KERNEL-BASED ONLINE GRADIENT DESCENT USING DISTRIBUTED APPROACH

XIAMING CHEN*  
Shantou University, No. 243 Daxue Rd.  
Shantou, Guangdong, China  

(Communicated by Yunlong Feng)

Abstract. In this paper we study the kernel-based online gradient descent with least squares loss without an explicit regularization term. Our approach is novel by controlling the expectation of the K-norm of \( f_t \) using an iterative process. Then we use distributed learning to improve our result.

1. Introduction. Different from the classical batch learning which learns from the entire data set, online learning seeks to learn from a data set with an increasing size. The gradient descent method is a powerful algorithm designed to find the optimal value of a function, and online gradient descent is an adaptation to the online scheme. This kind of stochastic approximation procedures can date back to [8, 5]. The online gradient descent algorithm has been studied in [9, 15] recently. In [14], the early stopping approach for batch learning is studied. In [9], the author studied an online gradient descent algorithm with a regularized term \( \lambda f_t \), which can be formulated as follows:

\[
\begin{aligned}
    f_1 &= 0, \\
    f_{t+1} &= f_t - \eta_t ((f_t(x_t) - y_t)K_{x_t} + \lambda f_t).
\end{aligned}
\]  

We call \( \lambda \) the regularization parameter and when \( \lambda > 0 \), the algorithm is called online regularized learning and it has been well studied in [7, 9, 16]. In [14], the regularized term is replaced by some early stopping rule, and in [15], the author studied (1) without an explicit regularized term (i.e. \( \lambda = 0 \)). Our algorithm is the same as that in [15], but we prove the risk bound by using a novel method which involve an iterative process. In [15], the constant \( \mu \) must be large enough in order to make the proof works. The size of \( \mu \) should be set more freely apart from being a proving technique. In our approach, we can have a much looser definition of \( \mu \).

Recently, researchers are interested in online algorithms in various situations. By abandoning the identical restriction while retaining the independence, [10] studied online learning with Markov sampling. In [4], the online regression algorithm with Gaussian kernels with changing variance is introduced and analysed. For the situation of unbounded sampling, the online minimum error entropy was proposed [12]. In industrial applications, some new algorithms are proposed such as [11, 3, 13].

2010 Mathematics Subject Classification. Primary: 68Q32, 68T05; Secondary: 62J02.  
Key words and phrases. Learning theory, distributed learning, online gradient descent.

The first author is supported by STU Scientific Research Foundation for Talents grant (NTF-18022).

* Corresponding author: Xiaming Chen.
Data are being used with unprecedented size and complexity recently, which raises problems such as the storage bottleneck and the algorithm scalability. To overcome these challenges, some distributed approaches are being used based on the divide and conquer strategy [17, 6]. In this paper, we use the distributed approach to eliminate the variance produced by the noise and obtain a better result. In [6], the authors use second order decomposition to estimate the learning error. In our paper, the approach is simpler and does not involve second order decomposition.

Let $X$ be a compact subset of Euclidean space $\mathbb{R}^d$ and $Y$ be a subset of $\mathbb{R}$. Define $Z = X \times Y$. Let $\rho$ be an unknown probability measure on $Z$ and $\rho_{Y|x}$ be a conditional probability measure for a fixed $x \in X$. We define the generalization error of a function $f : X \to Y$ as

$$
\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho(z).
$$

The regression function is defined as

$$
f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X,
$$

which minimizes the generalization error over all measurable functions. Our proposed goal is to approximate the regression function from a set of sample $Z = \{z_1, ..., z_n \}$ drawn independently from the unknown probability measure $\rho$. We set our learning scheme in Reproducing Kernel Hilbert Spaces (RKHS). Here, the regularity of a function is characterized by the integral operator.

