambda_t}^{\epsilon_t}(x_t)-y_t]-[f_t(x_t)-y_t]\}\\ \le&\psi_\tau^{\epsilon_t}(f_{\lambda_t}^{\epsilon_t}(x_t)-y_t)-\psi_\tau^{\epsilon_t}(f_t(x_t)-y_t).\end{aligned}$$

Also,

$$\lambda_t\langle f_{\lambda_t}^{\epsilon_t}-f_t,f_t\rangle_K\le\lambda_t\|f_{\lambda_t}^{\epsilon_t}\|_K\|f_t\|_K-\lambda_t\|f_t\|_K^2\le\frac{\lambda_t}{2}\|f_{\lambda_t}^{\epsilon_t}\|_K^2-\frac{\lambda_t}{2}\|f_t\|_K^2.$$

Thus,

$$\langle f_{\lambda_t}^{\epsilon_t}-f_t,G_t\rangle_K\le\psi_\tau^{\epsilon_t}(f_{\lambda_t}^{\epsilon_t}(x_t)-y_t)-\psi_\tau^{\epsilon_t}(f_t(x_t)-y_t)+\frac{\lambda_t}{2}\|f_{\lambda_t}^{\epsilon_t}\|_K^2-\frac{\lambda_t}{2}\|f_t\|_K^2.$$

Taking expectation with respect to z_t, we get by Lemma 3 in [15]

$$\mathbb{E}_{z_t}\langle f_{\lambda_t}^{\epsilon_t}-f_t,G_t\rangle_K\le\Delta_t-\frac{\lambda_t}{2}\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2.$$

Together with (3.7)

$$\mathbb{E}_{z_t}\|f_{t+1}-f_{\lambda_t}^{\epsilon_t}\|_K^2\le(1-\lambda_t\eta_t)\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2+\eta_t^2\|G_t\|^2+2\eta_t\Delta_t.$$

Note that $\|(\psi_\tau^{\epsilon_t})'_-\|_\infty\le1$ and $\|G_t\|_K\le2\kappa$. We get

$$\mathbb{E}_{z_t}\|f_{t+1}-f_{\lambda_t}^{\epsilon_t}\|_K^2\le(1-\lambda_t\eta_t)\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2+4\kappa^2\eta_t^2+2\eta_t\Delta_t.\tag{3.8}$$

Decompose $\|f_t-f_\lambda^{\epsilon_t}\|_K^2$ as $\|f_t-f_{\lambda_{t-1}}^{\epsilon_t}+f_{\lambda_{t-1}}^{\epsilon_t}-f_{\lambda_t}^{\epsilon_t}\|_K^2$. Using the elementary inequality $2ab\le Aa^2b^q+b^{2-q}/A$ with $0<q<2, A>0$ to the case of $a=\|f_t-f_{\lambda_{t-1}}^{\epsilon_t}\|_K, b=d_t, A=A_1, q=q_1$, we obtain

$$\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2\le\|f_t-f_{\lambda_{t-1}}^{\epsilon_t}\|_K^2+A_1\|f_t-f_{\lambda_{t-1}}^{\epsilon_t}\|_K^2d_t^{q_1}+d_t^{2-q_1}/A_1+d_t^2.$$

Applying the same inequality to the case $a=\|f_t-f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K, b=h_t, A=A_2, q=q_2$, we see that

$$\|f_t-f_{\lambda_{t-1}}^{\epsilon_t}\|_K^2\le\|f_t-f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2+A_2\|f_t-f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2h_t^{q_2}+h_t^{2-q_2}/A_2+h_t^2.$$

Combining the two estimates, we obtain

$$\begin{aligned}&\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2\\ \le&(1+A_1d_t^{q_1})(1+A_2h_t^{q_2})\|f_t-f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2+(1+A_1d_t^{q_1})(h_t^{2-q_2}/A_2+h_t^2)+d_t^{2-q_1}/A_1+d_t^2.\end{aligned}$$

Putting it into (3.8), we get the desired bound (3.6).

3.3 Bounding the Sample Error

Now we can state our estimate for the sample error as follows.

Proposition 3.5 Suppose (2.5), (2.7), (2.8) and (2.9) hold. Take the parameters $\eta_t, \lambda_t, \epsilon_t$ as the same form in Theorem 2.6, then we have

$$\mathbb{E}_{z_1,\ldots,z_T} \left\| f_{T+1} - f_{\lambda_T}^{\epsilon_t} \right\|_K^2 \leq C'' T^{-\theta^*} \tag{3.9}$$

where C'' is a constant independent of T and θ^* is given in (2.13).

