AIMS Mathematics, 7(11): 20066–20084.
DOI:10.3934/math.20221098
Received: 19 July 2022
Revised: 27 August 2022
Accepted: 06 September 2022
Published: 13 September 2022

http://www.aimspress.com/journal/Math

Research article

Convergence of online learning algorithm with a parameterized loss

Shuhua Wang*

School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen, China

* **Correspondence:** Email: w614sh@126.com.

Abstract: The research on the learning performance of machine learning algorithms is one of the important contents of machine learning theory, and the selection of loss function is one of the important factors affecting the learning performance. In this paper, we introduce a parameterized loss function into the online learning algorithm and investigate the performance. By applying convex analysis techniques, the convergence of the learning sequence is proved and the convergence rate is provided in the expectation sense. The analysis results show that the convergence rate can be greatly improved by adjusting the parameter in the loss function.

Keywords: online learning; parameterized loss; convergence rate; reproducing Hilbert space; convex analysis
Mathematics Subject Classification: 41A25, 68Q32, 68T40, 90C25

1. Introduction

Let $X \subset \mathbf{R}^d$ be the input space , $Y = [-M, M]$ be the output space for some $M > 0$. $z = \{(x_t, y_t)\}_{t=1}^T$ are the random samples i.i.d. (independently and identically drawn) according to a Borel probability measure $\rho(x, y) = \rho(y|x)\rho_X(x)$ on $Z = X \times Y$. Based on these samples, the goal of regression problems is to look for a predictor $f : X \to \mathbf{R}$ from some hypothesis space such that $f(x)$ is a "good" approximation of y. The quality of the predictor f is measured by the generalization error

$$\mathcal{E}(f) := \int_Z V(x, y, f) d\rho(x, y),$$

where $V(r) : \mathbf{R} \to \mathbf{R}_+$ is a prescribed loss function.

The hypothesis space considered in this paper is the reproducing kernel Hilbert space(RKHS) $\mathcal{H}_K$. This means that there exists a unique symmetric and positive definite continuous function $K : X \times X \to \mathbf{R}$, called the reproducing kernel of $\mathcal{H}_K$, or Mercer kernel, and an inner product $\langle \cdot, \cdot \rangle_K$ such that $f(x) = \langle K(x, \cdot), f \rangle_K$ which is the reproducing property of the kernel, and all $f \in \mathcal{H}_K$ are linear

combinations of kernel functions. In other words, the RKHS $\mathcal{H}_K$ is the closure of the linear span of the set of functions $\{K_x(\cdot) = K(x,\cdot) : x \in X\}$ with the inner product $\langle\cdot,\cdot\rangle_K$. For each $x \in X$ and $f \in \mathcal{H}_K$ the evaluation functional $e_x(f) := f(x)$ is continuous (i.e.bounded) in the topology of $\mathcal{H}_K$, and $|f(x)| \le \kappa\|f\|_K$ with $\kappa := \sup_{x\in X}\sqrt{K(x,x)}$ (see [1]).

Traditional off-line learning is also called batch learning, all sample points need to be tested in each training. When the amount of data is large or new sample points are added, the learning ability of batch learning decreases significantly. Online learning is one effective approach raised for analyzing and processing big data in various applications, such as communication, electronics, medicine, biology, and other fields (see e.g., [2–6]). The performance of the kernel-based regularized online learning algorithms have been researched and their effectiveness has been verified (see e.g., [7–11] and references therein). Unlike the off-line learning algorithms, online learning algorithms process the observations one by one, and the output is adjusted in time according to the results of the last learning.

With the observations $z = \{(x_t, y_t)\}_{t=1}^T$, the kernel regularized online learning algorithm based on the stochastic descent method is given by (see e.g., [8–10])

$$\begin{cases} f_1 = 0, \\ f_{t+1} = f_t - \eta_t(\nabla_f V(x_t, y_t, f_t)K_{x_t} + \lambda f_t), \end{cases} \tag{1.1}$$

where η_t is called the stepsize, $\lambda > 0$ is a regularization parameter and the sequence $\{f_t : t = 1, \ldots, T+1\}$ is the learning sequence.

When the least-square loss function $V(x, y, f(x)) = (f(x) - y)^2$ is selected, the specific iteration format of the online learning algorithm is given by

$$\begin{cases} f_1 = 0, \\ f_{t+1} = f_t - \eta_t((f_t(x_t) - y_t)K_{x_t} + \lambda f_t). \end{cases} \tag{1.2}$$

To study the learning performance of online learning algorithms we need to bound the convergence rate of the iterative sequence $\{f_t : t = 1, \cdots, T+1\}$. The convergence of the online learning algorithm (1.2) has been extensively studied in the literature (see e.g., [8, 9, 12]). The research results in [12] show that under mild conditions the regularized online learning algorithm can converge comparably fast as the off-line learning algorithm.

The least square-loss is the most widely utilized criterion for regression in practice. However, from a robustness point of view, the least square loss is not a good choice. In many practical applications, outliers or heavy-tailed distributions often occur in real data sets. In recent years, how to improve the robustness of algorithms has become one of the hot topics in the field of machine learning. In the literature on learning theory, a lot of efforts have been made and there have been some results on generalization analysis (see e.g., [13–17]) and empirical experiments (see e.g., [18, 19]) of learning algorithms when outliers or heavy-tailed noise are allowed.

One of the main strategies to improve robustness is to use some robust loss function with a scale parameter(see e.g., [20, 21]). Based on the quadratic function $\sqrt{1+t^2}, t \in \mathbf{R}$ which plays an important role in constructing shape preserving quasi-interpolation and solving partial differential equations with mesh-free method since its strong nonlinear property and its convexity, [21] constructed a robust loss function $V_\sigma(r)$ with a scale parameter $\sigma \in (0, 1]$. For $\sigma \in (0, 1]$, the parameterized loss function

$V_\sigma(r) : \mathbf{R} \to [0, \infty)$ is defined as

$$V_\sigma(r) := \sigma^2\left(\frac{\sqrt{\sigma^2 + r^2}}{\sigma} - 1\right), \quad r \in \mathbf{R}.$$

The analysis results in [21] show that the learning algorithm based on the parameterized loss function $V_\sigma(r)$ has a good generalization ability.

Encouraged by these researches, we want to further improve the performance and applicability of the online algorithm. In this paper, we introduce the parameterized loss function $V_\sigma(r)$ into the online learning algorithm, and analyze the influence of the scale parameter on the convergence rate of the algorithm.

To give the specific format of the learning algorithm with the parameterized loss function, we give the following notations correspondingly. We denote

$$\mathcal{E}_\sigma(f) = \int_Z V_\sigma(y - f(x))d\rho(x, y), \tag{1.3}$$

and

$$f_\sigma = \arg\min_{f \in L^2(\rho_X)} \mathcal{E}_\sigma(f). \tag{1.4}$$

The kernel-based regularized off-line algorithm is the following optimization problem:

$$f_\lambda^\sigma = \arg\min_{f \in \mathcal{H}_K} \mathcal{E}_\sigma(f) + \frac{\lambda}{2}\|f\|_K^2. \tag{1.5}$$

Based on the random observations $z = \{(x_t, y_t)\}_{t=1}^T$, the approximate solution of the problem (1.5) is obtained through the following learning process

$$f_{z,\lambda}^\sigma = \arg\min_{f \in \mathcal{H}_K} \frac{1}{|T|}\sum_{t=1}^{T} V_\sigma(y_t - f(x_t)) + \frac{\lambda}{2}\|f\|_K^2.$$

It is easy to see that the gradient of the loss function V_σ is given by

$$\nabla_f V_\sigma(y_t - f_t(x_t)) = -\frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}.$$

Along with the online algorithm (1.1), the kernel regularized online learning algorithm with the parameterized loss function $V_\sigma(r)$ is defined by

$$\begin{cases} f_1 = 0, \\ f_{t+1} = f_t - \eta_t(-\frac{y_t - f_t(x_t)}{\sqrt{1+(\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t} + \lambda f_t). \end{cases} \tag{1.6}$$

In this paper, we focus on the performance of the sequence $\{f_t : t = 1, ..., T + 1\}$ produced by the algorithm (1.6).

The remaining parts of this paper are organized as follows: we present the main results of this paper in Section 2. The proofs of the main results are given in Section 3. Discussions and comparisons with related works are given in Section 4. Conclusions and some questions for further study are mentioned in Section 5.

In the present paper, we write $A = O(B)$ if there is a constant $C \geq 0$ such that $A \leq CB$. We use $\mathbb{E}_z[\cdot]$ to denote the expectation with respect to z. When its meaning is clear from the context, we use the shorthand notation $\mathbb{E}[\cdot]$.

2. The main results

In this section, we present our main results about the performance of the algorithm (1.6), proofs are given in Section 3.

2.1. The convergence of the learning sequence

Our first main result establishes the convergence of the sequence $\{f_t : t = 1, ..., T+1\}$ in expectation, under mild conditions on the stepsize.

