Chapter 24
Quantile Regression with Gaussian Kernels

Baobin Wang, Ting Hu, and Hong Yin

Abstract This paper aims at the error analysis of stochastic gradient descent (SGD) for quantile regression, which is associated with a sequence of varying ϵ-insensitive pinball loss functions and flexible Gaussian kernels. Analyzing sparsity and learning rates will be provided when the target function lies in some Sobolev spaces and a noise condition is satisfied for the underlying probability measure. Our results show that selecting the variance of the Gaussian kernel plays a crucial role in the learning performance of quantile regression algorithms.

Keywords Quantile regresion · Gaussian kernels · Reproducing kernel Hilbert spaces · Insensitive pinball loss · Learning rate

24.1 Introduction

Quantile regression has been investigated in machine learning and statistics, see [3, 4, 13–15] and references therein. Compared with the least squares regression, quantile regression provides more information about the conditional distributions of output variables such as stretching or compressing tails and multimodality [5, 6]. In the setting of learning problems, let X be a multivariate random variable with

B. Wang
School of Mathematics and Statistics, South-Central University for Nationalities,
Wuhan 430074, People's Republic of China
e-mail: wbb@scuec.edu.cn

T. Hu
School of Mathematics and Statistics, Wuhan University,
Wuhan 430072, People's Republic of China
e-mail: tinghu@whu.edu.cn

H. Yin (✉)
School of Mathematics, Renmin University of China,
Beijing 100872, People's Republic of China
e-mail: yinhong@ruc.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_24

values in a compact subset of $\mathbb{R}^n$ and $Y \subset \mathbb{R}$ be a real valued response variable. The purpose of quantile regression is to study the quantile regression functions from a sample of T observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^T$ drawn independently according to the identical distribution ρ on $Z = X \times Y$. With a quantile parameter $0 < \tau < 1$, a *quantile regression function* $f_{\rho,\tau} : X \to Y$ is defined by its value $f_{\rho,\tau}(x)$ to be a τ-quantile of the conditional distribution $\rho(\cdot|x)$ of ρ at $x \in X$, that is, a value $u \in Y$ satisfying

$$\rho\left(\{y \in Y, y \leq u\}|x\right) \geq \tau, \ \textit{and} \ \rho\left(\{y \in Y, y \geq u\}|x\right) \geq 1 - \tau.$$

Gaussian kernels are one of the most often used kernels in modern machine learning methods such as support vector machines (SVMs) [12, 15]. The Gaussian kernel with variance $\sigma > 0$ is the function on $X \times X$ defined by

$$K_\sigma(x, u) := \exp\left\{-\frac{|x-u|^2}{2\sigma^2}\right\}.$$

Let $\mathscr{H}_\sigma(X)$ be the RKHS [1] on X associated with the kernel K_σ and the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}_\sigma(X)}$. Its reproducing property takes the form

$$\langle K_\sigma(x, \cdot), f(\cdot)\rangle_{\mathscr{H}_\sigma(X)} = f(x), \forall\, x \in \mathscr{X}, \ f \in \mathscr{H}_\sigma(X). \tag{24.1}$$

Quantile regression has been studied by means of kernel-based regularization schemes in a vast literature, see [7, 11, 15]. Its associated loss function is the pinball loss ϕ_τ defined by

$$\phi_\tau(u) = \begin{cases} (1-\tau)u, & \text{if} \quad u \geq 0, \\ -\tau u, & \text{if} \quad u < 0, \end{cases}$$

and the regularization scheme takes the form

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathscr{H}_\sigma(X)} \frac{1}{T} \sum_{i=1}^{T} \phi_\tau(y_i - f(x_i)) + \frac{\lambda}{2} \|f\|^2_{\mathscr{H}_\sigma(X)}. \tag{24.2}$$

In this paper, SGD method (or called online learning) is taken to solve the scheme (24.2) for its low complexity and good practical performance. Inspired by the work in [15, 19], we consider the below SGD algorithm for quantile regression associated with a varying ϵ-insensitive pinball loss $\phi_\tau^\epsilon(u)$ with an insensitive parameter $\epsilon \geq 0$, given as

$$\phi_\tau^\epsilon(u) = \begin{cases} (1-\tau)(u-\epsilon), & \text{if} \quad u > \epsilon, \\ -\tau(u+\epsilon), & \text{if} \quad u < -\epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 24.1 The SGD algorithm for (24.2) is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t \left\{ \left(\phi_\tau^{\epsilon_t}\right)'_- (f_t(x_t) - y_t) K_\sigma(x_t, \cdot) + \lambda f_t \right\} \tag{24.3}$$

where $\{\eta_t\} > 0$ is the positive stepsize sequence, $\lambda = \lambda(T)$ is the regularization parameter, $\{\epsilon_t\} > 0$ is the varying insensitive parameters and $\left(\phi_\tau^{\epsilon_t}\right)'_-$ is the left (one-side) derivative of $\phi_\tau^{\epsilon_t}$.

This algorithm is a generalization for the pinball loss ϕ_τ with $\epsilon = 0$ and the ϵ-insensitive loss with $\tau = \frac{1}{2}$ (median). The initial form of quantile regression with a fixed insensitive parameter $\epsilon > 0$ was introduced by [11, 16], which aims at producing possible sparsity of support vectors for the median. Then this idea was developed to τ-quantile regression with any $0 < \tau < 1$ and the ϵ-insensitive pinball loss $\phi_\tau^\epsilon(u)$ was proposed in [15, 17]. In the previous work [2, 17], the corresponding mathematical analysis in the batch learning has been conducted when ϵ change with the sample size T and $\epsilon = \epsilon(T) \to 0$ as the sample size T goes to infinity.

