CONVERGENCE OF ONLINE PAIRWISE REGRESSION LEARNING WITH QUADRATIC LOSS

SHUHUA WANG
School of Statistics and Mathematics, Zhejiang Gongshang University
Hangzhou, 310018, China
Department of Applied Statistics, Shaoxing University, Shaoxing, 312000, China

ZHENLONG CHEN
School of Statistics and Mathematics, Zhejiang Gongshang University
Hangzhou, 310018, China

BAOHUAI SHENG*
Department of Applied Statistics, Shaoxing University, Shaoxing, 312000, China

Abstract. Recent investigations on the error analysis of kernel regularized pairwise learning initiate the theoretical research on pairwise reproducing kernel Hilbert spaces (PRKHSs). In the present paper, we provide a method of constructing PRKHSs with classical Jacobi orthogonal polynomials. The performance of the kernel regularized online pairwise regression learning algorithms based on a quadratic loss function is investigated. Applying convex analysis and Rademacher complexity techniques, the bounds for the generalization error are provided explicitly. It is shown that the convergence rate can be greatly improved by adjusting the scale parameters in the loss function.

1. Introduction. In recent years, online learning algorithms have been attracting the attentions of a lot of researchers in statistical learning theory (see e.g., [21, 58, 71, 73, 74] and references therein). In present paper, we shall give an investigation on the convergence analysis of kernel-based regularized online pairwise learning algorithm associating with the quadratic loss.

1.1. Online learning algorithms. Let $X$ be a given compact set in the $d$-dimensional Euclidean space $\mathbb{R}^d$, $Y \subset \mathbb{R}$. Let $\rho$ be a fixed but unknown probability distribution on $Z = X \times Y$ which yields its marginal distribution $\rho_x$ on $X$ and its conditional distribution $\rho(\cdot \mid x)$ at $x \in X$. Then we denote by $\{z_t = (x_t, y_t)\}_{t=1}^T$ the sample drawn i.i.d. (independently and identically distribution) according to distribution $\rho$. The aim of regression learning is to find a predictor $f : X \to \mathbb{R}$ from a hypothesis space such that $f(x)$ is a ”good” approximation of $y$. Let $V(r) : \mathbb{R} \to \mathbb{R}_+$

2010 Mathematics Subject Classification. Primary: 41A25, 90C25; Secondary: 68Q32.

Key words and phrases. Online pairwise learning, quadratic loss, learning rate, pairwise reproducing kernel Hilbert space, convex analysis.

This work is supported by the National Natural Science Foundation of China under Grants Nos. 61877039, 11971432 and the Humanities and Social Sciences Research Project of Ministry of Education (18YJA910001).

* Corresponding author.
be an even loss function. It is well known that $V'(r)$ is an odd function. The quality of the predictor $f$ is measured by the generalization error

$$E(f) := \int Z V(f(x) - y) d\rho(x, y), \quad (1.1)$$

In the theory of kernel regularized learning, RKHSs are often chosen as the hypothesis spaces. The batch algorithms based on a RKHS $\mathcal{H}_K$ is defined as ([20])

$$f_{z, \lambda} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{T} \sum_{t=1}^{T} V(f(x_t) - y_t) + \frac{\lambda}{2} \|f\|_K^2, \quad (1.2)$$

and the data-free analogue

$$f_{\lambda} = \arg\min_{f \in \mathcal{H}_K} E(f) + \frac{\lambda}{2} \|f\|_K^2, \quad (1.3)$$

where $\lambda > 0$ is a regularization parameter.

Contrast to the batch learning which tackles the whole samples in a batch, online learning processes the samples one by one and update the output in time. For any $f_t \in \mathcal{H}_K$, by the results of functional gradient (see [42, 50]), we know that the gradient of $E(f) + \frac{\lambda}{2} \|f\|_K^2$ at the point $f_t$ with respect to $f$ is

$$\nabla f(E(f) + \frac{\lambda}{2} \|f\|_K^2)|_{f=f_t} = \int Z V'(f_t(x) - y) K_x(\cdot) d\rho(x, y) + \lambda f_t.$$

Using the classical gradient descent method [9], we can get the sequence $\{g_t : g_t \in \mathcal{H}_K\}$, which provides an approximation to $f_{\lambda}$, by the following iterative formula

$$\begin{cases}
g_1 = 0, \\
g_{t+1} = g_t - \eta_t \left( \int Z V'(g_t(x) - y) K_x(\cdot) d\rho(x, y) + \lambda g_t \right). \quad (1.4)
\end{cases}$$

However, since the distribution $\rho$ is unknown, the algorithm (1.4) can not be implemented directly. Based on the Stochastic Gradient Descent Method (see e.g. [23, 29, 48]), the integral in (1.4) is replaced by the value $V'(f_t(x_t) - y_t) K_{x_t}(\cdot)$, the kernel-based online learning algorithm is given by (see e.g. [58])

$$\begin{cases}
f_1 = 0, \\
f_{t+1} = f_t - \eta_t \left( V'(f_t(x_t) - y_t) K_{x_t}(\cdot) + \lambda f_t \right), \quad (1.5)
\end{cases}$$

where $\eta_t$ is called the step size and the sequence $\{f_t : t = 1, \ldots, T + 1\}$ is the learning sequence.

The convergence rate of the online learning algorithm (1.5) has been extensively studied in the literature (see e.g. [38, 52, 58, 69, 71, 75]). The research results in [38] show that under mild conditions the regularized online learning algorithm (1.5) can converge comparably fast as batch learning algorithm (1.2).

1.2. The kernel regularized online pairwise learning algorithm. Recently, the kernel-based regularization pairwise learning algorithms have been considered in [16, 17], whose background comes from many problems such as the ranking problem (see e.g. [1, 12, 13, 18, 35, 47, 76]), the similarity learning (see e.g. [11, 38]), et al. An important innovation of pairwise learning from the usual online learning is the use of the pairwise reproducing kernel Hilbert space (PRKHS) $\mathcal{H}_K$, which is often reproduced by an unique symmetric and positive definite continuous function $K : X^2 \times X^2 \to \mathbb{R}$, called pairwise reproducing kernel or pairwise Mercer kernel (see e.g. [10, 13, 17, 27, 45]).
The PRKHS $\mathcal{H}_K$ is defined as the completion of the linear span of the set of functions \( \{K(x,x')(\cdot,\cdot) = K((x,x')(\cdot,\cdot)) : (x,x') \in X^2\} \) with the inner product \( \langle \cdot, \cdot \rangle_K \) satisfying the following reproducing property
\[
f(x,x') = \langle K(x,x')(\cdot,\cdot), f \rangle_K, \quad \forall (x,x') \in X^2, f \in \mathcal{H}_K
\]
and the embedding inequality
\[
|f(x,x')| \leq \kappa \|f\|_K, \quad \forall (x,x') \in X^2, f \in \mathcal{H}_K, \quad (1.6)
\]
where \( \kappa := \sup_{x,x' \in X} \sqrt{K((x,x'),(x,x'))} \).

Let \( V(r) : \mathbb{R} \to [0, +\infty) \) be an even regression loss function, the batch regression algorithms based on the Tikhonov regularization scheme for pairwise learning is defined as (see e.g. [8, 15, 64, 72])
\[
f_{z,\lambda} = \arg \min_{f \in \mathcal{H}_K} \frac{1}{T(T-1)} \sum_{i,t=1,i \neq j} V(f(x_i,x_j) - \delta(y_i,y_j)) + \frac{\lambda}{2} \|f\|_K^2,
\]
The reducing function \( \delta : Y \times Y \to \mathbb{R} \) chosen according to the learning task, is a new terminology making pairwise learning essentially different from that of pointwise learning. For any hypothesis \( f \in \mathcal{H}_K \), if the inducing function \( \delta(y,y') \) is a symmetric function, we naturally hope that \( f \) is a symmetric function, i.e. \( f(x,x') = f(x',x), \forall x,x' \in X \), and if the inducing function \( \delta(y,y') \) is an anti-symmetric function, for example \( \delta(y,y') = y - y' \), we hope that \( f \) is an anti-symmetric function as well, i.e. \( f(x,x') = -f(x',x), \forall x,x' \in X \) (see e.g. [10, 13, 18]). So it is valuable if we can provide a way of constructing PRKHS whose functions have the symmetric or anti-symmetric property. This is the first motivation for us to write this paper.