To prove this proposition, we need the following lemma, whose proof can be found in [12].

Lemma 3.6 If K satisfies the kernel condition of order s, then we have

$$\|g\|_{C^s(X)} \leq (\kappa + \kappa_{2s})\|g\|_K, \quad \forall g \in \mathcal{H}_K.$$

Now we proceed proving Proposition 3.5.

Proof To apply the estimate in Lemma 3.4, we need to explicit bounds for d_t and h_t. According to Lemma 3 in [5], we find

$$d_t \leq d_1 t^{-\min\{1-p/2+p\gamma/2, 1-p/2+\beta/2\}}, \quad \forall t \in \mathbb{N}.$$

where $d_1 = p2^{p+1}\sqrt{(2\mathcal{D}_0\lambda_1^\gamma + 4\epsilon_1)/\lambda_1}$. Using Proposition 1 in [5] with $\lambda = \lambda_{t-1}$, we obtain

$$h_t \leq h_1 t^{p-(\beta+1)s}, \quad \forall t \in \mathbb{N}.$$

where $h_1 = C_\rho \kappa \epsilon_1^s \beta^s 2^{(\beta+1)s}/\lambda_1$.

Now we apply Lemma 3.4. Take

$$q_1 = \frac{\alpha + p}{\min\{1 - p/2 + p\gamma/2, 1 - p/2 + \beta/2\}}, \quad q_2 = \frac{\alpha + p}{(\beta+1)s - p}$$

and

$$A_1 = d_1^{-q_1}\frac{\lambda_1\eta_1}{6} > 0, \quad A_2 = h_1^{-q_2}\min\left\{\frac{\lambda_1\eta_1}{6}, 1\right\} > 0.$$

From the restrictions (2.10) and (2.11), we see that $0 < q_1 < 2$ and $0 < q_2 < 2$. Then the coefficient of the first term of bound (3.6) can be bounded as

$$\begin{aligned}(1 - \lambda_t\eta_t)(1 + A_1 d_t^{q_1})(1 + A_2 h_t^{q_2}) \leq & 1 + (A_1 d_t^{q_1} + A_2 h_t^{q_2} + A_1 A_2 d_t^{q_1} h_t^{q_2})t^{(-\alpha+p)} - \lambda_t\eta_t \\ \leq & 1 - \frac{\eta_1\lambda_1}{2}t^{-\alpha-p}.\end{aligned}$$

Thus by Lemma 3.4, we have

$$\begin{aligned}&\mathbb{E}_{z_1,\ldots,z_t}\|f_{t+1} - f_{\lambda_t}^{\epsilon_t}\|_K^2 \\ \leq &\left(1 - \frac{\eta_1\lambda_1}{2}t^{-\alpha-p}\right)\mathbb{E}_{z_1,\ldots,z_{t-1}}\|f_t - f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2 + A_3 t^{-\theta_1} + 2\eta_t\Delta_t,\end{aligned} \tag{3.10}$$

where

$$\theta_1 = \min\{2 + p\gamma - 2p - \alpha, 2 + \beta - 2p - \alpha, 2(\beta+1)s - 3p - \alpha, 2\alpha\}$$

and

$$A_3 = (1 + A_1 d_1^{q_1})(h_1^{2-q_2}/A_2 + h_1^2) + d_1^{2-q_1}/A_1 + d_1^2 + 4\kappa^2\eta_1^2.$$

Next we bound Δ_t. From the Lemma 3 in [5], we have that

$$\|f_{\lambda_t}^{\epsilon_t}\|_K \le \sqrt{(2\mathcal{D}(\lambda_t) + 4\epsilon_t)/\lambda_t}, \quad t = 1, \dots, T,$$

and $\|f_t\|_K \le \frac{\kappa}{\lambda_t}$. By Lemma 3.2 and Lemma 3.6,

$$\Delta_t \le B_t^* := \Big\{(\kappa + \kappa_{2s})(\sqrt{(2\mathcal{D}(\lambda_t) + 4\epsilon_t)/\lambda_t} + \kappa/\lambda_t) + 2C_\rho/\lambda_t\Big\}\|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*}.$$

