Online Kernel-Based Quantile Regression Using Huberized Pinball Loss

Takumi Ichinose, Masahiro Yukawa
Department of Electronics and Electrical Engineering
Keio University, Yokohama, JAPAN
ichinose@ykw.elec.keio.ac.jp, yukawa@elec.keio.ac.jp

Renato L. G. Cavalcante
Fraunhofer Heinrich Hertz Institute
Berlin, GERMANY
renato.cavalcante@hhi.fraunhofer.de

***Abstract*—We present an efficient online kernel-based quantile regression scheme based on the Moreau envelope of the pinball loss, which we call the Huberized pinball loss. The use of the Moreau envelope is motivated by the popular Huber loss, which is the Moreau envelope of the least absolute deviation in robust estimation. We show that the smooth Huberized pinball loss exhibits more robust learning behaviours than the ordinary pinball loss in some scenarios, while the discrepancy of its minimizer from the true quantile is bounded by constants dependent on the Moreau-envelope parameter. Numerical examples show that the proposed scheme achieves better and more stable performances than a pinball-loss-based online method.**

I. INTRODUCTION

Quantile regression [1]–[3] is a task of identifying an interval in which a prespecified fraction of data reside. It has been gaining growing attention because it tends to provide remarkably robust estimates. Estimation of "uncertainty (interval)" instead of an accurate estimate (point), is crucial in many applications including credit risk prediction [4], [5], wind power forecasting [1], [6], [7], and survival analysis [8]. There exist three popular batch methods: quantile regression forest [2], kernel quantile regression [9], and quantile regression neural network [3]. Those methods are known to work well in the sense of yielding an interval that is both reasonably narrow and contains the desired fraction of data. However, if data are contaminated by outliers, the distribution of the output spreads, making the intervals estimated by those methods undesirably wide. In this regard, the robustness of those methods is limited.

The "uncertainty" in the current context depends on the distribution of perturbations (disturbance, noise, outlier, etc.). In signal processing applications, the distribution may change over time. In such a case, the performance of batch methods for quantile regression is degraded seriously. In the present study, we therefore address the online quantile regression problem where the relation between input and output is nonlinear.

The study in [1] proposes an online quantile regression method that uses kernels and a subgradient method to minimize the so-called pinball loss [10]. Here, the pinball loss is known to give an "empirical" counterpart of the αth quantile for $\alpha \in (0, 1)$ [9] (see Lemma 1 in Section II-B), where the αth quantile is the smallest possible value for which the conditional probability of the output given an input measurement is at least α. Its use in the online case, however, may cause unstable performance in certain scenarios where the noise distribution changes during some time slots (see Section III-A). A smooth relaxation of the pinball loss has also been proposed [11] so that the standard gradient method can be applied. Unfortunately, it lacks theoretical verification of producing quantiles, and its performance with respect to the coverage accuracy leaves room for improvements.

In this study, we propose an online quantile regression method based on the multikernel adaptive filtering framework. To alleviate the instability mentioned above, we replace the pinball loss function by its Moreau envelope [12], which we call Huberized pinball loss. The idea of this relaxation comes from the analogy to the relation between the least absolute deviation (LAD) and the Huber loss (which is the Moreau envelope of the LAD) in robust statistics [13]. The Huberized pinball loss is a smooth function, whereas the pinball loss is nonsmooth. This smoothness property is advantageous in online scenarios because the computationally efficient gradient method can be applied. We also derive upper and lower bounds of the discrepancy between the αth quantile and the minimizer of the Huberized pinball loss, where the bounds depend on α and the Moreau-envelope parameter. Simulations show the superiority of the proposed loss compared to the ordinary pinball loss in the online case.

II. PRELIMINARIES

The sets of real numbers and nonnegative integers are denoted by $\mathbb{R}$ and $\mathbb{N}$, respectively. Scalars, vectors, and matrices are denoted by lowercase letters, lowercase boldfaced letters, and uppercase boldfaced letters, respectively. The transpose of a matrix or a vector is denoted by $(\cdot)^\top$.

A. System model and problem statement

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ denote the input space and the output space, respectively. Let $(\boldsymbol{x}_i)_{i=1}^{\infty}$ be a sequence of random input vectors taking values in $\mathcal{X}$, and the contaminated outputs

$$y_i := \psi(\boldsymbol{x}_i) + \epsilon_i + o_i \tag{1}$$

arrive sequentially, taking values in $\mathcal{Y}$. Here, $\psi\colon \mathbb{R}^d \to \mathbb{R}$ is an unknown nonlinear function, ϵ_i is Gaussian noise taking values in $\mathbb{R}$, and o_i is the outlier which is sparse in time but has large magnitude.