Online pairwise learning depends upon a sequence of samples \( z = \{z_t\}_{t=1}^T = \{(x_t,y_t)\}_{t=1}^T \). At each time step \( t = 2 \ldots T \), the algorithm posits a hypothesis \( f_t \in \mathcal{H}_K \) upon which the next sample \( z_t \) is revealed. And the algorithm incurs the following local regularized empirical error, on which the quality of the hypothesis \( f_t \) is assessed (see e.g. [15, 72]),
\[
\hat{E}_t^\lambda(f) = \frac{1}{t-1} \sum_{j=1}^{t-1} V(f(x_t,x_j) - \delta(y_t,y_j)) + \frac{\lambda}{2} \|f\|_K^2.
\]
where \( \lambda > 0 \) is a regularization parameter.

To update the current predictor \( f_t \), an iterative step is made based on the negative gradient of the above local regularized empirical error \( \hat{E}_t^\lambda(f) \) at \( f = f_t \), which is exactly defined by (see [42, 46])
\[
- \nabla \hat{E}_t^\lambda(f) |_{f=f_t} = -\left( \frac{1}{t-1} \sum_{j=1}^{t-1} V'(f(x_t,x_j) - \delta(y_t,y_j)\big)K(x_t,x_j)(\cdot,\cdot) + \lambda f_t \right).
\]

Using the above gradient descent method, the general form of online kernel-based regularization pairwise learning algorithm is defined as follows.
\[
\begin{aligned}
f_1 &= f_2 = 0, \\
f_{t+1} &= f_t - \eta_t \left( \frac{1}{t-1} \sum_{j=1}^{t-1} V'(f_t(x_t,x_j) - \delta(y_t,y_j))K(x_t,x_j)(\cdot,\cdot) + \lambda f_t \right),
\end{aligned}
\]
where \( \{\eta_t > 0\} \) is a stepsize sequence.
For a hypothesis \( f \), we define the generalization error or risk associated with the loss function \( V \) as

\[
\mathcal{E}(f) := \int_Z \int_Z V(f(x, x') - \delta(y, y')) d\rho(x, y) d\rho(x', y'),
\]

and the regularization generalization error or risk is defined as

\[
f_\lambda = \arg \min_{f \in \mathcal{H}} \left( \mathcal{E}_\lambda(f) = \mathcal{E}(f) + \frac{\lambda}{2} \|f\|_K^2 \right).
\]

To study the learning performance of online pairwise learning algorithms we need to bound the convergence rate of the iterative sequence \( \{f_t : t = 1, \ldots, T + 1\} \). Online pairwise learning in a linear space was investigated in [64], and the generalization bounds for the average of the iterates were established requiring the uniform boundedness of the loss function. Instead of the average of its iterates, recent literatures considered the last iterate of online pairwise learning algorithms, and the research results show that the performance of the last iterate is competitive to that of the average of iterates (\cite{15, 26, 39, 72}). [72] studied the convergence rates of online pairwise learning with the least squares loss

\[
V(f(x, x') - \delta(y, y')) = (y - y' - f(x, x'))^2,
\]
a novel error decomposition was given and the explicit error bounds of \( O(T^{-\beta/2}) \) were presented with the pairwise regression function \( f_\rho \in L^2_K(L^2_\rho) \). The convergence rate of online pairwise learning algorithm for binary classification was established in [26] by using strong convexity and re-weighting techniques. [39] investigated the unregularized online pairwise learning algorithms with a general convex loss satisfying an increment condition, the convergence rate \( O(T^{-\beta/2} \log T) \) was established with the step size \( \eta_t = \eta_1 t^{-\theta}, \theta = \frac{2\beta+1}{2\beta+2} \).

It is known that the learning rates of a kernel-based regularization algorithm are influenced by the geometry properties, e.g., the capacity, the covering number, the uniformly convexity (\cite{4, 5, 20, 40}). Some other parameters with respect to the RKHS also influence the learning rates. For example, it is shown by [37, 65, 68, 70] that the flexible variance \( \sigma \) in the Gaussian kernels greatly influence the learning rates. The recent researches on the semi-supervised learning show that the parameters in the loss function also influence the learning rates (see\cite{54}). [63] studied the convergence of the online pairwise algorithm with varying regularization parameters.

On the other hand, we find that the quadratic function \( \sqrt{1 + t^2}, t \in \mathbb{R} \) plays an important role in constructing shape preserving quasi-interpolation and solving partial differential equations with meshfree method since its strong nonlinear property and its convexity (see \cite{14, 25, 66, 67}), and it has been used by \cite{56} as the loss function which shows some advantages in forming semi-supervised learning algorithms. Encouraged by these researches, we want to make an investigation on how these parameters influence the learning rates of the online pairwise learning. This is the second reason for writing this paper.

The rest of this paper is organized as follows. In Section 2 we state the theory of bivariate orthogonal polynomials and give some examples of bivariate pairwise Mercer kernels. In Section 3, we consider the online pairwise learning algorithm associated with the quadratic loss function \( V_\sigma(r) := \sigma^2(\sqrt{1 + (\frac{r}{\sigma})^2} - 1) \) with parameters \( \sigma \in (0, 1] \). Unlike pointwise learning, pairwise learning usually involves pairs of training samples that are not independently and identically distributed.
(i.i.d.). By using tools from convex analysis and Rademacher complexity, we give some investigations on the performance of the learning algorithm and give an explicit convergence rate bound. The analysis results show that the learning rate can be improved by choosing the scale parameter $\sigma$ properly. We give in Section 5 the proofs.

Here and later we write $A = O(B)$ if there is a constant $C \geq 0$ such that $A \leq CB$.

2. **Pairwise Mercer kernels and PRKHSs.** Cucker and Smale gave a description of a usual RKHS with orthonormal systems (see Proposition 4 of [19], see also [51, 60, 61]), which itself is a way of constructing RKHSs. Along this line, we give here a method of constructing PRKHS with bivariate orthogonal polynomials.

Let $\Omega \subset \mathbb{R}^2$ be a simply connected bounded domain (having a nonempty interior) and $w(x, y)$ be a non-negative and nonzero integrable function defined on $\Omega$ and

$$L^2(w) := \left\{ f(x, y) : \|f\|_{L^2(w)} = \left( \int_\Omega |f(x, y)|^2 w(x, y) dxdy \right)^{\frac{1}{2}} < +\infty \right\}.$$ 

For any $f, g \in L^2(w)$ we define an inner product

$$\langle f, g \rangle_w := \int_\Omega f(x, y)g(x, y)w(x, y) dxdy$$

Denoted by $\Pi^2_n$ the linear space of real bivariate polynomials of total degree at most $n$, i.e.,

$$\Pi^2_n = \{ p_{m,l}(x, y) = x^m y^l + \text{lower power terms} : m + l \leq n \},$$

and denoted by $\Pi^2$ the collection of all bivariate polynomials.

A polynomial $p \in \Pi^2_n$ is called an orthogonal polynomial with respect to $w(x, y)$ if

$$\langle p, q \rangle_w = \int_\Omega p(x, y)q(x, y)w(x, y) dxdy = 0,$$

for all $q \in \Pi^2_{n-1}$.

Let $\mathcal{V}^2_n$ denote the set of polynomials of exact degree $n$ orthogonal with respect to the weight function $w(x, y)$. Then by [34] we know $\mathcal{V}^2_n$ is a linear space of polynomials of dimension $n + 1$. Also, let $\{P_{n-k,k} \}_{k=0}^n$ denote a basis of $\mathcal{V}^2_n$. Then

$$\{P_{n-k,k}(x, y) : (x, y) \in \Omega, 0 \leq k \leq n, n = 0, 1, 2, \ldots \}$$

forms an orthogonal polynomial system (see [34]).

Take $\mathbb{P}_n = (P_{n-k,k})_{k=0}^n = (P_{n,0}, P_{n-1,1}, \ldots, P_{0,n})^T$. Then $\mathbb{P}_n^T$ is a row vector. Using these notions, the orthogonality of $\mathbb{P}_n$ can be expressed as (see [24])

$$\langle \mathbb{P}_n, \mathbb{P}_m^T \rangle_w = \int_\Omega \mathbb{P}_n \mathbb{P}_m^T w(x, y) dxdy = \mathcal{O}_{(m+1) \times (n+1)}, \ n \neq m, m, n \geq 0,$$

$$\langle \mathbb{P}_n, \mathbb{P}_n^T \rangle_w = \int_\Omega \mathbb{P}_n \mathbb{P}_n^T w(x, y) dxdy = \mathbb{H}_n,$$

where $\mathbb{H}_n$ is a symmetric and positive-definite matrix of size $(n + 1) \times (n + 1)$ is a $(m + 1) \times (n + 1)$ matrix in which all elements are zero. If $\mathbb{H}_n$ is the identity matrix, then $\{P_{n-k,k} : 0 \leq k \leq n, n = 0, 1, 2, \ldots \}$ is an orthonormal polynomial system (see [24]).