A function $K : X \times X \to \mathbb{R}$ is called a Mercer Kernel, when it is continuous, symmetric and semidefinite [2], where semidefinite is defined by requiring the matrix $M = (K(x_i, x_j))_{m \times m}$ is semidefinite for any finite subset $\{x_1, ..., x_m\}$ of $X$. We define $V_K = \{K_t : t \in X\}$ where $K_t(x) := K(t, x)$. The closure of all the linear combination of $\{K_t : t \in X\}$ under the following inner product forms the Reproducing Kernel Hilbert Space induced by the kernel $K$. For $f(\cdot) = \sum_{i=1}^n c_i K(x_i, \cdot)$ and $g(\cdot) = \sum_{j=1}^m d_j K(x_j, \cdot)$, we define the inner product $\langle \cdot, \cdot \rangle_K$ as

$$
\langle f, g \rangle_K = \sum_{i,j=1}^{n,m} c_i d_j K(x_i, x_j),
$$

and $\|f\|_K = \sqrt{\langle f, f \rangle_K}$. A constant related to the RKHS, $\kappa$, is defined as

$$
\kappa = \sup_{x \in X} \sqrt{K(x, x)}.
$$

One of the important properties of RKHS is the reproducing property which is characterized by $f(x) = \langle f, K_x \rangle$. A linear map $L_K$ from $L^2_X$ to $H_K$ is defined as $L_K f(x) = \int_X f(t) K(x, t) d\rho_X$.

The online gradient descent algorithm is defined in the following way:

$$
\begin{align*}
    f_1 &= 0, \\
    f_{t+1} &= f_t - \eta_t (f_t(x_t) - y_t) K_{x_t},
\end{align*}
$$

This is an instance of (1) when setting $\lambda$ as 0. The sequence $\{\eta_t : t \in \mathbb{N}\}$ is called the step size or learning rate. Our machine receives a sequence of data $\{z_t : t \in \mathbb{N}\}$ one by one where $z_t = (x_t, y_t)$. The data $z_t$ are drawn independently. In the distributed learning we divide our source of data into $J$ different subsets and we use the online gradient descent algorithm for each subset of data. The output function after $t$-th
iteration on the $j$-th machine is denote as $f_j^t$. The algorithm can be formulated as follows:

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

(3)

By taking the average $f_t = \frac{1}{J} \sum_{j=1}^{J} f_t^j$, we improve the result of the learning error $\mathcal{E}(f_t)$. We define $f_\mathcal{H}$ as the optimizer of the error $\mathcal{E}(\cdot)$ over the RKHS $\mathcal{H}_K$

$$
  f_\mathcal{H} = \text{argmin}_{f \in \mathcal{H}} \mathcal{E}(f).
$$

(4)

2. Main result. By using an iterative process similar to [1], we can prove that for a sufficiently large $C$, the expectation of $K$-norm of $f_t$ can be controlled by $C$ uniformly.

**Theorem 2.1.** Let $f_\rho \in L_\rho^\beta(L_\rho^2)$ with $\beta > \frac{1}{2}$ and $\eta_t = \frac{1}{\mu} t^{-\theta}$ with $\mu \geq \kappa^2$ and $\frac{1}{2} < \theta < \frac{3}{2}$. Let the Online Gradient Descent Algorithm be defined by (2), we have

$$
\mathbb{E}(\|f_t\|_K) \leq C,
$$

where $C = MC_1^{\frac{1-\theta}{2\theta+1}}$ and $C_1 = \frac{3\kappa^2}{\mu^2 \sqrt{c_0}}$ and $c_0 = \min\{(1-2\theta)(\frac{1-\theta}{2\theta+1})+(1-\theta), -(1-2\theta)\frac{1-\theta}{2\theta+1}-(1-\theta)\}$.

**Corollary 1.** Let $f_\rho \in L_\rho^\beta(L_\rho^2)$ with $\beta > \frac{1}{2}$ and $\eta = \frac{1}{\mu} t^{-\frac{2\beta}{2\beta+1}}$ with $\mu \geq \kappa^2$. Let the Online Gradient Descent Algorithm defined in (2), we have

$$
\mathbb{E}(\|f_t\|_K^2) \leq C,
$$

where $C = MC_1^{\frac{1-\theta}{2\theta+1}}$ and $C_1 = \frac{3\kappa^2 \rho}{\mu^2 (1-\theta)^2}$.

**Proof.** Notice that

$$
\mathbb{E}(\|f_{t+1}\|_K^2) \leq 2\mathbb{E}(\|f_{t+1} - f_{\mathcal{H}}\|_K^2) + 2\mathbb{E}(\|f_{\mathcal{H}}\|_K^2),
$$

and by using the process in the previous proof for $\lceil \frac{1-\theta}{2\theta+1} \rceil$, the result is derived. \qed

With this estimation, we can prove the following theorem.

