Journal of Statistical Planning and Inference 142 (2012) 3107–3122

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Online learning for quantile regression and support vector regression

Ting Hu [a], Dao-Hong Xiang [b], Ding-Xuan Zhou [c,*]

[a] School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China
[b] Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang 321004, China
[c] Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 16 August 2011
Received in revised form
18 May 2012
Accepted 13 June 2012
Available online 21 June 2012

Keywords:
Quantile regression
Support vector regression
Insensitive pinball loss
Online learning
Reproducing kernel Hilbert space
Error analysis

ABSTRACT

We consider for quantile regression and support vector regression a kernel-based online learning algorithm associated with a sequence of insensitive pinball loss functions. Our error analysis and derived learning rates show quantitatively that the statistical performance of the learning algorithm may vary with the quantile parameter τ. In our analysis we overcome the technical difficulty caused by the varying insensitive parameter introduced with a motivation of sparsity.

1. Introduction

We consider online learning algorithms associated with a sequence of ϵ-insensitive pinball loss functions for the purposes of *quantile regression* (involving a pinball loss) and *support vector regression* (involving an ϵ-insensitive loss).

Quantile regression extends the classical least squares regression (Koenker and Bassett, 1978) and provides richer information about the distributions of response variables such as stretching or compressing tails and multimodality (Koenker, 2005). As a learning problem, it aims at learning from samples quantile regression functions on a complete separable metric space X (the input space). Let $Y=\mathbb{R}$ be the output space and ρ be a Borel probability measure on $Z := X \times Y$. With a quantile parameter $0<\tau<1$, a *quantile regression function* $f_{\rho,\tau}$ in the quantile regression setting is defined by its value $f_{\rho,\tau}(x)$ to be a τ-quantile of the conditional distribution ρ_x of ρ at $x \in X$. Here a τ-quantile of ρ_x means a value $u \in Y$ satisfying

$$\rho_x(\{y \in Y : y \le u\}) \ge \tau \quad \text{and} \quad \rho_x(\{y \in Y : y \ge u\}) \ge 1-\tau. \tag{1.1}$$

Quantile regression has been investigated in a learning theory literature (e.g. Hwang and Shim, 2005; Takeuchi et al., 2006; Rosset, 2009; Steinwart and Christmann, 2011, 2008; Xiang, in press) by means of regularization schemes in reproducing kernel Hilbert spaces. Here with a continuous, symmetric and positive semidefinite function $K : X \times X \to \mathbb{R}$ (called a Mercer kernel), the *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ is defined as the completion of the span of $\{K_x = K(x,\cdot) : x \in X\}$ with the inner product $\langle \cdot,\cdot \rangle_K$ satisfying $\langle K_x,K_u \rangle_K = K(x,u)$. A regularization scheme in learning theory is often associated with a convex loss function $\psi : \mathbb{R} \to \mathbb{R}_+$. For quantile regression, the loss function is the pinball

[*] Corresponding author. Tel.: +852 3442 9708; fax: +852 3442 0249.
E-mail addresses: tinghu@whu.edu.cn (T. Hu), daohongxiang@gmail.com (D.-H. Xiang), mazhou@cityu.edu.hk (D.-X. Zhou).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.06.010

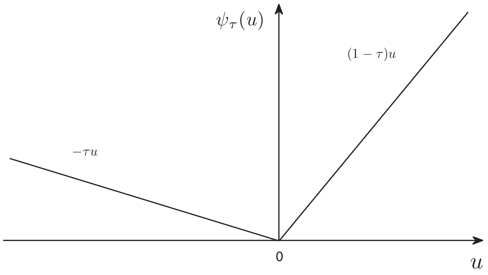

Fig. 1

loss $\psi=\psi_\tau:\mathbb{R}\to\mathbb{R}_+$ shown in Fig. 1 defined (Schoelkopf et al., 2000; Steinwart and Christmann, 2008) by

$$\psi_\tau(u)=\begin{cases}(1-\tau)u & \text{if } u>0,\\ -\tau u & \text{if } u\le 0,\end{cases} \tag{1.2}$$

and the regularization scheme takes the form

$$f_{\mathbf{z},\lambda}=\arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{T}\sum_{t=1}^{T}\psi(f(x_t)-y_t)+\lambda\|f\|_K^2\right\}. \tag{1.3}$$

Here $\lambda>0$ is a regularization parameter, and $\mathbf{z}=\{(x_i,y_i)\}_{i=1}^T\in Z^T$ is a sample for learning which is assumed throughout the paper to be independently drawn according to ρ.

Support vector regression is a classical learning algorithm introduced in Vapnik (1998). It is a regularization scheme (1.3) associated with an ϵ-insensitive *loss* $\psi(u)=\psi^{(\epsilon)}(u)=\max\{|u|-\epsilon,0\}$ to produce possible sparsity of support vectors (Vapnik, 1998; Steinwart and Christmann, 2009). When $\epsilon>0$ is fixed, convergence of the support vector regression algorithm was analyzed in Tong et al. (2009). Notice from the original motivation in Vapnik (1998) (for balancing the approximation and sparsity) that the insensitive parameter ϵ should change with the sample size and usually $\epsilon=\epsilon(m)\to 0$ as the sample size m increases. Mathematical analysis for this original algorithm is still open.

One advantage of regularization in RKHSs (1.3) is the reduction of optimization in a possibly infinite dimensional space $\mathcal{H}_K$ (in order to have powerful approximation ability) to that in a finite dimensional subspace span $\{K_{x_t}\}_{t=1}^T$. The computational complexity of the corresponding optimization problem might still be too high when the sample size T becomes large. Online learning algorithms can be used to reduce the computational complexity. They will be considered in this paper for both quantile regression and support vector regression.

An initial form of online algorithms was introduced in Kiefer and Wolfowitz (1952). It is a stochastic gradient descent method with another advantage of adaptivity to the situation when the samples are not available at the same time. Online learning algorithms associated with regularization schemes in RKHSs were considered for classification in Kivinen et al. (2004) with a hinge loss, for regression in Kivinen et al. (2004) and Smale and Yao (2006) with the least squares loss, and for classification with a general convex loss in Ying and Zhou (2006) which was extended to a non-iid setting in Smale and Zhou (2009), Hu and Zhou (2009), and Hu (2011). The online learning algorithm for quantile regression and support vector regression studied in this paper takes the following form with a sequence of convex loss functions $\{\psi^{(t)}:\mathbb{R}\to\mathbb{R}_+\}$.

Definition 1. The online algorithm for quantile regression is defined by $f_1=0$ and

$$f_{t+1}=f_t-\eta_t\{(\psi^{(t)})'_-(f_t(x_t)-y_t)K_{x_t}+\lambda_t f_t\},\quad t=1,2,\ldots, \tag{1.4}$$

where $\lambda_t>0$ is a regularization parameter, $\eta_t>0$ is a step size and $(\psi^{(t)})'_-$ is the left (one-side) derivative of $\psi^{(t)}$.

To deal with both quantile regression and support vector regression, we consider the above online algorithm associated with a varying ϵ-insensitive *pinball loss* $\psi_\tau^\epsilon:\mathbb{R}\to\mathbb{R}_+$ with an insensitive parameter $\epsilon\ge 0$ shown in Fig. 2 as

$$\psi_\tau^\epsilon(u)=\begin{cases}(1-\tau)(u-\epsilon) & \text{if } u>\epsilon,\\ -\tau(u+\epsilon) & \text{if } u\le-\epsilon,\\ 0 & \text{otherwise}.\end{cases} \tag{1.5}$$

This loss function is a generalization of the pinball loss (with $\epsilon=0$) and the ϵ-insensitive loss (with $\tau=\frac{1}{2}$) after scaling. The insensitive parameters $\{\epsilon_t\}_t$ in the sequence of flexible ϵ_t-insensitive pinball loss functions $\{\psi_\tau^{\epsilon_t}\}$ used in online algorithm (1.4) form a decreasing sequence converging to zero when the learning step t increases. So after putting the explicit formula (1.5) into (1.4), the learning sequence $\{f_t\}$ with the flexible pinball loss functions $\{\psi_\tau^{\epsilon_t}\}$ can be expressed as

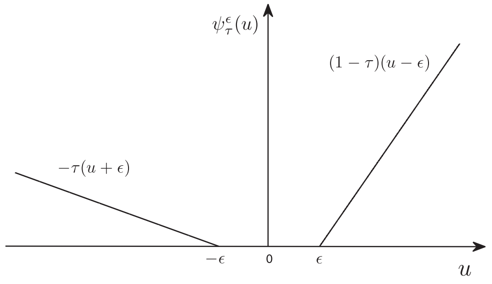

Fig. 2

$f_1=0$ and

$$f_{t+1}=\begin{cases}(1-\lambda_t\eta_t)f_t-(1-\tau)\eta_t K_{x_t} & \text{if } f_t(x_t)-y_t>\epsilon_t,\\ (1-\lambda_t\eta_t)f_t+\tau\eta_t K_{x_t} & \text{if } f_t(x_t)-y_t\le-\epsilon_t,\\ (1-\lambda_t\eta_t)f_t & \text{if } -\epsilon_t<f_t(x_t)-y_t\le\epsilon_t.\end{cases} \tag{1.6}$$

This paper aims at establishing error bounds for the convergence of the above online algorithm. We shall show in the next section that different quantile parameter τ may correspond to different learning rates. In particular, the learning rates given in Example 1 below take the form $\mathbb{E}_{z_1,\ldots,z_T}\|f_{T+1}-f_{\rho,\tau}\|_{L^2_{\rho_X}}=O(T^{-1/10})$ for $\tau\neq\frac{1}{2}$ and $O(T^{-1/5(2+\zeta)})$ for $\tau=\frac{1}{2}$ with ζ being arbitrarily large. Note that when $\tau=\frac{1}{2}$, the quantile regression function $f_{\rho,1/2}$ becomes the median function taking the median function value $f_{\rho,1/2}(x)$ of the conditional distribution ρ_x of ρ at x.

The main difference between our analysis and the previous work on online learning (e.g. Smale and Yao, 2006; Ying and Zhou, 2006; Smale and Zhou, 2009; Yao et al., 2007) is the technical difficulty caused by the insensitive parameter ϵ_t which will be overcome in our insensitive analysis in Section 3 and in our total error estimates in Section 5. Analyzing sparsity of algorithm (1.6) quantitatively will be considered elsewhere.

2. Error analysis and effects of parameters

Performance of learning algorithms generated by regularization schemes and their online versions in RKHSs is often measured by generalization errors. For the quantile regression learning problem, the *generalization error* $\mathcal{E}(f)$ of a function $f:X\to Y$ is defined by means of the pinball loss ψ_τ as

$$\mathcal{E}(f)=\int_Z \psi_\tau(f(x)-y)\,d\rho. \tag{2.1}$$

Throughout the paper, we assume that $\int_Z|y|\,d\rho<\infty$ and that the quantile regression function value $f_{\rho,\tau}(x)$ is uniquely determined for each point $x\in X$. With this assumption, we know from the bound $\psi_\tau(u)\le|u|$ that $\mathcal{E}(f)$ is finite if f is bounded on X or $f\in L^2_{\rho_X}$ where ρ_X is the marginal distribution of ρ on X. By decomposing the measure ρ into ρ_X and the conditional distributions $\{\rho_x\}$, we see the well-known fact that $f_{\rho,\tau}$ is a minimizer of $\mathcal{E}(f)$ among all measurable functions on X. Hence $f_{\rho,\tau}$ is the only minimizer of $\mathcal{E}(f)$.