Applying condition (2.5), we can bound B_t^* as

$$B_t^* \le A_4 t^{p-b} \quad where \quad A_4 = C\Big\{(\kappa + \kappa_{2s})(\sqrt{(2\mathcal{D}_0 + 4\epsilon_1)\lambda_1} + \kappa/\lambda_1) + 2C_\rho/\lambda_1\Big\}.$$

Therefore, for the one-step iteration, we have for each $t = 1, \dots, T$,

$$\mathbb{E}_{z_1,\dots,z_t}(\|f_{t+1} - f_{\lambda_t}^{\epsilon_t}\|_K^2) \le \left(1 - \frac{\eta_1\lambda_1}{2}t^{-\alpha-p}\right)\mathbb{E}_{z_1,\dots,z_{t-1}}(\|f_{t+1} - f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2) + A_5 t^{-\theta_2}$$

where $A_5 = A_3 + 2\eta_1 A_4$ and

$$\theta_2 = \min\{2 + p\gamma - 2p - \alpha, 2 + \beta - 2p - \alpha, 2(\beta+1)s - 3p - \alpha, 2\alpha, \alpha - p + b\}.$$

Applying this bound iteratively for $t = 1, \dots, T$ implies

$$\mathbb{E}_{z_1,\dots,z_T}(\|f_{T+1} - f_{\lambda_T}^{\epsilon_t}\|_K^2) \le A_5 \sum_{t=1}^{T}\prod_{j=t+1}^{T}(1 - \frac{\eta_1\lambda_1}{2}j^{-\alpha-p})t^{-\theta_2}.$$

Applying the following elementary inequality in [12] with $0 < a_1 < 1, c, a_2 > 0$ and $t \in \mathbb{N}$

$$\sum_{i=1}^{t-1} i^{-a_2}\exp\Big\{-c\sum_{j=i+1}^{t} j^{-a_1}\Big\} \le \left\{\frac{2^{a_1+a_2}}{c} + \left(\frac{1+a_2}{ec(1-2^{a_1-1})}\right)^{(1+a_2)/(1+a_1)}\right\}t^{a_1+a_2}$$

to the case of $a_1 = \alpha + p < 1, a_2 = \theta_2$ and $c = \eta_1\lambda_1/2$, we see that

$$\sum_{t=1}^{T}\prod_{j=t+1}^{T}(1 - \frac{\eta_1\lambda_1}{2}j^{-\alpha-p})t^{-\theta_2} \le A_6 T^{p+\alpha-\theta_2},$$

where

$$A_6 = \frac{2^{\alpha+p+\theta_2+1}}{\eta_1\lambda_1} + 1 + (\frac{2+2\theta_2}{e\eta_1\lambda_1(1-2^{\alpha+p-1})})^{\frac{1+\theta_2}{1-p-\alpha}}.$$

With the above estimate, we can get the desired bound (3.9) with $\theta^* = \theta_2 - p - \alpha$ and the constant $C'' = A_5 A_6$.

3.4 Estimating Total Error

This section is devoted to proving the main results in Section 2.3.

Proof of Theorem 2.6 By (3.1), we can get the statement by applying Propositions 3.1, 3.5 and (2.7).

To prove Theorem 2.8, we shall make use of the following comparison theorem [16].

Lemma 3.7 Let $0 < \varphi \leq \infty$ and $\xi > 1$. Denote $r = \varphi\xi/(\varphi + 1) > 0$. Assume the measure ρ has a τ- quantile of φ-average type ξ, then for any measurable function on X, we have

$$\|f - f_{\rho,\tau}\|_{L^r_{\rho_X}} \leq 2^{1-1/\xi}\xi^{1/\xi}\|\{b_\tau w_\tau^{\xi-1}\}^{-1}\|^{1/\xi}_{L^\varphi_{\rho_X}}\{\varepsilon(f) - \varepsilon(f_{\rho,\tau})\}^{1/\xi}.$$

Proof of Theorem 2.8 It is trivial to get the desired conclusion by Lemma 3.7 and Theorem 2.8.

4 Simulations

In this section we further discuss and demonstrate our theoretical results by illustrative examples.