We provide here a method for constructing Mercer kernel $K_{(x,y)}(x', y')$ on $\Omega^2$ with orthogonal polynomials on $\Omega$. 

\hspace{1cm}
Let \( \{ P_{n-k,k}(x, y) : (x, y) \in \Omega, 0 \leq k \leq n; n = 0, 1, 2, \cdots \} \) be an orthogonal polynomial system in \( Y^2_n \). Define
\[
\Phi^{(n)}_{(x,y)}(x', y') = \sum_{k=0}^{n} h_{n,k}^{-2} P_{n-k,k}(x, y) P_{n-k,k}(x', y')
\]
and
\[
1 = h_{n,k}^{-2} \int_{\Omega} |P_{n-k,k}(x, y)|^2 w(x, y) dxdy. 
\tag{2.1}
\]
Then
\[
K_{(x,y)}(x', y') = \sum_{n=0}^{\infty} \lambda_n \Phi^{(n)}_{(x,y)}(x', y'), \quad \lambda_n > 0
\]
is a pairwise Mercer kernel on \( \Omega^2 \) under the assumption that \( \lambda_n \) has a decay such that the convergence of the right side of (2.2) is absolute and uniform.

Assume the series
\[
\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} h_{n,k}^{-2} |P_{n-k,k}(x', y')|^2 
\tag{2.3}
\]
uniformly converges on \( \Omega \), we now show that \( K_{(x,y)}(x', y') \) is a Mercer kernel on \( \Omega^2 \).

In fact, by Hölder inequality we have
\[
|K_{(x,y)}(x', y')| \leq \sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} h_{n,k}^{-2} |P_{n-k,k}(x, y)||P_{n-k,k}(x', y')| \leq \left( \sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} h_{n,k}^{-2} \right)^{\frac{1}{2}} \left( \sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} h_{n,k}^{-2} |P_{n-k,k}(x, y)|^2 \right)^{\frac{1}{2}}.
\]

To show the convergence of (2.3), one needs to bound \( \sum_{k=0}^{n} h_{n,k}^{-2} |P_{n-k,k}(x', y')|^2 \) which itself is a problem touching upon the estimation of the Christoffel-Darboux formula (see e.g.\([22, 28, 32, 49]\)).

**Remark 1.** The following two cases show that the assumption (2.3) is reasonable.

- When \( \Omega \) is a compact set and \( w(x, y) \) is a bounded function on \( \Omega \), e.g. \( 0 \leq w(x, y) \leq 1 \). Since \( \sum_{k=0}^{n} h_{n,k}^{-2} |P_{n-k,k}(x', y')|^2 \) is a continuous function on \( \Omega \) for a given \( n \), it can attain its maximum value \( M_n \) on \( \Omega \). We can choose \( \lambda_n > 0 \) such that \( \sum_{n=0}^{\infty} \lambda_n M_n < +\infty \).
- When \( \Omega \) is a compact set and \( w(x, y)^{-1} \) is a bounded function on \( \Omega \), e.g. \( 0 \leq w(x, y)^{-1} \leq 1 \). For given \( n \) and \( k \), we have by (2.1) that \( h_{n,k}^{-2} |P_{n-k,k}(x', y')|^2 \) is bounded. There exists \( M_n \in \mathbb{R} \) such that \( \sum_{k=0}^{n} h_{n,k}^{-2} |P_{n-k,k}(x', y')|^2 \) \( w(x, y) \leq M_n \). It follows that
\[
\sum_{k=0}^{n} h_{n,k}^{-2} |P_{n-k,k}(x', y')|^2 \leq M_n w(x, y)^{-1} \leq M_n.
\]

Also, we can choose \( \lambda_n > 0 \) such that \( \sum_{n=0}^{\infty} \lambda_n M_n < +\infty \).

Take \( r_{n,k}(f) = \langle f, h_{n,k}^{-1} P_{n-k,k} \rangle_w, k = 0, 1, 2, \ldots, n, \ n = 0, 1, 2, \cdots \) to be the Fourier coefficients of \( f \) with respect to the orthonormal basis
\[
\left\{ h_{n,k}^{-1} P_{n-k,k}(x, y) : (x, y) \in \Omega, 0 \leq k \leq n, n = 1, 2, \cdots \right\}.
\]
Then

\[ f \sim \sum_{n=0}^{\infty} \sum_{k=0}^{n} r_{n,k}(f)h_{n,k}^{-1}P_{n-k,k}. \]

Define

\[ \mathcal{H}_K = \{ f \in L^2(w) : \sum_{n=0}^{\infty} \frac{1}{\lambda_n} \sum_{k=0}^{n} r_{n,k}^2(f) < +\infty, f = \sum_{n=0}^{\infty} \sum_{k=0}^{n} r_{n,k}(f)h_{n,k}^{-1}P_{n-k,k} \}. \]

Then we have the following proposition.

**Proposition 1.** \( \mathcal{H}_K \) is a PRKHS with reproducing kernel (2.2) and the inner product defined by

\[ (f, g)_K = \sum_{n=0}^{\infty} \frac{1}{\lambda_n} \sum_{k=0}^{n} r_{n,k}(f)r_{n,k}(g) \] (2.4)

for

\[ f = \sum_{n=0}^{\infty} \sum_{k=0}^{n} r_{n,k}(f)h_{n,k}^{-1}P_{n-k,k} \in \mathcal{H}_K, \quad g = \sum_{n=0}^{\infty} \sum_{k=0}^{n} r_{n,k}(g)h_{n,k}^{-1}P_{n-k,k} \in \mathcal{H}_K. \]

**Proof.** Rewrite \( K_{(x,y)}(x', y') \) as

\[ K_{(x,y)}(x', y') = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \left( \lambda_n h_{n,k}^{-1}P_{n-k,k}(x, y) \right) h_{n,k}^{-1}P_{n-k,k}(x', y'). \]

Then

\[ r_{n,k}(K_{(x,y)}(\cdot, \cdot)) = \left( K_{(x,y)}(\cdot, \cdot), h_{n,k}^{-1}P_{n-k,k} \right) = \lambda_n h_{n,k}^{-1}P_{n-k,k}(x, y). \]

By (2.3) we have

\[ \sum_{n=0}^{\infty} \frac{1}{\lambda_n} \sum_{k=0}^{n} r_{n,k}^2(K_{(x,y)}(\cdot, \cdot)) = \sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} h_{n,k}^{-2}(P_{n-k,k}(x', y'))^2 < +\infty. \]

Therefore, for any given \((x, y) \in \Omega\) we have \( K_{(x,y)}(x', y') \in \mathcal{H}_K\). By (2.4) we know for

\[ f = \sum_{n=0}^{\infty} \sum_{k=0}^{n} r_{n,k}(f)h_{n,k}^{-1}P_{n-k,k} \in \mathcal{H}_K \]

there holds

\[ (f, K_{(x,y)})_K = \sum_{n=0}^{\infty} \frac{1}{\lambda_n} \sum_{k=0}^{n} r_{n,k}(f) \left( \lambda_n h_{n,k}^{-1}P_{n-k,k}(x, y) \right) = f(x, y). \]

So \( \mathcal{H}_K \) is a PRKHS with reproducing kernel \( K_{(x,y)}(x', y') \).

**2.1. Some pairwise Mercer kernels.** By Proposition 1 we know that to construct a PRKHS, one only needs to construct the pairwise Mercer kernels \( K_{(x,y)}(x', y') \) satisfying (2.3). We give some examples.

Take the classical Jacobi polynomials as

\[ p_{n}(x, y) = \frac{(\alpha + 1)_n}{n!} \binom{n}{n} F_1(-n, \alpha + \beta + n + 1; \alpha + 1; \frac{1-x}{2}) \]

\[ = \frac{(-1)^n(1-x)^{-\alpha}(1+x)^{-\beta}}{2^n n!} D_x^n \left\{ (1-x)^{n+\alpha}(1+x)^{n+\beta} \right\} (D_x := \frac{d}{dx}, \alpha > -1, \beta > -1, \]

Then

\[ f \sim \sum_{n=0}^{\infty} \sum_{k=0}^{n} r_{n,k}(f)h_{n,k}^{-1}P_{n-k,k}. \]
where
\[
\mu F_q(\alpha_1, \ldots, \alpha_p; \beta_1, \ldots, \beta_q; x) = \sum_{n=0}^{\infty} \frac{(\alpha_1)_n \cdots (\alpha_p)_n}{(\beta_1)_n \cdots (\beta_q)_n} \frac{x^n}{n!}, \quad |x| < 1,
\]
\[
(\lambda)_0 = 1, (\lambda)_k = \lambda(\lambda + 1) \cdots (\lambda + k - 1) = \frac{\Gamma(\lambda + k)}{\Gamma(\lambda)}, k \in \mathbb{N}_0 = \{0\} \cup \mathbb{N},
\]
and \(\Gamma(\lambda)\) is the Euler’s Gamma function defined by
\[
\Gamma(\lambda) = \int_0^{+\infty} t^{\lambda-1} e^{-t} dt, \quad \text{for Re}(\lambda) > 0.
\]
Now, we give some examples.