Since the RKHS $\mathcal{H}_K$ is used as a hypothesis space for quantile regression, the learning ability of algorithm (1.6) depends on the approximation power of the space $\mathcal{H}_K$ with respect to the target function $f_{\rho,\tau}$. As in Smale and Zhou (2003, 2009), Wu et al. (2007), Yao (2010), Cucker and Zhou (2007), Steinwart and Christmann (2008), we can quantify this approximation power by the following concept.

Definition 2. The approximation error $\mathcal{D}(\lambda)$ of the triple (ρ,K,τ) is defined by

$$\mathcal{D}(\lambda)=\inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f)-\mathcal{E}(f_{\rho,\tau})+\frac{\lambda}{2}\|f\|_K^2\right\},\quad \lambda>0. \tag{2.2}$$

A minimizer f_λ of (2.2) is called a regularizing function.

The existence and uniqueness of the regularizing function f_λ can be seen from Corollary 5.3 of Steinwart and Christmann (2008).

For our error analysis, we also need a condition about the continuity of the conditional distributions $\{\rho_x\}_{x\in X}$ introduced in Hu and Zhou (2009).

Definition 3. Let $s>0$. We say that the family of conditional distributions $\{\rho_x\}_{x\in X}$ is Lipschitz-s if there exists a constant $C_\rho>0$ such that

$$\rho_x(\{y\in Y : u<y\le v\})\le C_\rho|u-v|^s, \quad \forall x\in X, \quad u<v\in Y. \tag{2.3}$$

Remark 1. If we denote F_x the probability distribution function of the conditional distribution ρ_x at $x\in X$, we see that Definition 3 is equivalent to the uniform Lipschitz-s continuity of the family of functions $\{F_x\}_{x\in X}$ in the sense that for a constant $C_\rho>0$,

$$|F_x(v)-F_x(u)|\le C_\rho|u-v|^s, \quad \forall x\in X,\ u<v\in Y.$$

In particular, if each ρ_x has a density function $d\rho_x/dy$ and these density functions are uniformly bounded on Y by a constant C_ρ, then (2.3) holds true with $s=1$.

Now we can state an error bound for online algorithm (1.6) which will be proved in Section 5. Throughout the paper, the kernel K is assumed to be bounded in the sense that $\kappa=\sup_{x\in X}\sqrt{K(x,x)}$ is finite. A Gaussian kernel $K(x,u)=\exp\{-|x-u|^2/\sigma^2\}$ on $X\subseteq\mathbb{R}^n$ with $\sigma>0$ is a typical example.

Theorem 1. *Assume* (2.3) *with* $s>0$ *and with two constants* $\mathcal{D}_0$ *and* $0<\gamma\le 1$,

$$\mathcal{D}(\lambda)\le\mathcal{D}_0\lambda^\gamma, \quad \forall\lambda>0. \tag{2.4}$$

Let the parameters $\eta_t,\lambda_t,\epsilon_t$ *be of the form*

$$\eta_t=t^{-(2+4\gamma)/(5+5\gamma)}, \quad \lambda_t=t^{-2/(5+5\gamma)}, \quad \epsilon_t=t^{-\beta}, \tag{2.5}$$

where

$$\beta\ge\max\left\{\frac{2\gamma}{5+5\gamma},\ \frac{6}{5s}-1,\ \frac{2}{5s}\right\}. \tag{2.6}$$

If there exist two positive constants w_τ *and* b_τ *such that for any* $w\in(0,w_\tau]$ *and* $x\in X$, *there holds*

$$\begin{aligned}&\rho_x(\{y: f_{\rho,\tau}(x)<y<f_{\rho,\tau}(x)+w\})\ge b_\tau w,\\ &\quad\rho_x(\{y: f_{\rho,\tau}(x)-w<y<f_{\rho,\tau}(x)\})\ge b_\tau w,\end{aligned} \tag{2.7}$$

then

$$\mathbb{E}_{z_1,\ldots,z_T}\|f_{T+1}-f_{\rho,\tau}\|_{L^2_{\rho_X}}\le\frac{C^*}{\sqrt{b_\tau w_\tau}}T^{-\gamma/(5+5\gamma)}, \tag{2.8}$$

where C^* *is a constant independent of* T *or* τ.

Remark 2. Condition (2.4) with $\gamma>1$ may hold true only if $f_{\rho,\tau}(x)=0$ almost everywhere. To prove this, we recall the reproducing property of the RKHS

$$f(x)=\langle f,K_x\rangle_K, \quad \forall x\in X,\ f\in\mathcal{H}_K \tag{2.9}$$

which implies

$$\|f\|_\infty\le\kappa\|f\|_K, \quad \forall f\in\mathcal{H}_K. \tag{2.10}$$

Then condition (2.4) with $\gamma>1$ implies $\mathcal{E}(f_\lambda)-\mathcal{E}(f_{\rho,\tau})\le\mathcal{D}_0\lambda^\gamma$ and

$$\frac{\lambda}{2}\|f_\lambda\|_K^2\le\mathcal{D}_0\lambda^\gamma\Rightarrow\|f_\lambda\|_\infty\le\kappa\|f_\lambda\|_K\le\kappa\sqrt{2\mathcal{D}_0}\lambda^{(\gamma-1)/2}.$$

This together with the Lipschitz property $|\psi_\tau(x)-\psi_\tau(u)|\le|x-u|$ of the loss function yields

$$\begin{aligned}\mathcal{E}(0)-\mathcal{E}(f_{\rho,\tau})&=\mathcal{E}(0)-\mathcal{E}(f_\lambda)+\mathcal{E}(f_\lambda)-\mathcal{E}(f_{\rho,\tau})\le\int_Z\psi_\tau(0-y)-\psi_\tau(f_\lambda(x)-y)\,d\rho+\mathcal{D}_0\lambda^\gamma\\ &\le\int_Z|0-f_\lambda(x)|\,d\rho+\mathcal{D}_0\lambda^\gamma\le\kappa\sqrt{2\mathcal{D}_0}\lambda^{(\gamma-1)/2}+\mathcal{D}_0\lambda^\gamma\to 0 \quad (\text{as } \lambda\to 0)\end{aligned}$$

and thereby $\mathcal{E}(0)=\mathcal{E}(f_{\rho,\tau})$ meaning that $f_{\rho,\tau}(x)=0$ almost everywhere. The above argument can be found in Chen et al. (2004).

It can also be seen from Corollary 5.18 of Steinwart and Christmann (2008) that condition (2.4) with $\gamma=1$ holds if and only if $f_{\rho,\tau}\in\mathcal{H}_K$.

The parameters given by (2.5) and the restriction imposed on β by (2.6) in Theorem 1 take a special form. Error bounds for more general forms of parameters will be given in Section 5. Assumption (2.7) is a noise condition about conditional distributions ρ_x near the quantile regression function values $f_{\rho,\tau}(x)$ which is a special case of a concept in Steinwart and Christmann (2011). A more general noise condition will be given in Definition 6 in Section 5.

To illustrate our error analysis, we give a simple example to be proved in Section 6. Recall that the regression function f_ρ of the probability measure ρ is the function on X defined by $f_\rho(x)=\int_Y y\, d\rho_x$.

Example 1. Let $\zeta>0$ and $f_\rho\in\mathcal{H}_K$. Assume that $\mathcal{H}_K$ contains the constant 1 function and that the conditional distributions $\{\rho_x\}_{x\in X}$ have density functions given by

$$\frac{d\rho_x}{dy}(y)=\begin{cases}\frac{\zeta+1}{2}|y-f_\rho(x)|^\zeta & \text{if } |y-f_\rho(x)|<1,\\ 0 & \text{otherwise}.\end{cases} \tag{2.11}$$

Set

$$\eta_t=t^{-3/5},\quad \lambda_t=t^{-1/5},\quad \epsilon_t=t^{-\beta}\quad \text{with } \beta\ge\frac{2}{5}.$$

Then we have

$$\mathbb{E}_{z_1,\dots,z_T}\|f_{T+1}-f_{\rho,\tau}\|_{L^2_{\rho_X}}\le C^*T^{-\theta}, \tag{2.12}$$

where

$$\theta=\begin{cases}\frac{1}{10} & \text{if } \tau\neq\frac{1}{2},\\ \frac{1}{5(2+\zeta)} & \text{if } \tau=\frac{1}{2},\end{cases}$$

and C^* is a constant independent of T.

Remark 3. By taking ζ to be large enough, we see that the power exponent in (2.12) can be arbitrarily small for $\tau=\frac{1}{2}$ while remaining to be $\frac{1}{10}$ for $\tau\neq\frac{1}{2}$. It tells us that for the quantile regression online algorithm (1.4), different quantile parameters τ may correspond to various learning rates. This is caused by different concentrations of the conditional distributions around the median, hence for an arbitrary $\tau_0\in(0,1)$, similar constructions leading to different learning rates corresponding to $\tau=\tau_0$ and $\tau\neq\tau_0$ can be made.

3. Insensitive analysis

Approximation or learning ability of a learning algorithm for quantile regression can usually be studied by estimating the *excess generalization error* $\mathcal{E}(f)-\mathcal{E}(f_{\rho,\tau})$. To bound this for the output function f_{T+1} produced by online algorithm (1.6), we apply an error decomposition, developed well for regularization schemes in the literature of learning theory (e.g. Wu and Zhou, 2008) as

$$\begin{aligned}\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\rho,\tau})&\le\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\lambda_T})+\left\{\mathcal{E}(f_{\lambda_T})-\mathcal{E}(f_{\rho,\tau})+\frac{\lambda_T}{2}\|f_{\lambda_T}\|_K^2\right\}\\ &=\{\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\lambda_T})\}+\mathcal{D}(\lambda_T)\le\kappa\|f_{T+1}-f_{\lambda_T}\|_K+\mathcal{D}(\lambda_T).\end{aligned} \tag{3.1}$$

Here we have used the fact that $|(\psi_\tau)'_-(u)|\le 1$ which yields

$$\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\lambda_T})=\int_Z\psi_\tau(f_{T+1}(x)-y)-\psi_\tau(f_{\lambda_T}(x)-y)\,d\rho\le\|f_{T+1}-f_{\lambda_T}\|_\infty$$

and by (2.10)

$$\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\lambda_T})\le\kappa\|f_{T+1}-f_{\lambda_T}\|_K.$$

The above procedure is standard in error analysis (e.g. Wu et al., 2007) and does not involve the insensitive parameter (except the output function f_{T+1}). Now we turn to our insensitive analysis concerning insensitive regularizing functions.