Consider the models as follows. Let $\mathcal{X} = [0, 1]^{10}$, ρ_X be the Lebesgue measure on $[0, 1]^{10}$, then the marginal distribution sequence $\{\rho_{\mathcal{X}}^{(t)}\}$ satisfies $d\rho_{\mathcal{X}}^{(t)} = d\rho_{\mathcal{X}} + Ct^{-b}d\rho_{\mathcal{X}}$, and for each $x \in \mathcal{X}$, the conditional distribution $\rho_{\mathcal{X}}$ is noised by the uniform distribution on $[-0.5, 0.5]$ around the regression function value where the parameters are described in Table 1.

$$f_\rho(x) = \sum_{i=1}^{3} A_i \exp(-\frac{|x - P_i|^2}{2v_i^2}).$$

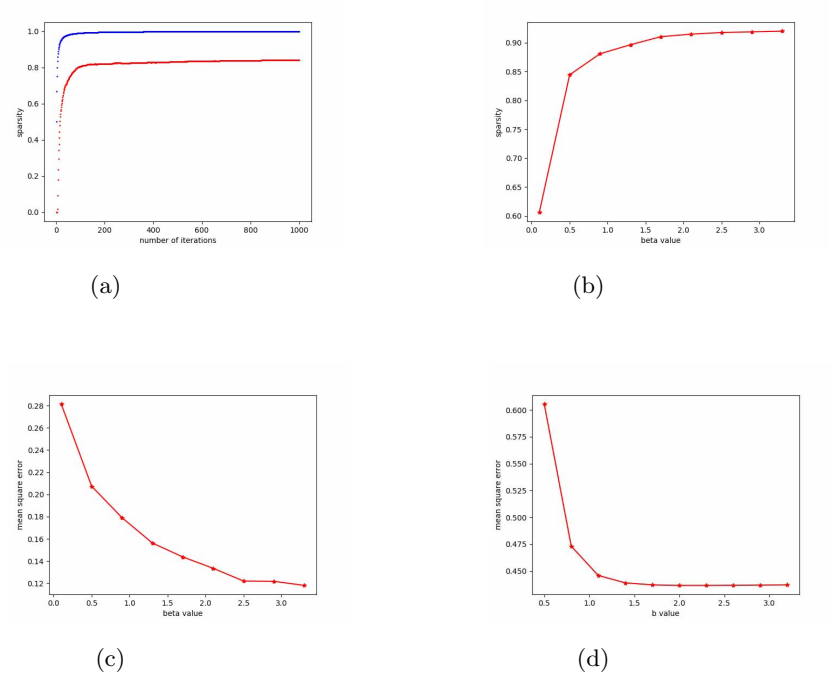
Table 1 Parameters

i	Coefficient A_i	Variation v_i^2	Center P_i
1	2.0	0.62^2	$(0.3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)$
2	3.5	0.64^2	$(0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6)$
3	0.7	0.65^2	$\frac{1}{9}(0.9, 1.7, 2.5, 3.3, 4.1, 4.9, 5.7, 6.5, 7.3, 8.1)$

We take the Gaussian kernel $K(x, u) = \exp\{-|x - u|^2/2\sigma^2\}$ with variance $\sigma^2 = 0.6^2$. When $\tau = 0.5$, $s = 1$ is valid. Meantime, the measure ρ has a $\frac{1}{2}$-quantile of ∞-average type 2. In our simulations, we compare mean square error in each numerical experiment.

$$MSE(T) = (\frac{1}{M}\sum_{j=1}^{M}(f_{T+1}(\xi_j) - f_{\rho,\tau}(\xi_j))^2)^{1/2}$$

where M is the sample size and $\{\xi_j\}$ is an unlabelled sample set drawn from non-identical distribution.



(a) (b)

(c) (d)

Figure 1 Simulation

For the sparsity caused by varying ϵ_t-insensitive loss, by (2.4), we can express the output function f_T as $f_T = \sum_{i=1}^{T} a_i K_{x_i}, \mathbf{a} = \{a_i\}_{i=1}^{T} \in \mathbb{R}^T$. Here, the degree of the sparsity of the online learning algorithm is measured by $\|\mathbf{a}\|_{\mathbf{0}}$, the proportion of non-zero coefficients in $\mathbf{a}$. Take $\tau = 0.5$, $\eta_1 = 0.4$, $\lambda_1 = 0.001$, $\epsilon_1 = 7.1$, $\alpha = 0.1$, $\beta = 0.8$, $p = 0.04$. Note that $\epsilon_t \equiv 0$ corresponds to the online quantile regression without threshold. We compare their sparsity and mean squared errors in Figure 1(a). Obviously, the red curve of $\epsilon_t = 7.1t^{-0.8}$ has more sparsity than the blue one of $\epsilon_t \equiv 0$.