Here the insensitive parameters $\{\epsilon_t\} > 0$ used in the algorithm (24.3) form a decreasing sequence converging to zero when the learning step t increases. In the work [15], Hu et al. derived the learning rate of (24.3) with flexible insensitive parameters $\{\epsilon_t\}$ under the suitable choices of the parameters (λ, η_t) for balancing the approximation and sparsity. Their convergence rate is closely related to the strong assumption on the approximation power of RKHS. Actually, for a Gaussian RKHS $\mathscr{H}_\sigma$ with the fixed variance $\sigma > 0$, the approximation error decays logarithmically with respect to the range of $\mathscr{H}_\sigma$, which has been proved in [8]. So, putting this decay into their analysis leads that the learning rate for quantile regression is rather slow, which is unaccepted in real applications. In simulations, the variance σ of $\mathscr{H}_\sigma$ usually serves as a tuned parameter for a good learning performance in training processes and can be chosen in a data-dependent way such as cross-validation. Since the variance of a Gaussian kernel reflexes the specific structure of RKHS induced by the Gaussian or other important features of learning problems such as the frequency of function components, choosing the variance σ of $\mathscr{H}_\sigma$ is related to the model selection problem, which adjusts the complexity or the capacity of learning problems according to the learning time or sample size. The selecting rule of σ has been studied in various learning settings [7, 12, 18], SVM, least squares, etc.

The goal of this paper is to study the convergence behavior of the algorithm (24.3) with flexible Gaussians and investigate the effects of parameters in keeping sparsity and nice learning power for quantile problems. Our results show that the online quantile regression is feasible in the framework of the Gaussian RKHS, in which the variance of Gaussian serves as a trade-off between the approximation ability and sparsity of the algorithm. We present a selection rule for the variance $\sigma = \sigma(T)$ to avoid over-fitting or under-fitting in the iteration process. The performance of the iterates $\{f_t\}$ is usually measured by the convergence in terms of the excess generalization error. In this work, under the noise condition, we can obtain the convergence result in Banach spaces, which implies that $\{f_t\}$ is closed to the target function $f_{\rho,\tau}$ in a strong sense.

24.2 Main Results and Effects of Parameters

For conceptual simplicity, we assume throughout this paper that the support of the conditional distribution $\rho(\cdot|x)$ is $[-1, 1]$ and our results below is applicable for the support $[-M, M]$ with any $M > 0$. Moreover, let the value of $f_{\rho,\tau}(x)$ be unique at each $x \in X$. To demonstrate our main result in the general case, we first shall give the following learning rate in the special case if the quantile regression function $f_{\rho,\tau}$ lies in some smooth functional space. Its regularity is usually measured in terms of Sobolev spaces. Recall the Sobolev space $H^r(\mathbb{R}^n)$ with index $r > 0$ consisting of all functions in $L^2(\mathbb{R}^n)$ with the semi-norm $|f|_{H^r(\mathbb{R}^n)} = \left\{(2\pi)^{-n} \int_{\mathbb{R}^n} |\xi|^{2r} |\widehat{f}(\xi)|^2\right\}^{\frac{1}{2}}$ finite where $\widehat{f}$ is the Fourier transform of f defined as $\widehat{f}(\xi) = \int_{\mathbb{R}^n} f(x)e^{-i\xi\cdot x}dx$. In the sequel, ρ_X denotes the marginal distribution of ρ on X and $\widehat{f}$ denotes the projection operation on any measurable function $f : X \to \mathbb{R}$, given as

$$\widehat{f}(x) = \begin{cases} 1, & \text{if} \quad f(x) \geq 1, \\ f(x), & \text{if} \quad -1 < f(x) < 1, \\ -1, & \text{if} \quad f(x) \leq -1. \end{cases}$$

Theorem 24.1 *Let $X \subset \mathbb{R}^n$ be a domain with Lipschitz boundary and ρ_X be the uniform distribution on X. Assume that $f_{\rho,\tau} \in H^r(X)$ for some $r > 0$, $\|f_{\rho,\tau}\|_\infty \leq 1$ and the conditional distributions $\{\rho(\cdot|x), x \in X\}$ have density functions given with $\zeta > 0$,*

$$\frac{d\rho(y|x)}{dy} = \begin{cases} \frac{\zeta+1}{2}|y - f_{\rho,\tau}(x)|^\zeta, & \textit{if} \quad |y - f_{\rho,\tau}(x)| \leq 1; \\ 0, & \textit{otherwise}. \end{cases}$$

Take $\eta_t = -\frac{n+3r}{2n+5r}$, $\lambda = T^{-\frac{n+r}{2n+5r}}$, $\sigma = T^{-\frac{1}{2n+5r}}$ and $\epsilon_t = t^{-\beta}$ with $\beta \geq \frac{1}{2}$ then

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|\widehat{f}_{T+1} - f_{\rho,\tau}\|_{L^2_{\rho_X}}\right] \leq C^* T^{-\frac{r}{(2n+5r)(\zeta+2)}} \tag{24.4}$$

where C^* is a constant independent of T, and will be given in the proof.

Remark 24.1 Notice that the larger the index r is, the faster the projected function f_{T+1} in (24.3) converges to $f_{\rho,\tau}$. In addition, the choice of parameters λ, σ, η_t is closely related to r. Thus, the regularity of the quantile function $f_{\rho,\tau}$ is important in the learning process. The index β of the insensitive parameter characterizes the sparsity and the learning rate will not be affected if $\beta \geq \frac{1}{2}$. As the index β increases, the value of the insensitive parameter ϵ_t will decrease at each iteration t. So, it is suitable to choose $\beta = \frac{1}{2}$ in this case. Here the variance σ of the Gaussian kernel K_σ changes with the learning time T. This is reasonable since a small σ will lead to over-fitting and a large σ to under-fitting. In the above example, we are considering the quantile regression problems on a domain of $\mathbb{R}^n$, so the learning rate is poor if the dimension n is large. However, in many situations, the input space X is a low-dimensional manifold embedded in the large-dimensional space $\mathbb{R}^n$. In such a situation, the learning rates may be greatly improved.

Now we are in a position of stating our main result in the general case. First, a noise condition on the measure ρ is given, which was introduced in [13].