Definition 4. For $\epsilon\ge 0$ and $\lambda>0$, the insensitive regularizing function f_λ^ϵ is defined by

$$f_\lambda^\epsilon=\arg\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}^{(\epsilon)}(f)+\frac{\lambda}{2}\|f\|_K^2\right\}, \tag{3.2}$$

where $\mathcal{E}^{(\epsilon)}(f)$ is the ϵ-insensitive generalization error given by

$$\mathcal{E}^{(\epsilon)}(f)=\int_Z\psi_\tau^\epsilon(f(x)-y)\,d\rho.$$

Algorithm (1.6) is implemented iteratively in terms of the loss function $\psi_\tau^{\epsilon_t}$ involving the insensitive parameter ϵ_t. So it is natural to expect from the previous study on online algorithms (Ying and Zhou, 2006; Smale and Zhou, 2009; Hu and Zhou, 2009) that f_{T+1} approximates the insensitive regularizing function $f_{\lambda_T}^{\epsilon_T}$. This leads us to continue error decomposition (3.1) as

$$\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\rho,\tau}) \le \kappa\|f_{T+1}-f_{\lambda_T}^{\epsilon_T}\|_K+\kappa\|f_{\lambda_T}^{\epsilon_T}-f_{\lambda_T}\|_K+\mathcal{D}(\lambda_T). \tag{3.3}$$

While the first term in the above bound will be analyzed in the next section, the second term, purely caused by the insensitive parameter ϵ_T introduced here, can be estimated immediately by the following general result, inequality (3.4) with $v=0$ and $\mu=\epsilon_T$. Note that $\mathcal{E}^{(\epsilon)}(f)=\mathcal{E}(f)$ and $f_\lambda^\epsilon=f_\lambda$ when $\epsilon=0$.

Proposition 1. *If the family of conditional distributions* $\{\rho_x\}_{x\in X}$ *is Lipschitz-s for some* $s>0$ *satisfying* (2.3), *then for any* $0\le v<\mu$, *we have*

$$\|f_\lambda^\mu-f_\lambda^v\|_K \le \frac{C_\rho\kappa|\mu-v|^s}{\lambda}. \tag{3.4}$$

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

$$\|f_\lambda^{\epsilon_{t-1}}-f_\lambda^{\epsilon_t}\|_K \le \frac{C_\rho\kappa\epsilon_1^s\beta^s 2^{(\beta+1)s}}{\lambda}t^{-(\beta+1)s}, \quad \forall t\ge 2. \tag{3.5}$$

The proof of Proposition 1 is postponed to the end of this section due to the technical difficulty that the loss function ψ_τ^ϵ defined by (1.5) is not differential and differentiating the functional in (3.2) involves subdifferentials of ψ_τ^ϵ. To overcome the difficulty, we borrow some ideas from Ying and Zhou (2006) and approximate the non-differential loss function ψ_τ^ϵ by a family of convex, differentiable ones $\{(\psi_\tau^\epsilon)_h : 0<h\le 1\}$ defined on $\mathbb{R}$ as

$$(\psi_\tau^\epsilon)_h(x)=\int_0^1 \psi_\tau^\epsilon(x+h\theta)\,d\theta=\frac{1}{h}\int_x^{x+h}\psi_\tau^\epsilon(u)\,du, \quad x\in\mathbb{R}. \tag{3.6}$$

The approximation satisfies

$$|(\psi_\tau^\epsilon)_h(x)-\psi_\tau^\epsilon(x)|=\left|\int_0^1 \psi_\tau^\epsilon(x+h\theta)-\psi_\tau^\epsilon(x)\,d\theta\right|\le h, \quad \forall x\in\mathbb{R},\ 0<h\le 1. \tag{3.7}$$

Let us first prove relation (3.4) for the approximating regularizing functions $(f_\lambda^\epsilon)_h$ defined by

$$(f_\lambda^\epsilon)_h=\arg\min_{f\in\mathcal{H}_K}\left\{\int_Z(\psi_\tau^\epsilon)_h(f(x)-y)\,d\rho+\frac{\lambda}{2}\|f\|_K^2\right\}, \quad 0<h\le 1,\ \epsilon\ge 0,\ \lambda>0. \tag{3.8}$$

Lemma 1. *If the family of conditional distributions* $\{\rho_x\}_{x\in X}$ *is Lipschitz-s for some* $s>0$ *satisfying* (2.3), *then we have*

$$\|(f_\lambda^\mu)_h-(f_\lambda^v)_h\|_K \le \frac{C_\rho\kappa|\mu-v|^s}{\lambda}, \quad 0\le v<\mu,\ 0<h\le 1,\ \lambda>0. \tag{3.9}$$

Proof. Taking the functional derivative of the functional in (3.8), we know that the regularizing function $(f_\lambda^\epsilon)_h\in\mathcal{H}_K$ satisfies

$$\lambda(f_\lambda^\epsilon)_h+\int_Z(\psi_\tau^\epsilon)'_h((f_\lambda^\epsilon)_h(x)-y)K_x\,d\rho=0. \tag{3.10}$$

Putting the corresponding expression for $(f_\lambda^\epsilon)_h$ with $\epsilon=\mu,v$ into $\|(f_\lambda^\mu)_h-(f_\lambda^v)_h\|_K^2=\langle(f_\lambda^\mu)_h-(f_\lambda^v)_h,(f_\lambda^\mu)_h-(f_\lambda^v)_h\rangle_K$, we know that $\|(f_\lambda^\mu)_h-(f_\lambda^v)_h\|_K^2$ equals

$$-\frac{1}{\lambda}\langle(f_\lambda^\mu)_h-(f_\lambda^v)_h,\int_Z\{(\psi_\tau^\mu)'_h((f_\lambda^\mu)_h(x)-y)-(\psi_\tau^v)'_h((f_\lambda^v)_h(x)-y)\}K_x d\rho\rangle_K.$$

Applying the reproducing property (2.9) yields

$$\|(f_\lambda^\mu)_h-(f_\lambda^v)_h\|_K^2=-\frac{1}{\lambda}\int_Z\{(f_\lambda^\mu)_h(x)-(f_\lambda^v)_h(x)\}\{(\psi_\tau^\mu)'_h((f_\lambda^\mu)_h(x)-y)-(\psi_\tau^v)'_h((f_\lambda^v)_h(x)-y)\}\,d\rho=:I_1+I_2, \tag{3.11}$$

where

$$I_1=-\frac{1}{\lambda}\int_Z\{(f_\lambda^\mu)_h(x)-(f_\lambda^v)_h(x)\}\{(\psi_\tau^\mu)'_h((f_\lambda^\mu)_h(x)-y)-(\psi_\tau^\mu)'_h((f_\lambda^v)_h(x)-y)\}\,d\rho,$$

$$I_2=\frac{1}{\lambda}\int_Z\{(f_\lambda^\mu)_h(x)-(f_\lambda^v)_h(x)\}\{(\psi_\tau^v)'_h((f_\lambda^v)_h(x)-y)-(\psi_\tau^\mu)'_h((f_\lambda^v)_h(x)-y)\}\,d\rho.$$

By the convexity of the loss function $(\psi_\tau^\mu)_h$, we know that $I_1\le 0$.

To bound I_2, we observe from definition (3.6) of the differentiable approximating loss $(\psi_\tau^\epsilon)_h$ that $(\psi_\tau^\epsilon)'_h(u) = (1/h)\{\psi_\tau^\epsilon(u+h) - \psi_\tau^\epsilon(u)\}$ and then

$$(\psi_\tau^\nu)'_h(u) - (\psi_\tau^\mu)'_h(u) = \frac{1}{h}\{\psi_\tau^\nu(u+h) - \psi_\tau^\mu(u+h) + \psi_\tau^\mu(u) - \psi_\tau^\nu(u)\} = \frac{1}{h}\int_0^h (\psi_\tau^\nu)'_-(u+\upsilon) - (\psi_\tau^\mu)'_-(u+\upsilon)\, d\upsilon.$$

But from definition (1.5) of the ϵ-insensitive pinball loss ψ_τ^ϵ we find that

$$(\psi_\tau^\epsilon)'_-(\upsilon) = \begin{cases} 1-\tau & \text{if } \upsilon > \epsilon, \\ -\tau & \text{if } \upsilon \le -\epsilon, \\ 0 & \text{if } -\epsilon < \upsilon \le \epsilon, \end{cases}$$

which implies

$$(\psi_\tau^\nu)'_-(\upsilon) - (\psi_\tau^\mu)'_-(\upsilon) = \begin{cases} 1-\tau & \text{if } \nu < \upsilon \le \mu, \\ -\tau & \text{if } -\mu < \upsilon \le -\nu, \\ 0 & \text{otherwise.} \end{cases}$$

It follows by denoting the characteristic function of a set J as χ_J that

$$\begin{aligned} I_2 &= \frac{1}{\lambda}\int_Z \{(f_\lambda^\mu)_h(x) - (f_\lambda^\nu)_h(x)\}\left\{\frac{1}{h}\int_0^h (\psi_\tau^\nu)'_-((f_\lambda^\nu)_h(x) - y + \upsilon) - (\psi_\tau^\mu)'_-((f_\lambda^\nu)_h(x) - y + \upsilon)\, d\upsilon\right\} d\rho \\ &= \frac{1}{\lambda h}\int_0^h \int_X \{(f_\lambda^\mu)_h(x) - (f_\lambda^\nu)_h(x)\}\int_Y (\psi_\tau^\nu)'_-((f_\lambda^\nu)_h(x) - y + \upsilon) - (\psi_\tau^\mu)'_-((f_\lambda^\nu)_h(x) - y + \upsilon)\, d\rho_x(y)\, d\rho_X(x)\, d\upsilon \\ &= \frac{1}{\lambda h}\int_0^h \int_X \{(f_\lambda^\mu)_h(x) - (f_\lambda^\nu)_h(x)\}\int_Y (1-\tau)\chi_{\{\nu < (f_\lambda^\nu)_h(x) - y + \upsilon \le \mu\}} - \tau\chi_{\{-\mu < (f_\lambda^\nu)_h(x) - y + \upsilon \le -\nu\}}\, d\rho_x(y)\, d\rho_X(x)\, d\upsilon \\ &= \frac{1}{\lambda h}\int_0^h \int_X \{(f_\lambda^\mu)_h(x) - (f_\lambda^\nu)_h(x)\}((1-\tau)\rho_x\{y \in Y:\ (f_\lambda^\nu)_h(x) + \upsilon - \mu \le y < (f_\lambda^\nu)_h(x) + \upsilon - \nu\} \\ &\quad -\tau\rho_x\{y \in Y:\ (f_\lambda^\nu)_h(x) + \upsilon + \nu \le y < (f_\lambda^\nu)_h(x) + \upsilon + \mu\})\, d\rho_X(x)\, d\upsilon. \end{aligned}$$

Since the family of conditional distributions $\{\rho_x\}_{x\in X}$ is Lipschitz-s, by (2.3) we have

$$\begin{aligned} I_2 &\le \frac{1}{\lambda h}\int_0^h \int_X |(f_\lambda^\mu)_h(x) - (f_\lambda^\nu)_h(x)|\, C_\rho |\mu-\nu|^s\, d\rho_X(x)\, d\upsilon \\ &\le \frac{C_\rho \kappa |\mu-\nu|^s}{\lambda}\|(f_\lambda^\mu)_h - (f_\lambda^\nu)_h\|_K. \end{aligned}$$

This in connection with (3.11) verifies (3.4), and proves Lemma 1. □

We are now in a position to prove Proposition 1. Denote $M = \int_Z |y|\, d\rho < \infty$.