This work was supported in part by JST SICORP under Grant JPMJSC20C6, Japan. R. L. G. Cavalcante acknowledges the financial support by the Federal Ministry of Education and Research of Germany in the programme of "Souveran. Digital. Vernetzt." Joint project 6G-RIC, project identification number: 16KISK020K. The authors alone are responsible for the content of the paper.

The goal of quantile regression is to find an interval in which the random observations y_i, perturbed by noise and outliers, occur with a prespecified probability $\beta \in (0,1)$. We suppose for simplicity that the interval is desired to be located in the center of the distribution, so that data residing outside the interval are equally distributed below and above the interval. To be more specific, we define the αth quantile for some given $\alpha \in (0,1)$ by

$$q_\alpha(\boldsymbol{x}) \coloneqq \inf\left\{y \in \mathcal{Y} \colon F(y \mid \boldsymbol{x}) \geq \alpha\right\}, \quad \boldsymbol{x} \in \mathcal{X}, \tag{2}$$

where $F(y \mid \boldsymbol{x})$ is the conditional distribution of y given $\boldsymbol{x}$. Then, the objective is to estimate the upper and lower quantiles q_{α_+} and q_{α_-} for $\alpha_+ := (1+\beta)/2 \in (0,1)$ and $\alpha_- := (1-\beta)/2 \in (0, \alpha_+)$ to obtain the desired interval $[q_{\alpha_-}, q_{\alpha_+}]$. The difference $\beta (= \alpha_+ - \alpha_-)$ is referred to as the target coverage rate.

B. Pinball loss function

As common in regression methods, practical approaches to the quantile regression task rely on an empirical loss defined with observed data. Specifically, given $n \in \mathbb{N}$ samples, let us consider the regularized *pinball loss* [9], [10], [14]:

$$R_{\mathrm{reg}}(g,b) := \sum_{i=1}^{n} \rho_\alpha(y_i - g(\boldsymbol{x}_i) - b) + \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2, \ g \in \mathcal{H}, \ b \in \mathbb{R}, \tag{3}$$

where $(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|_{\mathcal{H}})$ is a reproducing kernel Hilbert space, $\lambda > 0$ is the regularization parameter, and

$$\rho_\alpha : \mathbb{R} \to [0,+\infty) : z \mapsto \begin{cases} \alpha z & \text{if } z \geq 0, \\ -(1-\alpha)z & \text{if } z < 0. \end{cases} \tag{4}$$

Note here that the offset term b is not regularized. The particular choice $\alpha = 0.5$ reduces the pinball loss to the LAD, yielding the median. In this sense, the pinball loss is a generalization of LAD, and it induces robust estimates in the presence of outliers. For readers' convenience, we present a theoretical result known in the literature below.

Lemma 1 ([9]). *Let $(g_\star, b_\star) \in \mathcal{H} \times \mathbb{R}$ be the minimizer of the $R_{\mathrm{reg}}(g,b)$. Then, the following statements hold ($|S|$ denotes the cardinality of a set S).*

1) $\left|\left\{i \in \overline{1,n} := \{1,2,\cdots,n\} : y_i < g_\star(\boldsymbol{x}_i) + b_\star\right\}\right| \leq \alpha n$.
2) $\left|\left\{i \in \overline{1,n} : y_i > g_\star(\boldsymbol{x}_i) + b_\star\right\}\right| \leq (1-\alpha) n$.
3) *Assume that $\boldsymbol{x}_i$ and y_i are i.i.d. with $F(\boldsymbol{x} \mid y)$ continuous and the expectation of the modulus of absolute continuity of its density satisfying $\lim_{\delta \to 0} E(\epsilon(\delta)) = 0$. Then, $\lim_{n\to\infty} \left|\left\{i \in \overline{1,n} : y_i < g_\star(\boldsymbol{x}_i) + b_\star\right\}\right| / n = \alpha$ with probability 1.*

III. Proposed Online Method

We start by introducing a smooth relaxation of the pinball loss by using its Moreau envelope, and we show that the smoothed loss yields an approximation of quantile. This relaxation not only simplifies the optimization process, but it also makes the online method robust against changes in the noise distribution that may happen during some time slots. We then present the proposed kernel-based method using the relaxed pinball function.