Theorem 2.1. *Let f_λ^σ be defined as in (1.5), and let $\{f_t : t = 1, ..., T+1\}$ be the sequence produced by the algorithm (1.6), $\lambda > 0$. If $\{\eta_t\}$ satisfies $\sum_{t=1}^{\infty} \eta_t = +\infty$, $\lim_{t\to+\infty} \eta_t = 0$ and $\eta_t \le \frac{1}{\lambda}$, then it holds that*

$$\mathbb{E}_{z_1,...,z_T}\left[\|f_{T+1} - f_\lambda^\sigma\|_K\right] \to 0 \quad (T \to +\infty).$$

2.2. Error analysis

Our second main result gives the explicit convergence rate of the last iterate by specifying the step sizes in the algorithm (1.6).

Theorem 2.2. *Let f_λ^σ be defined as in (1.5), and let $\{f_t : t = 1, ..., T+1\}$ be the sequence produced by the algorithm (1.6). For any $0 < \lambda \le 1$, $0 < \theta \le 1$, take $\eta_t = \frac{1}{C}t^{-\theta}$ with $C \ge \lambda + 4(\kappa^2+1)$. Then, it holds that*

$$\mathbb{E}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right] \le \begin{cases} \left(D_\sigma(\lambda) + \frac{9C_\sigma T^{1-\theta}}{C^2(1-\theta)2^{1-\theta}}\right)\exp\left\{-\frac{\lambda(1-2^{\theta-1})}{2C(1-\theta)}(T+1)^{1-\theta}\right\} + \frac{36C_\sigma}{\lambda C}T^{-\theta}, & if\ 0 < \theta < 1, \\ \left(D_\sigma(\lambda) + \frac{5C_\sigma T^{1-\theta}}{C^2(C-\lambda)}\right)T^{-\frac{\lambda}{2C}}, & if\ \theta = 1, \end{cases}$$

where $D_\sigma(\lambda) = \frac{2M\sigma}{\lambda}$, $C_\sigma = 4M\sigma(\kappa^2+1)$.

Corollary 2.1. *Let f_λ^σ be defined as in (1.5), and let $\{f_t : t = 1, ..., T+1\}$ be the sequence produced by the algorithm (1.6). For any $\theta \in (0,1)$, $0 < \alpha \le \min\{1-\theta, \theta\}$, take $\lambda = T^{-\alpha}$. If the stepsize is chosen as $\eta_t = \frac{1}{C}t^{-\theta}$, then it holds that*

$$\mathbb{E}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right] \le C_{M,C,\kappa,\theta} \times \frac{\sigma}{T^{\theta-\alpha}}, \tag{2.1}$$

where $C_{M,C,\kappa,\theta}$ is a constant depending only on θ, κ, C and M.

Remark 1. The results given in Theorem 2.2 and Corollary 2.1 show that the scale parameter σ can effectively control the convergence rate of $\|f_{T+1} - f_\lambda^\sigma\|_K$, which is usually referred to as the sample error. Depending on the circumstances, the sample error bound can be greatly improved by choosing the parameter σ properly. In fact, take $\sigma = \lambda = T^{-\alpha}$. Then by (2.1), we have $\mathbb{E}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right] = O(T^{-\theta})$ which is better than the sample error bound $O(T^{-(\theta-\alpha)})$ given in [9].

The results provided above mainly describe the convergence rate of the sample error. However, in the studying the learning performance of learning algorithms, we are often interested in the excess generalization error $\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma)$. Define

$$\mathcal{K}(f,\lambda) = \inf_{g\in\mathcal{H}_\mathcal{K}}\left\{\mathcal{E}_\sigma(g) - \mathcal{E}_\sigma(f) + \frac{\lambda}{2}\|g\|_K^2\right\},$$

which is often used to denote the approximation error, whose convergence is determined by the capacity of $\mathcal{H}_K$. We assume the $\mathcal{K}$-functional satisfies the following decay

$$\mathcal{K}(f_\sigma, \lambda) = O(\lambda^\beta), \quad \lambda \to 0^+, \tag{2.2}$$

with $0 < \beta \leq 1$.

By combining the sample error with the approximation error, we obtain the overall learning rate stated as follows.

Corollary 2.2. *Let $\mathcal{E}_\sigma(f)$ be the generalization error defined as in (1.3), $\{f_t : t = 1, ..., T+1\}$ be the sequence produced by the algorithm (1.6). For $\theta \in (0, 1), 0 < \alpha \leq \min\{1-\theta, \theta\}$, take $\lambda = T^{-\alpha}$. If the stepsize is chosen as $\eta_t = \frac{1}{C} t^{-\theta}$, then it holds that*

$$\mathbb{E}\left[\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma)\right] = O\left(\frac{\sigma^{\frac{3}{2}}}{T^{\frac{\theta-\alpha}{2}}} + \frac{1}{T^{\beta\alpha}}\right). \tag{2.3}$$

Now, we compare the performance of the algorithm (1.6) with that of the online kernel regularized learning algorithm based on the least square loss.

In [22], the online kernel regularized learning algorithm with the least-square loss is researched, and the learning rates are established under some assumptions. Namely, for $0 < \beta \leq 1$, $0 < \delta < \frac{\beta}{\beta+1}$, there holds

$$\mathbb{E}\left[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho)\right] = O(T^{\delta - \frac{\beta}{\beta+1}}) \tag{2.4}$$

and for $\frac{1}{2} < \beta \leq 1$, $0 < \delta < \frac{2\beta-1}{2\beta+1}$, there holds

$$\mathbb{E}\left[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho)\right] = O(T^{\delta - \frac{2\beta-1}{2\beta+1}}). \tag{2.5}$$

For $0 < \beta \leq 1$, we choose $\sigma = T^{-\frac{\mu}{3}}$ with $0 < \mu \leq \frac{2\beta-1}{2\beta+1}$. Take $\lambda = T^{\frac{\delta}{2\beta} - \frac{1}{2(\beta+1)} - \frac{\mu}{2\beta+1}}$ and $\eta_t = \frac{1}{4\kappa^2+5} t^{\frac{2\beta+1}{2\beta}\delta - \frac{2\beta+1}{2(\beta+1)}}$ with $\frac{\beta}{\beta+1} - \frac{2\beta}{2\beta+1}\mu < \delta < \frac{\beta}{\beta+1} + \frac{2\beta}{2\beta+1}$. By Corollary 2.2, we have

$$\mathbb{E}\left[\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma)\right] = O(T^{\frac{1}{2}(\delta - \frac{\beta}{\beta+1}) - \frac{\beta}{2\beta+1}\mu}). \tag{2.6}$$

And for $\frac{1}{2} < \beta \leq 1$, we choose $\sigma^3 = \lambda = T^{\frac{2\beta}{2\beta-1}\delta - \frac{2\beta}{2\beta+1}}$ with $0 < \delta < \frac{2\beta-1}{2\beta+1}$. Take $\eta_t = \frac{1}{4\kappa^2+5} t^{\frac{2\beta}{2\beta-1}\delta - \frac{2\beta}{2\beta+1}}$. By Corollary 2.2, we have that

$$\mathbb{E}\left[\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma)\right] = O(T^{\frac{\beta}{2\beta-1}\delta - \frac{1}{2\beta+1}}). \tag{2.7}$$

The rate (2.6) is better than (2.4), and (2.7) is better than (2.5). The analysis results illustrate that the convergence rate of the algorithm (1.6) can be improved by choosing the parameter σ appropriately.

3. Proofs

Lemma 3.1. *Let $\{f_t : t = 1, ..., T+1\}$ be the sequence produced by the algorithm (1.6). If $\eta_t \leq \frac{1}{\lambda}$, then, for any $t = 1, ..., T+1$, it holds that*

$$\|f_t\|_K \leq \frac{\kappa\sigma}{\lambda}. \tag{3.1}$$

Proof. We prove (3.1) by induction on t. The inequality (3.1) is true for $t = 1$ because of the initialization condition $f_1 = 0$. Suppose the bound (3.1) holds for any t, and consider f_{t+1}. The iteration (1.6) can be rewritten as

$$\begin{aligned} f_{t+1} &= f_t + \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}\eta_t K_{x_t} - \lambda\eta_t f_t \\ &= (1 - \lambda\eta_t)f_t + \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}\eta_t K_{x_t}. \end{aligned} \tag{3.2}$$

This implies that

$$\|f_{t+1}\|_K \le (1 - \lambda\eta_t)\|f_t\|_K + \kappa\eta_t\Big|\frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}\Big|. \tag{3.3}$$

Since

$$\Big|\frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}\Big| \le \sigma. \tag{3.4}$$

Combined with the assumption $\|f_t\|_K \le \frac{\kappa\sigma}{\lambda}$, we have

$$\begin{aligned} \|f_{t+1}\|_K &\le (1 - \lambda\eta_t)\|f_t\|_K + \kappa\sigma\eta_t \\ &\le (1 - \lambda\eta_t)\frac{\kappa\sigma}{\lambda} + \kappa\sigma\eta_t \\ &= \frac{\kappa\sigma}{\lambda}. \end{aligned} \tag{3.5}$$

This completes the proof. □

Lemma 3.2. *Assume f_λ^σ is defined as in (1.5). Then, for any $f \in \mathcal{H}_K$, it holds that*

$$\int_Z \frac{y - f_\lambda^\sigma(x)}{\sqrt{1 + (\frac{y - f_\lambda^\sigma(x)}{\sigma})^2}}(f_\lambda^\sigma(x) - f(x))d\rho = \lambda\langle f_\lambda^\sigma, f_\lambda^\sigma - f\rangle_K.$$

Proof. By Taylor formula, for any $u, v \in \mathbf{R}$, we have

$$V_\sigma(u) - V_\sigma(v) = V_\sigma'(v)(u - v) + \frac{1}{2}V_\sigma''(\xi)(u - v)^2, \tag{3.6}$$

where $\xi \in \mathbf{R}$ is between u and v.