Proof of Proposition 1. We first derive a sequence $\{h_j \to 0\}$ such that the sequences $\{(f_\lambda^\mu)_{h_j}\}$ and $\{(f_\lambda^\nu)_{h_j}\}$ are weakly convergent, respectively. To this end, we observe that for $\epsilon \ge 0$ and $0 < h \le 1$, taking $f=0$ in (3.8) yields

$$\frac{\lambda}{2}\|(f_\lambda^\epsilon)_h\|_K^2 \le \int_Z (\psi_\tau^\epsilon)_h(0-y)\, d\rho \le \int_Z |y| + 1\, d\rho = M+1.$$

Hence the set of functions $\{(f_\lambda^\epsilon)_h : 0 < h \le 1\}$ lies in the ball $B_R = \{f \in \mathcal{H}_K : \|f\|_K \le R\}$ with radius $R := \sqrt{2(M+1)/\lambda}$ of the Hilbert space $\mathcal{H}_K$.

Since the ball B_R of the Hilbert space $\mathcal{H}_K$ is weakly compact, for $\epsilon = \mu$, we can first find a sequence $\{\widehat{h}_j \in (0,1]\}_{j\in\mathbb{N}}$ such that $\lim_{j\to\infty}\widehat{h}_j = 0$ and that $(f_\lambda^\mu)_{\widehat{h}_j}$ converges to some $\widehat{f}_\lambda^\mu$ weakly in the Hilbert space $\mathcal{H}_K$. Then for $\epsilon = \nu$, we consider the sequence $\{(f_\lambda^\nu)_{\widehat{h}_j}\}_{j\in\mathbb{N}}$ of functions in the closed ball B_R. By the weak compactness again, we can find a subsequence $\{h_j \in (0,1]\}_{j\in\mathbb{N}}$ of the sequence $\{\widehat{h}_j \in (0,1]\}_{j\in\mathbb{N}}$ such that $(f_\lambda^\nu)_{h_j}$ converges to some $\widehat{f}_\lambda^\nu$ weakly in the Hilbert space $\mathcal{H}_K$. The sequence $\{h_j \in (0,1]\}_{j\in\mathbb{N}}$ now satisfies $\lim_{j\to\infty}h_j = 0$ and a weak convergence relation:

$$\lim_{j\to\infty}\langle (f_\lambda^\mu)_{h_j}, f\rangle_K = \langle \widehat{f}_\lambda^\mu, f\rangle_K \quad \text{and} \quad \lim_{j\to\infty}\langle (f_\lambda^\nu)_{h_j}, f\rangle_K = \langle \widehat{f}_\lambda^\nu, f\rangle_K, \quad \forall f \in \mathcal{H}_K. \tag{3.12}$$

Next we show that $\widehat{f}_\lambda^\mu = f_\lambda^\mu$ and $\widehat{f}_\lambda^\nu = f_\lambda^\nu$. To prove this, we take $\epsilon = \mu$ or ν and notice from (3.12) that

$$\|\widehat{f}_\lambda^\epsilon\|_K^2 = \langle \widehat{f}_\lambda^\epsilon, \widehat{f}_\lambda^\epsilon\rangle_K = \lim_{j\to\infty}\langle (f_\lambda^\epsilon)_{h_j}, \widehat{f}_\lambda^\epsilon\rangle_K \le \|\widehat{f}_\lambda^\epsilon\|_K \underline{\lim}_{j\to\infty}\|(f_\lambda^\epsilon)_{h_j}\|_K$$

and

$$\|\widehat{f}^{\epsilon}_{\lambda}\|_K \le \underline{\lim}_{j\to\infty}\|(f^{\epsilon}_{\lambda})_{h_j}\|_K \le R. \tag{3.13}$$

Let $f=K_x$ in (3.12). Then the reproducing property (2.9) yields

$$\lim_{j\to\infty}(f^{\epsilon}_{\lambda})_{h_j}(x)=\lim_{j\to\infty}\langle (f^{\epsilon}_{\lambda})_{h_j},K_x\rangle_K=\langle \widehat{f}^{\epsilon}_{\lambda},K_x\rangle_K=\widehat{f}^{\epsilon}_{\lambda}(x).$$

This in connection with the continuity of the loss function ψ^{ϵ}_{τ}, the uniform bounds for $(f^{\epsilon}_{\lambda})_{h_j},\widehat{f}^{\epsilon}_{\lambda}$ and the Lebesgue Dominant Theorem gives $\mathcal{E}^{(\epsilon)}(\widehat{f}^{\epsilon}_{\lambda})=\lim_{j\to\infty}\int_Z\psi^{\epsilon}_{\tau}((f^{\epsilon}_{\lambda})_{h_j}(x)-y)\,d\rho$. The uniform bound for $(f^{\epsilon}_{\lambda})_{h_j}$ in connection with the error bound (3.7) for $(\psi^{\epsilon}_{\tau})_h$ to approximate ψ^{ϵ}_{τ} ensures that

$$\lim_{j\to\infty}\int_Z\psi^{\epsilon}_{\tau}((f^{\epsilon}_{\lambda})_{h_j}(x)-y)\,d\rho=\lim_{j\to\infty}\int_Z(\psi^{\epsilon}_{\tau})_{h_j}((f^{\epsilon}_{\lambda})_{h_j}(x)-y)\,d\rho.$$

Therefore, by (3.13), we have

$$\mathcal{E}^{(\epsilon)}(\widehat{f}^{\epsilon}_{\lambda})+\frac{\lambda}{2}\left\|\widehat{f}^{\epsilon}_{\lambda}\right\|_K^2\le\underline{\lim}_{j\to\infty}\left\{\int_Z(\psi^{\epsilon}_{\tau})_{h_j}((f^{\epsilon}_{\lambda})_{h_j}(x)-y)d\rho+\frac{\lambda}{2}\left\|(f^{\epsilon}_{\lambda})_{h_j}\right\|_K^2\right\}.$$

Since the function $(f^{\epsilon}_{\lambda})_{h_j}$ minimizes the regularized functional in (3.8) with $h=h_j$, by taking $f=f^{\epsilon}_{\lambda}$ we see that

$$\int_Z(\psi^{\epsilon}_{\tau})_{h_j}((f^{\epsilon}_{\lambda})_{h_j}(x)-y)\,d\rho+\frac{\lambda}{2}\|(f^{\epsilon}_{\lambda})_{h_j}\|_K^2\le\int_Z(\psi^{\epsilon}_{\tau})_{h_j}(f^{\epsilon}_{\lambda}(x)-y)\,d\rho+\frac{\lambda}{2}\|f^{\epsilon}_{\lambda}\|_K^2.$$

Then it follows from (3.7) that

$$\mathcal{E}^{(\epsilon)}(\widehat{f}^{\epsilon}_{\lambda})+\frac{\lambda}{2}\|\widehat{f}^{\epsilon}_{\lambda}\|_K^2\le\underline{\lim}_{j\to\infty}\left\{\int_Z(\psi^{\epsilon}_{\tau})_{h_j}(f^{\epsilon}_{\lambda}(x)-y)\,d\rho+\frac{\lambda}{2}\|f^{\epsilon}_{\lambda}\|_K^2\right\}$$

$$=\int_Z\psi^{\epsilon}_{\tau}(f^{\epsilon}_{\lambda}(x)-y)\,d\rho+\frac{\lambda}{2}\|f^{\epsilon}_{\lambda}\|_K^2=\mathcal{E}^{(\epsilon)}(f^{\epsilon}_{\lambda})+\frac{\lambda}{2}\|f^{\epsilon}_{\lambda}\|_K^2.$$

It means that $\widehat{f}^{\epsilon}_{\lambda}$ is also a minimizer of the functional $\mathcal{E}^{(\epsilon)}(f)+(\lambda/2)\|f\|_K^2$. The strict convexity of this functional on $\mathcal{H}_K$ verifies the uniqueness of its minimizer, hence $\widehat{f}^{\epsilon}_{\lambda}=f^{\epsilon}_{\lambda}$.

Finally we prove the desired bound (3.4) for $\|f^{\mu}_{\lambda}-f^{\nu}_{\lambda}\|_K$. The previous step shows that (3.12) holds with $\widehat{f}^{\epsilon}_{\lambda}$ replaced by f^{ϵ}_{λ}. By taking $f=f^{\mu}_{\lambda}-f^{\nu}_{\lambda}$ we see that

$$\|f^{\mu}_{\lambda}-f^{\nu}_{\lambda}\|_K^2=\langle f^{\mu}_{\lambda}-f^{\nu}_{\lambda},f^{\mu}_{\lambda}-f^{\nu}_{\lambda}\rangle_K=\lim_{j\to\infty}\langle (f^{\mu}_{\lambda})_{h_j}-(f^{\nu}_{\lambda})_{h_j},f^{\mu}_{\lambda}-f^{\nu}_{\lambda}\rangle_K$$

$$\le\underline{\lim}_{j\to\infty}\|(f^{\mu}_{\lambda})_{h_j}-(f^{\nu}_{\lambda})_{h_j}\|_K\|f^{\mu}_{\lambda}-f^{\nu}_{\lambda}\|_K.$$

Then we apply Lemma 1 and obtain

$$\|f^{\mu}_{\lambda}-f^{\nu}_{\lambda}\|_K^2\le\frac{C_{\rho}\kappa|\mu-\nu|^s}{\lambda}\|f^{\mu}_{\lambda}-f^{\nu}_{\lambda}\|_K.$$

Hence (3.4) is verified.

By taking $\mu=\epsilon_1(t-1)^{-\beta}$ and $\nu=\epsilon_1t^{-\beta}$, we see from the elementary inequality $(t-1)^{-\beta}-t^{-\beta}\le\beta(t-1)^{-\beta-1}\le\beta2^{\beta+1}t^{-\beta-1}$ that (3.5) holds true. □

4. One-step iteration and key error analysis

In this section, we estimate the first term $\|f_{T+1}-f^{\epsilon_T}_{\lambda}\|_K$ of the bound in the error decomposition (3.3). This will be conducted by iterating one-step analysis with $t=1,\dots,T$. Here one-step analysis aims at bounding $\|f_{t+1}-f^{\epsilon_t}_{\lambda_t}\|_K$ in terms of $\|f_t-f^{\epsilon_{t-1}}_{\lambda_{t-1}}\|_K$. To this end, we need two types of errors, caused by the changing parameters ϵ_t and λ_t. Denote $f^{\epsilon_1}_{\lambda_0}=f^{\epsilon_0}_{\lambda_0}=0$.