**Theorem 2.2.** Let $f_\rho \in L_\rho^\beta(L_\rho^2)$ with $\beta > \frac{1}{2}$ and $\eta = \frac{1}{\mu} t^{-\frac{2\beta}{2\beta+1}}$ with $\mu \geq \kappa^2$. Let the Online Gradient Descent Algorithm defined in (2), we have the result at $t$ iteration

$$
\mathbb{E}(\|f_{t+1} - f_\rho\|_2^2) \leq C t^{-\frac{2\beta}{2\beta+1}}
$$

By using distributed online gradient descent, the following result is derived.

**Theorem 2.3.** Assume $f_\rho \in L_\rho^\beta(L_\rho^2)$ with $\beta > \frac{1}{2}$, let $f_t^j$ be the resulting function after $t$-th iteration at the $j$-th machine by using the Distributed Online Gradient Descent (3) with $\eta_t = \frac{1}{\mu} t^{-\theta}$ and $1 \leq J \leq t^{(\beta-1)(1-\theta)}$, then the following holds

$$
\mathbb{E}(\|f_t - f_\rho\|_2^2) \leq C J^{-\frac{1}{2} t^{\frac{1}{2} (\theta-1)}},
$$

where $C$ is a constant independent of $J$ or $t$. 

3. Proof of main result. First, we prove a bound for the $K$-norm of step function $f_t$.

**Lemma 3.1.** Let the learning sequence $\{f_t : t \in \mathbb{N}\}$ be given by (2). Assume $|y| \leq M$ almost surely and, for any $t \in \mathbb{N}$, that $\eta_t \kappa^2 \leq 1$. Then we have

$$\|f_t\|_K \leq M \sqrt{\sum_{j=2}^{t-1} \eta_j}, \quad \forall t \in \mathbb{N}.$$  

**Proof.** For $t = 1$, $f_1 = 0$ clearly satisfies the stated bound. For $t > 1$, we have from (2) that

$$\|f_{t+1}\|^2_K = \langle f_t - \eta_t(f_t(x_t) - y_t), f_t - \eta_t(f_t(x_t) - y_t), K_{x_t} \rangle_K$$

$$= \|f_t\|^2_K - 2\eta_t \langle f_t, (f_t(x_t) - y_t)K_{x_t} \rangle + \eta_t^2 \|f_t(x_t) - y_t\|^2 K(x_t, x_t)$$

$$\leq \|f_t\|^2_K - 2\eta_t \langle f_t, (f_t(x_t) - y_t) \rangle + \eta_t^2 \|f_t(x_t) - y_t\|^2 \kappa^2.$$  

Define a quadratic function $F$ by $F(s) = \eta_t(s - y_t)^2 \kappa^2 - 2s(s - y_t)$. The derivative of $F$ is $F'(s) = 2\eta_t \kappa^2 (s - y_t) - 4s + 2y_t$. Let $F'(s^*) = 0$ to obtain $s^* = (\frac{1 - \eta_t \kappa^2}{2\eta_t \kappa^2}) y_t$ and with $\eta_t \kappa^2 \leq 1$ the maximum value of $F$ is $F(s^*) = \frac{y_t^2}{2 - \eta_t \kappa^2} \leq M^2$. Hence, we have $\|f_{t+1}\|^2_K \leq \|f_t\|^2_K + M^2 \eta_t$. By summing up all the iterations, we have $\|f_{t+1}\|^2_K \leq M^2 \sum_{j=1}^{t-1} \eta_j$. By taking square roots, we have

$$\|f_t\|_K \leq M \sqrt{\sum_{j=2}^{t-1} \eta_j}.$$  

This verifies the desired bound. \hfill \Box

**Lemma 3.2.** Let $f_H$ be defined as above, we have

$$L_K f_H = L_K f_p,$$

where $f_H$ is defined in (4).