Note that, $V_\sigma''(\xi) = \frac{1}{\sqrt{\left(1+(\frac{\xi}{\sigma})^2\right)^3}} > 0$. Then,

$$V_\sigma(u) - V_\sigma(v) \ge V_\sigma'(v)(u - v) = \frac{v(u - v)}{\sqrt{1 + (\frac{v}{\sigma})^2}}. \tag{3.7}$$

Therefore, for any $f, g \in \mathcal{H}_K$, we have

$$\begin{aligned}\mathcal{E}_\sigma(f) - \mathcal{E}_\sigma(g) &= \int_Z V_\sigma(y - f(x))d\rho - \int_Z V_\sigma(y - g(x))d\rho \\ &\geq -\int_Z \frac{y - g(x)}{\sqrt{1 + (\frac{y-g(x)}{\sigma})^2}}(f(x) - g(x))d\rho \\ &= \langle f - g, -\int_Z \frac{y - g(x)}{\sqrt{1 + (\frac{y-g(x)}{\sigma})^2}} K_x d\rho\rangle_K \\ &= \langle f - g, \nabla_g \mathcal{E}_\sigma(g)\rangle_K. \end{aligned} \tag{3.8}$$

By (i) of Lemma 5.1 in [23], we know $\mathcal{E}_\sigma(f)$ is a convex function on $\mathcal{H}_K$. And $\|f\|_K^2$ is a strictly convex function on $\mathcal{H}_K$. Then, $\Omega_\sigma(f) = \mathcal{E}_\sigma(f) + \frac{\lambda}{2}\|f\|_K^2$ is a convex function on $\mathcal{H}_K$. Based on (ii) of Lemma 5.1 in [23], we have

$$\begin{aligned} 0 &= \nabla_f \Omega_\sigma(f) \mid_{f=f_\lambda^\sigma} \\ &= \nabla_f \mathcal{E}_\sigma(f) \mid_{f=f_\lambda^\sigma} + \lambda f_\lambda^\sigma \\ &= -\int_Z \frac{y - f_\lambda^\sigma(x)}{\sqrt{1 + (\frac{y-f_\lambda^\sigma(x)}{\sigma})^2}} K_x d\rho + \lambda f_\lambda^\sigma. \end{aligned} \tag{3.9}$$

Taking inner product with $f - f_\lambda^\sigma$ on both sides of the above formula, we get

$$\begin{aligned} 0 &= \langle -\int_Z \frac{y - f_\lambda^\sigma(x)}{\sqrt{1 + (\frac{y-f_\lambda^\sigma(x)}{\sigma})^2}} K_x d\rho + \lambda f_\lambda^\sigma, f - f_\lambda^\sigma\rangle_K \\ &= \int_Z -\frac{y - f_\lambda^\sigma(x)}{\sqrt{1 + (\frac{y-f_\lambda^\sigma(x)}{\sigma})^2}} \langle K_x, f - f_\lambda^\sigma\rangle_K d\rho + \lambda\langle f_\lambda^\sigma, f - f_\lambda^\sigma\rangle_K \\ &= \int_Z \frac{y - f_\lambda^\sigma(x)}{\sqrt{1 + (\frac{y-f_\lambda^\sigma(x)}{\sigma})^2}} (f_\lambda^\sigma(x) - f(x))d\rho + \lambda\langle f_\lambda^\sigma, f - f_\lambda^\sigma\rangle_K. \end{aligned}$$

This proves our conclusion. □

Lemma 3.3. *Let f_λ^σ be defined as in (1.5). For any $f \in \mathcal{H}_K$, we denote $\Omega_\sigma(f) = \mathcal{E}_\sigma(f) + \frac{\lambda}{2}\|f\|_K^2$. Then, it holds that*

$$\frac{\lambda}{2}\|f - f_\lambda^\sigma\|_K^2 \leq \Omega_\sigma(f) - \Omega_\sigma(f_\lambda^\sigma).$$

Proof. For any $f \in \mathcal{H}_K$, we define a function$f_{(\theta)} = f_\lambda^\sigma + \theta(f - f_\lambda^\sigma), \theta \in [0, 1]$. Then, $f_{(0)} = f_\lambda^\sigma$ and $f_{(1)} = f$. Denote $F(\theta) = \Omega_\sigma(f_{(\theta)}) = \int_Z V_\sigma(y - f_{(\theta)}(x))d\rho + \frac{\lambda}{2}\|f_{(\theta)}\|_K^2$, then $F(1) = \Omega_\sigma(f), F(0) = \Omega_\sigma(f_\lambda^\sigma)$. Since V_σ is differentiable, as a function of θ, $F(\theta)$ is differentiable. And for any $\theta \in [0, 1]$, we have

$$F'(\theta) = \lim_{\triangle\theta \to 0} \frac{F(\theta + \triangle\theta) - F(0)}{\triangle\theta}$$

$$
\begin{aligned}
&= \lim_{\triangle\theta\to 0}\frac{\Omega_\sigma(f_{\theta+\triangle\theta})-\Omega_\sigma(f_\theta)}{\triangle\theta}\\
&= \lim_{\triangle\theta\to 0}\frac{1}{\triangle\theta}\left(\int_Z\left(V_\sigma(y-f_{(\theta+\triangle\theta)}(x))-V_\sigma(y-f_{(\theta)}(x))\right)d\rho+\frac{\lambda}{2}\|f_{(\theta+\triangle\theta)}\|_K^2-\frac{\lambda}{2}\|f_{(\theta)}\|_K^2\right)\\
&= \lim_{\triangle\theta\to 0}\frac{1}{\triangle\theta}\Big(\int_Z\left(V_\sigma(y-f_{(\theta)}(x))-\triangle\theta(f(x)-f_\lambda^\sigma(x))-V_\sigma(y-f_{(\theta)}(x))\right)d\rho\\
&\qquad +\frac{\lambda}{2}(\|f_{(\theta)}+\triangle\theta(f-f_\lambda^\sigma)\|_K^2-\|f_{(\theta)}\|_K^2)\Big). \qquad (3.10)
\end{aligned}
$$

By the median value theorem, there holds

$$
V_\sigma\left(y-f_{(\theta)}(x)\right)-\triangle\theta(f(x)-f_\lambda^\sigma(x)))-V_\sigma(y-f_{(\theta)}(x))=\triangle\theta V_\sigma^{'}(\xi)(f_\lambda^\sigma(x)-f(x)), \qquad (3.11)
$$

where $\xi\in(y-f_{(\theta)}(x))-\triangle\theta(f(x)-f_\lambda^\sigma(x),y-f_{(\theta)}(x))$.

This in connection with $\|f_{(\theta)}+\triangle\theta(f-f_\lambda^\sigma)\|_K^2-\|f_{(\theta)}\|_K^2=2\triangle\theta\langle f-f_\lambda^\sigma,f_{(\theta)}\rangle_K+(\triangle\theta)^2\|f-f_\lambda^\sigma)\|_K^2$, according to (3.10), tells us that

$$
\begin{aligned}
F^{'}(\theta) &= \lim_{\triangle\theta\to 0}\Big(\int_Z V_\sigma^{'}(\xi)(f_\lambda^\sigma(x)-f(x))d\rho+\lambda\langle f-f_\lambda^\sigma,f_{(\theta)}\rangle_K+\frac{\lambda}{2}\triangle\theta\|f-f_\lambda^\sigma\|_K^2\Big)\\
&= \int_Z V_\sigma^{'}(y-f_{(\theta)}(x))(f_\lambda^\sigma(x)-f(x))d\rho+\lambda\langle f-f_\lambda^\sigma,f_{(\theta)}\rangle_K\\
&= \int_Z V_\sigma^{'}\left((y-f_\lambda^\sigma(x))+\theta(f_\lambda^\sigma(x)-f(x))\right)(f_\lambda^\sigma(x)-f(x))d\rho+\lambda\langle f-f_\lambda^\sigma,f+\theta(f-f_{(\theta)})\rangle_K\\
&= \int_Z V_\sigma^{'}\left((y-f_\lambda^\sigma(x))+\theta(f_\lambda^\sigma(x)-f(x))\right)(f_\lambda^\sigma(x)-f(x))d\rho\\
&\qquad +\lambda\langle f-f_\lambda^\sigma,f\rangle_K+\theta\lambda\|f-f_\lambda^\sigma)\|_K^2. \qquad (3.12)
\end{aligned}
$$