Definition 5. The insensitive error is defined as

$$h_t=\|f^{\epsilon_{t-1}}_{\lambda_{t-1}}-f^{\epsilon_t}_{\lambda_{t-1}}\|_K,\quad t\in\mathbb{N}. \tag{4.1}$$

The drift error is defined as

$$d_t=\|f^{\epsilon_t}_{\lambda_{t-1}}-f^{\epsilon_t}_{\lambda_t}\|_K,\quad t\in\mathbb{N}. \tag{4.2}$$

The insensitive error can be easily estimated by (3.5). The drift error will be bounded by Lemma 3 below.

Now we turn to the one-step analysis. By the convexity of the loss function, the following relation about the minimizer f_λ^ϵ of the regularized generalization error $\mathcal{E}^{(\epsilon)}(f)+(\lambda/2)\|f\|_K^2$ is seen from (Ying and Zhou, 2006)

$$\left\{\mathcal{E}^{(\epsilon)}(f)+\frac{\lambda}{2}\|f\|_K^2\right\}-\left\{\mathcal{E}^{(\epsilon)}(f_\lambda^\epsilon)+\frac{\lambda}{2}\|f_\lambda^\epsilon\|_K^2\right\}\geq\frac{\lambda}{2}\|f-f_\lambda^\epsilon\|_K^2,\quad \forall f\in\mathcal{H}_K,\ \lambda>0,\ \epsilon\geq 0. \tag{4.3}$$

Lemma 2. *Define* $\{f_t\}$ *by* (1.6) *and* h_t,d_t *by* (4.1) *and* (4.2), *respectively. Let* $0<q_1,q_2<2$, *and* $A_1,A_2>0$. *If* $\lambda_t\geq\lambda_{t+1}>0$ *for each* t, *then we have*

$$\begin{aligned}\mathbb{E}_{z_t}\|f_{t+1}-f_{\lambda_t}^{\epsilon_t}\|_K^2&\leq(1+A_1d_t^{q_1}+A_2h_t^{q_2}+A_1A_2d_t^{q_1}h_t^{q_2}-\lambda_t\eta_t)\|f_t-f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2\\&\quad+(1+A_1d_t^{q_1})(h_t^{2-q_2}/A_2+h_t^2)+d_t^{2-q_1}/A_1+d_t^2+4\kappa^2\eta_t^2.\end{aligned} \tag{4.4}$$

Proof. First, we claim that

$$\|f_t\|_K\leq\frac{\kappa}{\lambda_t},\quad\forall t. \tag{4.5}$$

It can be easily seen by induction from $f_1=0$ and the following estimate derived from definition (1.6):

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

Denote $G_t=(\psi_\tau^{\epsilon_t})'_-(f_t(x_t)-y_t)K_{x_t}+\lambda_tf_t$. Note that $\|(\psi_\tau^{\epsilon_t})'_-\|_\infty\leq 1$. Then $\|G_t\|_K\leq 2\kappa$ by (4.5).
From definition (1.4), we see by inner products that

$$\|f_{t+1}-f_{\lambda_t}^{\epsilon_t}\|_K^2=\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2+2\eta_t\langle f_{\lambda_t}^{\epsilon_t}-f_t,G_t\rangle_K+\eta_t^2\|G_t\|_K^2. \tag{4.6}$$

By the reproducing property,

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

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

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

Also,

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

Thus, taking expectation with respect to z_t, we find

$$\mathbb{E}_{z_t}\langle f_{\lambda_t}^{\epsilon_t}-f_t,G_t\rangle_K\leq\left[\mathcal{E}^{(\epsilon_t)}(f_{\lambda_t}^{\epsilon_t})+\frac{\lambda_t}{2}\|f_{\lambda_t}^{\epsilon_t}\|_K^2\right]-\left[\mathcal{E}^{(\epsilon_t)}(f_t)+\frac{\lambda_t}{2}\|f_t\|_K^2\right].$$

Combining with (4.3), this yields

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

Therefore, together with the bound for G_t and (4.6), we have

$$\mathbb{E}_{z_t}\|f_{t+1}-f_{\lambda_t}^{\epsilon_t}\|_K^2\leq(1-\lambda_t\eta_t)\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2+4\kappa^2\eta_t^2. \tag{4.7}$$

Decompose $\|f_t-f_{\lambda_t}^{\epsilon_t}\|_K^2$ 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 $2ab\leq Aa^2b^q+b^{2-q}/A$ 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

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

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

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

Combining the above two estimates, we obtain

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

Since $(1-\lambda_t\eta_t)(1+A_1d_t^{q_1})\big(1+A_2h_t^{q_2}\big)\leq 1+A_1d_t^{q_1}+A_2h_t^{q_2}+A_1A_2d_t^{q_1}h_t^{q_2}-\lambda_t\eta_t$, the above bound together with (4.7) gives our desired conclusion. □

The following perturbation results and drift error bound are derived by methods from Xiang et al. (2011), Ying and Zhou (2006), and Ye and Zhou (2007).

Lemma 3. *Let $\epsilon \ge 0$ and $f_{\rho,\tau}^{(\epsilon)}$ be a minimizer of $\mathcal{E}^{(\epsilon)}(f) = \int_Z \psi_\tau^\epsilon(f(x)-y)\, d\rho$.*

(a) *We have $\|f_{\rho,\tau}^{(\epsilon)} - f_{\rho,\tau}\|_\infty \le \epsilon$.*
(b) *For any measurable function f on X, we have*

$$\mathcal{E}(f) - \mathcal{E}(f_{\rho,\tau}) - 2\epsilon \le \mathcal{E}^{(\epsilon)}(f) - \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)}) \le \mathcal{E}(f) - \mathcal{E}(f_{\rho,\tau}) + 2\epsilon.$$

(c) *If $\lambda_{t-1} \ge \lambda_t > 0$, then there holds*

$$\|\epsilon_{\lambda_{t-1}} - f_{\lambda_t}^\epsilon\|_K \le \frac{1}{2}\left(\frac{\lambda_{t-1}}{\lambda_t} - 1\right)\left(\sqrt{(2\mathcal{D}(\lambda_{t-1}) + 4\epsilon)/\lambda_{t-1}} + \sqrt{(2\mathcal{D}(\lambda_t) + 4\epsilon)/\lambda_t}\right).$$

Proof. We first claim that for almost every $x \in X$,

$$(1-\tau)\rho_x\{y < f_\rho^{(\epsilon)}(x) - \epsilon\} \le \tau\rho_x\{y \ge f_\rho^{(\epsilon)}(x) + \epsilon\}, \quad (1-\tau)\rho_x\{y \le f_\rho^{(\epsilon)}(x) - \epsilon\} \ge \tau\rho_x\{y > f_\rho^{(\epsilon)}(x) + \epsilon\}. \tag{4.8}$$

To prove this, we denote $\mathcal{E}^{(\epsilon)}(\cdot|x)$ a function on $\mathbb{R}$ given by $\mathcal{E}^{(\epsilon)}(u|x) = \int_Y \psi_\tau^\epsilon(u-y)\, d\rho_x(y)$. Its left derivative $(\mathcal{E}^{(\epsilon)}(\cdot|x))'_-(u)$ equals

$$\int_{-\infty}^{+\infty} -\tau\chi_{(-\infty, y-\epsilon]} + (1-\tau)\chi_{(y+\epsilon, +\infty)}\, d\rho_x(y) = (1-\tau)\rho_x\{y < u - \epsilon\} - \tau\rho_x\{y \ge u + \epsilon\}$$

while its right derivative is $(\mathcal{E}^{(\epsilon)}(\cdot|x))'_+(u) = (1-\tau)\rho_x\{y \le u - \epsilon\} - \tau\rho_x\{y > u + \epsilon\}$. Observe that $\mathcal{E}^{(\epsilon)}(f) = \int_X \mathcal{E}^{(\epsilon)}(f(x)|x)\, d\rho_X(x)$. Then for almost every $x \in X$, $f_\rho^{(\epsilon)}(x)$ is a minimum point of the function $\mathcal{E}^{(\epsilon)}(\cdot|x)$ satisfying $(\mathcal{E}^{(\epsilon)}(\cdot|x))'_-(f_\rho^{(\epsilon)}(x)) \le 0 \le (\mathcal{E}^{(\epsilon)}(\cdot|x))'_+(f_\rho^{(\epsilon)}(x))$. Hence (4.8) follows.

(a) We prove that $f_\rho^{(\epsilon)}(x) \le f_{\rho,\tau}(x) + \epsilon$ for almost every $x \in X$. Suppose to the contrary that $f_\rho^{(\epsilon)}(x) > f_{\rho,\tau}(x) + \epsilon$. Then by the first inequality of (4.8) and the trivial relation $\rho_x\{y \ge f_\rho^{(\epsilon)}(x) + \epsilon\} + \rho_x\{y < f_\rho^{(\epsilon)}(x) - \epsilon\} \le 1$, we see that

$$(1-\tau)\rho_x\{y < f_\rho^{(\epsilon)}(x) - \epsilon\} \le \tau(1 - \rho_x\{y < f_\rho^{(\epsilon)}(x) - \epsilon\}) \Rightarrow \rho_x\{y < f_\rho^{(\epsilon)}(x) - \epsilon\} \le \tau.$$

But the definition of τ-quantile tells us $\rho_x\{y \le f_{\rho,\tau}(x)\} \ge \tau$. So $\rho_x\{f_{\rho,\tau}(x) < y < f_\rho^{(\epsilon)}(x) - \epsilon\} = 0$. Since $f_\rho^{(\epsilon)}(x) - \epsilon = f_{\rho,\tau}(x) + \varDelta$ with $\varDelta := f_\rho^{(\epsilon)}(x) - f_{\rho,\tau}(x) - \epsilon > 0$, we have

$$\rho_x\left\{y \ge f_{\rho,\tau}(x) + \frac{\varDelta}{2}\right\} = \rho_x\{y \ge f_{\rho,\tau}(x)\} \ge \tau.$$

But $\rho_x\{y \le f_{\rho,\tau}(x) + \varDelta/2\} \ge \rho_x\{y \ge f_{\rho,\tau}(x)\} \ge \tau$. It follows that $f_{\rho,\tau}(x) + \varDelta/2$ is also a τ-quantile of ρ_x, which is a contradiction to our assumption on the uniqueness of the quantile regression function value $f_{\rho,\tau}(x)$.

In the same way, we can show by means of the second inequality of (4.8) that $f_\rho^{(\epsilon)}(x) \ge f_{\rho,\tau}(x) - \epsilon$ for almost every $x \in X$. This proves the statement in (a).