In Figure 1(b) and (c), we show how the sparsity power β affects the mean square error and sparsity. As we see, if β increases, the mean square error will decrease while $\|\mathbf{a}\|_{\mathbf{0}}$ will become larger. Thus, the choice of β should balance the mean square error and sparsity. It confirms our theoretical results in Theorems 2.6 and 2.8.

In Figure 1(d), we report the change of the mean squared error as the power index b increases. We set the sample size $M = 200$, number of iterations $T = 3000$ and $\eta_1 = 0.4$, $\lambda = 0.001$, $\epsilon_1 = 7.1$, $\alpha = 0.1$, $\beta = 0.8$ and $p = 0.02$. We plot mean squared errors from 0.5 to 3.2. At the beginning, the error decreases as b increases, but when b is larger than 1.7, the error does not change. This phenomenon coincides with our theoretical results that the non-identical sampling does not affect on the learning ability if the dependence between the samples is weak.

References

[1] Buchinsky M, Hahn J. An alternative estimator for the censored quantile regression model[J]. Econometrica, 1998: 653–671.

[2] Koenker R, Hallock K F. Quantile regression[J]. Journal of Economic Perspectives, 2001, 15(4): 143–156.

[3] Yu K, Lu Z, Stander J. Quantile regression: applications and current research areas[J]. Journal of the Royal Statistical Society: Series D (The Statistician), 2003, 52(3): 331–350.

[4] Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function[J]. The Annals of Mathematical Statistics, 1952, 23(3): 462–466.

[5] Hu T, Xiang D H, Zhou D X. Online learning for quantile regression and support vector regression[J]. Journal of Statistical Planning and Inference, 2012, 142(12): 3107–3122.

[6] Hwang C, Shim J. A simple quantile regression via support vector machine[C]// International Conference on Natural Computation. Springer, Berlin, Heidelberg, 2005: 512–520.

[7] Kivinen J, Smola A J, Williamson R C. Online learning with kernels[J]. IEEE Transactions on Signal Processing, 2004, 52(8): 2165–2176.

[8] Rosset S. Bi-Level path following for cross validated solution of kernel quantile regression[J]. Journal of Machine Learning Research, 2009, 10(11): 840–847

[9] Smale S, Yao Y. Online learning algorithms[J]. Foundations of Computational Mathematics, 2006, 6(2): 145–170.

[10] Takeuchi I, Le Q V, Sears T D, et al. Nonparametric quantile estimation[J]. Journal of Machine Learning Research, 2006, 7(Jul): 1231–1264.

[11] Hu T, Zhou D X. Online learning with samples drawn from non-identical distributions[J]. Journal of Machine Learning Research, 2009, 10(12): 2873–2898

[12] Smale S, Zhou D X. Online learning with Markov sampling[J]. Analysis and Applications, 2009, 7(01): 87–113.

[13] Christmann A, Steinwart I. How SVMs can estimate quantiles and the median[C]// Advances in Neural Information Processing Systems. 2008: 305–312.

[14] Xiang D H, Hu T, Zhou D X. Learning with varying insensitive loss[J]. Applied Mathematics Letters, 2011, 24(12): 2107–2109.

[15] Ying Y, Zhou D X. Online regularized classification algorithms[J]. IEEE Transactions on Information Theory, 2006, 52(11): 4775–4788.

[16] Steinwart I, Christmann A. Estimating conditional quantiles with the help of the pinball loss[J]. Bernoulli, 2011, 17(1): 211–225.

多阈值和非独立同分布的在线分位数学习算法

蒋铭勤

(武汉大学数学与统计学院, 湖北 武汉 430072)

摘要: 本文研究了多阈值和非一致分布下的在线分位数回归算法, 在每一次迭代中, 样本会来自不同的分布和取不同的阈值. 利用边缘分布在对偶空间中多项式收敛的性质, 我们得到了算法的学习速度, 并且做了相应的数值模拟来支持我们的结论.

关键词: 非一致分布; 在线学习; 分位数回归; 再生核希尔伯特空间

MR(2010)**主题分类号**: 62J99 **中图分类号**: O29