By Lemma 3.2, we see that

$$
\begin{aligned}
\lambda\langle f-f_\lambda^\sigma,f\rangle_K &= -\int_Z\frac{y-f_\lambda^\sigma(x)}{\sqrt{1+(\frac{y-f_\lambda^\sigma(x)}{\sigma})^2}}(f_\lambda^\sigma(x)-f(x))d\rho\\
&= -\int_Z V_\sigma^{'}(y-f_\lambda^\sigma(x))(f_\lambda^\sigma(x)-f(x))d\rho. \qquad (3.13)
\end{aligned}
$$

On the other hand, since $V_\sigma(u)$ is a convex function in $\mathbf{R}$, by discussions in [24], we know that

$$
V_\sigma^{'}(y-f_\lambda^\sigma(x))+\theta(f_\lambda^\sigma(x)-f(x))-V_\sigma^{'}(y-f_\lambda^\sigma(x))(f_\lambda^\sigma(x)-f(x))\geq 0. \qquad (3.14)
$$

Therefore, for $\theta\in(0,1)$, we have

$$
\begin{aligned}
F^{'}(\theta) &= \int_Z\Big(V_\sigma^{'}(y-f_\lambda^\sigma(x))+\theta(f_\lambda^\sigma(x)-f(x))-V_\sigma^{'}(y-f_\lambda^\sigma(x))\Big)(f_\lambda^\sigma(x)-f(x))d\rho+\lambda\theta\|f-f_\lambda^\sigma\|_K^2\\
&\geq \lambda\theta\|f-f_\lambda^\sigma\|_K^2. \qquad (3.15)
\end{aligned}
$$

By the definition of f_λ^σ, we know that $F(\theta)\geq F(0)=\Omega_\sigma(f_\lambda^\sigma),\theta\in[0,1]$. Therefore, (3.14) implies that

$$
\Omega_\sigma(f)-\Omega_\sigma(f_\lambda^\sigma) = F(1)-F(0)=\int_0^1 F^{'}(\theta)d\theta
$$

$$\begin{aligned} &\geq \int_0^1 \lambda\theta\|f - f_\lambda^\sigma\|_K^2 d\theta \\ &= \lambda\|f - f_\lambda^\sigma\|_K^2 \int_0^1 \theta d\theta \\ &= \frac{\lambda}{2}\|f - f_\lambda^\sigma\|_K^2. \end{aligned}$$

The proof is completed. □

Lemma 3.4. *Let f_λ^σ be defined as in (1.5), $\{f_t : t = 1, ..., T+1\}$ be the sequence produced by the algorithm (1.6), if $\lambda > 0$, $\eta_t \leq \frac{1}{\lambda}$, then*

$$\begin{aligned} &\mathbb{E}_{z_1,...,z_T}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right] \\ &\leq 4\kappa^2\sigma^2 \sum_{t=1}^T \eta_t^2 \prod_{j=t+1}^T (1 - \lambda\eta_j) + \frac{2M\sigma}{\lambda}\prod_{t=1}^T (1 - \lambda\eta_t). \end{aligned} \tag{3.16}$$

Furthermore, there holds

$$\begin{aligned} &\mathbb{E}_{z_1,...,z_T}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right] \\ &\leq 4\kappa^2\sigma^2 \sum_{t=1}^T \eta_t^2 \exp\left\{-\lambda \sum_{j=t+1}^T \eta_j\right\} + \frac{2M\sigma}{\lambda}\exp\left\{-\lambda\sum_{t=1}^T \eta_t\right\}. \end{aligned} \tag{3.17}$$

Proof. According to the algorithm (1.6), we know

$$\begin{aligned} \|f_{t+1} - f_\lambda^\sigma\|_K^2 &= \|f_t - f_\lambda^\sigma\|_K^2 + \eta_t^2\|\lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2 \\ &\quad + 2\eta_t\left\langle \lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}, f_\lambda^\sigma - f_t\right\rangle_K \\ &= \|f_t - f_\lambda^\sigma\|_K^2 + \eta_t^2\|\lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2 \\ &\quad + 2\eta_t A, \end{aligned} \tag{3.18}$$

where $A = \left\langle \lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1+(\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}, f_\lambda^\sigma - f_t\right\rangle_K$.

By using the inequality $\langle a, b - a\rangle_K \leq \frac{1}{2}(\|b\|_K^2 - \|a\|_K^2), a, b \in \mathcal{H}_K$, with $a = f_t, b = f_\lambda^\sigma$, we have

$$\begin{aligned} A &= \lambda\langle f_t, f_\lambda^\sigma - f_t\rangle_K - \left\langle \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}, f_\lambda^\sigma - f_t\right\rangle_K \\ &\leq \frac{\lambda}{2}(\|f_\lambda^\sigma\|_K^2 - \|f_t\|_K^2) - \left\langle \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}, f_\lambda^\sigma - f_t\right\rangle_K \end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda}{2}(\|f_\lambda^\sigma\|_K^2 - \|f_t\|_K^2) - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}(f_\lambda^\sigma(x_t) - f_t(x_t)) \\
&\leq \frac{\lambda}{2}(\|f_\lambda^\sigma\|_K^2 - \|f_t\|_K^2) + V_\sigma(y_t - f_\lambda^\sigma(x_t)) - V_\sigma(y_t - f_t(x_t)) \\
&= \left(V_\sigma(y_t - f_\lambda^\sigma(x_t)) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2\right) - \left(V_\sigma(y_t - f_t(x_t)) + \frac{\lambda}{2}\|f_t\|_K^2\right). \qquad (3.19)
\end{aligned}
$$

Since f_t depends on $\{z_1, ..., z_{t-1}\}$ but not on z_t, it follows that

$$
\begin{aligned}
\mathbb{E}_{z_1,...,z_t}(A) &\leq \mathbb{E}_{z_1,...,z_{t-1}}\left[\mathbb{E}_{z_t}\left[\left(V_\sigma(y_t - f_\lambda^\sigma(x_t)) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2\right) - \left(V_\sigma(y_t - f_t(x_t)) + \frac{\lambda}{2}\|f_t\|_K^2\right)\right]\right] \\
&= \mathcal{E}_\sigma(f_\lambda^\sigma) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2 - \mathbb{E}_{z_1,...,z_{t-1}}\left[\mathcal{E}_\sigma(f_t) + \frac{\lambda}{2}\|f_t\|_K^2\right]. \qquad (3.20)
\end{aligned}
$$

Combining (3.18) with (3.20), we have

$$
\begin{aligned}
\mathbb{E}_{z_1,...,z_t}[\|f_{t+1} - f_\lambda^\sigma\|_K^2] &\leq \mathbb{E}_{z_1,...,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] + \eta_t^2\mathbb{E}_{z_1,...,z_t}[\|\lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2] \\
&\quad + 2\eta_t\mathbb{E}_{z_1,...,z_{t-1}}[(\mathcal{E}_\sigma(f_\lambda^\sigma) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2) - (\mathcal{E}_\sigma(f_t) + \frac{\lambda}{2}\|f_t\|_K^2)] \\
&= \mathbb{E}_{z_1,...,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] + \eta_t^2\mathbb{E}_{z_1,...,z_t}[\|\lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2] \\
&\quad + 2\eta_t(\Omega_\sigma(f_\lambda^\sigma) - \Omega_\sigma(f_t)). \qquad (3.21)
\end{aligned}
$$

According to Lemma 3.3, we know

$$
\Omega_\sigma(f_\lambda^\sigma) - \Omega_\sigma(f_t) \leq -\frac{\lambda}{2}\|f_t - f_\lambda^\sigma\|_K^2. \qquad (3.22)
$$

Therefore, (3.21) implies that

$$
\begin{aligned}
&\mathbb{E}_{z_1,...,z_t}[\|f_{t+1} - f_\lambda^\sigma\|_K^2] \\
&\leq \mathbb{E}_{z_1,...,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] + \eta_t^2\mathbb{E}_{z_1,...,z_t}[\|\lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2] \\
&\quad - 2\eta_t\mathbb{E}_{z_1,...,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] \\
&= (1 - \lambda\eta_t)\mathbb{E}_{z_1,...,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] + \eta_t^2\mathbb{E}_{z_1,...,z_t}[\|\lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2]. \qquad (3.23)
\end{aligned}
$$

By Lemma 3.1, we know

$$
\left\|\lambda f_t - \frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}K_{x_t}\right\|_K^2 \leq \left(\lambda\|f_t\|_K + \kappa\left|\frac{y_t - f_t(x_t)}{\sqrt{1 + (\frac{y_t - f_t(x_t)}{\sigma})^2}}\right|\right)^2 \leq (\lambda \times \frac{\kappa\sigma}{\lambda} + \kappa\sigma)^2 = 4\kappa^2\sigma^2. \qquad (3.24)
$$