**Proof.** We know that $E(\cdot)$ is differentiable and convex, hence we know that $E(\cdot)$ achieves the minimum value when its gradient is equal to zero and for $g \in \mathcal{H}_K$,

$$\lim_{h \to 0} \frac{E(f_H + hg) - E(f_H)}{h} = \int_Z \langle (f_H(x) - y)K_x, g \rangle d\rho(z)$$

$$= \langle \int_Z (f_H(x) - f_p(x))K_x d\rho_X(x), g \rangle_K$$

$$= \langle L_K(f_H - f_p), g \rangle_K = 0.$$  

Since $g \in \mathcal{H}_K$ is arbitrary, we have $L_K(f_H - f_p) = 0$. \hfill \Box

**Proposition 1.** The operator $L_1$ on $\mathcal{H}_K$ defined by $L_1 = \langle \cdot, K_{x_1} \rangle_K K_{x_1}$ has the operator norm bounded as $\|L_1\| \leq \kappa^2$.

**Proof.** For $f \in \mathcal{H}_K$, we have

$$\|L_1 f\|^2_K = \langle (f(x_1))K_{x_1}, f(x_1)K_{x_1} \rangle_K = f(x_1)^2 \|K(x_1, x_1)\| \leq \kappa^2 \|f\|_{\infty}^2 \leq \kappa^4 \|f\|^2_K.$$  

Hence, we have $\|L_1\|^2 \leq \kappa^4$ and the proposition is proved. \hfill \Box
Proof of Theorem 2.1. From (2) we have
\[ f_{t+1} - f_H = f_t - f_H - \eta_t(f_t(x_t) - y_t)K_{x_t} \]
\[ = (I - \eta_tL_H)(f_t - f_H) + \eta_tL_H(f_t - f_H) - \eta_t(f_t(x_t) - y_t)K_{x_t} \]
\[ = (I - \eta_tL_H)(f_t - f_H) + \eta_t(L_Hf_t - f_t(x_t)K_{x_t}) + \eta_t(y_tK_{x_t} - L_Hf_H). \]
Hence, by summing up \( t \) iterations we have
\[ f_{t+1} - f_H = -\omega_1^t(L_H)f_H + \sum_{j=1}^t \eta_j\omega_{j+1}^t(L_H)[(L_Hf_j - L_jf_j) - (L_Hf_H - S_j)]. \quad (5) \]

Denote \( \mathcal{B}_t := (L_Hf_t - L_tf_t) - (L_Hf_H - S_t) \). Notice that \( f_t \) is independent of \( z_t \). Hence
\[ \mathbb{E}(L_tf_t|z_1, \ldots, z_{t-1}) = \int_Z f_t(x)K(x, \cdot)d\rho_X(z) = L_Hf_t, \quad (6) \]
and similarly
\[ \mathbb{E}(S_t|z_1, \ldots, z_{t-1}) = \int_Z yK(x, \cdot)d\rho(z) = L_Hf_p. \quad (7) \]

From (6), (7) and Lemma 3.2 we have
\[ \mathbb{E}(\mathcal{B}_t|z_1, \ldots, z_{t-1}) = 0, \quad (8) \]
for each \( t \). Then we have
\[ \mathbb{E}(\|f_{t+1} - f_H\|_K^2) = \mathbb{E}(\|\omega_1^t(L_H)f_H\|_K^2) + \sum_{j=1}^t \eta_j^2\mathbb{E}(\|\omega_{j+1}^t(L_H)\mathcal{B}_j\|_K^2) \]
\[ + \sum_{j=1}^t \eta_j\mathbb{E}(\langle \omega_1^t(L_H)f_H, \omega_{j+1}^t(L_H)\mathcal{B}_j \rangle_K) \]
\[ - \sum_{j=1}^t \eta_j\mathbb{E}(\langle \omega_1^t(L_H)f_H, \omega_{j+1}^t(L_H)\mathcal{B}_j \rangle_K) \]
\[ =: I_1 + I_2 + I_3 + I_4. \]

For \( I_3 \), assume \( i < j \), from (8) we have
\[ \mathbb{E}(\langle \omega_{i+1}^t(L_H)\mathcal{B}_i, \omega_{j+1}^t(L_H)\mathcal{B}_j \rangle_K) \]
\[ = \mathbb{E}(\langle \omega_{i+1}^t(L_H)\mathcal{B}_i, \omega_{j+1}^t(L_H)\mathcal{B}_j \rangle_K|z_1, \ldots, z_{j-1}) = 0. \]
Hence we derive that \( I_3 = 0 \). Similarly
\[ \mathbb{E}(\langle \omega_1^t(L_H)f_H, \omega_{j+1}^t(L_H)\mathcal{B}_j \rangle_K) = 0, \]
and \( I_4 = 0 \). From Lemma 1 we have
\[ \|\mathcal{B}_t\|_K \leq \|L_t - L_H\|_K\|f_t\|_K + \|L_Hf_H - S_t\|_K \leq \kappa^2(2M + \|f_t\|_K) \]
\[ \leq 3\kappa^2 \max\{\|f_t\|_K, M\}. \quad (9) \]