(b) Observe that $\|\psi_\tau - \psi_\tau^\epsilon\|_\infty \le \epsilon$ which implies

$$|\mathcal{E}(f) - \mathcal{E}^{(\epsilon)}(f)| = \left|\int_Z \psi_\tau(f(x)-y) - \psi_\tau^\epsilon(f(x)-y)\, d\rho\right| \le \epsilon.$$

It follows that

$$\begin{aligned}\mathcal{E}(f) - \mathcal{E}(f_{\rho,\tau}) &= \mathcal{E}(f) - \mathcal{E}^{(\epsilon)}(f) + \mathcal{E}^{(\epsilon)}(f) - \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)}) + \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)}) - \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}) + \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}) - \mathcal{E}(f_{\rho,\tau}) \\ &\le \epsilon + \mathcal{E}^{(\epsilon)}(f) - \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)}) + \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)}) - \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}) + \epsilon.\end{aligned}$$

But $\mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)}) \le \mathcal{E}^{(\epsilon)}(f_{\rho,\tau})$. So the first inequality in statement (b) holds true. The second statement can be proved in the same way by expressing $\mathcal{E}^{(\epsilon)}(f) - \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)})$ into

$$\{\mathcal{E}^{(\epsilon)}(f) - \mathcal{E}(f)\} + \{\mathcal{E}(f) - \mathcal{E}(f_{\rho,\tau})\} + \{\mathcal{E}(f_{\rho,\tau}) - \mathcal{E}(f_{\rho,\tau}^{(\epsilon)})\} + \{\mathcal{E}(f_{\rho,\tau}^{(\epsilon)}) - \mathcal{E}^{(\epsilon)}(f_{\rho,\tau}^{(\epsilon)})\}.$$

(c) We first give some estimates involving the differentiable loss functions $(\psi_\tau^\epsilon)_h$ with $h > 0$ and the regularizing function $(f_\lambda^\epsilon)_h$ defined by (3.8). Applying (3.9) in the proof of Lemma 1, we know that

$$(f_{\lambda_{t-1}}^\epsilon)_h - (f_{\lambda_t}^\epsilon)_h = \frac{1}{\lambda_t}\int_Z (\psi_\tau^\epsilon)'_h((f_{\lambda_t}^\epsilon)_h(x) - y)K_x\, d\rho - \frac{1}{\lambda_{t-1}}\int_Z (\psi_\tau^\epsilon)'_h((f_{\lambda_{t-1}}^\epsilon)_h(x) - y)K_x\, d\rho.$$

This in connection with the reproducing property (2.9) tells us that

$$\begin{aligned}\|(f_{\lambda_{t-1}}^\epsilon)_h - (f_{\lambda_t}^\epsilon)_h\|_K^2 = \langle (f_{\lambda_{t-1}}^\epsilon)_h - (f_{\lambda_t}^\epsilon)_h, (f_{\lambda_{t-1}}^\epsilon)_h - (f_{\lambda_t}^\epsilon)_h \rangle_K = \frac{1}{\lambda_t}\int_Z (\psi_\tau^\epsilon)'_h((f_{\lambda_t}^\epsilon)_h(x) - y)\{(f_{\lambda_{t-1}}^\epsilon)_h(x) - (f_{\lambda_t}^\epsilon)_h(x)\}\, d\rho \\ - \frac{1}{\lambda_{t-1}}\int_Z (\psi_\tau^\epsilon)'_h((f_{\lambda_{t-1}}^\epsilon)_h(x) - y)\{(f_{\lambda_{t-1}}^\epsilon)_h(x) - (f_{\lambda_t}^\epsilon)_h(x)\}\, d\rho.\end{aligned}$$

As $(\psi^\epsilon_\tau)_h$ is a convex function on $\mathbb{R}$, it satisfies

$$(\psi^\epsilon_\tau)'_h(a)(b-a) \le (\psi^\epsilon_\tau)_h(b)-(\psi^\epsilon_\tau)_h(a), \quad \forall a,b \in \mathbb{R}.$$

So we can continue our estimation by taking a,b to be $(f^\epsilon_{\lambda_t})_h(x)-y, b=(f^\epsilon_{\lambda_{t-1}})_h(x)-y$ and obtain

$$\begin{aligned}\|(f^\epsilon_{\lambda_{t-1}})_h-(f^\epsilon_{\lambda_t})_h\|_K^2 &\le \frac{1}{\lambda_t}\int_Z (\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_{t-1}})_h(x)-y)-(\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_t})_h(x)-y)\, d\rho+\frac{1}{\lambda_{t-1}}\int_Z (\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_t})_h(x)-y)-(\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_{t-1}})_h(x)-y)\, d\rho\\ &= \left(\frac{1}{\lambda_t}-\frac{1}{\lambda_{t-1}}\right)\left\{\int_Z (\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_{t-1}})_h(x)-y)\, d\rho-\int_Z (\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_t})_h(x)-y)\, d\rho\right\}.\end{aligned}$$

From definition (3.8) of $(f^\epsilon_\lambda)_h$, we see that

$$\int_Z (\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_{t-1}})_h(x)-y)\, d\rho+\frac{\lambda_{t-1}}{2}\|(f^\epsilon_{\lambda_{t-1}})_h\|_K^2 \le \int_Z (\psi^\epsilon_\tau)_h((f^\epsilon_{\lambda_t})_h(x)-y)\, d\rho+\frac{\lambda_{t-1}}{2}\|(f^\epsilon_{\lambda_t})_h\|_K^2.$$

Therefore,

$$\|(f^\epsilon_{\lambda_{t-1}})_h-(f^\epsilon_{\lambda_t})_h\|_K^2 \le \left(\frac{1}{\lambda_t}-\frac{1}{\lambda_{t-1}}\right)\frac{\lambda_{t-1}}{2}(\|(f^\epsilon_{\lambda_t})_h\|_K^2-\|(f^\epsilon_{\lambda_{t-1}})_h\|_K^2).$$

But $\|(f^\epsilon_{\lambda_t})_h\|_K^2-\|(f^\epsilon_{\lambda_{t-1}})_h\|_K^2=(\|(f^\epsilon_{\lambda_t})_h\|_K+\|(f^\epsilon_{\lambda_{t-1}})_h\|_K)\|(f^\epsilon_{\lambda_{t-1}})_h-(f^\epsilon_{\lambda_t})_h\|$. So we have the following inequality which is valid for every $h>0$,

$$\|(f^\epsilon_{\lambda_{t-1}})_h-(f^\epsilon_{\lambda_t})_h\|_K \le \frac{\lambda_{t-1}}{2}\left(\frac{1}{\lambda_t}-\frac{1}{\lambda_{t-1}}\right)(\|(f^\epsilon_{\lambda_t})_h\|_K+\|(f^\epsilon_{\lambda_{t-1}})_h\|_K), \quad \forall h>0. \tag{4.9}$$

Now we turn to estimating $\|f^\epsilon_{\lambda_{t-1}}-f^\epsilon_{\lambda_t}\|_K$ by taking limits over (4.9). We apply the procedure in the proof of Proposition 1 and get a sequence $\{h_j \to 0\}$ such that the sequence $\{(f^\epsilon_\lambda)_{h_j}\}$ converges weakly to the function f^ϵ_λ for both $\lambda=\lambda_{t-1}$ and $\lambda=\lambda_t$. The weak convergence yields

$$\|f^\epsilon_{\lambda_{t-1}}-f^\epsilon_{\lambda_t}\|_K^2=\langle f^\epsilon_{\lambda_{t-1}}-f^\epsilon_{\lambda_t}, f^\epsilon_{\lambda_{t-1}}-f^\epsilon_{\lambda_t}\rangle_K=\lim_{j\to\infty}\langle f^\epsilon_{\lambda_{t-1}}-f^\epsilon_{\lambda_t}, (f^\epsilon_{\lambda_{t-1}})_{h_j}-(f^\epsilon_{\lambda_t})_{h_j}\rangle_K$$

which gives

$$\|f^\epsilon_{\lambda_{t-1}}-f^\epsilon_{\lambda_t}\|_K \le \underline{\lim}_{j\to\infty}\|(f^\epsilon_{\lambda_{t-1}})_{h_j}-(f^\epsilon_{\lambda_t})_{h_j}\|_K. \tag{4.10}$$

From the proof of Proposition 1, we also see that

$$\mathcal{E}^{(\epsilon)}(f^\epsilon_\lambda)+\frac{\lambda}{2}\|f^\epsilon_\lambda\|_K^2 \le \underline{\lim}_{j\to\infty}\left\{\mathcal{E}^{(\epsilon)}(f^\epsilon_\lambda)+\frac{\lambda}{2}\|(f^\epsilon_\lambda)_{h_j}\|_K^2\right\} \le \mathcal{E}^{(\epsilon)}(f^\epsilon_\lambda)+\frac{\lambda}{2}\|f^\epsilon_\lambda\|_K^2,$$

which implies $\|f^\epsilon_\lambda\|_K=\underline{\lim}_{j\to\infty}\|(f^\epsilon_\lambda)_{h_j}\|_K$. Combining this with (4.9) and (4.10) we obtain

$$\begin{aligned}\|f^\epsilon_{\lambda_{t-1}}-f^\epsilon_{\lambda_t}\|_K &\le \frac{\lambda_{t-1}}{2}\left(\frac{1}{\lambda_t}-\frac{1}{\lambda_{t-1}}\right)\underline{\lim}_{j\to\infty}(\|(f^\epsilon_{\lambda_t})_{h_j}\|_K+\|(f^\epsilon_{\lambda_{t-1}})_{h_j}\|_K)\\ &= \frac{\lambda_{t-1}}{2}\left(\frac{1}{\lambda_t}-\frac{1}{\lambda_{t-1}}\right)(\|f^\epsilon_{\lambda_t}\|_K+\|f^\epsilon_{\lambda_{t-1}}\|_K).\end{aligned}$$

To bound the norms further, we notice from definition (3.2) of f^ϵ_λ and the minimizer $f^{(\epsilon)}_{\rho,\tau}$ of $\mathcal{E}^{(\epsilon)}(f)=\int_Z \psi^\epsilon_\tau(f(x)-y)\, d\rho$ that

$$\frac{\lambda}{2}\|f^\epsilon_\lambda\|_K^2 \le \mathcal{E}^{(\epsilon)}(f^\epsilon_\lambda)-\mathcal{E}^{(\epsilon)}(f^{(\epsilon)}_{\rho,\tau})+\frac{\lambda}{2}\|f^\epsilon_\lambda\|_K^2=\inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}^{(\epsilon)}(f)-\mathcal{E}^{(\epsilon)}(f^{(\epsilon)}_{\rho,\tau})+\frac{\lambda}{2}\|f\|_K^2\right\}.$$

By the inequality in Part (b), we have

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

Hence $\|f^\epsilon_\lambda\|_K \le \sqrt{(2\mathcal{D}(\lambda)+4\epsilon)\lambda)}$. Using this bound with $\lambda=\lambda_{t-1}$ and $\lambda=\lambda_t$, the desired bound in Part (c) follows. This completes the proof of Lemma 3. □

We are in a position to present the key analysis in our study.