(1) A bivariate pairwise Mercer kernel on the unit disk.
For \(\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}\) and \(w(x, y) = (1 - x^2 - y^2)^{\alpha}, \alpha > -1\), there are following bivariate orthogonal polynomials (see [24, 32, 44])
\[
P^{(\alpha)}_{n-k,k}(x, y) = \sum_{\lambda=0}^{\infty} \int_{\Omega} w(x, y) P^{(\alpha)}_{n-k}(x, y) P^{(\alpha)}_{n-k}(x', y') dx dy,
\]
where
\[
\left( h^{(\alpha)}_{n,k} \right)^2 = \int_{\Omega} \left| P^{(\alpha)}_{n-k}(x, y) \right|^2 w(x, y) dx dy, \quad \lambda_n > 0.
\]
The \(\lambda_n\) \(\in\) \(\mathbb{R}\) are chosen such that the series
\[
\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( h^{(\alpha)}_{n,k} \right)^{-2} \left| P^{(\alpha)}_{n-k}(x, y) \right|^2
\]
converges uniformly on \(\Omega\). Since \(\Omega\) is a compact set, for \(\alpha \geq 0\), \(w(x, y)\) satisfies \(0 \leq w(x, y) \leq 1\) and for \(-1 < \alpha < 0\), \(w(x, y)^{-1} \leq 1\), by the remarks we know the assumption (2.5) is reasonable.

(2) A bivariate pairwise Mercer kernel over the triangle.
Let \(\Omega = \{(x, y) \in \mathbb{R}^2 : 0 < y < x < 1\}\) and \(w(x, y) = (1-x)^\alpha (x-y)^\beta y^\gamma, \alpha, \beta, \gamma > -1\). Koornwinder defines an orthogonal polynomial system on \(\Omega\) (see [24, 44])
\[
P^{(\alpha, \beta, \gamma)}_{n, k}(x, y) = \sum_{\lambda=0}^{\infty} \int_{\Omega} w(x, y) P^{(\alpha, \beta, \gamma)}_{n-k}(x, y) P^{(\alpha, \beta, \gamma)}_{n-k}(x', y') dx dy,
\]
where
\[
\left( h^{(\alpha, \beta, \gamma)}_{n,k} \right)^2 = \int_{\Omega} \left| P^{(\alpha, \beta, \gamma)}_{n-k}(x, y) \right|^2 w(x, y) dx dy, \quad \lambda_n > 0.
\]
The \(\lambda_n\) are chosen such that the series
\[
\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( h^{(\alpha, \beta, \gamma)}_{n,k} \right)^{-2} \left| P^{(\alpha, \beta, \gamma)}_{n-k}(x, y) \right|^2
\]
converges uniformly on $\Omega$. If $-1 < \alpha, \beta, \gamma < 0$, then $0 \leq w(x, y)^{-1} \leq 1$, by the remarks we can see that this assumption is reasonable.

(3) A bivariate pairwise Mercer kernel over the biangle.

For $\Omega = \{(x, y) \in \mathbb{R}^2 : y^2 \leq x \leq 1\}$ and $w(x, y) = (1 - x)^\alpha (x - y^2)^\beta$, $\alpha, \beta > -1$, there are the following Koornwinder orthogonal polynomials over the parabolic biangle (see [44])

$$P_{n,k}^{(\alpha, \beta)}(x, y) = p_{n-k}^{(\alpha, \beta + k + \frac{1}{2})} (2x - 1)x^k p_k^{(\beta)} \left( \frac{y}{\sqrt{x}} \right), \quad 0 \leq k \leq n.$$

We may define a bivariate pairwise Mercer kernel as

$$K_{(x,y)}^{(\alpha, \beta)}(x', y') = \sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( h_{n,k}^{(\alpha, \beta)} \right)^{-2} P_{n,k}^{(\alpha, \beta)}(x, y) P_{n,k}^{(\alpha, \beta)}(x', y'), \quad (x, y), (x', y') \in \Omega,$$

where

$$\left( h_{n,k}^{(\alpha, \beta)} \right)^2 = \int_{\Omega} |P_{n,k}^{(\alpha, \beta)}(x, y)|^2 w(x, y)dx dy, \quad \lambda_n > 0.$$

The $\lambda_n$ are chosen such that the series

$$\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( h_{n,k}^{(\alpha, \beta)} \right)^{-2} \left| P_{n,k}^{(\alpha, \beta)}(x, y) \right|^2$$

converges uniformly on $\Omega$. Since $\Omega$ is a compact set, for $\alpha, \beta \geq 0$, $w(x, y)$ satisfies $0 \leq w(x, y) \leq 1$ and for $-1 < \alpha, \beta < 0$, $w(x, y)^{-1} \leq 1$, by the remarks we know this assumption is reasonable.

(4) A bivariate pairwise Mercer kernel over the square.

According to the results of [36] we know if $\{p_n(x)\}_{n=0}^{\infty}$ and $\{q_n(x)\}_{n=0}^{\infty}$ are orthogonal polynomial systems in one variable relative to $d\mu(x)$ and $d\nu(x)$ respectively, then the product polynomials

$$\{p_{n-k}(x)p_k(y) : 0 \leq k \leq n, n = 0, 1, 2, \ldots\}$$

forms an orthogonal polynomial system relative to the product measure $d\mu(x)d\nu(y)$.

For $\Omega = \{(x, y) \in \mathbb{R}^2 : -1 \leq x \leq 1, -1 \leq y \leq 1\}$ and $w(x, y) = (1 - x)^\alpha (1 + x)^\beta (1 - y)^\gamma (1 + y)^\delta$, $\alpha, \beta, \gamma, \delta > -1$, there are the following Koornwinder orthogonal polynomials (see [24, 44])

$$P_{n,k}^{(\alpha, \beta, \gamma, \delta)}(x, y) = P_{n-k}^{(\alpha, \beta)}(x)p_k^{(\gamma, \delta)}(y), \quad 0 \leq k \leq n.$$

For $(x, y), (x', y') \in \Omega$, we may define a bivariate pairwise Mercer kernel as

$$K_{(x,y)}^{(\alpha, \beta, \gamma, \delta)}(x', y') = \sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( h_{n,k}^{(\alpha, \beta, \gamma, \delta)} \right)^{-2} P_{n,k}^{(\alpha, \beta, \gamma, \delta)}(x, y) P_{n,k}^{(\alpha, \beta, \gamma, \delta)}(x', y'),$$

where

$$\left( h_{n,k}^{(\alpha, \beta, \gamma, \delta)} \right)^2 = \int_{\Omega} |P_{n,k}^{(\alpha, \beta, \gamma, \delta)}(x, y)|^2 w(x, y)dx dy, \lambda_n > 0.$$

The $\lambda_n$ are chosen such that the series

$$\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( h_{n,k}^{(\alpha, \beta, \gamma, \delta)} \right)^{-2} \left| P_{n,k}^{(\alpha, \beta, \gamma, \delta)}(x, y) \right|^2$$

converges uniformly on $\Omega$. Since $\Omega$ is a compact set, for $\alpha, \beta, \gamma, \delta \geq 0$, $w(x, y)$ satisfies $0 \leq w(x, y) \leq 1$ and for $-1 < \alpha, \beta, \gamma, \delta < 0$, $w(x, y)^{-1} \leq 1$, by the remarks we know this assumption is reasonable.