A. Huberized pinball loss

Simple algorithms are desirable in online implementations, so the subgradient method is typically used to minimize a cost function based on the pinball loss [1]. The pinball loss, however, increases linearly when one deviates from the minimizer, and therefore it is sensitive to small deviations. To illustrate a potential issue due to this property, let us consider the case when the noise distribution changes in a way that noise happens in the positive region more frequently than in the negative region during some time slots. In this case, the quantile estimates are updated more in the upward direction than in the downward direction. These updates may cause serious estimation errors.

To avoid the above situation, we focus on the similarity between the pinball loss and LAD (the ℓ_1 norm). The popular Huber loss is less sensitive to small perturbations than LAD, because it is a quadratic function in the vicinity of the minimizer. Based on this analogy, we use the Moreau envelope of the pinball loss $\rho_\alpha(x)$, defined as follows:

$$\begin{aligned} {}^{\gamma}\rho_\alpha(z) &\coloneqq \min_u \left(\rho_\alpha(u) + \frac{1}{2\gamma}\|u - z\|_2^2\right) \\ &= \begin{cases} \alpha z - \dfrac{\alpha^2\gamma}{2} & z \geq \alpha\gamma, \\ \dfrac{1}{2\gamma}z^2 & -(1-\alpha)\gamma < z < \alpha\gamma, \\ -(1-\alpha)z - \dfrac{(1-\alpha)^2\gamma}{2} & z \leq -(1-\alpha)\gamma, \end{cases} \end{aligned} \tag{5}$$

which we call *the Huberized pinball loss.*

As explained above, the Huberized pinball loss is motivated by online algorithms using (sub)gradient. Now, considering the asymptotic behaviour of online algorithms, a natural question would be the following: *does the Huberized pinball loss ${}^{\gamma}\rho_\alpha$ give an efficient estimate of the αth quantile?* The following proposition gives an answer to this fundamental question.

Proposition 1 (Huberized pinball loss and αth quantile)**.** *Assume that y is an integrable random variable, where the integrability is defined with respect to a probability measure. Assume also that the cumulative distribution F of y is a strictly-increasing continuous function. Let $\hat{q}_\alpha \in \operatorname*{argmin}_{q\in\mathbb{R}} E_y[{}^{\gamma}\rho_\alpha(y - q)]$. Then, it holds that*

$$-(1-\alpha)\gamma \leq q_\alpha - \hat{q}_\alpha \leq \alpha\gamma. \tag{6}$$

Proof. We first show the following equality:

$$\frac{1}{\gamma}\int_{\hat{q}_\alpha - (1-\alpha)\gamma}^{\hat{q}_\alpha + \alpha\gamma} F(y)dy = \alpha. \tag{7}$$

Let $\mathcal{I} \subset \mathcal{Y}$ be a bounded open interval that is supposed to be sufficiently wide. Let $m_{\mathcal{Y}}$ be the probability distribution, i.e., $E_y(h(y)) = \int_{\mathcal{Y}} h(y) m_{\mathcal{Y}}(dy)$ for any measurable function h. The function $g(y,q) := {}^{\gamma}\rho_\alpha(y-q)$ on $\mathcal{Y} \times \mathcal{I}$ is integrable in terms of y for an arbitrarily fixed $q \in \mathcal{I}$, and it is partially differentiable with respect to q given any y. In addition, it is not difficult to verify that $\left|\frac{\partial g}{\partial q}(y,q)\right| \leq 1 =: M(y)$ for all $q \in \mathcal{I}$, where $\int_{\mathcal{Y}} |M(y)| \, m_{\mathcal{Y}}(dy) = 1 < +\infty$. Hence,

$G(q) := E_y(g(y,q)) = \int_{\mathcal{Y}} g(y,q) m_{\mathcal{Y}}(dy)$ is differentiable w.r.t. q, and it holds [15, Theorem 14.2] that $\frac{\partial}{\partial q} E_y(g(y,q)) = E_y(\frac{\partial g}{\partial q}(y,q))$. Because the expectation $E_y[{}^{\gamma}\rho_\alpha(y-q)]$ of the measurable convex function ${}^{\gamma}\rho_\alpha(y-q)$ is a convex function [16, Proposition 8.24], the function G given y is minimized when the following equality holds:

$$E_y\left(\frac{\partial g}{\partial q}(y,q)\right) = 0. \tag{8}$$

Note here that ${}^{\gamma}\rho_\alpha(y-q) \geq 0$ over $\mathcal{Y} \times \mathcal{I}$ to ensure the convexity as the zero function is absolutely integrable. With some manipulations, one can show that $E_y\left(\frac{\partial g}{\partial q}(y,q)\right) = -\alpha + \frac{1}{\gamma}\int_{\hat{q}_\alpha-(1-\alpha)\gamma}^{\hat{q}_\alpha+\alpha\gamma} F(y)dy$, which together with (8) implies the equality in (7).