Definition 24.2 Let $0 < p \leq \infty$ and $w > 0$. We say that ρ has a τ-quantile of p-average type w if there exist two functions b and a from X to $\mathbb{R}$ such that $\{ba^w\}^{-1} \in L^p_{\rho_X}$ and for any $x \in X$ and $q \in (0, a(x)]$, there hold

$$\rho(\{y : f_{\rho,\tau}(x) < y < f_{\rho,\tau}(x) + q\}|x) \geq b(x)q^w$$

and

$$\rho(\{y : f_{\rho,\tau}(x) - q < y < f_{\rho,\tau}(x)\}|x) \geq b(x)q^w. \tag{24.5}$$

This assumption can be satisfied with many common conditional distributions such as Guassian, students' t distributions and uniform distributions. In the following, we will give an example to illustrate it. More examples can be found in [2, 13].

Example 24.1 Let the conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$ be a family of Gaussian distributions with a uniform variance $\tilde{\sigma} > 0$, i.e. $\frac{d\rho(y|x)}{dy} = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left\{-\frac{(y-\mu_x)^2}{2\tilde{\sigma}^2}\right\}$ where $\{\mu_x\}_{x\in X}$ are expectations of the Gaussian distributions $\{\rho(\cdot|x)\}_{x\in X}$. It is direct to calculate that $f_{\rho,\tau}(x)$ can take the value of μ_x at each $x \in X$. We also find that for any $q \in (0, \tilde{\sigma}]$, there holds

$$\rho(\{y : f_{\rho,\tau}(x) < y < f_{\rho,\tau}(x) + q\}|x) = \rho(\{y : \mu_x < y < \mu_x + q\}|x)$$
$$= \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_{\mu_x}^{\mu_x+q} \exp\left\{-\frac{(y-\mu_x)^2}{2\tilde{\sigma}^2}\right\} dy = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_0^q \exp\left\{-\frac{y^2}{2\tilde{\sigma}^2}\right\} dy \geq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\tilde{\sigma}} q.$$

Similarly, we have that $\rho(\{y : f_{\rho,\tau}(x) - q < y < f_{\rho,\tau}(x)\}|x) \geq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\tilde{\sigma}} q$. Thus, the measure ρ has a ∞-average type 1.

In addition, we need a condition about the continuity of the conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$.

Definition 24.3 Let $s > 0$. We say that the family of conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$ is Lipschitz-s if there exists a constant C_ρ such that

$$\rho(\{y : u \leq y \leq v\}|x) \leq C_\rho |u - v|^s, \quad \forall u < v \in Y,\ x \in X. \tag{24.6}$$

With these preliminaries in place, we present the following learning rates whose proof will be provided in the next section.

Theorem 24.2 *Suppose that for some $r > 0$, the quantile regression function $f_{\rho,\tau}$ is the restriction of some $\tilde{f}_{\rho,\tau} \in H^r(\mathbb{R}^n) \bigcap L^\infty(\mathbb{R}^n)$ over X, and the density function $\frac{d\rho_X}{dx}$ lies in $L^2(X)$. Let the parameters $\eta_t, \epsilon_t, \lambda, \sigma$ be of the form*

$$\eta_t = t^{-\frac{n+3r}{2n+5r}}, \ \epsilon_t = t^{-\beta}, \ \lambda = T^{-\frac{n+r}{2n+5r}}, \ \sigma = T^{-\frac{1}{2n+5r}} \tag{24.7}$$

with $\beta \geq \max\left\{\frac{3(n+2r)}{s(2n+5r)} - 1, \frac{n+2r}{s(2n+5r)}\right\}$.

Denote $\mu := \frac{p(w+1)}{p+1}$. *If the measure* ρ *satisfies (24.5) and (24.6), then*

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|\widehat{f}_{T+1} - f_{\rho,\tau}\|_{L^{\mu}_{\rho_X}}\right] \leq C^* T^{-\frac{r}{(2n+5r)(w+1)}}. \tag{24.8}$$

Here the constant C^* *is independent of* T *and will be given in the proof.*

This theorem investigates the learning ability of the learned function $\widehat{f}_{T+1}$ that approximates the quantile regression function $f_{\rho,\tau}$ with suitable chosen parameters including the variance parameter σ and the insensitive parameters $\{\epsilon_t\}$. It shows how to adapt the variance σ in the learning process while keeping the sparsity and the learning power for the algorithm (24.3). It is also worth noticing that our leaning rate is given in a weighted L^{μ}-space by the noise condition (24.5). Our rate still holds for the generalization error (see Sect. 24.3) if the condition (24.5) is not imposed on ρ. At the end of this section, we would like to remark that the quantile regression problem considered here is fully nonparametric, so the parameters in (3) are usually unknown in advance and tuned in training processes according to various quantile regression problems. They can be chosen by a data-dependent way in training processes, e.g. cross-validation.

24.3 Error Analysis and Proofs of Main Results

In learning theory, the performance of learning algorithms is often measured by the generalization error. For the quantile regression, we define the *generalization error* for $f : X \to \mathbb{R}$ associated with the pinball loss ϕ_τ as

$$\mathscr{E}(f) = \int_Z \phi_\tau(f(x) - y)d\rho$$

and the quantile regression function $f_{\rho,\tau}$ is a minimizer of $\mathscr{E}(f)$. Meanwhile, we define the ϵ-insensitive generalization error $\mathscr{E}^{\epsilon}(f)$, given as $\mathscr{E}^{\epsilon}(f) := \int_Z \phi_\tau^{\epsilon}(f(x) - y)d\rho$. Our error analysis is conducted based on an error decomposition for the *excess generalization error* $\mathscr{E}(f_{T+1}) - \mathscr{E}(f_{\rho,\tau})$. To this end, we introduce the below approximation error with respect to the approximation ability of $\mathscr{H}_\sigma(X)$. In the sequel, we denote the norm $\|\cdot\|_{\mathscr{H}_\sigma(X)}$ by $\|\cdot\|_\sigma$ and $\mathscr{H}_\sigma(X)$ by $\mathscr{H}_\sigma$ for simplicity.