(5) A bivariate pairwise Mercer kernel over elliptic region.
For $\Omega = \{(x, y) \in \mathbb{R}^2 : \frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1\}$ and $w(x, y) = (1 - \frac{x^2}{a^2} - \frac{y^2}{b^2})^\gamma$, $a, b > 0$, $\gamma > -1$, Malave defined the following the orthogonal polynomials (see [2, 41])

$$S_{p,m}^{(\gamma)}(x, y) = \frac{p_{p-m}^{(\gamma+m+\frac{1}{2}, \gamma+m+\frac{1}{2})}(x)}{a^{p-\gamma}} \left(1 - \frac{x^2}{a^2}\right)^{\frac{\gamma}{2}} \frac{w}{b\sqrt{1 - \frac{x^2}{a^2}}}, \quad p \geq m \geq 0.$$

We may define a bivariate pairwise Mercer kernel by

$$K_{(x,y)}^{(\gamma)}(x', y') = \sum_{p=0}^{\infty} \lambda_p \sum_{k=0}^{p} (h_{p,m}^{(\gamma)})^{-2} S_{p,m}^{(\gamma)}(x, y) S_{p,m}^{(\gamma)}(x', y'), \quad (x, y), (x', y') \in \Omega,$$

where

$$(h_{p,m}^{(\gamma)})^2 = \int_{\Omega} |S_{p,m}^{(\gamma)}(x, y)|^2 w(x, y) dx dy, \quad \lambda_p > 0.$$

The $\lambda_p$ are chosen such that the series

$$\sum_{p=0}^{\infty} \lambda_p \sum_{k=0}^{p} (h_{p,m}^{(\gamma)})^{-2} |S_{p,m}^{(\gamma)}(x, y)|^2$$

converges uniformly on $\Omega$.

Other bivariate polynomials can be seen from [2, 31, 59] and Chapter 20 of [22]. The multivariables extension of above bivariate polynomials can be found from [3, 32].

### 2.2. Symmetric and antisymmetric pairwise Mercer kernels.

In this subsection, we give two methods for constructing symmetric and antisymmetric pairwise Mercer kernels with orthogonal polynomial systems.

(1) To construct symmetric and antisymmetric pairwise Mercer kernels with Jacoby orthogonal algebraic polynomials.

Let

$$p_{n}^{(\alpha, \beta)}(x) = \frac{2^n n!}{(n + \alpha + \beta + 1)_n} p_{n}^{(\alpha, \beta)}(x).$$

Define

$$f_{n,k}^{(\alpha, \beta)}(x, y) = p_{n+1}^{(\alpha, \beta)}(x) p_{k}^{(\alpha, \beta)}(y) - p_{n}^{(\alpha, \beta)}(x) p_{n+1}^{(\alpha, \beta)}(y), \quad 0 \leq k \leq n$$

and

$$g_{n,k}^{(\alpha, \beta)}(x, y) = \begin{cases} p_{n}^{(\alpha, \beta)}(x) p_{k}^{(\alpha, \beta)}(y) \quad &n > k, \\ p_{n+1}^{(\alpha, \beta)}(x) p_{k}^{(\alpha, \beta)}(y) \quad &n = k. \end{cases}$$

Then we know by page 504 of [59] or page 55 of [33] that both $\{f_{n,k}^{(\alpha, \beta)}(x, y) : k = 0, 1, \cdots, n; n = 0, 1, \cdots\}$ and $\{g_{n,k}^{(\alpha, \beta)}(x, y) : k = 0, 1, \cdots, n; n = 0, 1, \cdots\}$ are bivariate orthogonal polynomial systems on simplex $\Delta_1 = \{(x, y) : -1 < y < x < 1\}$ with respect to the weight function $w^{(\alpha, \beta)}(x, y) = (1 - x)^{\alpha}(1 + x)^{\beta}(1 - y)^{\alpha}(1 + y)^{\beta}, \alpha > -1, \beta > -1$. Denote $\Delta_2 = \{(x, y) : -1 < x < y < 1\}$ and $\Omega = [-1,1] \times [-1,1]$. Define bivariate polynomial function systems respectively as

$$f_{n,k}^{(\alpha, \beta)}(x, y) = \begin{cases} f_{n,k}^{(\alpha, \beta)}(x, y), \quad & (x, y) \in \Delta_1, \\ -f_{n,k}^{(\alpha, \beta)}(y, x), \quad & (x, y) \in \Delta_2 \end{cases}$$

and

$$g_{n,k}^{(\alpha, \beta)}(x, y) = \begin{cases} g_{n,k}^{(\alpha, \beta)}(x, y), \quad & (x, y) \in \Delta_1, \\ g_{n,k}^{(\alpha, \beta)}(y, x), \quad & (x, y) \in \Delta_2. \end{cases}$$
Also they are bivariate orthogonal polynomial systems on $\Omega$. With these notions in
hand we define a bivariate symmetric pairwise Mercer kernel as

$$\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( w_{n,k}^{(\alpha,\beta)} \right)^2 g_{n,k}^{(\alpha,\beta)}(x,y) g_{n,k}^{(\alpha,\beta)}(x',y), \ (x,y), (x',y') \in \Omega,$$

where

$$\left( w_{n,k}^{(\alpha,\beta)} \right)^2 = \int_{\Omega} \left| g_{n,k}^{(\alpha,\beta)}(x,y) \right|^2 w^{(\alpha,\beta)}(x,y) dx dy.$$  

The $\lambda_n$ are chosen such that the series

$$\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( w_{n,k}^{(\alpha,\beta)} \right)^2 \left| g_{n,k}^{(\alpha,\beta)}(x,y) \right|^2$$

converges uniformly on $\Omega$.

And a bivariate antisymmetric pairwise Mercer kernel

$$\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( d_{n,k}^{(\alpha,\beta)} \right)^2 f_{n,k}^{(\alpha,\beta)}(x,y) f_{n,k}^{(\alpha,\beta)}(x',y'), \ (x,y), (x',y') \in \Omega,$$

where

$$\left( d_{n,k}^{(\alpha,\beta)} \right)^2 = \int_{\Omega} \left| f_{n,k}^{(\alpha,\beta)}(x,y) \right|^2 w^{(\alpha,\beta)}(x,y) dx dy.$$  

The $\lambda_n$ are chosen such that the series

$$\sum_{n=0}^{\infty} \lambda_n \sum_{k=0}^{n} \left( d_{n,k}^{(\alpha,\beta)} \right)^2 \left| f_{n,k}^{(\alpha,\beta)}(x,y) \right|^2$$

converges uniformly on $\Omega$.

(2) Symmetric and anti-symmetric bivariate kernels associating with exponential functions

Define $\mathbb{F}(S_2^{aff})_1 = \{(x,y) \in (0,1) \times (0,1), x > y\}$, and $\mathbb{F}(S_2^{aff})_2 = (0,1) \times (0,1) - \mathbb{F}(S_2^{aff})_1$. Then [7] defined a symmetric kernel

$$E^+_{(k,l)}(x,y) = e^{2\pi i (kx+ly)} + e^{2\pi i (ky+lx)}, (x,y) \in \mathbb{F}(S_2^{aff})_1,$$

and an anti-symmetric kernel

$$E^-_{(k,l)}(x,y) = e^{2\pi i (kx+ly)} - e^{2\pi i (ky+lx)}, (x,y) \in \mathbb{F}(S_2^{aff})_1.$$  

where $k, l \in \mathbb{Z}$ and $k \geq l$.

Take

$$E^s_{(k,l)}(x,y) = \begin{cases} E^+_{(k,l)}(x,y), & (x,y) \in \mathbb{F}(S_2^{aff})_1, \\ E^+_{(k,l)}(y,x), & (x,y) \in \mathbb{F}(S_2^{aff})_2. \end{cases}$$

and

$$E^{anti}_{(k,l)}(x,y) = \begin{cases} E^-_{(k,l)}(x,y), & (x,y) \in \mathbb{F}(S_2^{aff})_1, \\ -E^-_{(k,l)}(y,x), & (x,y) \in \mathbb{F}(S_2^{aff})_2. \end{cases}$$

Then $E^s_{(k,l)}(x,y)$ is a symmetric kernel on $(0,1) \times (0,1)$ and $E^{anti}_{(k,l)}(x,y)$ is an anti-symmetric bivariate kernel on $(0,1) \times (0,1)$.
With above notions, for \((x, y), (x', y') \in (0, 1) \times (0, 1)\), we now define a bivariate exponential symmetric Mercer kernel as

\[
K^s_{(x,y)}(x', y') = \sum_{n=0}^{+\infty} \lambda_n \sum_{k+l=n} \left(h^{(s)}_{(n,k,l)}\right)^{-2} E^{s}_{(k,l)}(x,y) E^{s}_{(k,l)}(x', y'),
\]

where \(E^{s}_{(k,l)}\) denote the conjugate function of \(E^{s}_{(k,l)}\), \(\lambda_n > 0\), \(\sum_{n=0}^{+\infty} n \lambda_n < +\infty\) and \(h^{(s)}_{(n,k,l)}\) is a constant, the quadratic pairwise loss function is defined as

\[
V_{\sigma}(f(x, x') - \delta(y, y')) = \sigma^2 \left(1 + \left(\frac{f(x, x') - \delta(y, y')}{\sigma}\right)^2 - 1\right),
\]

where \(\sigma \in (0, 1]\) is a scale parameter, and \(\sup_{y,y' \in Y} |\delta(y, y')| \leq M\) with some constant \(M > 0\). Some properties of the quadratic pairwise loss function are given below.