From Lemma 3.1, we have \( \|f_t\|_K \leq M \frac{1}{\sqrt{\mu(1-\theta)}}t^{\frac{1}{2}(1-\theta)} \). It holds that
\[ I_2 \leq \sum_{j=1}^t \eta_j^2\|\omega_{j+1}^t(L_H)\|_K^2\mathbb{E}(\|\mathcal{B}_j\|_K^2) \leq \sum_{j=1}^t \frac{1}{\mu^3(1-\theta)}j^{-2\theta}9\kappa^4M^2 \max\{j^{1-\theta}, 1\} \]
\[ \leq C' \max\{t^{1-2\theta + (1-\theta)}, 1\}, \]
where $C'$ is the constant $C' = \frac{9s^4M^2}{\mu^3(1-\theta)^2}$. With $I_1 = \|\omega_1^t(L_K)f_\mathcal{H}\|_K^2 \leq \|f_\mathcal{H}\|_K^2$, we have

$$
E(\|f_{t+1} - f_\mathcal{H}\|_K) \leq E(\|f_{t+1} - f_\mathcal{H}\|_K^2)^{1/2}
$$

$$
= \sqrt{I_1 + I_2} \leq \sqrt{C'} \max\{t^{\frac{1}{2} - \theta + \frac{1}{2}(1-\theta)}, 1\}
$$

$$
\leq \frac{M}{\mu(1-\theta)}C_1 \max\{t^{\frac{1}{2} - \theta + \frac{1}{2}(1-\theta)}, 1\},
$$

(10)

where $C_1 = \frac{3s^2}{\mu\sqrt{c_0}}$ and $c_0 = \min\{(1-2\theta)(\frac{1-\theta}{2\theta - 1} + (1-\theta), -(1-2\theta)(\frac{1-\theta}{2\theta - 1} - (1-\theta))\}$.

From (10), we have

$$
E(\|f_{t+1}\|_K) \leq \frac{M}{\sqrt{\mu(1-\theta)}}C_1 \max\{t^{\frac{1}{2} - \theta + \frac{1}{2}(1-\theta)}, 1\}
$$

holds for a sufficiently large $C$.

In general, we assume

$$
E(\|f_{t+1}\|_K) \leq \frac{M}{\sqrt{\mu(1-\theta)}}C_1 s \max\{t^{s(\frac{1}{2} - \theta) + \frac{1}{2}(1-\theta)}, 1\}
$$

holds with $s \leq \lfloor \frac{1-\theta}{2\theta - 1} \rfloor$. By (9), we have

$$
I_2 \leq \sum_{j=1}^{t} \eta_j^2 \|\omega_{j+1}^t(L_K)\|^2 \mathbb{E}(\|\mathcal{B}_j\|_K^2)
$$

$$
\leq \sum_{j=1}^{t} \frac{1}{\mu^3(1-\theta)}j^{-2\eta_j^2}M^2C_1^2 \max\{t^{s(1-2\theta) + (1-\theta)}, 1\}
$$

$$
\leq \frac{M^2}{\mu(1-\theta)}C_1^2(s+1) \max\{t^{(s+1)(1-2\theta) + (1-\theta)}, 1\}.
$$

Hence, we have

$$
E(\|f_{t+1}\|_K) \leq \frac{M}{\sqrt{\mu(1-\theta)}}C_1^{s+1} \max\{t^{(s+1)(\frac{1}{2} - \theta) + \frac{1}{2}(1-\theta)}, 1\}.
$$

By repeating this process until the case of $j = \lfloor \frac{1-\theta}{2\theta - 1} \rfloor$, we have

$$
E(\|f_{t+1}\|_K) \leq \frac{M}{\mu(1-\theta)}C_1^{\lfloor \frac{1-\theta}{2\theta - 1} \rfloor}.
$$