We conclude from (7) that the (strictly) increasing function F must satisfy $F(\hat{q}_\alpha - (1-\alpha)\gamma) \leq \alpha \leq F(\hat{q}_\alpha + \alpha\gamma)$. Hence, due to the continuity of F, there exists a unique $q_c \in [\hat{q}_\alpha - (1-\alpha)\gamma, \hat{q}_\alpha + \alpha\gamma]$ such that $F(q_c) = \alpha$. It is clear that the q_c coincides with the αth quantile q_α, and we thus obtain the desired inequality (6). □

Proposition 1 states that the minimizer $\hat{q}_\alpha$ of the Huberized pinball loss is in the neighbor of the α quantile q_α with radius vanishing as γ tends to zero.[1]

B. Online quantile kernel regression (OQKR)

The Gaussian kernel with width $\sigma > 0$ is defined by

$$k : \mathcal{X} \times \mathcal{X} \to [0,+\infty) : (\boldsymbol{x}, \hat{\boldsymbol{x}}) \mapsto \exp\left(-\frac{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2}{2\sigma^2}\right), \tag{9}$$

where $\|\cdot\|_2$ is the Euclidean norm. Let $\{k(\cdot,\boldsymbol{\xi}_1),\ k(\cdot,\boldsymbol{\xi}_2),\ \cdots,\ k(\cdot,\boldsymbol{\xi}_r)\}$ be the *dictionary* with $\boldsymbol{\xi}_1,\ \boldsymbol{\xi}_2,\ \cdots, \boldsymbol{\xi}_r \in \mathcal{X}$ selected from the observed input vectors based on some novelty criterion in an online manner [17], [18]. Note that the size r and the vectors $\boldsymbol{\xi}_i$ themselves may change over time in practice (see [17], [18] for details).

Define the kernelized input vector $\boldsymbol{k}(\boldsymbol{x}) := [k(\boldsymbol{x},\boldsymbol{\xi}_1), k(\boldsymbol{x},\boldsymbol{\xi}_2),\ \cdots, k(\boldsymbol{x},\boldsymbol{\xi}_r)]^\top \in \mathbb{R}^r$. Given any $\boldsymbol{x} \in \mathcal{X}$, the kernel adaptive filter is given by

$$\hat{\psi}_{\boldsymbol{w}}(\boldsymbol{x}) := \boldsymbol{w}^\top \tilde{\boldsymbol{k}}(\boldsymbol{x}), \tag{10}$$

where $\boldsymbol{w} \in \mathbb{R}^{r+1}$ is the weight vector, and $\tilde{\boldsymbol{k}}(\boldsymbol{x}) := [\boldsymbol{k}(\boldsymbol{x})^\top\ c]^\top \in \mathbb{R}^{r+1}$ is the kernelized input vector, where $c := \sqrt{E(\|\boldsymbol{k}(\boldsymbol{x}_i)\|_2^2)/r}$ is a constant corresponding to the offset variable. Note here that $\boldsymbol{k}(\boldsymbol{x}_i)$ is a random vector as $\boldsymbol{x}_i$ is so. The scalar c reduces the eigenvalue spread of the autocorrelation matrix of $\tilde{\boldsymbol{k}}(\boldsymbol{x})$. As such, the last entry of $\boldsymbol{w}$ (i.e., the offset variable) converges at nearly the same speed as the other entries.

The smooth cost function is then given by

$$\Theta_\alpha(\boldsymbol{w}) := E\left({}^{\gamma}\rho_\alpha(y_i - \boldsymbol{w}^\top \tilde{\boldsymbol{k}}(\boldsymbol{x}_i))\right) \tag{11}$$

[1]If $y \in \mathcal{Y}$ follows a uniform distribution, it is immediate to verify from Proposition 1 that $\hat{q}_\alpha(\boldsymbol{x}) = q_\alpha(\boldsymbol{x}) + \gamma\left(\frac{1}{2} - \alpha\right)$.