Substituting (3.24) into (3.23), we obtain

$$\mathbb{E}_{z_1,\ldots,z_t}[\|f_{t+1}-f_\lambda^\sigma\|_K^2] \le (1-\lambda\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|f_t-f_\lambda^\sigma\|_K^2] + 4\kappa^2\sigma^2\eta_t^2.$$

Applying the relation iteratively for $t = T, T-1, \ldots, 1$, we have

$$\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_T}[\|f_{T+1}-f_\lambda^\sigma\|_K^2] \\
&\le (1-\lambda\eta_T)\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|f_T-f_\lambda^\sigma\|_K^2] + 4\kappa^2\sigma^2\eta_T^2 \\
&\le (1-\lambda\eta_T)\Big(((1-\lambda\eta_{T-1}))\mathbb{E}_{z_1,\ldots,z_{T-2}}[\|f_{T-1}-f_\lambda^\sigma\|_K^2] + 4\kappa^2\sigma^2\eta_{T-1}^2\Big) + 4\kappa^2\sigma^2\eta_T^2 \\
&\le \ldots\ldots \\
&\le 4\kappa^2\sigma^2\sum_{t=1}^{T}\eta_t^2\prod_{j=t+1}^{T}(1-\lambda\eta_j) + \prod_{t=1}^{T}(1-\lambda\eta_t)\mathbb{E}[\|f_1-f_\lambda^\sigma\|_K^2].
\end{aligned}$$

Denote $\prod_{j=T+1}^{T}(1-\lambda\eta_j) = 1$. From the definition of f_λ^σ, we see that

$$\frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2 \le \mathcal{E}_\sigma(0) \le M\sigma. \tag{3.25}$$

This in connection with the initialization condition $f_1 = 0$, it follows that

$$\begin{aligned}
\mathbb{E}_{z_1,\ldots,z_T}[\|f_{T+1}-f_\lambda^\sigma\|_K^2] &\le 4\kappa^2\sigma^2\sum_{t=1}^{T}\eta_t^2\prod_{j=t+1}^{T}(1-\lambda\eta_j) + \prod_{t=1}^{T}(1-\lambda\eta_t)\mathbb{E}\left[\|f_\lambda^\sigma\|_K^2\right] \\
&\le 4\kappa^2\sigma^2\sum_{t=1}^{T}\eta_t^2\prod_{j=t+1}^{T}(1-\lambda\eta_j) + \frac{2M\sigma}{\lambda}\prod_{t=1}^{T}(1-\lambda\eta_t).
\end{aligned}$$

This shows (3.16). (3.17) follows from (3.16) and the inequality $1-u \le e^{-u}$ for any $u \ge 0$. □

3.1. Proof of Theorem 2.1

Proof. It is easy to see that

$$\prod_{t=1}^{T}(1-\lambda\eta_t) \le \exp\{-\lambda\sum_{t=1}^{T}\eta_t\} \to 0 \quad (T \to +\infty).$$

It implies that, for any $\varepsilon > 0$, there exists some $T_1 \in \mathbb{N}$ such that

$$\prod_{t=1}^{T}(1-\lambda\eta_t) \le \varepsilon,$$

whenever $T \ge T_1$.

And according to the assumption $\lim_{t\to+\infty}\eta_t = 0$, we know that there exists some $t(\varepsilon) \in \mathbb{N}$ such that $\eta_t \le \lambda\varepsilon$, for every $t \ge t(\varepsilon)$. Furthermore, we have

$$\sum_{t=t(\varepsilon)+1}^{T}\eta_t^2\prod_{j=t+1}^{T}(1-\lambda\eta_j) \;\le\; \lambda\varepsilon\sum_{t=t(\varepsilon)+1}^{T}\eta_t\prod_{j=t+1}^{T}(1-\lambda\eta_j)$$

$$
\begin{aligned}
&= \varepsilon \sum_{t=t(\varepsilon)+1}^{T} \lambda\eta_t \prod_{j=t+1}^{T}(1-\lambda\eta_j) \\
&= \varepsilon \sum_{t=t(\varepsilon)+1}^{T} \left((1-(1-\lambda\eta_t)) \prod_{j=t+1}^{T}(1-\lambda\eta_j) \right) \\
&= \varepsilon \sum_{t=t(\varepsilon)+1}^{T} \left(\prod_{j=t+1}^{T}(1-\lambda\eta_j) - \prod_{j=t}^{T}(1-\lambda\eta_j) \right) \\
&= \varepsilon\Big((\prod_{j=t(\varepsilon)+2}^{T}(1-\lambda\eta_j) - \prod_{j=t(\varepsilon)+1}^{T}(1-\lambda\eta_j)) \\
&+ (\prod_{j=t(\varepsilon)+3}^{T}(1-\lambda\eta_j) - \prod_{j=t(\varepsilon)+2}^{T}(1-\lambda\eta_j)) \\
&+ \cdots \\
&+ (\prod_{j=T+1}^{T}(1-\lambda\eta_j) - \prod_{T}^{T}(1-\lambda\eta_j))\Big) \\
&= \varepsilon\Big(1 - \prod_{j=t(\varepsilon)+1}^{T}(1-\lambda\eta_j)\Big) \\
&\leq \varepsilon. \qquad (3.26)
\end{aligned}
$$

Since $t(\varepsilon)$ is fixed, there exists some $T_2 \in \mathbb{N}$ such that, for every $T \geq T_2$, it holds that

$$\sum_{j=t(\varepsilon)+1}^{T} \eta_j \geq \sum_{j=t(\varepsilon)+1}^{T_2} \eta_j \geq \frac{1}{\lambda}\log\frac{t(\varepsilon)}{\lambda^2\varepsilon}.$$

So, for any $1 \leq t \leq t(\varepsilon)$, we have

$$\prod_{j=t+1}^{T}(1-\lambda\eta_j) \leq \exp\{-\sum_{j=t+1}^{T}\lambda\eta_j\} \leq \exp\{-\sum_{j=t(\varepsilon)+1}^{T}\lambda\eta_j\} \leq \frac{\lambda^2\varepsilon}{t(\varepsilon)}.$$

Hence

$$\sum_{t=1}^{t(\varepsilon)} \eta_t^2 \prod_{j=t+1}^{T}(1-\lambda\eta_j) \leq \frac{\lambda^2\varepsilon}{t(\varepsilon)} \sum_{t=1}^{t(\varepsilon)} \eta_t^2 \leq \varepsilon. \qquad (3.27)$$

From (3.26) and (3.27), we know that for any $\varepsilon > 0$, there exists some $T_2 \in \mathbb{N}$ such that

$$\sum_{t=1}^{T} \eta_t^2 \prod_{j=t+1}^{T}(1-\lambda\eta_j) = \sum_{t=1}^{t(\varepsilon)} \eta_t^2 \prod_{j=t+1}^{T}(1-\lambda\eta_j) + \sum_{t=t(\varepsilon)+1}^{T} \eta_t^2 \prod_{j=t+1}^{T}(1-\lambda\eta_j) \leq \varepsilon + \varepsilon = 2\varepsilon, \qquad (3.28)$$

whenever $T \geq T_2$. Let $T' = \max\{T_1, T_2\}$, then by (3.16), (3.26) and (3.27) we have

$$\mathbb{E}_{z_1,\ldots,z_T}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right] \leq (8\kappa^2\sigma^2 + \frac{2M\sigma}{\lambda})\varepsilon,$$

when $T \geq T'$. Thus we complete the proof of Theorem 2.1. □

3.2. Proof of Theorem 2.2

Proof. By (3.21), we know

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_t}[\|f_{t+1}-f_\lambda^\sigma\|_K^2] \;\le\; & \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t-f_\lambda^\sigma\|_K^2\right]+\eta_t^2\mathbb{E}_{z_1,\dots,z_t}[\|\lambda f_t-\frac{y_t-f_t(x_t)}{\sqrt{1+(\frac{y_t-f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2]\\ &+2\eta_t\mathbb{E}_{z_1,\dots,z_{t-1}}\left[\left(\mathcal{E}_\sigma(f_\lambda^\sigma)+\frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2\right)-\left(\mathcal{E}_\sigma(f_t)+\frac{\lambda}{2}\|f_t\|_K^2\right)\right]. \qquad (3.29)\end{aligned}$$

From the inequality $\|a-b\|_K^2\le 2\|a\|_K^2+2\|b\|_K^2$, we have

$$\begin{aligned}\|\lambda f_t-\frac{y_t-f_t(x_t)}{\sqrt{1+(\frac{y_t-f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2 &\le 2\lambda^2\|f_t\|_K^2+2\Big|\frac{y_t-f_t(x_t)}{\sqrt{1+(\frac{y_t-f_t(x_t)}{\sigma})^2}}\Big|^2\|K_{x_t}(\cdot)\|_K^2\\ &\le 2\lambda^2\|f_t\|_K^2+2\kappa^2\sigma^2(\sqrt{1+(\frac{y_t-f_t(x_t)}{\sigma})^2}-1) \qquad (3.30)\end{aligned}$$