Proof of Theorem 2.2. From (5) we have

$$
f_{t+1} - f_\rho = -\omega_1^t(L_K)f_\rho + \sum_{j=1}^{t} \eta_j \omega_{j+1}^t(L_K)\mathcal{B}_j,
$$

by taking the square norm we have

$$
\|f_{t+1} - f_\rho\|_2^2
$$

$$
= \langle -\omega_1^t(L_K)f_\rho + \sum_{j=1}^{t} \eta_j \omega_{j+1}^t(L_K)\mathcal{B}_j, -\omega_1^t(L_K)f_\rho + \sum_{j=1}^{t} \eta_j \omega_{j+1}^t(L_K)\mathcal{B}_j \rangle
$$

$$
= \|\omega_1^t(L_K)f_\rho\|_2^2 + \sum_{j=1}^{t} \eta_j^2 \|\omega_{j+1}^t(L_K)\mathcal{B}_j\|_2^2 + 2 \sum_{j=1}^{t} \eta_j \langle \omega_1^t(L_K)f_\rho, \omega_{j+1}^t(L_K)\mathcal{B}_j \rangle,
$$
where $B$ of Proof of Theorem 2.3. First, it holds that

$$\text{Hence, it holds that}$$

$$E(\|f_{t+1} - f_{\rho}\|_2^2) \leq E(\|\omega^t_1(L_K) f_{\rho}\|_2^2) + E(\sum_{j=1}^{t} \eta_j^2 \|\omega^t_{j+1}(L_K) \mathfrak{B}_j\|_2^2)$$

and

$$E(\|\omega^t_1(L_K) L^2_K \|_2^2 L^{-\beta} f_{\rho}\|_2^2) = C t^{-\beta(1 - \theta)} + Ct^{\theta - 1} \sum_{j=1}^{t} j^{-2\theta} \leq Ct^{-\min\left(\frac{\theta}{2 + \theta}, \frac{\beta}{2 + \beta}\right)}.$$

**Proof of Theorem 2.3.** First, it holds that

$$f_{t+1}^i - f_{\rho} = (I - \eta_t L_K) (f_t^i - f_{\rho}) + \eta_t (L_K f_t^i - f_t^i(x_t) K x_t) - \eta_t (L_K f_\rho - y_t^i K x_t)$$

$$= (I - \eta_t L_K) (f_t^i - f_{\rho}) + \eta_t \mathfrak{B}_t^i,$$

where $\mathfrak{B}_t^i = (L_K f_t^i - f_t^i(x_t) K x_t) + (L_K f_\rho - y_t^i K x_t)$, notice that $f_t^i$ is independent of $z_t$, we have $E(\mathfrak{B}_t^i) = 0$ and

$$E(f_{t+1}^i - f_{\rho}) = (I - \eta_t L_K) E(f_t^i - f_{\rho}).$$

Applying the above equations $t$ times, we obtain $E(f_{t+1}^i - f_{\rho}) = -\omega^t_1(L_K) f_{\rho}$, and it holds that

$$\|E(f_{t+1}^i - f_{\rho})\|_2 = \|\omega^t_1(L_K) f_{\rho}\|_2 \leq \|\omega^t_1(L_K) L^2_K \|_2 L^{-\beta} f_{\rho}\|_2 \leq Ct^{-\beta(1 - \theta)}.$$

Hence

$$E(\|f_t^i - f_{\rho}\|_2) = E(\|\frac{1}{J} \sum_{j=1}^{J} (f_t^j - f_{\rho}) - E(\|f_t^i - f_{\rho}\|_2) + \|E(f_t^i - f_{\rho})\|_2$$

$$\leq E(\|\frac{1}{J} \sum_{j=1}^{J} (f_t^j - f_{\rho}) - E(\|f_t^i - f_{\rho}\|_2) + Ct^{-\beta(1 - \theta)}). \quad (11)$$