of which the minimizer $\hat{\boldsymbol{w}} = \operatorname{argmin}_{\boldsymbol{w}} \Theta_\alpha(\boldsymbol{w})$ gives the estimate $\hat{q}_\alpha(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \tilde{\boldsymbol{k}}(\boldsymbol{x})$ of the αth quantile. Here, $E(\cdot)$ stands for expectation taken with respect to the input vector $\boldsymbol{x}_i \in \mathcal{X}$, the noise ϵ_i, and the outlier o_i. Given an initial guess $\boldsymbol{w}_0 \in \mathbb{R}^r$, the stochastic gradient descent method to minimize the smooth cost function ${}^{\gamma}\Theta_\alpha$ is given by

$$\begin{aligned}\boldsymbol{w}_{t+1} &:= \boldsymbol{w}_t - \mu \nabla_{\boldsymbol{w}} {}^{\gamma}\rho_\alpha(y_t - \boldsymbol{w}_t^\top \tilde{\boldsymbol{k}}(\boldsymbol{x}_t)) \\ &= \boldsymbol{w}_t + \mu\ \mathrm{clip}_\alpha\left(\frac{y_t - \boldsymbol{w}_t^\top \tilde{\boldsymbol{k}}(\boldsymbol{x}_t)}{\gamma}\right)\tilde{\boldsymbol{k}}(\boldsymbol{x}_t),\end{aligned} \tag{12}$$

where $\mu > 0$ is the step size, and

$$\mathrm{clip}_\alpha : \mathbb{R} \to [-(1-\alpha), \alpha] : x \mapsto \begin{cases}\alpha, & x > \alpha, \\ x, & x \in [-(1-\alpha), \alpha], \\ -(1-\alpha), & x < -(1-\alpha).\end{cases}$$

C. Online quantile multikernel regression (OQMkR)

We now consider the use of multiple kernels to enhance the accuracy and the convergence speed. The Gaussian kernels with different widths $\sigma_1 > \sigma_2 > \cdots > \sigma_Q > 0$ are defined by

$$k_q : \mathcal{X} \times \mathcal{X} \to [0,+\infty) : (\boldsymbol{x}, \hat{\boldsymbol{x}}) \mapsto \exp\left(-\frac{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2}{2\sigma_q^2}\right). \tag{13}$$

For each $q \in \overline{1,Q}$, we define the dictionary $\{k_q(\cdot,\boldsymbol{\xi}_{q,1}),\ k_q(\cdot,\boldsymbol{\xi}_{q,2}),\ \cdots,\ k_q(\cdot,\boldsymbol{\xi}_{q,r})\}$ with $\boldsymbol{\xi}_{q,1},\ \boldsymbol{\xi}_{q,2},\ \cdots, \boldsymbol{\xi}_{q,r} \in \mathcal{X}$. Given an input vector $\boldsymbol{x} \in \mathcal{X}$, we define the kernelized input vectors

$$\boldsymbol{k}_q(\boldsymbol{x}) := [k_q(\boldsymbol{x},\boldsymbol{\xi}_{q,1}), k_q(\boldsymbol{x},\boldsymbol{\xi}_{q,2}), \cdots, k_q(\boldsymbol{x},\boldsymbol{\xi}_{q,r})]^\top \in \mathbb{R}^r. \tag{14}$$

We also define its augmented version $\tilde{\boldsymbol{k}}_q(\boldsymbol{x}) \in \mathbb{R}^r$ in an analogous way to the previous subsection. Our estimate of the αth quantile is then given by

$$\hat{\psi}_{\boldsymbol{w}}(\boldsymbol{x}) := \sum_{q=1}^{Q} \boldsymbol{w}_q^\top \tilde{\boldsymbol{k}}_q(\boldsymbol{x}), \tag{15}$$

where $\boldsymbol{w}_q \in \mathbb{R}^{r+1}$ is the weight vector. By defining $\boldsymbol{w} := [\boldsymbol{w}_1^\top\ \boldsymbol{w}_2^\top\ \cdots\ \boldsymbol{w}_Q^\top]^\top \in \mathbb{R}^{(r+1)Q}$ and $\tilde{\boldsymbol{k}}(\boldsymbol{x}) = [\tilde{\boldsymbol{k}}_1^\top(\boldsymbol{x})\ \tilde{\boldsymbol{k}}_2^\top(\boldsymbol{x})\ \cdots\ \tilde{\boldsymbol{k}}_q^\top(\boldsymbol{x})]^\top \in \mathbb{R}^{(r+1)Q}$, (15) reduces to (10). The cost function is then given by the same form as in (11) but for $\boldsymbol{w} \in \mathbb{R}^{rQ}$, and the estimate $\boldsymbol{w}_t \in \mathbb{R}^{(r+1)Q}$ is updated by the same recursion as given in (12) accordingly.