Definition 24.4 For any regularization parameter $\lambda > 0$, the approximation error $\mathscr{D}(\sigma, \lambda)$ of the triple (K_σ, ρ, τ) is defined by

$$\mathscr{D}(\sigma, \lambda) = \min_{f \in \mathscr{H}_\sigma} \left\{\mathscr{E}(f) - \mathscr{E}(f_{\rho,\tau}) + \frac{\lambda}{2}\|f\|_\sigma^2\right\}.$$

The regularization function is defined as

$$f_\lambda = \arg\min_{f\in\mathscr{H}_\sigma} \left\{\mathscr{E}(f) - \mathscr{E}(f_{\rho,\tau}) + \frac{\lambda}{2}\|f\|_\sigma^2\right\} \quad or \quad f_\lambda = \arg\min_{f\in\mathscr{H}_\sigma} \left\{\mathscr{E}(f) + \frac{\lambda}{2}\|f\|_\sigma^2\right\}. \tag{24.9}$$

Then its associated insensitive regularization function for any $\epsilon > 0$ is

$$f_\lambda^\epsilon = \arg\min_{f\in\mathscr{H}_\sigma} \left\{\mathscr{E}^\epsilon(f) + \frac{\lambda}{2}\|f\|_\sigma^2\right\}. \tag{24.10}$$

Now, the error decomposition for $\mathscr{E}(f_{T+1}) - \mathscr{E}(f_{\rho,\tau})$ can be displayed as

$$\mathscr{E}(f_{T+1}) - \mathscr{E}(f_{\rho,\tau}) = \left\{\mathscr{E}(f_{T+1}) - \mathscr{E}(f_\lambda)\right\} + \left\{\mathscr{E}(f_\lambda) - \mathscr{E}(f_{\rho,\tau})\right\} \le \left\{\mathscr{E}(f_{T+1}) - \mathscr{E}(f_\lambda)\right\} + \mathscr{D}(\sigma,\lambda). \tag{24.11}$$

Notice the Lipschitz continuity of ϕ_τ and the property of RKHS with $\|f\|_\infty \le \|f\|_\sigma,\ \forall f \in \mathscr{H}_\sigma$. It yields that $|\mathscr{E}(f_{T+1}) - \mathscr{E}(f_\lambda)| \le \|f_{T+1} - f_\lambda\|_\infty \le \|f_{T+1} - f_\lambda\|_\sigma$. So, the first term on the right-hand side of (24.11) will be handled in the sequel by means of the *sample error* $\|f_{T+1} - f_\lambda\|_\sigma$.

24.3.1 Approximation Error

For the second term $\mathscr{D}(\sigma,\lambda)$, it is associated with the approximation powers of the RKHSs induced by Gaussians with variance $\sigma > 0$. The following polynomial decay of $\mathscr{D}(\sigma,\lambda)$ under some Sobolev smoothness conditions on the function $f_{\rho,\tau}$ can be found in [18].

Lemma 24.1 *Suppose that for some $r > 0$, the quantile regression function $f_{\rho,\tau}$ is the restriction of some $\tilde{f}_{\rho,\tau} \in H^r(\mathbb{R}^n)\bigcap L^\infty(\mathbb{R}^n)$ over X, and the density function $\frac{d\rho_X}{dx}$ lies in $L^2(X)$. Then*

$$\mathscr{D}(\sigma,\lambda) \le C'(\sigma^r + \lambda\sigma^{-n}), \quad \forall\, 0 < \sigma < 1, \quad \lambda > 0, \tag{24.12}$$

where C' is a constant independent of σ, λ.

24.3.2 Insensitive Analysis

According to the above error analysis, we need to estimate $\|f_{t+1} - f_\lambda\|_\sigma$ by iteration on $t = 1, \cdots, T$. In the iteration procedure, the function f_{t+1} is generated by updating f_t according to the sample (x_t, y_t). Here, the technical difficulty lies in the change of the insensitive parameters ϵ_t. This can be handled by the following lemma in [15] for varying $\{\epsilon_t\}$.

Lemma 24.2 *Suppose that the family of conditional distributions* $\{\rho(\cdot|x)\}_{x\in X}$ *is Lipschitz-s satisfying (24.6). Then for any* $0 \le u < v$, *we have*

$$\|f_\lambda^u - f_\lambda^v\|_\sigma \le C\lambda^{-1}|u-v|^s. \tag{24.13}$$

If the insensitive parameters $\epsilon_t = \epsilon_1 t^{-\beta}$ *with* $\epsilon_1, \beta > 0$, *then*

$$\|f_\lambda^{\epsilon_{t-1}} - f_\lambda^{\epsilon_t}\|_\sigma \le C\lambda^{-1}t^{-(\beta+1)s}, \ \forall t \ge 2. \tag{24.14}$$

Here C *is a constant independent of* λ *and insensitive parameters.*

24.3.3 One Step-Iteration

Denote $h_t := \|f_\lambda^{\epsilon_{t-1}} - f_\lambda^{\epsilon_t}\|_\sigma$. We can get the one step iteration result as follows. To obtain optimal error bounds, we shall use the flexibility caused by some free parameters $0 < d < 2$ and $c_1 > 0$.

Lemma 24.3 *Define* $\{f_t\}$ *by (24.3). Let some constants* $0 < d < 2$ *and* $c_1 > 0$, *then*

$$\mathbb{E}_{z_t}\left(\|f_{t+1} - f_\lambda^{\epsilon_t}\|_\sigma^2\right) \le \left(1 + c_1 h_t^d - \lambda\eta_t\right)\|f_t - f_\lambda^{\epsilon_{t-1}}\|_\sigma^2 + h_t^{2-d}/c_1 + h_t^2 + 4\eta_t^2. \tag{24.15}$$

Proof First, we claim that $\|f_t\|_\sigma \le \frac{1}{\lambda}, \forall t \ge 2$. It can be easily derived from $f_1 = 0$ and the following induction by (24.3) that

$$\|f_{t+1}\|_\sigma \le (1-\lambda\eta_t)\|f_t\|_\sigma + \eta_t \le (1-\lambda\eta_t)\frac{1}{\lambda} + \eta_t = \frac{1}{\lambda}. \tag{24.16}$$

Denote $B_t := \left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t) - y_t)K_\sigma(x_t,\cdot) + \lambda f_t$. The online algorithm (24.3) can be written as $f_{t+1} = f_t - \eta_t B_t$. Then

$$\|f_{t+1} - f_\lambda^{\epsilon_t}\|_\sigma^2 = \|f_t - f_\lambda^{\epsilon_t}\|_\sigma^2 + \eta_t^2\|B_t\|_\sigma^2 - 2\eta_t\langle f_t - f_\lambda^{\epsilon_t}, B_t\rangle_\sigma \tag{24.17}$$