On the other hand, for any $r\in\mathbf{R}$, it holds that $\Big|\frac{r}{\sqrt{1+(\frac{r}{\sigma})^2}}\Big|^2\le 2\sigma^2(\frac{\sqrt{\sigma^2+r^2}}{\sigma}-1)=2V_\sigma(r)$. This implies that

$$\Big|\frac{y_t-f_t(x_t)}{\sqrt{1+(\frac{y_t-f_t(x_t)}{\sigma})^2}}\Big|^2\le 2V_\sigma(y_t-f_t(x_t)). \qquad (3.31)$$

Combining (3.30) with (3.31), we get

$$\begin{aligned}\|\lambda f_t-\frac{y_t-f_t(x_t)}{\sqrt{1+(\frac{y_t-f_t(x_t)}{\sigma})^2}}K_{x_t}\|_K^2 &\le 4\kappa^2V_\sigma(y_t-f_t(x_t))+2\lambda^2\|f_t\|_K^2\\ &\le 4\kappa^2V_\sigma(y_t-f_t(x_t))+2\lambda\|f_t\|_K^2\\ &= 4\kappa^2V_\sigma(y_t-f_t(x_t))+4\times\frac{\lambda}{2}\|f_t\|_K^2\\ &\le 4(\kappa^2+1)(V_\sigma(y_t-f_t(x_t))+\frac{\lambda}{2}\|f_t\|_K^2). \qquad (3.32)\end{aligned}$$

Substituting (3.32) into (3.29), we get

$$\begin{aligned}&\mathbb{E}_{z_1,\dots,z_t}[\|f_{t+1}-f_\lambda^\sigma\|_K^2]\\ &\le \mathbb{E}_{z_1,\dots,z_{t-1}}[\|f_t-f_\lambda^\sigma\|_K^2]+4(\kappa^2+1)\eta_t^2\mathbb{E}_{z_1,\dots,z_t}[V_\sigma(y_t-f_t(x_t))+\frac{\lambda}{2}\|f_t\|_K^2]\\ &\quad+2\eta_t\mathbb{E}_{z_1,\dots,z_{t-1}}[(\mathcal{E}_\sigma(f_\lambda^\sigma)+\frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2)-(\mathcal{E}_\sigma(f_t)+\frac{\lambda}{2}\|f_t\|_K^2)]\\ &= \mathbb{E}_{z_1,\dots,z_{t-1}}[\|f_t-f_\lambda^\sigma\|_K^2]+4(\kappa^2+1)\eta_t^2\mathbb{E}_{z_1,\dots,z_t}[\mathcal{E}_\sigma(f_t)+\frac{\lambda}{2}\|f_t\|_K^2]\\ &\quad+2\eta_t\mathbb{E}_{z_1,\dots,z_{t-1}}[(\mathcal{E}_\sigma(f_\lambda^\sigma)+\frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2)-(\mathcal{E}_\sigma(f_t)+\frac{\lambda}{2}\|f_t\|_K^2)]\\ &\le \mathbb{E}_{z_1,\dots,z_{t-1}}[\|f_t-f_\lambda^\sigma\|_K^2]+4(\kappa^2+1)\eta_t^2\mathbb{E}_{z_1,\dots,z_t}[\mathcal{E}_\sigma(f_t)+\frac{\lambda}{2}\|f_t\|_K^2-(\mathcal{E}_\sigma(f_\lambda^\sigma)+\frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2)]\end{aligned}$$

$$
\begin{aligned}
&+ 2\eta_t \mathbb{E}_{z_1,\ldots,z_{t-1}}[(\mathcal{E}_\sigma(f_\lambda^\sigma) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2) - (\mathcal{E}_\sigma(f_t) + \frac{\lambda}{2}\|f_t\|_K^2)] + 4(\kappa^2+1)\eta_t^2\mathcal{E}_\sigma(0) \\
&\le \mathbb{E}_{z_1,\ldots,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] + 2\eta_t(1 - 2(\kappa^2+1)\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[(\mathcal{E}_\sigma(f_\lambda^\sigma) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2) - (\mathcal{E}_\sigma(f_t) + \frac{\lambda}{2}\|f_t\|_K^2)] \\
&\quad + 4(\kappa^2+1)M\sigma\eta_t^2 = \mathbb{E}_{z_1,\ldots,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] + B + 4(\kappa^2+1)M\sigma\eta_t^2, \qquad (3.33)
\end{aligned}
$$

where

$$
B = 2\eta_t(1 - 2(\kappa^2+1)\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[(\mathcal{E}_\sigma(f_\lambda^\sigma) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2) - (\mathcal{E}_\sigma(f_t) + \frac{\lambda}{2}\|f_t\|_K^2)].
$$

Based on the assumptions about η, we know that $1 - 2(\kappa^2+1)\eta_t \ge \frac{1}{2}$. And by Lemma 3.3, we know

$$
\mathbb{E}_{z_1,\ldots,z_{t-1}}[(\mathcal{E}_\sigma(f_\lambda^\sigma) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_K^2) - (\mathcal{E}_\sigma(f_t) + \frac{\lambda}{2}\|f_t\|_K^2)] \le -\frac{\lambda}{2}\|f_t - f_\lambda^\sigma\|_K^2.
$$

This implies that

$$
B \le -\frac{\lambda\eta_t}{2}\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2]. \qquad (3.34)
$$

Combining (3.33) with (3.34), we obtain

$$
\mathbb{E}_{z_1,\ldots,z_t}[\|f_{t+1} - f_\lambda^\sigma\|_K^2] \le (1 - \frac{\lambda\eta_t}{2})\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|f_t - f_\lambda^\sigma\|_K^2] + 4M\sigma^2(\kappa^2+1)\eta_t^2. \qquad (3.35)
$$

Denote $C_\sigma = 4M\sigma^2(\kappa^2+1)$. For $t = T, T-1, \ldots, 1$, we apply the relation iteratively. Then we have

$$
\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_T}[\|f_{T+1} - f_\lambda^\sigma\|_K^2] \\
&\le (1 - \frac{\lambda}{2}\eta_T)\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|f_T - f_\lambda^\sigma\|_K^2] + C_\sigma\eta_T^2 \\
&\le (1 - \frac{\lambda}{2}\eta_T)\Big((1 - \frac{\lambda}{2}\eta_{T-1})\mathbb{E}_{z_1,\ldots,z_{T-2}}[\|f_{T-1} - f_\lambda^\sigma\|_K^2] + C_\sigma\eta_{T-1}^2\Big) + C_\sigma\eta_T^2 \\
&= (1 - \frac{\lambda}{2}\eta_T)(1 - \frac{\lambda}{2}\eta_{T-1})\mathbb{E}_{z_1,\ldots,z_{T-2}}[\|f_{T-1} - f_\lambda^\sigma\|_K^2] + (1 - \frac{\lambda}{2}\eta_T)C_\sigma\eta_{T-1}^2 + C_\sigma\eta_T^2 \\
&\le (1 - \frac{\lambda}{2}\eta_T)(1 - \frac{\lambda}{2}\eta_{T-1})\Big((1 - \frac{\lambda}{2}\eta_{T-2})\mathbb{E}_{z_1,\ldots,z_{T-3}}[\|f_{T-2} - f_\lambda^\sigma\|_K^2] + C_\sigma\eta_{T-2}^2\Big) + (1 - \frac{\lambda}{2}\eta_T)C_\sigma\eta_{T-1}^2 + C_\sigma\eta_T^2 \\
&\le \cdots \\
&\le (1 - \frac{\lambda}{2}\eta_T)(1 - \frac{\lambda}{2}\eta_{T-1})\cdots(1 - \frac{\lambda}{2}\eta_1)(\mathbb{E}[\|f_1 - f_\lambda^\sigma\|_K^2] + C_\sigma\eta_1^2) \\
&= C_\sigma\sum_{t=1}^{T}\eta_t^2\prod_{j=t+1}^{T}(1 - \frac{\lambda}{2}\eta_j) + \prod_{t=1}^{T}(1 - \frac{\lambda}{2}\eta_t)\mathbb{E}[\|f_\lambda^\sigma\|_K^2] \\
&\le C_\sigma\sum_{t=1}^{T}\eta_t^2\prod_{j=t+1}^{T}(1 - \frac{\lambda}{2}\eta_j) + \frac{2M\sigma}{\lambda}\prod_{t=1}^{T}(1 - \frac{\lambda}{2}\eta_t). \qquad (3.36)
\end{aligned}
$$

For any $u \ge 0$, we know that the inequality $1 - u \le e^{-u}$ holds. And (3.36) implies that

$$
\mathbb{E}_{z_1,\ldots,z_T}[\|f_{T+1} - f_\lambda^\sigma\|_K^2] \le C_\sigma\sum_{t=1}^{T}\eta_t^2\exp\{-\frac{\lambda}{2}\sum_{j=t+1}^{T}\eta_j\} + \frac{2M\sigma}{\lambda}\exp\{-\frac{\lambda}{2}\sum_{t=1}^{T}\eta_t\}. \qquad (3.37)
$$