**Proposition 2.** For any \(f, g \in \mathcal{H}_K\), there holds

\[
| V_{\sigma}(f(x, x') - \delta(y, y')) - V_{\sigma}(g(x, x') - \delta(y, y')) | \leq \sigma | f(x, x') - g(x, x') |.
\]

**Proof.** According to the median value theorem, for any \(f, g \in \mathcal{H}_K\), there exists \(\xi \) between \(f(x, x')\) and \(g(x, x')\), such that

\[
| V_{\sigma}(f(x, x') - \delta(y, y')) - V_{\sigma}(g(x, x') - \delta(y, y')) | = | V_{\sigma}(\xi) | | (f(x, x') - g(x, x')) | = \frac{| \xi |}{\sqrt{1 + | \frac{\xi}{\sigma} |^2}} | (f(x, x') - g(x, x')) |.
\]

And since for any \(\xi \in \mathbb{R}\), we have \(\sqrt{1 + | \frac{\xi}{\sigma} |^2} \geq | \frac{\xi}{\sigma} |\), so that,

\[
| V_{\sigma}(f(x, x') - \delta(y, y')) - V_{\sigma}(g(x, x') - \delta(y, y')) | \leq \sigma | f(x, x') - g(x, x') |. \quad \Box
\]

Take \( | V_{\sigma} |_0 := \sup_{y,y' \in Y} V_{\sigma}(\delta(y, y'))\). It is easy to see that

\[
| V_{\sigma} |_0 = \sigma^2 \left(\sqrt{1 + \left(\frac{M}{\sigma}\right)^2} - 1\right) \leq M \sigma.
\]

For a hypothesis \(f \in \mathcal{H}_K\), the local regularized empirical error with the quadratic loss function is

\[
\tilde{E}_{\lambda, \sigma}^l(f) = \frac{1}{l-1} \sum_{j=1}^{l-1} V_{\sigma}(f(x_t, x_j) - \delta(y_t, y_j)) + \frac{\lambda}{2} \|f\|^2_K.
\]
and the gradient of the local regularized empirical error $\hat{E}_{\lambda,\sigma}^t(f)$ at $f = f_t$ is explicitly given by

$$
\nabla \hat{E}_{\lambda,\sigma}^t(f) |_{f=f_t} = \frac{1}{t-1} \sum_{j=1}^{t-1} V_\sigma' \left( f_t(x_t, x_j) - \delta(y_t, y_j) \right) K_{(x_t,x_j)}(\cdot,\cdot) + \lambda f_t \\
= \frac{1}{t-1} \sum_{j=1}^{t-1} \left( \frac{f_t(x_t, x_j) - \delta(y_t, y_j)}{\sqrt{1 + \left( \frac{K_{(x_t,x_j)} - \delta(y_t, y_j)}{\sigma} \right)^2}} \right) K_{(x_t,x_j)}(\cdot,\cdot) + \lambda f_t.
$$

At the current iteration point $f_t$, updating the predictor by using the negative gradient direction $-\nabla \hat{E}_{\lambda,\sigma}^t(f) |_{f=f_t}$ as the search direction, the online pairwise learning algorithm with quadratic loss function $V_\sigma(r)$ is given as

$$
\begin{align*}
\begin{cases}
  f_1 = f_2 = 0, \\
  f_{t+1} = f_t - \eta_t \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \left( \frac{f_t(x_t, x_j) - \delta(y_t, y_j)}{\sqrt{1 + \left( \frac{K_{(x_t,x_j)} - \delta(y_t, y_j)}{\sigma} \right)^2}} \right) K_{(x_t,x_j)}(\cdot,\cdot) + \lambda f_t \right),
\end{cases}
\end{align*}
$$

(3.1)

For any predictor $f$, $z = (x, y) \in Z, z' = (x', y') \in Z$, define the corresponding general error by

$$
E_\sigma(f) := \int_Z \int_Z V_\sigma(f(x,x') - \delta(y,y')) \, d\rho(x,y) \, d\rho(x',y').
$$

(3.2)

Denote

$$
f_\sigma = \arg \min_{f \in H_K} E_\sigma(f),
$$

(3.3)

and

$$
f_{\lambda,\sigma} = \arg \min_{f \in H_K} E_{\lambda,\sigma}(f),
$$

(3.4)

where $E_{\lambda,\sigma}(f) = E_\sigma(f) + \frac{\lambda}{2} \| f \|_K^2$.

Furthermore, we define

$$
\hat{E}_\sigma^t(f) = \frac{1}{t-1} \sum_{j=1}^{t-1} \int_Z V_\sigma(f(x, x_j) - \delta(y, y_j)) \, d\rho(x,y),
$$

and $\hat{E}_{\lambda,\sigma}^t(f) = \hat{E}_\sigma^t(f) + \frac{\lambda}{2} \| f \|_K^2$. With these notions in hand, we can give the performance analysis for the algorithm (3.1).

4. **The Performance.** In this section, we present our main results about the performance of algorithm (3.1) developed in Section 3. Proofs are given in Section 5.2.

4.1. **Convergence analysis.** Now we provide a theorem which shows that under some conditions on the step sizes, the last iterate $\{f_T\}$ generated by algorithm (3.1) converges to the optimal function $f_{\lambda,\sigma}$ stated in (3.4). In this paper, we assume $T \geq 4$.

**Theorem 4.1.** For every fixed $\lambda > 0$ if the step sizes $\{\eta_t : t \in \mathbb{N}\}$ in algorithm (3.1) satisfy $\eta_t \lambda \leq 1$ for $t \geq 2$ and

$$
\lim_{t \to \infty} \eta_t = 0, \quad \sum_{t=2}^{\infty} \eta_t = \infty,
$$

then $\{f_T\}$ converges to the optimal function $f_{\lambda,\sigma}$.
then we have
\[ \lim_{T \to \infty} \mathbb{E}\left(\|f_T - f_{\lambda,\sigma}\|_K\right) = 0.\]

### 4.2. Learning rates.

Specifying the stepsize \( \eta \) in algorithm (3.1), the learning rate of the last iterate may be stated as follows.

**Theorem 4.2.** Let \( \lambda \in (0, 1) \) and \( \{f_t\}_{t=1}^{T+1} \) be the function sequence generated by algorithm (3.1). If the stepsize are chosen as \( \eta = \frac{1}{\lambda t^{-\theta}} \) and \( \theta \in (0, 1) \), then
\[
\mathbb{E}\left(\|f_{T+1} - f_{\lambda,\sigma}\|_K^2\right) \leq \frac{C_\theta \sigma}{\lambda^2 T^{\theta'}}\]
where \( \theta' = \min\{\theta, \frac{1}{2}\} \), \( C_\theta = 2M \left(\frac{\theta}{(1-\frac{\theta}{2})^2 - \frac{\theta}{2}}\right)^{\frac{\theta}{2}} + 8\kappa^2 \left(19 + \frac{3^2}{4\kappa^2}\varepsilon\right)\left(\frac{1}{(1/2-\frac{\theta}{2})^2}\right)^{\frac{1}{2}} \).

Furthermore if the stepsize are chosen as \( \eta = \frac{1}{\lambda T} \), then we have the following result.

**Theorem 4.3.** Let \( \lambda \in (0, 1) \) and \( \{f_t\}_{t=1}^{T+1} \) be the function sequence generated by algorithm (3.1). If the stepsize are chosen as \( \eta = \frac{1}{\lambda T} \), then
\[
\mathbb{E}\left(\|f_{T+1} - f_{\lambda,\sigma}\|_K^2\right) \leq c \frac{\sigma}{\lambda^2 \sqrt{T}},
\]
where \( c = 2M + (4 + 8\sqrt{2})\kappa^2 \).