IV. Numerical Examples

We evaluate the performance of the proposed method in comparisons with the existing methods for the following function:

$$\psi(x) := 4c_1 \exp\left(-\frac{(x-\eta_1)^2}{2\sigma_1^2}\right) + 2c_2 \exp\left(-\frac{(x-\eta_2)^2}{2\sigma_2^2}\right),$$

where $c_1, c_2, \eta_1, \eta_2 \geq 0$ are constants drawn from the uniform distribution $\mathcal{U}[0,1]$, and $\sigma_1 := 0.1$ and $\sigma_2 := 0.3$ ($d := 1$ so that $\mathcal{X} \subset \mathbb{R}$). The noise ϵ_i follows the zero-mean normal distribution with variance $\sigma_\epsilon^2(x) := (|x| + 0.3)^2$ at each point

$x \in \mathcal{X}$. The nonzero outliers obey the zero-mean normal distribution with variance $\sigma_o^2 = 1.0 \times 10^3$.

We use two performance measures. The first measure is the estimation errors of the upper quantile (those of the lower quantile can be defined analogously) Quantile Estimation Error $:= (\int (q_{\alpha_+}(x) - \hat{q}_{\alpha_+}(x))^2 dx)/(\int (q_{\alpha_+}(x))^2 dx)$. The second one is the actual coverage rate indicating the proportion of the observations that are included in the estimated interval relative to the total observations: Coverage $:= \frac{1}{n}\sum_{i=1}^{n} 1_C(y_i)$, where $1_C(y_i) = \begin{cases} 1 & \text{if } y_i - o_i \in C := [q_{\alpha_-}(x_i), q_{\alpha_+}(x_i)], \\ 0 & \text{otherwise.} \end{cases}$ The coverage rate is desired to be close to the prespecified rate β; Coverage $> \beta$ and Coverage $< \beta$ imply overcoverage and undercoverage, respectively.

A. Case of no distribution change of ϵ_i

The outlier rate is set to 0.02 ; i.e., the outlier occurs with probability 0.02. The proposed methods are compared with the online kernel-based method with the pinball loss [1] for the target coverage rate $\beta := 0.95$ ($\alpha_+ := 0.975$, $\alpha_- := 0.025$). For the proposed methods, $\gamma := 0.5$ and $\mu := 5.0 \times 10^{-3}$ are used by manual tuning. For the multikernel version, the number of kernels is set to $Q := 2$ with $\sigma_1^2 := 2.5 \times 10^{-3}$ and $\sigma_2^2 := 0.01$. For the pinball loss, the same step size $\mu := 5.0 \times 10^{-3}$ is used.

Figure 1 shows the learning curves. In the estimation error, (i) the proposed Huberized pinball loss function outperforms the ordinary pinball loss, and (ii) the use of multiple kernels further improves the convergence speed. The improvements come from the smoothness of the proposed loss. More precisely, the subgradient method for the pinball loss with fixed step size fluctuates largely in the vicinity of the true quantile, whereas the gradient of the proposed method vanishes as the estimate approaches the minimizer of the loss because of its smoothness. Moreover, the multikernel version of the proposed method works better than the single kernel version owing to its flexibility. In the actual coverage rate, the proposed method achieves comparable performance to the pinball loss. This fact indicates that the proposed methods yield efficient estimates of the αth quantile despite the relaxation, just as expected in light of Proposition 1.

B. Changing the distribution of ϵ_i at some points

To evaluate the robustness of the proposed loss function, we consider noise distributed asymmetrically during iterations 6000 – 6500 and 8000 – 8500, and it is normally distributed during other iterations. More precisely, the noise ϵ_i takes the positive sign at the rate 0.7 during iterations 6000 – 6500, while it takes the negative sign at the same rate 0.7 during iterations 8000 – 8500. In this subsection, we solely consider the single kernel version. We test $\gamma := 0.1, 0.2, 0.3, 0.4, 0.5$ as γ governs the sensitivity of the proposed method to noise.

Figure 2 shows the results. It is seen that the proposed method with a sufficiently large γ shows stable performance in the quantile estimation errors, because the gradient tends to be small around the minimizer of the loss. Note that a smaller γ makes the proposed loss closer to the pinball loss. The error in coverage rate increases slightly as γ increases; this is consistent with Proposition 1.