Denote

$$I_1 = \frac{2M\sigma}{\lambda}\exp\{-\frac{\lambda}{2}\sum_{t=1}^{T}\eta_t\} = D_\sigma(\lambda)\exp\{-\frac{\lambda}{2}\sum_{t=1}^{T}\eta_t\} = D_\sigma(\lambda)\exp\{-\frac{\lambda}{2C}\sum_{t=1}^{T}t^{-\theta}\}$$

and

$$I_2 = C_\sigma\sum_{t=1}^{T}(\frac{1}{C}t^{-\theta})^2\exp\{-\frac{\lambda}{2C}\sum_{j=t+1}^{T}j^{-\theta}\} = \frac{C_\sigma}{C^2}\sum_{t=1}^{T}t^{-2\theta}\exp\{-\frac{\lambda}{2C}\sum_{j=t+1}^{T}j^{-\theta}\}.$$

Then, by (3.37) and the assumptions about the stepsize η_t, we know

$$\mathbb{E}_{z_1,\ldots,z_T}[\|f_{T+1} - f_\lambda^\sigma\|_K^2] \le I_1 + I_2. \tag{3.38}$$

Now, we estimate I_1 and I_2 respectively. By Lemma 4 of [9], we obtain the following estimate of I_1

$$I_1 \le \begin{cases} D_\sigma(\lambda)\exp\{-\frac{\lambda}{2C}\frac{1-2^{\theta-1}}{1-\theta}(T+1)^{1-\theta}\}, & if 0<\theta<1, \\ D_\sigma(\lambda)(T+1)^{-\frac{\lambda}{2C}}, & if \theta = 1. \end{cases} \tag{3.39}$$

On the other hand, by Lemma 5.10 of [23] with $\nu = \frac{\lambda}{2C}$, $s = \theta$, we have

$$\sum_{t=1}^{T-1}t^{-2\theta}\exp\{-\frac{\lambda}{2C}\sum_{j=t+1}^{T}j^{-\theta}\} \le \begin{cases} \frac{18}{\frac{\lambda}{2C}T_\theta} + \frac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}}\exp\{-\frac{\lambda(1-2^{\theta-1})}{2C(1-\theta)}(T+1)^{1-\theta}\}, if\ 0<\theta<1, \\ \frac{8}{1-\frac{\lambda}{2C}}(T+1)^{-\frac{\lambda}{2C}}, \quad if\ \theta = 1. \end{cases}$$

So,

$$\begin{aligned} \sum_{t=1}^{T}t^{-2\theta}\exp\{-\frac{\lambda}{2C}\sum_{j=t+1}^{T}j^{-\theta}\} &\le \sum_{t=1}^{T-1}t^{-2\theta}\exp\{-\frac{\lambda}{2C}\sum_{j=t+1}^{T}j^{-\theta}\} + T^{-2\theta}\exp\{-\frac{\lambda}{2C}\sum_{j=T+1}^{T}j^{-\theta}\} \\ &= \sum_{t=1}^{T-1}t^{-2\theta}\exp\{-\frac{\lambda}{2C}\sum_{j=t+1}^{T}j^{-\theta}\} + T^{-2\theta} \\ &\le \begin{cases} \frac{36C}{\lambda T_\theta} + \frac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}}\exp\{-\frac{\lambda(1-2^{\theta-1})}{2C(1-\theta)}(T+1)^{1-\theta}\} + T^{-2\theta}, & if\ 0<\theta<1, \\ \frac{8C}{2C-\lambda}(T+1)^{-\frac{\lambda}{2C}} + T^{-2}, \quad if\ \theta = 1. \end{cases} \end{aligned}$$

Furthermore, we have the following estimate of I_2

$$I_2 \le \begin{cases} \frac{C_\sigma}{C^2}\left(\frac{36C}{\lambda T_\theta} + \frac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}}\exp\{-\frac{\lambda(1-2^{\theta-1})}{2C(1-\theta)}(T+1)^{1-\theta}\} + T^{-2\theta}\right), & if\ 0<\theta<1, \\ \frac{C_\sigma}{C^2}\left(\frac{8C}{2C-\lambda}(T+1)^{-\frac{\lambda}{2C}} + T^{-2}\right), \quad if\ \theta = 1. \end{cases} \tag{3.40}$$

Since $T^{-2\theta} \le T^{-\theta}$ and $\lambda \le C$, the conclusion can be established by combining (3.38) with (3.39) and (3.40). □

3.3. Proof of Corollary 2.1

Proof. For $\theta \in (0,1), 0 < \alpha \leq \min\{1-\theta, \theta\}$, by Theorem 2.2 with $\lambda = T^{-\alpha}$, we have

$$\begin{aligned}&\mathbb{E}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right]\\ &\leq \sigma\left(2MT^\alpha + \frac{36M(\kappa^2+1)}{C^2(1-\theta)2^{1-\theta}}T^{1-\theta}\exp\left\{-\frac{\lambda(1-2^{\theta-1})}{2C(1-\theta)}T^{(1-\theta)-\alpha}\right\} + \frac{36M(\kappa^2+1)}{C}T^{\alpha-\theta}\right)\end{aligned} \tag{3.41}$$

Since for any $\nu > 0, c > 0, \eta > 0$, there exists $L > 0$ such that $\exp\{-cT^\nu\} \leq LT^{-\eta}$, and hence the first term on the right-hand side of (3.41) decays in the form of $O(\frac{\sigma}{T^\eta})$ for any large $\eta > 0$. However, the second term on the right-hand side of (3.41) is bounded by $O(\frac{\sigma}{T^{\theta-\alpha}})$. Consequently, there exists a constant $C_{M,C,\kappa,\theta}$ depending only on θ, κ, C and M such that

$$\mathbb{E}\left[\|f_{T+1} - f_\lambda^\sigma\|_K^2\right] \leq C_{M,C,\kappa,\theta} \times \frac{\sigma}{T^{\theta-\alpha}}.$$

The proof is completed. □

3.4. Proof of Corollary 2.2

Proof. According to the median value theorem, there exists ξ between $y - f_{T+1}(x)$ and $y - f_{\lambda,\sigma}(x)$ such that

$$\begin{aligned}V_\sigma(y - f_{T+1}(x)) - V_\sigma(y - f_{\lambda,\sigma}(x)) =& V_\sigma'(\xi)|f_{T+1}(x) - f_{\lambda,\sigma}(x)|\\ =& \frac{|\xi|}{\sqrt{1+(\frac{\xi}{\sigma})^2}}\left|f_{T+1}(x) - f_{\lambda,\sigma}(x)\right|\\ \leq& \sigma|f_{T+1}(x) - f_{\lambda,\sigma}(x)|.\end{aligned}$$

Then, we have

$$\begin{aligned}\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_{\lambda,\sigma}) &\leq \int_Z \left|V_\sigma(y - f_{T+1}(x)) - V_\sigma(y - f_{\lambda,\sigma}(x))\right| d\rho(x,y)\\ &\leq \sigma\int_Z \left|f_{T+1}(x) - f_{\lambda,\sigma}(x)\right| d\rho(x,y)\\ &\leq \kappa\sigma\|f_{T+1} - f_{\lambda,\sigma}\|_K.\end{aligned} \tag{3.42}$$

And we get

$$\begin{aligned}\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma) &\leq (\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_{\lambda,\sigma})) + \mathcal{E}_\sigma(f_{\lambda,\sigma}) - \mathcal{E}_\sigma(f_\sigma)\\ &\leq \kappa\sigma\|f_{T+1} - f_{\lambda,\sigma}\|_K + \mathcal{E}_\sigma(f_{\lambda,\sigma}) - \mathcal{E}_\sigma(f_\sigma)\\ &\leq \kappa\sigma\|f_{T+1} - f_{\lambda,\sigma}\|_K + \mathcal{E}_\sigma(f_{\lambda,\sigma}) - \mathcal{E}_\sigma(f_\sigma) + \frac{\lambda}{2}\|f_{\lambda,\sigma}\|_K^2\\ &= \kappa\sigma\|f_{T+1} - f_{\lambda,\sigma}\|_K + \inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}_\sigma(f) - \mathcal{E}_\sigma(f_\sigma) + \frac{\lambda}{2}\|f_{\lambda,\sigma}\|_K^2\right\}\\ &= \kappa\sigma\|f_{T+1} - f_{\lambda,\sigma}\|_K + \mathcal{K}(f_\sigma, \lambda).\end{aligned}$$

Combined which with Corollary 2.1 and the assumption 2.2, the desired result follows.

□

4. Discussions

• Most studies of online learning algorithms focus on the convergence in expectation (for example [8–10, 25]). However, these results were established based on some fixed loss functions, such as the least-square loss function (see e.g., [9, 22]). Our results are established based on a parameterized loss function with a scale parameter σ. The analysis results in Section 2 show that the scale parameter σ can effectively control the convergence rate of the learning algorithm, and a better convergence rate is obtained. On the other hand, the previous researches on online learning algorithms rely on integral operator theory (see [25]), this paper establishes the error bounds for the learning sequence by applying the convex analysis method. Convex analysis method has been widely used in various research fields, for example, in the analysis of machine learning algorithms (see e.g., [21, 23, 26]) and the studies of discrete fractional operators (see e.g., [27, 28]), and it has been proved to be a very effective analysis method.