The theorems provided above mainly describe the convergence rate of \( \|f_{T+1} - f_{\lambda,\sigma}\|_K \) which is usually referred as the sample error. However, in the studying on the learning performance of learning algorithms, we are often interested in the excess generalization error \( \mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma) \) (see [20, 26]). Define
\[
\mathcal{K}(f, \lambda) = \inf_{g \in \mathcal{H}_K} \left(\mathcal{E}_\sigma(g) - \mathcal{E}_\sigma(f) + \frac{\lambda}{2}\|g\|_K^2\right),
\]
which is often used to denote the approximation error, whose convergence is determined by the capacity of \( \mathcal{H}_K \). By combining the sample error and approximation error, we obtain the overall learning rate stated as follows.

**Corollary 1.** Let \( \lambda \in (0, 1] \), \( \mathcal{E}_\sigma(f) \) be the generalization error defined as (3.2). Denoted by \( \{f_t\}_{t=1}^{T+1} \) the function sequence generated by algorithm (3.1). If the stepsize are chosen as \( \eta = \frac{1}{\lambda t^{-\theta}} \) with \( \theta \in (0, 1) \), then
\[
\mathbb{E}[\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma)] \leq \frac{\kappa\sqrt{C_\theta \sigma^2}}{\lambda T^{\theta'}} + \mathcal{K}(f_\sigma, \lambda),
\]
where \( \theta' = \min\{\theta, \frac{1}{2}\} \), \( C_\theta = 2M \left(\frac{\theta}{(1-\frac{\theta}{2})^2 - \frac{\theta}{2}}\right)^{\frac{\theta}{2}} + 8\kappa^2 \left(19 + \frac{3^2}{4\kappa^2}\varepsilon\right)\left(\frac{1}{(1/2-\frac{\theta}{2})^2}\right)^{\frac{1}{2}} \).

In learning theory, we often assume the \( \mathcal{K} \)-functionals satisfy certain decay, for example, we assume (see e.g. [53, 55, 56, 62]).
\[
\mathcal{K}(f_\sigma, \lambda) = O(\lambda^\beta), \quad \lambda \to 0^+,
\]
where \( 0 < \beta \leq 1 \). In fact, under the assumption (4.2), we now have the following Corollary, which follows directly from Theorem 4.3.

**Corollary 2.** Let \( \beta \in (0, 1] \). If \( \mathcal{K}(f_\sigma, \lambda) = O(\lambda^\beta) \) and \( \eta = \frac{1}{\lambda T} \), then
\[
\mathbb{E}[\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_\sigma)] = O\left(\frac{\sigma^2}{\lambda^{\beta} \sqrt{T}} + \lambda^\beta\right), \lambda \to 0^+.
\]
4.3. **Further discussions.** We now give some discussions on the main results of the paper.

- In this paper, we analyze the convergence rate of the kernel-based regularization online pairwise learning algorithm with a quadratic loss function. The results show that the scale parameter $\sigma$ can effectively control the convergence rate of learning algorithm. Depending on the circumstances, the learning rates can be greatly improved by choosing the parameter $\sigma$ properly, which are better than existing error bounds.

  In fact take $\sigma = \lambda^2$. Then by (4.1) we have
  $$
  E[E_\sigma(f_T) - E_\sigma(f_\sigma)] \leq \frac{\kappa \sqrt{C}}{T} \cdot \frac{\theta}{2} + \mathcal{K}(f_\sigma, \lambda),
  $$
  which shows that the parameter $\lambda$ may be removed.

- By the Lipschitzian of $V_\sigma$, we know there exits a constant $C > 0$ such that the $K-$functional $\mathcal{K}(f_\sigma, \lambda)$ satisfies
  $$
  \mathcal{K}(f_\sigma, \lambda) \leq C \inf_{g \in \mathcal{H}_K} ||g - f_\sigma||_{L^2(\rho_X)} + \lambda ||g||^2_{K}, \quad \lambda > 0.
  $$
  So the assumption (4.2) is reasonable if $\mathcal{H}_K$ is density in $L^2(\rho_X)$ (see e.g. [57]).

- We establish the error bounds for the learning sequence using convex analysis and Rademacher complexity, while previous analysis rely on integral operator theory [72] or covering numbers approach [64]. And the results in [72] are in probability while our results are obtained in expectation. The quadratic loss function is not a strongly convex function, we have relaxed the strong convexity assumption in the literature [26]. And our method may be extended to some online pairwise learning algorithms with non-convex loss functions, e.g., the robust loss function in [62].

- For the special case $V(f(x, x') - \delta(y, y')) = (y - y' - f(x, x'))^2$, [72] studied the relations between the pairwise Mercer kernel and pointwise Mercer kernel. We consider the more general situation in this paper. [45] provided a way to construct pairwise kernels by considering a projection from the set of all kernels to the set of permutation invariant kernels. Unlike [45], we provide a method of constructing pairwise Mercer kernels with classical bivariate Jacobi orthogonal polynomials.

5. **Proofs.** In this section, we first show some important lemmas, and then give the proofs of the main results. We use $E_z(\cdot)$ to denote the expectation with respect to $z$. When underlying random variable in expectation is clear from the context, we will simply write $E(\cdot)$.

5.1. **Some lemmas.** To prove the results in section 4, we need some lemmas as follows.

**Lemma 5.1 ([6]).** Let $(\mathcal{H}, \| \cdot \|_{\mathcal{H}})$ be a Hilbert space, and $F(f) : \mathcal{H} \to \mathbb{R} \cup \{\mp \infty\}$ a function defined on $\mathcal{H}$. Then

(i) If $F(f)$ is a Gâteaux differentiable function, then $F(f)$ is a convex function on $\mathcal{H}$ if and only if for any $f, g \in \mathcal{H}$ there holds

$$
F(g) - F(f) \geq \langle g - f, \nabla F(f) \rangle_{\mathcal{H}}.
$$

(ii) If $F(f)$ is a differentiable and convex function, then $F(f)$ attains minimal value at $f^*$ if and only if $\nabla F(f^*) = 0$.

(iii) For any $f, g \in \mathcal{H}$ there holds

$$
\|f\|^2_{\mathcal{H}} - \|g\|^2_{\mathcal{H}} = \langle f - g, 2g \rangle_{\mathcal{H}} + \|f - g\|^2_{\mathcal{H}}.
$$
Following Lemma 5.2 gives some properties of the generalization error $\mathcal{E}_\sigma(f)$.

**Lemma 5.2.** (i) $\mathcal{E}_\sigma(f)$ satisfies the following equality

$$\nabla f \mathcal{E}_\sigma(f)(\cdot) = - \int_Z \int_Z \frac{\delta(y,y') - f(x,x')}{\sqrt{1 + (\frac{\delta(y,y') - f(x,x')}{\sigma})^2}} K(x,x')(\cdot,\cdot) d\rho(x,y) d\rho(x',y'), f \in \mathcal{H}_K. \tag{5.1}$$

(ii) $\mathcal{E}_\sigma(f)$ is a convex function on $\mathcal{H}_K$.

**Proof.** We prove (i) firstly. For any $f, g \in \mathcal{H}_K$ simple computations show that

$$\lim_{t \to 0} \frac{\mathcal{E}_\sigma(f + tg) - \mathcal{E}_\sigma(f)}{t} = \sigma^2 \lim_{t \to 0} \int_Z \int_Z \left\{ \frac{1}{\sqrt{1 + (\frac{\delta(y,y') - f(x,x')}{\sigma})^2}} - \frac{1}{\sqrt{1 + (\frac{\delta(y,y') - f(x,x')}{\sigma})^2}} \right\} d\rho(x,y) d\rho(x',y'). \tag{5.2}$$

By the Taylor formula we have

$$\sqrt{1 + \left( \frac{\delta(y,y') - f(x,x')}{\sigma} \right)^2} = \sqrt{1 + \left( \frac{\delta(y,y') - f(x,x')}{\sigma} \right)^2} + \frac{t}{\sigma} g(x,x') + \frac{1}{2} \frac{1}{\sqrt{1 + (\frac{\delta(y,y') - f(x,x')}{\sigma})^2}} \left( \frac{t}{\sigma} g(x,x') \right)^2 + \cdots \tag{5.3}$$

Substituting (5.3) into (5.2), we have

$$\lim_{t \to 0} \frac{\mathcal{E}_\sigma(f + tg) - \mathcal{E}_\sigma(f)}{t} = \int_Z \int_Z \left( \frac{g(x,x')}{(\frac{\delta(y,y') - f(x,x')}{\sigma})^2} \right) d\rho(x,y) d\rho(x',y').$$

Using the reproducing property $g(x,x') = \langle K(x,x')(\cdot,\cdot), g \rangle_K$, we get

$$\lim_{t \to 0} \frac{\mathcal{E}_\sigma(f + tg) - \mathcal{E}_\sigma(f)}{t} = \langle \int_Z \int_Z \frac{\delta(y,y') - f(x,x')}{\sqrt{1 + (\frac{\delta(y,y') - f(x,x')}{\sigma})^2}} K(x,x')(\cdot,\cdot) d\rho(x,y) d\rho(x',y'), g \rangle_K.$$

Thus

$$\nabla f \mathcal{E}_\sigma(f)(\cdot) = - \int_Z \int_Z \frac{\delta(y,y') - f(x,x')}{\sqrt{1 + (\frac{\delta(y,y') - f(x,x')}{\sigma})^2}} K(x,x')(\cdot,\cdot) d\rho(x,y) d\rho(x',y').$$

(i) is proved.