Theorem 2. *Let the parameters $\eta_t, \lambda_t, \epsilon_t$ be of the form*

$$\eta_t=\eta_1 t^{-\alpha}, \quad \lambda_t=\lambda_1 t^{-p}, \quad \epsilon_t=\epsilon_1 t^{-\beta} \tag{4.11}$$

with $\eta_1, \lambda_1, \alpha, p, \beta>0$ and $\epsilon_1 \ge 0$. Assume (2.3) *with $s>0$ and* (2.4) *with $0<\gamma\le 1$. If*

$$0<p<\min\left\{\frac{2+\beta}{5}, \frac{2}{5-\gamma}, \frac{\beta+1}{3}s\right\} \tag{4.12}$$

and

$$p<\alpha<\min\{(2-3p+p\gamma)/2,(2-3p+\beta)/2,(\beta+1)s-2p\}, \tag{4.13}$$

then

$$\mathbb{E}_{z_1,\dots,z_T}\|f_{T+1}-f_{\lambda_T}^{\epsilon_T}\|_K^2\le C'T^{-\theta^*}, \tag{4.14}$$

where

$$\theta^*=\min\{2+p\gamma-3p-2\alpha,2+\beta-3p-2\alpha,2(\beta+1)s-4p-2\alpha,\alpha-p\}>0 \tag{4.15}$$

and C' is a constant independent of T or τ (given explicitly in the proof).

Proof. To apply the estimate in Lemma 2, we need explicit bounds for d_t and h_t.

Putting the expression (4.11) for the parameters and approximation error (2.4) into Part (c) of Lemma 3, we find

$$d_t\le d_1t^{-\min\{1-p/2+p\gamma/2,1-p/2+\beta/2\}},\quad \forall t,$$

where $d_1=p2^{p+1}\sqrt{(2\mathcal{D}_0\lambda_1^{\gamma}+4\epsilon_1)/\lambda_1}$.

From (3.5) with $\lambda=\lambda_{t-1}$, we obtain

$$h_t\le h_1t^{p-(\beta+1)s},\quad \forall t,$$

where $h_1=C_\rho\kappa\epsilon_1^s\beta^s2^{(\beta+1)s}/\lambda_1$.

Now we apply Lemma 2. 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 (4.12) and (4.13), we see that $0<q_1<2$ and $0<q_2<2$. Then the coefficient of the first term of bound (4.4) can be bounded as

$$1+A_1d_t^{q_1}+A_2h_t^{q_2}+A_1A_2d_t^{q_1}h_t^{q_2}-\lambda_t\eta_t\le 1+(A_1d_1^{q_1}+A_2h_1^{q_2}+A_1A_2d_1^{q_1}h_1^{q_2})t^{-(\alpha+p)}-\lambda_t\eta_t\le 1-\frac{\eta_1\lambda_1}{2}t^{-\alpha-p}.$$

Thus by Lemma 2, we have

$$\mathbb{E}_{z_1,\dots,z_t}\|f_{t+1}-f_{\lambda_t}^{\epsilon_t}\|_K^2\le\left(1-\frac{\eta_1\lambda_1}{2}t^{-\alpha-p}\right)\mathbb{E}_{z_1,\dots,z_{t-1}}\|f_t-f_{\lambda_{t-1}}^{\epsilon_{t-1}}\|_K^2+A_3t^{-\tilde{\theta}}, \tag{4.16}$$

where

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

and

$$A_3=(1+A_1d_1^{q_1})(h_1^{2-q_2}/A_2+h_1^2)+d_1^{2-q_1}/A_1+d_1^2+4\kappa^2\eta_1^2.$$

The restrictions (4.12) and (4.13) on the parameters tell us that θ^* defined by (4.15) is positive. Hence $\tilde{\theta}=\theta^*+\alpha+p>0$. Moreover, the first term of the upper bound of (4.13) implies $\alpha+p<1+(\gamma-1)p/2\le 1$.

Applying relation (4.16) iteratively for $t=1,\dots,T$, we obtain that

$$\mathbb{E}_{z_1,\dots,z_T}\|f_{T+1}-f_{\lambda_T}^{\epsilon_T}\|_K^2\le A_3\sum_{t=1}^{T}\prod_{j=t+1}^{T}\left(1-\frac{\eta_1\lambda_1}{2}j^{-\alpha-p}\right)t^{-\tilde{\theta}}. \tag{4.17}$$

Applying the following elementary inequality (Smale and Zhou, 2009) with $0<a_1<1$, $c,a_2>0$ and $t\in\mathbb{N}$,

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

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

$$\sum_{t=1}^{T}\prod_{j=t+1}^{T}\left(1-\frac{\eta_1\lambda_1}{2}j^{-\alpha-p}\right)t^{-\tilde{\theta}}\le\sum_{t=1}^{T}t^{-\tilde{\theta}}\exp\left\{-\frac{\eta_1\lambda_1}{2}\sum_{j=t+1}^{T}j^{-\alpha-p}\right\}\le A_4T^{\alpha+p-\tilde{\theta}},$$

where

$$A_4=2^{\alpha+p+\tilde{\theta}+1}/(\lambda_1\eta_1)+\{2(1+\tilde{\theta})/[e\lambda_1\eta_1(1-2^{\alpha+p-1})]\}^{(1+\tilde{\theta})/(1+\alpha+p)}+1.$$

This together with (4.17) gives our desired conclusion (4.14) with the constant $C'=A_3A_4$. □

5. Estimating total error

Putting the bound in Theorem 2 into error decomposition (3.3), we can estimate the excess generalization error $\mathcal{E}(f)-\mathcal{E}(f_{\rho,\tau})$ and then convergence of the learning algorithm after applying a comparison theorem for ρ satisfying the following condition introduced in Steinwart and Christmann (2011).

Definition 6. Let $0<\varphi\le\infty$ 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 $\{b_\tau w_\tau^{\xi-1}\}^{-1}\in L^{\varphi}_{\rho_X}$ and for any $x\in X$ and $w\in(0,w_\tau(x)]$, there hold

$$\rho_x(\{y: f_{\rho,\tau}(x)<y<f_{\rho,\tau}(x)+w\})\ge b_\tau(x)w^{\xi-1} \tag{5.1}$$

and

$$\rho_x(\{y: f_{\rho,\tau}(x)-w<y<f_{\rho,\tau}(x)\})\ge b_\tau(x)w^{\xi-1}. \tag{5.2}$$

In our analysis we shall make use of the following comparison theorem (Steinwart and Christmann, 2011).

Lemma 4. *Let $0<\varphi\le\infty$ and $\xi>1$. Denote $r=\varphi\xi/(\varphi+1)>0$. If the measure ρ has a τ-*quantile *of φ-average type ξ, then for any measurable function f on X, we have*

$$\|f-f_{\rho,\tau}\|_{L^r_{\rho_X}}\le 2^{1-1/\xi}\xi^{1/\xi}\|\{b_\tau w_\tau^{\xi-1}\}^{-1}\|^{1/\xi}_{L^{\varphi}_{\rho_X}}\{\mathcal{E}(f)-\mathcal{E}(f_{\rho,\tau})\}^{1/\xi}. \tag{5.3}$$

Now we can present our total error estimate for the convergence of online algorithms (1.6) in a general form.

Theorem 3. *Let $0<\varphi\le\infty$ 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 conditions of Theorem 2, we have*

$$\mathbb{E}_{z_1,\dots,z_T}\|f_{T+1}-f_{\rho,\tau}\|_{L^r_{\rho_X}}\le C^*\|\{b_\tau w_\tau^{\xi-1}\}^{-1}\|^{1/\xi}_{L^{\varphi}_{\rho_X}}T^{-\theta}, \tag{5.4}$$

where θ is given by

$$\theta=\min\left\{\frac{\theta^*}{2\xi},\frac{\beta s-p}{\xi},\frac{p\gamma}{\xi}\right\} \tag{5.5}$$

with θ^* given by (4.15) and C^* is a constant independent of T or τ.

Proof. Applying Theorem 2 and (3.4) to the error decomposition expression (3.3), we see that

$$\mathbb{E}_{z_1,\dots,z_T}\{\mathcal{E}(f_{T+1})-\mathcal{E}(f_{\rho,\tau})\}\le\kappa\sqrt{C'}T^{-\theta^*/2}+\frac{C_\rho\kappa^2\epsilon_1^s}{\lambda_1}T^{p-\beta s}+\mathcal{D}_0\lambda_1^{\gamma}T^{-p\gamma}.$$

This together with Lemma 4 yields our desired conclusion where $C^*=2^{1-1/\xi}\xi^{1/\xi}(\kappa\sqrt{C'}+C_\rho\kappa^2\epsilon_1^s/\lambda_1+\mathcal{D}_0\lambda_1^{\gamma})^{1/\xi}$ is a constant independent of T or τ. □

We are in a position to prove Theorem 1 stated in Section 2.

Proof of Theorem 1. We apply Theorem 3. Let us first determine the power parameters in the expression (2.5). By requiring half of the last term $\alpha-p$ in definition (4.15) for θ^* equal to the term $p\gamma$ involved in (5.5), we choose $\alpha=(1+2\gamma)p$. By further requiring the first term $2+p\gamma-3p-2\alpha$ and the last term $\alpha-p$ in (4.15) to be equal, we take

$$p=\frac{2}{5+5\gamma},\quad \alpha=\frac{2+4\gamma}{5+5\gamma}.$$

These are the power parameters set in the expression (2.5). When β satisfies (2.6), we see that the middle two terms of (4.15) are at least $\alpha-p$, and that the middle term of (5.5) is not less than the last term. Therefore, we have $\theta^*=4\gamma/(5+5\gamma)$ and $\theta=2\gamma/((5+5\gamma)\xi)$.

Next we check conditions of Theorem 3. From the choice of p, we see that $p<\frac{2}{5}$. The restriction (2.6) on β tells us that $((\beta+1)/3)s\ge\frac{2}{5}$. So (4.12) holds true.

Obviously, $p<\alpha$. By the first restriction in (2.6) we find $\beta\ge p\gamma$. The second restriction $\beta\ge 6/5s-1$ yields $(\beta+1)s-2p\ge(2+6\gamma)/(5+5\gamma)$. Also, $(2-3p+p\gamma)/2=(2+6\gamma)/(5+5\gamma)$. Hence (4.13) is also valid. Thus we can apply Theorem 3 and find that (5.4) holds with the power exponent θ given by (5.5).

Finally we determine ξ. By condition (2.7), we know that ρ has a τ-quantile of ∞-average type $\xi=2$ with two positive constant functions w_τ and b_τ on X and $\{b_\tau w_\tau^{\xi-1}\}^{-1}\in L^{\infty}_{\rho_X}=1/b_\tau w_\tau$. Moreover, r=2. Therefore, $\theta=\gamma/(5+5\gamma)$, and our desired bound (5.4) is proved. □

6. Examples and numerical simulation

In this section we discuss two concrete examples. The first example stated in Section 2 is proved here by applying Theorem 3.