Applying the reproducing property (24.1) to part of the last term of (24.17), we have that

$$\langle f_t - f_\lambda^{\epsilon_t}, \left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t) - y_t)K_\sigma(x_t,\cdot)\rangle_\sigma = \left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t) - y_t)\left(f_t(x_t) - f_\lambda^{\epsilon_t}(x_t)\right).$$

The convexity of $\phi_\tau^{\epsilon_t}$ implies that

$$\left(\phi_\tau^{\epsilon_t}\right)'_-\left(f_t(x_t)-y_t\right)\left(f_t(x_t)-f_\lambda^{\epsilon_t}(x_t)\right)\geq\phi_\tau^{\epsilon_t}(f_t(x_t)-y_t)-\phi_\tau^{\epsilon_t}(f_\lambda^{\epsilon_t}(x_t)-y_t).$$

For the other part of the last term of (24.17), we have that

$$\langle f_t-f_\lambda^{\epsilon_t},f_t\rangle_\sigma\geq\|f_t\|_\sigma^2-\frac{1}{2}\|f_t\|_\sigma^2-\frac{1}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2=\frac{1}{2}\|f_t\|_\sigma^2-\frac{1}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2.$$

Thus, the last term of (24.17) can be bounded as

$$\langle f_t-f_\lambda^{\epsilon_t},B_t\rangle_\sigma\geq\left[\phi_\tau^{\epsilon_t}(f_t(x_t)-y_t)+\frac{\lambda}{2}\|f_t\|_\sigma^2\right]-\left[\phi_\tau^{\epsilon_t}(f_\lambda^{\epsilon_t}(x_t)-y_t)\frac{\lambda}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2\right].$$

Since f_t only depends on $z_1,\cdots,z_{t-1}$, then

$$\mathbb{E}_{z_t}\langle f_t-f_\lambda^{\epsilon_t},B_t\rangle_\sigma\geq\left[\mathscr{E}(f_t)+\frac{\lambda}{2}\|f_t\|_\sigma^2\right]-\left[\mathscr{E}(f_\lambda^{\epsilon_t})+\frac{\lambda}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2\right]$$

This together with Theorem 2 in [19], implies that $\mathbb{E}_{z_t}\langle f_t-f_\lambda^{\epsilon_t},B_t\rangle_\sigma\geq\frac{\lambda}{2}\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2$. Putting it into (24.17), then

$$\mathbb{E}_{z_t}\left(\|f_{t+1}-f_\lambda^{\epsilon_t}\|_\sigma^2\right)\leq(1-\lambda\eta_t)\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2+\eta_t^2\mathbb{E}_{z_t}\|B_t\|_\sigma^2. \tag{24.18}$$

Now we estimate $\|f_t-f_\lambda^{\epsilon_t}\|_\sigma$. It is decomposed as

$$\|f_t-f_\lambda^{\epsilon_t}\|_\sigma=\|f_t-f_\lambda^{\epsilon_{t-1}}+f_\lambda^{\epsilon_{t-1}}-f_\lambda^{\epsilon_t}\|_\sigma\leq\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma+h_t.$$

Applying the elementary inequality $2xy\leq c_1x^2y^d+y^{2-d}/c_1$ with any $0<d<2$ and $c_1>0$, to $x=\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma$ and $y=h_t$, then

$$\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2\leq\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma^2+2\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma h_t+h_t^2\leq(1+c_1h_t^d)\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma^2+h_t^{2-d}/c_1+h_t^2.$$

Plugging it into (24.18) and noticing that $(1-\lambda\eta_t)(1+c_1h_t^d)\leq 1+c_1h_t^d-\lambda\eta_t$, we get

$$\mathbb{E}_{z_t}\left(\|f_{t+1}-f_\lambda^{\epsilon_t}\|_\sigma^2\right)\leq(1+c_1h_t^d-\lambda\eta_t)\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2+h_t^{2-d}/c_1+h_t^2+\eta_t^2\mathbb{E}_{z_t}\|B_t\|_\sigma^2.$$

We now only need to estimate $\|B_t\|_\sigma^2$. Note that $\|\left(\phi_\tau^{\epsilon_t}\right)'_-\|_\infty\leq 1$ and the bound (24.16)holds for the learning sequence $\{f_t\}$. Using the reproducing property $\|K_\sigma(x_t,\cdot)\|_\sigma^2=\langle K_\sigma(x_t,\cdot),K_\sigma(x_t,\cdot)\rangle_\sigma=K_\sigma(x_t,x_t)=1$, then

$$\|B_t\|_\sigma\leq\left\|\left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t)-y_t)K_\sigma(x_t-\cdot)\right\|_\sigma+\lambda\|f_t\|_\sigma\leq\|\left(\phi_\tau^{\epsilon_t}\right)'_-\|_\infty\|K_\sigma(x_t,\cdot)\|_\sigma++\lambda\|f_t\|_\sigma\leq 2.$$

Based on the above analysis, we can get the desired conclusion (24.15).

24.3.4 Sample Error Estimate

We are in a position to present the estimate of the sample error $\|f_{T+1} - f_\lambda\|_\sigma$, which is the key analysis in our study. For simplicity, denote $\prod_{j=T+1}^{T}\left(1 - \frac{1}{2}\lambda\eta_j\right) := 1$, $\sum_{j=T+1}^{T}\lambda\eta_j := 0$ and $f_\lambda^0 := f_\lambda$.