• In [23], the online pairwise regression problem with the quadratic loss is researched. Different from the reference [23], in this paper, we use the parameterized loss function for the pointwise learning model, which has a wider range of applications than the pairwise learning model. It is known that deep convolution networks can increase approximation order (see e.g., [29–32]),then it is hopeful that the convergence rate provided in this paper can be improved by choosing the deep neural network method.

5. Conclusions

In the present paper, we analyze the learning performance of the kernel regularized online algorithm with a parameterized loss. The convergence of the learning sequence is proved and the error bound is provided in the expectation sense by using the convex analysis method. There are some questions for further study. In this paper, we focus on the theoretical analysis of the kernel regularized online algorithm with a parameterized loss V_σ. However, there is still a gap between theoretical analysis and the optimization process of empirical risk minimization based on a parameterized loss. In the future study, it would be interesting to apply the online learning algorithm based on V_σ to solve some practical problems and construct an effective solution method. In addition, we mainly analyze the sample error in this paper, and the approximation error is represented by $\mathcal{K}$-functional. How to make a more accurate analysis of the approximation error and further study the influence of the scale parameter σ on the approximation error still need to be further studied.

Acknowledgments

This work is supported by the Special Project for Scientific and Technological Cooperation (Project No. 20212BDH80021) of Jiangxi Province, the Science and Technology Project in Jiangxi Province Department of Education (Project No. GJJ211334).

Conflict of interest

No potential conflict of interest was reported by the author.

References

1. N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*, **68** (1950), 337–404. http://dx.doi.org/10.2307/1990404
2. W. Dai, J. Hu, Y. Cheng, X. Wang, T. Chai, RVFLN-based online adaptive semi-supervised learning algorithm with application to product quality estimation of industrial processes, *J. Cent. South Univ.*, **26** (2019), 3338–3350. http://dx.doi.org/10.1007/s11771-019-4257-6
3. J. Gui, Y. Liu, X. Deng, B. Liu, Network capacity optimization for Cellular-assisted vehicular systems by online learning-based mmWave beam selection, *Wirel. Commun. Mob. Com.*, **2021** (2021), 8876186. http://dx.doi.org/10.1155/2021/8876186
4. M. Li, I. Sethi, A new online learning algorithm with application to image segmentation, *Image Processing: Algorithms and Systems IV*, **5672** (2005), 277–286. http://dx.doi.org/10.1117/12.586328
5. S. Sai Santosh, S. Darak, Intelligent and reconfigurable architecture for KL divergence based online machine learning algorithm, arXiv:2002.07713.
6. B. Yang, J. Yao, X. Yang, Y. Shi, Painting image classification using online learning algorithm, In: *Distributed, ambient and pervasive interactions*, Cham: Springer, 2017, 393–403. http://dx.doi.org/10.1007/978-3-319-58697-7_29
7. S. Das, Kuhoo, D. Mishra, M. Rout, An optimized feature reduction based currency forecasting model exploring the online sequential extreme learning machine and krill herd strategies, *Physica A*, **513** (2019), 339–370. http://dx.doi.org/10.1016/j.physa.2018.09.021
8. S. Smale, Y. Yao, Online learning algorithms, *Found. Comput. Math.*, **6** (2006), 145–170. http://dx.doi.org/10.1007/s10208-004-0160-z
9. Y. Ying, D. Zhou, Online regularized classification algorithms, *IEEE Trans. Inform. Theory*, **52** (2006), 4775–4788. http://dx.doi.org/10.1109/TIT.2006.883632
10. Y. Ying, D. Zhou, Unregularized online learning algorithms with general loss functions, *Appl. Comput. Harmon. Anal.*, **42** (2017), 224–244. http://dx.doi.org/10.1016/J.ACHA.2015.08.007
11. Y. Zeng, D. Klabjian, Online adaptive machine learning based algorithm for implied volatility surface modeling, *Knowl.-Based Syst.*, **163** (2019), 376–391. http://dx.doi.org/10.1016/j.knosys.2018.08.039
12. J. Lin, D. Zhou, Online learning algorithms can converge comparably fast as batch learning, *IEEE Trans. Neural Netw. Learn. Syst.*, **29** (2018), 2367–2378. http://dx.doi.org/10.1109/TNNLS.2017.2677970
13. P. Huber, E. Ronchetti, *Robust statistics*, Hoboken: John Wiley & Sons, 2009. http://dx.doi.org/10.1002/9780470434697
14. Y. Wu, Y. Liu, Robust truncated hinge loss support vector machine, *J. Am. Stat. Assoc.*, **102** (2007), 974–983. http://dx.doi.org/10.1198/016214507000000617
15. Y. Yu, M. Yang, L. Xu, M. White, D. Schuurmans, Relaxed clipping: a global training method for robust regression and classification, *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, **2** (2010), 2532–2540.
16. S. Huang, Y. Feng, Q. Wu, Learning theory of minimum error entropy under weak moment conditions, *Anal. Appl.*, **20** (2022), 121–139. http://dx.doi.org/10.1142/S0219530521500044

17. F. Lv, J. Fan, Optimal learning with Gaussians and correntropy loss, *Anal. Appl.*, **19** (2021), 107–124. http://dx.doi.org/10.1142/S0219530519410124
18. X. Zhu, Z. Li, J. Sun, Expression recognition method combining convolutional features and Transformer, *Math. Found. Compt.*, in press. http://dx.doi.org/10.3934/mfc.2022018
19. S. Suzumura, K. Ogawa, M. Sugiyama, M. Karasuyama, I. Takeuchi, Homotopy continuation approaches for robust SV classification and regression, *Mach. Learn.*, **106** (2017), 1009–1038. http://dx.doi.org/10.1007/s10994-017-5627-7
20. Z. Guo, T. Hu, L. Shi, Gradient descent for robust kernel-based regression, *Inverse Probl.*, **34** (2018), 065009. http://dx.doi.org/10.1088/1361-6420/aabe55
21. B. Sheng, H. Zhu, The convergence rate of semi-supervised regression with quadratic loss, *Appl. Math. Comput.*, **321** (2018), 11–24. http://dx.doi.org/10.1016/j.amc.2017.10.033
22. M. Pontil, Y. Ying, D. Zhou, Error analysis for online gradient descent algorithms in reproducing kernel Hilbert spaces, *Proceedings of Technical Report, University College London*, 2005, 1–20.
23. S. Wang, Z. Chen, B. Sheng, Convergence of online pairwise regression learning with quadratic loss, *Commun. Pur. Appl. Anal.*, **19** (2020), 4023–4054. http://dx.doi.org/10.3934/cpaa.2020178
24. H. Bauschke, P. Combettes, *Convex analysis and monotone operator theory in Hilber spaces*, Cham: Springer-Verlag, 2010. http://dx.doi.org/10.1007/978-3-319-48311-5
25. Z. Guo, L. Shi, Fast and strong convergence of online learning algorithms, *Adv. Comput. Math.*, **45** (2019), 2745–2770. http://dx.doi.org/10.1007/s10444-019-09707-8
26. Y. Lei, D. Zhou, Convergence of online mirror descent, *Appl. Comput. Harmon. Anal.*, **48** (2020), 343–373. http://dx.doi.org/10.1016/j.acha.2018.05.005
27. I. Baloch, T. Abdeljawad, S. Bibi, A. Mukheimer, G. Farid, A. Haq, Some new Caputo fractional derivative inequalities for exponentially $(\theta, h-m)$-convex functions, *AIMS Mathematics*, **7** (2022), 3006–3026. http://dx.doi.org/10.3934/math.2022166
28. P. Mohammed, D. O'Regan, A. Brzo, K. Abualnaja, D. Baleanu, Analysis of positivity results for discrete fractional operators by means of exponential kernels, *AIMS Mathematics*, **7** (2022), 15812–15823. http://dx.doi.org/10.3934/math.2022865
29. Y. Xia, J. Zhou, T. Xu, W. Gao, An improved deep convolutional neural network model with kernel loss function in image classifiaction, *Math. Found. Comput.*, **3** (2020), 51–64. http://dx.doi.org/10.3934/mfc.2020005
30. D. Zhou, Deep distributed convolutional neural networks: universality, *Anal. Appl.*, **16** (2018), 895–919. http://dx.doi.org/10.1142/S0219530518500124
31. D. Zhou, Universality of deep convolutional neural networks, *Appl. Comput. Harmon. Anal.*, **48** (2020), 787–794. http://dx.doi.org/10.1016/j.acha.2019.06.004
32. D. Zhou, Theory of deep convolutional neural networks: downsampling, *Neural Networks*, **124** (2020), 319–327. http://dx.doi.org/10.1016/j.neunet.2020.01.018

AIMS Press