Proof of Example 1. First we find an expression for the τ-quantile regression function. From (2.11), we see that

$$\rho_x(\{y \in Y : y \le f_\rho(x)+u\}) = \frac{1+\mathrm{sgn}(u)|u|^{\zeta+1}}{2}, \quad -1 \le u \le 1. \tag{6.1}$$

It follows that

$$f_{\rho,\tau}(x) = f_\rho(x)+\mathrm{sgn}(2\tau-1)|2\tau-1|^{1/(\zeta+1)}, \quad 0<\tau<1, \ x \in X.$$

Second we check the condition in Definition 6. For $\tau \ne \frac{1}{2}$, we take $\varphi=\infty$, $\xi=2$ and

$$w_\tau(x) \equiv \min\left\{|2\tau-1|^{1/(\zeta+1)}/2, (1-|2\tau-1|^{1/(\zeta+1)})/2\right\}, \quad b_\tau(x) \equiv \frac{|2\tau-1|^{\zeta/(\zeta+1)}}{2^{\zeta+1}}.$$

For $\frac{1}{2}<\tau<1$ and $0<w\le w_\tau(x)$, we have

$$\rho_x(\{y : f_{\rho,\tau}(x) < y < f_{\rho,\tau}(x)+w\}) = \frac{((2\tau-1)^{1/(\zeta+1)}+w)^{\zeta+1}}{2} - \frac{2\tau-1}{2} \ge \frac{(2\tau-1)^{\zeta/(\zeta+1)}}{2} w$$

and

$$\rho_x(\{y : f_{\rho,\tau}(x)-w < y < f_{\rho,\tau}(x)\}) = \frac{2\tau-1}{2} - \frac{((2\tau-1)^{1/(\zeta+1)}-w)^{\zeta+1}}{2} \ge \frac{(2\tau-1)^{\zeta/(\zeta+1)}}{2^{\zeta+1}} w.$$

Hence both (5.1) and (5.2) hold. The same is true when $0<\tau<\frac{1}{2}$.

When $\tau=\frac{1}{2}$, we see from (6.1) that (5.1) and (5.2) hold by taking $\varphi=\infty, \xi=\zeta+2, w_\tau(x)\equiv 1$ and $b_\tau(x)\equiv\frac{1}{2}$. Therefore, the measure ρ has a τ-quantile of ∞-average type ξ where

$$\xi = r = \begin{cases} 2 & \text{if } \tau \ne \dfrac{1}{2}, \\ \zeta+2 & \text{if } \tau = \dfrac{1}{2}. \end{cases}$$

Third we verify conditions (2.3) and (2.4). Since the density function of ρ_x is uniformly bounded by $(\zeta+1)/2$, we know that the family of conditional distributions $\{\rho_x\}_{x\in X}$ is Lipschitz-1 and (2.3) is satisfied with $s=1$ and $C_\rho=(\zeta+1)/2$. By our assumption, both f_ρ and the constant 1 function lie in $\mathcal{H}_K$. So for each τ, $f_{\rho,\tau} \in \mathcal{H}_K$ and (2.4) is valid with $\gamma=1$ and $\mathcal{D}_0 = \|f_\rho+\mathrm{sgn}(2\tau-1)|2\tau-1|^{1/(\zeta+1)}\|_K^2/2$.

Finally we draw our conclusion.

When $\tau\ne\frac{1}{2}$, our conclusion follows directly from Theorem 1. When $\tau=\frac{1}{2}$, our conclusion can be proved in the same way as for proving Theorem 1 by applying Theorem 3. The only difference is to replace $\xi=2$ by $\xi=2+\zeta$ and θ in (5.5) becomes $1/5(2+\zeta)$. This completes the proof of Example 1. □

Example 1 requires the condition that $\mathcal{H}_K$ contains the constant 1 function. It is not satisfied when K is a Gaussian kernel $K(x,u)=\exp\{-|x-u|^2/2\sigma^2\}$ on $X\subset\mathbb{R}^n$ with variance $\sigma^2>0$, though the condition is valid if we revise the kernel by adding a constant $K(x,u)=\exp\{-|x-u|^2/2\sigma^2\}+1$.

In the next example, we compare the learning rates derived in this paper with that from Smale and Yao (2006) for the purpose of least squares regression corresponding to the special case $\tau=1/2$.

Example 2. Let $X=[0,1]^{10}$, ρ_X be the Lebesgue measure on $[0,1]^{10}$, and for each $x\in X$, the conditional distribution ρ_x is noised by the uniform distribution on $[-0.5,0.5]$ around the regression function value

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

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)

Table 2
Numerical results.

Parameters	$T=3000$	$T=10{,}000$	$T=3000$	$T=10{,}000$
τ	0.5	0.5	0.95	0.95
β	0.44	0.40	0.86	0.88
γ	0.40	0.45	0.07	0.12
ϵ_1	2.1	7.1	6.9	7.0
η_1	1.6	2.1	4.6	6.0
λ_1	0.004	0.002	0.001	0.004
Error*	0.05957789	0.04631984	0.09879489	0.05867203

where the parameters are prescribed in Table 1 as in Guo and Zhou (2012). We take the Gaussian kernel $K(x,u)=\exp\{-|x-u|^2/2\sigma^2\}$ with variance $\sigma^2=0.6^2$. When $\tau=\frac{1}{2}$, conditions (2.3) with $s=1$ and (2.4) with $\gamma=1$ are valid. Also, the measure ρ has a $\frac{1}{2}$-quantile of ∞-average type 2. If the parameters $\eta_t,\lambda_t,\epsilon_t$ take the form (4.11) with $\alpha=0.6,\beta=0.4,p=0.2$, then restrictions (4.12) and (4.13) are satisfied, and by Theorem 3 we have

$$\mathbb{E}_{z_1,\ldots,z_T}\|f_{T+1}-f_\rho\|_{L^2_{\rho_X}} \le C^* T^{-1/10}, \tag{6.2}$$

where C^* is a constant independent of T. When the least squares loss is used, the online algorithm becomes

$$f_{t+1}=f_t-\eta_t\{(f_t(x_t)-y_t)K_{x_t}+\lambda_t f_t\}.$$

The optimized learning rate that can be derived from Smale and Yao (2006) is $O(T^{-1/9})$. This is slightly better than our learning rate (6.2).

For this example, we have conducted a numerical simulation to show the behavior of the online learning algorithm when the sample size T and the parameter τ change. Here the parameters $\eta_t,\lambda_t,\epsilon_t$ take the form (4.11) with $\alpha=(2+4\gamma)/(5+5\gamma)$, $p=2/(5+5\gamma)$ and the other parameters determined by coordinate descent methods as in Table 2. The learning error is estimated empirically by independently drawing another unlabelled sample set $\{\xi_j\}$ uniformly on $[0,1]^{10}$ of size 12,000 and computing

$$\text{Error}^*=\left(\frac{1}{12{,}000}\sum_{j=1}^{12{,}000}(f_{T+1}(\xi_j)-f_{\rho,\tau}(\xi_j))^2\right)^{1/2}.$$

Acknowledgments

The work described in this paper is supported by National Science Foundation of China under Grant 11001247 and by a Grant from the Research Grants Council of Hong Kong [Project No. CityU 103709].

References

Chen, D., Wu, Q., Ying, Y.M., Zhou, D.X., 2004. Support vector machine soft margin classifiers: error analysis. Journal of Machine Learning Research 5, 1143–1175.

Cucker, F., Zhou, D.X., 2007. Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge.

Guo, X., Zhou, D.-X., 2012. An empirical feature-based learning algorithm producing sparse approximations. Applied and Computational Harmonic Analysis 32, 389–400.

Hu, T., 2011. Online regression with varying Gaussians and non-identical distributions. Analysis and Applications 9, 395–408.

Hu, T., Zhou, D.X., 2009. Online learning with samples drawn from non-identical distributions. Journal of Machine Learning Research 10, 2873–2898.

Hwang, C., Shim, J., 2005. A simple quantile regression via support vector machine. In: Advances in Natural Computation: First International Conference (ICNC). Springer, pp. 512–520.

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. Annals of Mathematical Statistics 23, 462–466.

Kivinen, J., Smola, A.J., Williamson, R.C., 2004. Online learning with kernels. IEEE Transactions on Signal Processing 100, 2165–2176.

Koenker, R., 2005. Quantile Regression. Cambridge University Press.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Rosset, S., 2009. Bi-level path following for cross validated solution of kernel quantile regression. Journal of Machine Learning Research 10, 2473–2505.

Schoelkopf, B., Smola, A., Williamson, R., Bartlett, P., 2000. New support vector algorithms. Neural Computation 12, 1207–1245.

Smale, S., Yao, Y., 2006. Online learning algorithms. Foundations of Computational Mathematics 6, 145–170.

Smale, S., Zhou, D.X., 2003. Estimating the approximation error in learning theory. Analysis and Applications 1, 17–41.

Smale, S., Zhou, D.X., 2009. Online learning with Markov sampling. Analysis and Applications 7, 87–113.

Steinwart, I., Christmann, A., 2011. Estimating conditional quantiles with the help of the pinball loss. Bernoulli 17, 211–225.

Steinwart, I., Christmann, A., 2008. Support Vector Machines. Springer, New York.

Steinwart, I., Christmann, A., 2008. How support vector machines can estimate quantiles and the median. In: Advances in Neural Information Processing Systems, vol. 20. MIT Press, Cambridge, MA, pp. 305–312.

Steinwart, I., Christmann, A., 2009. Sparsity of SVMs that use the ϵ-insensitive loss. In: Advances in Neural Information Processing Systems, vol. 21. MIT Press, Cambridge, MA, pp. 1569–1576.

Takeuchi, I., Le, Q.V., Sears, T.D., Smola, A.J., 2006. Nonparametric quantile estimation. Journal of Machine Learning Research 7, 1231–1264.

Tong, H.Z., Chen, D.R., Peng, L.Z., 2009. Analysis of support vector machines regression. Foundations of Computational Mathematics 9, 243–257.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

Wu, Q., Ying, Y., Zhou, D.X., 2007. Multi-kernel regularized classifiers. Journal of Complexity 23, 108–134.

Wu, Q., Zhou, D.X., 2008. Analysis of support vector machine classification. Journal of Computational Analysis and Applications 8, 99–119.

Xiang, D.H., Conditional quantiles with varying Gaussians. Advances in Computational Mathematics, http://dx.doi.org/10.1007/s10444-011-9257-5, in press.

Xiang, D.H., Hu, T., Zhou, D.X., 2011. Learning with varying insensitive loss. Applied Mathematics Letters 24, 2107–2109.

Yao, Y., 2010. On complexity issue of online learning algorithms. IEEE Transactions on Information Theory 56, 6470–6481.

Yao, Y., Rosasco, L., Caponnetto, A., 2007. On early stopping in gradient descent learning. Constructive Approximation 26, 289–315.

Ye, G.B., Zhou, D.X., 2007. Fully online classification by regularization. Applied and Computational Harmonic Analysis 23, 198–214.

Ying, Y., Zhou, D.X., 2006. Online regularized classification algorithms. IEEE Transactions on Information Theory 52, 4775–4788.