Lemma 24.4 *Let the parameters $\eta_t, \epsilon_t, \lambda$ be of the form as $\eta_t = \eta_1 t^{-\alpha}$, $\epsilon_t = \epsilon_1 t^{-\beta}$ and $\lambda = T^{-(1-\alpha-\varepsilon)}$ for any $1 - 2\alpha < \varepsilon < 1 - \alpha$, $\eta_1 > 0, \epsilon_1 > 0$ satisfying*

$$\max\{1 - \beta s - \varepsilon, 2 - (\beta + 1)s - 2\varepsilon\} < \alpha < \min\{2(\beta + 1)s, 1\}. \tag{24.19}$$

Then we have

$$\mathbb{E}_{z_1,\cdots,z_T}\left(\|f_{T+1} - f_\lambda\|_\sigma\right) \leq \tilde{C}T^{-\min\left\{(\beta+1)s+\alpha-2+2\varepsilon,\alpha-\frac{1}{2}+\frac{\varepsilon}{2},\beta s-1+\alpha+\varepsilon\right\}} + \sqrt{\frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}}\exp\left\{-\frac{\lambda\eta_1}{8(1-\alpha)}(T+1)^{1-\alpha}\right\} \tag{24.20}$$

where $\tilde{C}$ is a constant independent of T, given in the proof.

Proof We split $\|f_{T+1} - f_\lambda\|_\sigma$ into two parts as $\|f_{T+1} - f_\lambda^{\epsilon_T}\|_\sigma$ and $\|f_\lambda - f_\lambda^{\epsilon_T}\|_\sigma$. For the first term $\|f_{T+1} - f_\lambda^{\epsilon_T}\|_\sigma$, we shall apply the conclusion in Lemma 24.3. By (24.14), $h_t \leq C\lambda^{-1}t^{-(\beta+1)s}$. We take $d = \frac{\alpha}{(\beta+1)s}$ and $c_1 = \frac{1}{2}\eta_1 C^{-d}T^{-(d+1)(1-\alpha-\varepsilon)}$ for any $1 - 2\alpha < \varepsilon < 1 - \alpha$. The restriction (24.19) of parameters implies that $c_1 h_t^d \leq \frac{1}{2}\lambda\eta_t$ and $1 + c_1 h_t^d - \lambda\eta_t \leq 1 - \frac{1}{2}\lambda\eta_t$. With (24.15), it yields that

$$\mathbb{E}_{z_t}\left(\|f_{t+1} - f_\lambda^{\epsilon_t}\|_\sigma^2\right) \leq \left(1 - \frac{1}{2}\lambda\eta_t\right)\|f_t - f_\lambda^{\epsilon_{t-1}}\|_\sigma^2 + 2h_t^{2-d}/c_1 + 4\eta_t^2.$$

Applying the relation above iteratively for $t = t_0, \cdots, T$, we obtain that

$$\begin{aligned}
&\mathbb{E}_{z_1,\cdots,z_T}\left(\|f_{T+1} - f_\lambda^{\epsilon_T}\|_\sigma^2\right) \leq \left(1 - \frac{1}{2}\lambda\eta_T\right)\left(1 - \lambda\eta_{T-1}\right)\mathbb{E}_{z_1,\cdots,z_{T-1}}\left(\|f_T - f_\lambda^{\epsilon_{T-1}}\|_\sigma^2\right) \\
&+ 2h_T^{2-d}/c_1 + 4\eta_T^2 + \left(1 - \frac{1}{2}\lambda\eta_T\right)\left(2h_{T-1}^{2-d}/c_1 + 4\eta_{T-1}^2\right) \\
&= \prod_{t=t_0}^{T}\left(1 - \frac{1}{2}\lambda\eta_t\right)\mathbb{E}_{z_1,\cdots,z_{t_0-1}}\left(\|f_{t_0} - f_\lambda^{t_0-1}\|_\sigma^2\right) + \sum_{t=t_0}^{T}\left(2h_t^{2-d}/c_1 + 4\eta_t^2\right)\prod_{j=t+1}^{T}\left(1 - \frac{1}{2}\lambda\eta_j\right).
\end{aligned}$$

Using the above inequality with $t_0 = 1$ and noting that $\|f_\lambda\|_\sigma^2 \leq 2\mathscr{D}(\sigma,\lambda)/\lambda$, with the elementary inequality $1 - x \leq e^{-x}$ for any $x > 0$, then we have

$$\begin{aligned}
&\mathbb{E}_{z_1,\cdots,z_T}\left(\|f_{T+1}-f_\lambda^{\epsilon T}\|_\sigma^2\right) \le \exp\left\{-\frac{\lambda}{2}\sum_{t=1}^{T}\eta_t\right\}\|f_\lambda\|_\sigma^2 + \sum_{t=1}^{T}\left(2h_t^{2-d}/c_1+4\eta_t^2\right)\exp\left\{-\frac{\lambda}{2}\sum_{j=t+1}^{T}\eta_t\right\}\\
&\le 2\exp\left\{-\frac{\lambda}{2}\sum_{t=1}^{T}\eta_t\right\}\mathscr{D}(\sigma,\lambda)/\lambda + \sum_{t=1}^{T}\left(2h_t^{2-d}/c_1+4\eta_t^2\right)\exp\left\{-\frac{\lambda}{2}\sum_{j=t+1}^{T}\eta_t\right\}\\
&= 2\exp\left\{-\frac{\lambda\eta_1}{2}\sum_{t=1}^{T}t^{-\alpha}\right\}\mathscr{D}(\sigma,\lambda)/\lambda + \sum_{t=1}^{T}\left(\frac{2C^{2-d}}{c_1\lambda^{2-d}}t^{-(2-d)(\beta+1)s}+4\eta_1^2t^{-2\alpha}\right)\exp\left\{-\frac{\lambda\eta_1}{2}\sum_{j=t+1}^{T}t^{-\alpha}\right\}\\
&:= I_1 + I_2.
\end{aligned}$$

For I_1, using the elementary inequality in Lemma 4 of [19], that for any $0<\alpha<1$, there holds $\sum_{t=1}^{T}t^{-\alpha}\ge\frac{(T+1)^{1-\alpha}-1}{1-\alpha}$, we have

$$I_1 \le \frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}\exp\left\{-\frac{\lambda\eta_1}{2(1-\alpha)}\left((T+1)^{1-\alpha}-1\right)\right\} \le \frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}\exp\left\{-\frac{\lambda\eta_1}{4(1-\alpha)}(T+1)^{1-\alpha}\right\}.$$