In contrast to the stable performance of the proposed loss, the pinball loss suffers from instability. More precisely, the estimation error of the upper quantile for the pinball loss increases during the first period of distribution change (6000 – 6500), and those of the lower quantile increase during the second period (8000 – 8500). As opposed to those phenomena, the former errors for the first period and the latter ones for the second period decrease. This can be explained as follows. In fact, owing to the asymmetry of the noise distribution in the first period, the quantile estimate is pushed in the upward direction, while it is pushed in the downward direction in the second period. As a result, the upper quantile is overestimated during the first period, and this overestimation is alleviated in the second period. On the other hand, the lower quantile tends to be underestimated; this could be because of the definition of the quantile using the infimum. Hence, the underestimated lower quantile is pushed upwards in the first period, and its error decreases accordingly. In summary, the "going up-and-down" phenomena in estimation error show that the subgradient algorithm using the pinball loss is sensitive to the distribution changes considered in this section.

V. Conclusion

We presented an efficient online quantile regression scheme with multiple kernels based on the Huberized pinball loss function. The Huberized pinball loss function provides an efficient estimate in the sense of a bounded error from the α quantile with the lower and upper bounds given by $-(1-\alpha)\gamma$ and $\alpha\gamma$, respectively. Numerical examples showed that the superior performance of the proposed scheme compared to the pinball-loss-based online method. We emphasize that, in contrast to the ordinary pinball loss, the Huberized pinball loss exhibited robust performance under the distribution changes of noise.

References

[1] C. Gallego-Castillo, R. Bessa, L. Cavalcante, and O. Lopez-Garcia, "On-line quantile regression in the RKHS (Reproducing Kernel Hilbert Space) for operational probabilistic forecasting of wind power," *Energy*, vol. 113, pp. 355–365, 2016.

[2] N. Meinshausen and G. Ridgeway, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, no. 6, 2006.

[3] J. W. Taylor, "A quantile regression neural network approach to estimating the conditional density of multiperiod returns," *Journal of Forecasting*, vol. 19, no. 4, pp. 299–311, 2000.

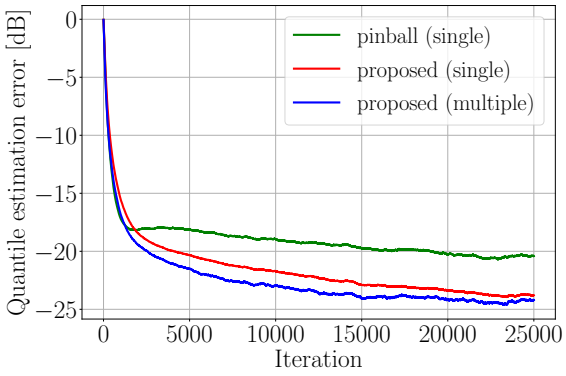
[4] M.-Y. L. Li and P. Miu, "A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach," *Journal of Empirical Finance*, vol. 17, no. 4, pp. 818–833, 2010.

[5] N. Naifar, "Do global risk factors and macroeconomic conditions affect global islamic index dynamics? a quantile regression approach," *The Quarterly Review of Economics and Finance*, vol. 61, pp. 29–39, 2016.

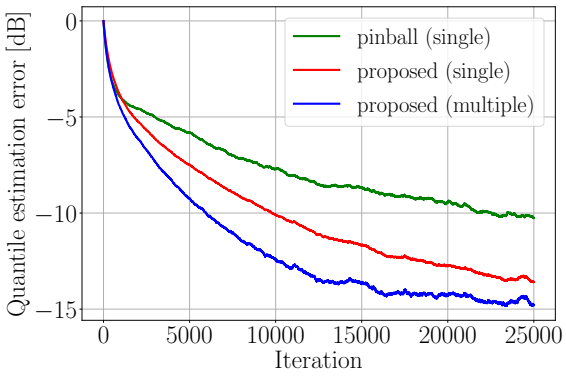
[6] J. B. Bremnes, "Probabilistic wind power forecasts using local quantile regression," *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, vol. 7, no. 1, pp. 47–54, 2004.

[7] A. U. Haque, M. H. Nehrir, and P. Mandal, "A hybrid intelligent model for deterministic and quantile regression approach for probabilistic wind power forecasting," *IEEE Transactions on Power Systems*, vol. 29, no. 4, pp. 1663–1672, 2014.