Since

$$E(\frac{1}{J} \sum_{j=1}^{J} (f_t^j - f_{\rho}) - E(f_t^i - f_{\rho})\|_2^2) \leq \frac{1}{J^2} \sum_{j=1}^{J} E(\|f_t^i - f_{\rho}\|_2^2) \quad (12)$$
To finish the proof, it suffices to estimate the term $E(\|f_t - f_\rho\|^2_2)$. Similar to the proof of Theorem 2.1 and the fact that $1 \leq (1 - 2\beta)(1 - \theta)$, we have

$$
E(\|f_t - f_\rho\|^2_2) 
\leq 
E(\|\omega_1(L_K) f_\rho\|^2_2) + \sum_{j=1}^t \eta_j^2 \|\omega_j(L_K) B_{j K}\|^2_2
$$

$$
\leq 
E(\|\omega_1(L_K) L_{K_1/2}\|^2_2 L_{K_1}^{-\beta} f_\rho\|^2_2) + \sum_{j=1}^t \eta_j^2 \|\omega_j(L_K) L_{K_1/2}\|^2_2 \|B_j\|^2_2
$$

$$
\leq 
Ct^{-2\beta(1-\theta)} + Ct^{\theta-1} \sum_{j=1}^t j^{2\theta} \leq C(t^{-2\beta(1-\theta)} + t^{\theta-1}) \leq C t^{\theta-1}. \quad (14)
$$

Let $J \leq t^{(2\beta-1)(1-\theta)}$, we have $t^{-\beta(1-\theta)} \leq J^{-\frac{1}{2}} t^{\frac{1}{2}(\theta-1)}$. Hence from (11), (12) and (13), we have

$$
E(\|\hat{f}_t - f_\rho\|^2_2) \leq CJ^{-\frac{1}{2}} t^{\frac{1}{2}(\theta-1)}.
$$

\[\square\]

**Acknowledgments.** The author is grateful to Prof. Ding-Xuan Zhou and Dr. Jinshan Zeng for their constructive comments. This paper is supported by STU Scientific Research Foundation for Talents (NTF18022).

**REFERENCES**

[1] X. Chen and Y. Lei, Refined Bounds for online pairwise learning algorithm, *Neurocomputing*, **275** (2018), 2656–2665.

[2] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.*, **39** (2001), 1–49.

[3] J. Duchi, E. Hazan and Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *Journal of Machine Learning Research*, **12** (2011), 2121–2159.

[4] T. Hu, Online regression with varying gaussians and non-identical distributions, *Analysis and Applications*, **9** (2011), 395–408.

[5] J. Kiefer and J. Wolfowitz, Stochastic estimation of the maximum of a regression function, *Ann. Math. Statist.*, **23** (1952), 462–466.

[6] S. B. Lin, X Guo and D. X. Zhou, Distributed learning with regularized least squares, *Journal of Machine Learning Research*, **18** (2017), Paper No. 92, 31 pp.

[7] M. Pontil, Y. Ying and D. X. Zhou, Error analysis for online gradient descent algorithms in reproducing kernel Hilbert space, *Technical Report, Department of Computer Science, University College London*, (2005).

[8] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Statist.*, **22** (1951), 400–407.

[9] S. Smale and Y. Yao, Online learning algorithms, *Found. Comput. Math.*, **6** (2006), 145–170.

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

[11] S. Vijayakumar, A. D’Souza and S. Schaal, Incremental online learning in high dimensions, *Neural Computation*, **17** (2005), 2692–2634.

[12] C. Wang and T Hu, Online minimum error entropy algorithm with unbounded sampling, *Analysis and Applications*, **17** (2019), 293–322.

[13] J. Xu, Z. Zheng, Z. Fan and W. Liu, Online personalized QoS prediction approach for cloud services, *4th International Conference on Cloud Computing and Intelligence Systems*, 2016.

[14] Y. Yao, L. Losasco and A. Caponnetto, Early stopping in gradient descent boosting, *Constr. Approx.*, **26** (2007), 289–315.

[15] Y. Ying and M. Pontil, Online pairwise learning algorithms, *Found. Comput. Math.*, **28** (2016), 743–777.
[16] Y. Ying and D. X. Zhou, Online regularized classification algorithm, IEEE, trans. Inform. Theory, 52 (2006), 4775–4788.

[17] Z. H. Zhou, N. V. Chawla, Y. Jin and G. J. Williams, Big data opportunities and challenges: Discussions from data analytics perspective, IEEE Computational Intelligence Magazine, 9 (2014), 62–74.

E-mail address: chenxm@stu.edu.cn