For I_2, we apply the following elementary inequality valid for $t\in\mathbb{N}$, $0<q_1<1$ and $c, q_2>0:$

$$\sum_{i=1}^{t-1}i^{-q_2}\exp\left\{-c\sum_{j=i+1}^{t}j^{-q_1}\right\} \le \frac{2^{q_1+q_2}}{c}t^{q_1-q_2}+\frac{t}{2}\exp\left\{-\frac{c(1-2^{q_1-1})}{1-q_1}(t+1)^{1-q_1}\right\}. \tag{24.21}$$

It can be derived in the proof procedure of Lemma 2 (b) of [9]. Here we omit it for simplicity.
Take $q_1=\alpha$, $q_2=(2-d)(\beta+1)s$ and $c=\frac{\lambda\eta_1}{2}$. Then the first part of I_2 is bounded as

$$\begin{aligned}
I_{21} &:= \sum_{t=1}^{T}\left(\frac{2C^{2-d}}{c_1\lambda^{2-d}}t^{-(2-d)(\beta+1)s}\right)\exp\left\{-\frac{\lambda\eta_1}{2}\sum_{j=t+1}^{T}t^{-\alpha}\right\} \le 2C^{2-d}\left[\frac{2^{(2-d)(\beta+1)s+\alpha+1}}{\eta_1c_1\lambda^{3-d}}T^{-(2-d)(\beta+1)s+\alpha}\right.\\
&\left.+\frac{T}{2c_1\lambda^{2-d}}\exp\left\{-\frac{\eta_1(1-2^{\alpha-1})\lambda}{2(1-\alpha)}(T+1)^{1-\alpha}\right\}+\frac{T^{-(2-d)(\beta+1)s}}{c_1\lambda^{2-d}}\right].
\end{aligned}$$

Note that $\lambda=T^{-(1-\alpha-\varepsilon)}$ implies that there exists a constant C_ε independent of T such that the middle term $\frac{T}{2c_1\lambda^{2-d}}\exp\left\{-\frac{\eta_1(1-2^{\alpha-1})\lambda}{2(1-\alpha)}(T+1)^{1-\alpha}\right\}\le C_\varepsilon T^{-(2(\beta+1)s+2\alpha-4+4\varepsilon)}$. Together with the choice of d, c_1, we have that

$$I_{21}\le A_1T^{-(2(\beta+1)s+2\alpha-4+4\varepsilon)}$$

where $A_1:=2C^{2-d}\left(\frac{2^{(2-d)(\beta+1)s+\alpha+2}}{\eta_1^2}C^d+C_\varepsilon+\frac{2C^d}{\eta_1}\right)$.

For the second part of I_2, by similarity, applying (24.21) with $q_1=\alpha$, $q_2=2\alpha$ and $c=\frac{\lambda\eta_1}{2}$, we have that

$$I_{22} := \sum_{t=1}^{T} 4\eta_1^2 t^{-2\alpha} \exp\left\{-\frac{\lambda\eta_1}{2}\sum_{j=t+1}^{T} t^{-\alpha}\right\} \leq A_2 T^{-2\alpha+1-\varepsilon}$$

where $A_2 := 4\eta_1^2\left(\frac{2^{3\alpha+1}}{\eta_1} + C_\varepsilon + 1\right)$. Based on the above analysis, we see that

$$\begin{aligned}&\mathbb{E}_{z_1,\cdots,z_T}\left(\| f_{T+1} - f_\lambda^{\epsilon_T}\|_\sigma^2\right)\\ \leq &\frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}\exp\left\{-\frac{\lambda\eta_1}{4(1-\alpha)}(T+1)^{1-\alpha}\right\} + (A_1 + A_2)T^{-\min\{2(\beta+1)s+2\alpha-4+4\varepsilon, 2\alpha-1+\varepsilon\}}.\end{aligned}$$

For the term $\| f_\lambda - f_\lambda^{\epsilon_T}\|_\sigma$, by (24.13), it can be bounded as $\| f_\lambda - f_\lambda^{\epsilon_T}\|_\sigma \leq C\epsilon_1^s T^{-s\beta}\lambda^{-1} = C\epsilon_1^s T^{-s\beta+1-\alpha-\varepsilon}$. Then we can get the conclusion (24.20) with $\tilde{C} = \sqrt{A_1 + A_2} + C\epsilon_1^s$.

24.3.5 Bounding the Total Error

In our analysis we shall make use of the following comparison theorem [2, 13]. Recall that $\mu := \frac{p(w+1)}{p+1}$.

Lemma 24.5 *Suppose that the measure ρ has a p-average type w satisfying (24.5). Then for any measurable function $f : X \to [-1, 1]$, we have*

$$\| f - f_{\rho,\tau}\|_{L^\mu_{\rho_X}} \leq C_\mu\big(\mathscr{E}(f) - \mathscr{E}(f_{\rho,\tau})\big)^{\frac{1}{w+1}} \tag{24.22}$$

where the constant $C_\mu = 2(w+1)^{\frac{1}{w+1}}\|(ba^w)^{-1}\|^{\frac{1}{w+1}}_{L^p_{\rho_X}}$.

Now we can present the proof of our error estimate for the convergence of online algorithm (24.3) in a general form.

Proof of Theorem 24.2 Putting the explicit form (24.7) of $\eta_t, \epsilon_t, \lambda, \sigma$ into (24.20), we know that there exists a constant C'_ϵ independent of T or τ such that

$$\sqrt{\frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}}\exp\left\{-\frac{\lambda\eta_1}{8(1-\alpha)}(T+1)^{1-\alpha}\right\} \leq C'_\epsilon T^{-\frac{r}{2n+5r}}$$

and

$$\min\left\{(\beta+1)s + \alpha - 2 + 2\varepsilon, \alpha - \frac{1}{2} + \frac{\varepsilon}{2}, \beta s - 1 + \alpha + \varepsilon\right\} = \frac{r}{2n+5r}.$$

This yields that

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|f_{T+1}-f_\lambda\|_\sigma\right] \le \left(\tilde{C}+C'_\epsilon\right)T^{-\frac{r}{2n+5r}}.$$

By (24.11), we know that

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\mathscr{E}(f_{T+1})-\mathscr{E}(f_{\rho,\tau})\right] \le \mathbb{E}_{z_1,\cdots,z_t}\left[\|f_{T+1}-f_\lambda\|_\sigma\right]+\mathscr{D}(\sigma,\lambda) \le \left(\tilde{C}+C'_\epsilon+2C'\right)T^{-\frac{r}{2n+5r}}.$$