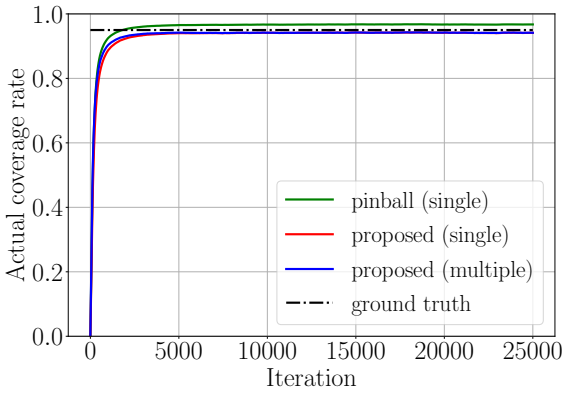
[8] K. Yu, Z. Lu, and J. Stander, "Quantile regression: applications and current research areas," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 3, pp. 331–350, 2003.



(a) upper quantile errors

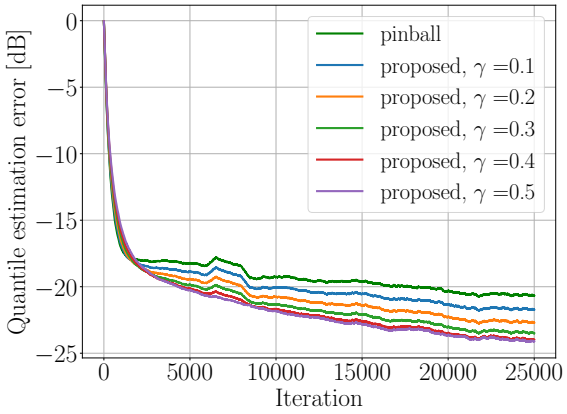


(b) lower quantile errors

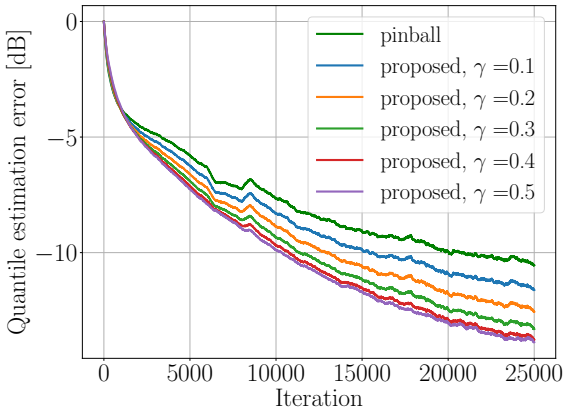


(c) coverage rate

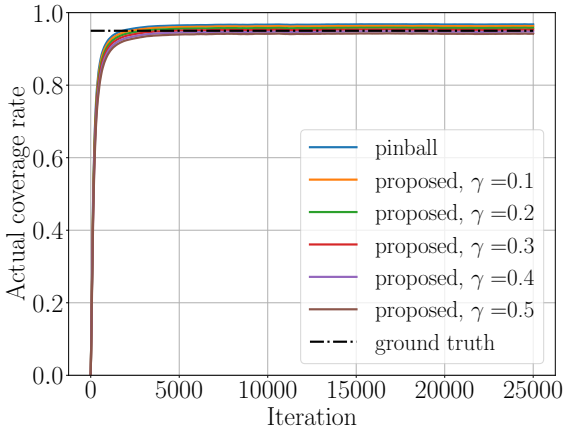
Fig. 1. Comparisons between the pinball loss and the proposed loss (Huberized pinball loss) for $\beta = 0.95$ and outlier rate 0.02.



(a) upper quantile errors



(b) lower quantile errors



(c) coverage rate

Fig. 2. Comparisons under noise-distribution-changes during 6000 – 6500 and 8000 – 8500 for $\beta = 0.95$ and outlier rate 0.02 (the single kernel version is used for the proposed method).

[9] I. Takeuchi, Q. Le, T. Sears, and A. Smola, "Nonparametric quantile estimation," 2006.

[10] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

[11] S. Zheng, "Gradient descent algorithms for quantile regression with smooth approximation," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 3, pp. 191–207, 2011.

[12] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.

[13] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[14] R. Koenker, *Quantile Regression*, Cambridge University Press, 2005.

[15] S. Ito, *Lebesgue Sekibun Nyumon*, Shokabo, 46th edition, 1963, in Japanese.

[16] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York: NY, 2nd edition, 2017.

[17] M. Yukawa, "Multikernel adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4672–4682, 2012.

[18] M. Yukawa, "Adaptive learning in Cartesian product of reproducing kernel Hilbert spaces," *IEEE Transactions on Signal Processing*, vol. 63, no. 22, pp. 6037–6048, 2015.