Since the support of $\rho(\cdot|x)$ is $[-1, 1]$, we have that $\phi_\tau(\widehat{f}(x)-y) \le \phi_\tau(f(x)-y)$ for any measurable function $f : X \to \mathbb{R}$. It yields that $\mathscr{E}(\widehat{f}_{T+1}) \le \mathscr{E}(f_{T+1})$ and

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\mathscr{E}(\widehat{f}_{T+1})-\mathscr{E}(f_{\rho,\tau})\right] \le \left(\tilde{C}+C'_\epsilon+2C'\right)T^{-\frac{r}{2n+5r}}.$$

Using the relation (24.22), we can complete the proof of Theorem 24.2 with $C^* = \left(\tilde{C}+C'_\epsilon+2C'\right)^{\frac{1}{w+1}}C_\mu$.

Proof of Theorem 24.1 We shall prove Theorem 24.1 by Theorem 24.2. Since X has a Lipschitz boundary, we know from [10] that there exists an extension function $\tilde{f}_{\rho,\tau} \in H^r(\mathbb{R}^n)$ such that $\tilde{f}_{\rho,\tau}|_X = f_{\rho,\tau}$. Next, we check the noise condition (24.5). Let the function $a(x) = 1$ and $b(x) = \frac{1}{2}$, we have that for any $q \in [0, 1]$

$$\rho(\{y : f_{\rho,\tau}(x) \le y \le f_{\rho,\tau}(x)+q\}|x) = \int_{f_{\rho,\tau}(x)}^{f_{\rho,\tau}(x)+q}\frac{d\rho(y|x)}{dy}dy = \frac{1}{2}q^{\zeta+1}.$$

By similarity, we have $\rho(\{y : f_{\rho,\tau}(x)-q \le y \le f_{\rho,\tau}(x)\}|x) = \frac{1}{2}q^{\zeta+1}$. Therefore, the measure ρ has a τ-quantile of ∞-average type $\zeta+1$. Meanwhile, we find that the family of conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$ is Lipschitz-1 and (24.6) is satisfied with $C_\rho = \frac{\zeta+1}{2}$ and $s = 1$ since the density function $\frac{d\rho(y|x)}{dy}$ is uniformly bounded by $\frac{\zeta+1}{2}$. Thus, we can apply (24.8) to get that

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|f_{T+1}-f_{\rho,\tau}\|_{L^2_{\rho_X}}\right] \le \mathbb{E}_{z_1,\cdots,z_T}\left[\|f_{T+1}-f_{\rho,\tau}\|_{L^{\zeta+2}_{\rho_X}}\right] \le C^*T^{-\frac{r}{(2n+5r)(\zeta+2)}}.$$

Then the proof is completed.

Acknowledgments The work described in this paper is partially supported by National Natural Science Foundation of China [Nos. 11671307 and 11571078], Natural Science Foundation of Hubei Province in China [No. 2017CFB523] and the Special Fund for Basic Scientific Research of Central Colleges, South-Central University for Nationalities [No. CZY18033].

References

1. Aronszajn, N.: Theory of reproducing kernels. Tran. Am. Math. Soc. **68**(3), 337–404 (1950)
2. Hu, T., Yuan, Y.: Learning rates of regression with q-norm loss and threshold. Anal. Appl. **14**(06), 809–827 (2016)
3. Hwang, c., Shim, J.: A simple quantile regression via support vector machine. In: International Conference on Natural Computation, Springer, pp. 512–520 (2005)
4. Koenker, R., Geling, O.: Reappraising medfly longevity: a quantile regression survival analysis. J. Am. Stat. Assoc. **96**(454), 458–468 (2001)
5. Koenker, R.: Quantile Regression. Cambridge University Press, New York (2005)
6. Rosset, S.: Bi-level path following for cross validated solution of kernel quantile regression. J. Mach. Learn. Res. **10**(11), 2473–2505 (2009)
7. Shi, L., Huang, X., Tian, Z., Suykens, J.A.: Quantile regression with $l1$-regularization and Gaussian kernels. Adv. Comput. Math. **40**(2), 517–551 (2014)
8. Smale, S., Zhou, D.X.: Estimating the approximation error in learning theory. Anal. Appl. **1**(01), 17–41 (2003)
9. Smale, S., Zhou, D.X.: Online Learning with Markov Sampling. Anal. Appl. **7**(01), 87–113 (2009)
10. Stein, E.M.: Singular Integrals and Differentiability Properties of Functions. In Bulletin of the London Mathematical Society (1973)
11. Steinwart, I., Christman, A.: Sparsity of SVMs that use the ϵ-insensitive loss. In: Advances in Neural Information Processing Systems, pp. 1569–1576 2008)
12. Steinwart, I., Scovel, C., et al.: Fast rates for support vector machines using Gaussian kernels. Ann. Stat. **35**(2), 575–607 (2007)
13. Steinwart, I., Christmann, A., et al.: Estimating conditional quantiles with the help of the pinball loss. Bernoulli. **17**(1), 211–225 (2011)
14. Takeuchi, I., Le, Q.V., Sears, T.D., Smola, A.J.: Nonparametric quantile estimation. J. Mach. Learn. Res. **7**, 1231–1264 (2006)
15. Ting, H., Xiang, D.H., Zhou, D.X.: Online learning for quantile regression and support vector regression. J. Stat. Plan. Inference **142**(12), 3107–3122 (2012)
16. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)
17. Xiang, D.H., Hu, T., Zhou, D.X.: Approximation analysis of learning algorithms for support vector regression and quantile regression. J. Appl. Math. (2012). https://doi.org/10.1155/2012/902139
18. Xiang, D.H., Zhou, D.X.: Classification with Gaussians and Convex Loss. J. Mach. Learn. Res. **10**(10), 1447–1468 (2009)
19. Ying, Y., Zhou, D.X.: Online regularized classification algorithms. IEEE Trans. Inf. Theory. **52**(11), 4775–4788 (2006)