Now we prove (ii).

For arbitrary $u, v \in \mathbb{R}$, by the second-order Taylor expansion of $V_\sigma(r)$ at $r = v$, there exists $\xi$ between $u$ and $v$, such that

$$V_\sigma(u) - V_\sigma(v) = \frac{v}{\sqrt{1 + (\frac{v}{\sigma})^2}} (u - v) + \frac{1}{2 \sqrt{(1 + (\frac{v}{\sigma})^2)^3}} (u - v)^2.$$
Since \( \frac{1}{\sqrt{1+(\frac{t}{j})^2}}>0 \), we know, for any \( f,g \in \mathcal{H}_K \),
\[
\mathcal{E}_\sigma(f) - \mathcal{E}_\sigma(g)
= \int_z \int_z \left( V_\sigma(f(x,x')) - g(y,y') \right) d\rho(x) d\rho(y)
\geq \int_z \int_z \left( f(x,x') - g(x,x') \right) d\rho(x) d\rho(y)
= \left\langle f - g, \int_z \int_z \frac{\delta(y,y') - g(x,x')}{1 + (\frac{\delta(y,y') - g(x,x')}{\sigma})^2} K_{x,x'}(\cdot,\cdot) d\rho(x) d\rho(y) \right\rangle
= \left\langle f - g, \nabla_x \mathcal{E}_\sigma(g) \right\rangle_K.
\]
(5.4) Using Lemma 5.1 (i), \( \mathcal{E}_\sigma(f) \) is a convex function on \( \mathcal{H}_K \). We get our desired result.

The following Lemma shows that the function sequence \( \{f_t\} \) is bounded.

**Lemma 5.3.** Let \( \lambda > 0 \), \( \{f_t\} \) be the function sequence generated by the algorithm 3.1. If the stepsize \( \{\eta_t\} \) satisfies \( \eta_t \lambda \leq 1, \forall t \geq 2 \), then for every \( t \in \mathbb{N} \) there holds
\[
\|f_t\|_K \leq \frac{\kappa \sigma}{\lambda}.
\]
(5.5)

**Proof.** We prove this conclusion by induction. The initial functions \( f_1 = f_2 = 0 \) certainly satisfy (5.5). Assuming \( f_t \) satisfies the inequality (5.5), we now prove \( f_{t+1} \) also satisfies the inequality (5.5). Since
\[
|V_{\eta_t}'(f_t(x_t, x_j) - \delta(y_t, y_j))| = \left| \frac{\delta(y_t, y_j) - f(x_t, x_j)}{1 + (\frac{\delta(y_t, y_j) - f(x_t, x_j)}{\sigma})^2} \right| \leq \sigma,
\]
(5.6)
we have
\[
\|f_{t+1}\|_K = (1 - \lambda \eta_t) \|f_t\|_K - \frac{\eta_t}{t-1} \sum_{j=1}^{t-1} |V_{\eta_t}'(f_t(x_t, x_j) - \delta(y_t, y_j))| K(x_t, x_j)(\cdot, \cdot)|_K
\leq (1 - \lambda \eta_t) \|f_t\|_K + \frac{\eta_t}{t-1} \sum_{j=1}^{t-1} |V_{\eta_t}'(f_t(x_t, x_j) - \delta(y_t, y_j))| \cdot \|K(x_t, x_j)(\cdot, \cdot)\|_K
\leq (1 - \lambda \eta_t) \|f_t\|_K + \eta_t \kappa \sigma.
\]
By the assumption \( \|f_t\|_K \leq \frac{\kappa \sigma}{\lambda} \) and \( \eta_t \lambda \leq 1 \) we get
\[
\|f_{t+1}\|_K \leq (1 - \lambda \eta_t) \frac{\kappa \sigma}{\lambda} + \eta_t \kappa \sigma = \frac{\kappa \sigma}{\lambda}.
\]
This completes the proof of the lemma.

**Lemma 5.4.** Let \( \lambda > 0 \), \( \{f_t\} \) be the function sequence generated by the algorithm 3.1. If the stepsize \( \{\eta_t\} \) satisfies \( \eta_t \lambda \leq 1, \forall t \geq 2 \), then there holds inequality
\[
\mathbb{E}\left( \|f_{t+1} - f_{\lambda,\sigma}\|_K^2 \right) \leq \mathbb{E}\left( \|f_t - f_{\lambda,\sigma}\|_K^2 \right) + 4\eta_t^2 \kappa^2 \sigma^2 + 2\eta_t \mathbb{E}\left( \mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\lambda,\sigma}(f_t) \right).
\]
Proof. Denote \( \hat{A}_{\lambda,\sigma} = \frac{1}{t} \sum_{j=1}^{t-1} V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) K(x_j, x_j)(\cdot, \cdot) + \lambda f_t \). Then \( f_{t+1} = f_t - \eta_t \hat{A}_{\lambda,\sigma} \). Therefore
\[
\mathbb{E}\left( \left\| f_{t+1} - f_{\lambda,\sigma} \right\|_K^2 \right) = \mathbb{E}\left( \left\| f_t - f_{\lambda,\sigma} \right\|_K^2 \right) + \eta_t^2 \mathbb{E}\left( \left\| \hat{A}_{\lambda,\sigma} \right\|_K^2 \right) + 2\eta_t \mathbb{E}\left( (f_{\lambda,\sigma} - f_t, \hat{A}_{\lambda,\sigma})_K \right)
= \mathbb{E}\left( \left\| f_t - f_{\lambda,\sigma} \right\|_K^2 \right) + \eta_t^2 \mathbb{E}(A) + 2\eta_t \mathbb{E}(B),
\]
where \( A = \left\| \hat{A}_{\lambda,\sigma} \right\|_K^2, \ B = (f_{\lambda,\sigma} - f_t, \hat{A}_{\lambda,\sigma})_K \).

(i) At first, we estimate \( A \). From (5.6) and Lemma 5.3, one can see that
\[
A = \left\| \hat{A}_{\lambda,\sigma} \right\|_K^2 = \left\| \frac{1}{t-1} \sum_{j=1}^{t-1} V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) K(x_j, x_j)(\cdot, \cdot) + \lambda f_t \right\|_K^2
\leq \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \left| V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) \right| \cdot \left\| K(x_j, x_j) \right\|_K + \lambda \left\| f_t \right\|_K \right)^2
\leq (\kappa \sigma + \lambda \left\| f_t \right\|_K)^2 \leq (\kappa \sigma + \lambda \frac{\kappa \sigma}{\lambda})^2 = 4 \kappa^2 \sigma^2.
\]
(ii) We now estimate \( B \).

By the definition of \( \hat{A}_{\lambda,\sigma} \) and the reproducing property, we know that
\[
B = (f_{\lambda,\sigma} - f_t, \hat{A}_{\lambda,\sigma})_K
= (f_{\lambda,\sigma} - f_t, \frac{1}{t-1} \sum_{j=1}^{t-1} V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) K(x_j, x_j)(\cdot, \cdot) + \lambda f_t)_K
= \frac{1}{t-1} \sum_{j=1}^{t-1} V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) (f_{\lambda,\sigma} - f_t, K(x_j, x_j))_K + \lambda (f_{\lambda,\sigma} - f_t, f_t)_K
= \frac{1}{t-1} \sum_{j=1}^{t-1} V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) (f_{\lambda,\sigma}(x_t, x_j) - f_t(x_t, x_j))
+ \lambda (f_{\lambda,\sigma}, f_t)_K - \lambda \left\| f_t \right\|_K^2.
\]
By the inequality \( (f_{\lambda,\sigma}, f_t)_K \leq \frac{1}{2}(\left\| f_{\lambda,\sigma} \right\|_K^2 + \left\| f_t \right\|_K^2) \) we have
\[
B \leq \frac{1}{t-1} \sum_{j=1}^{t-1} V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) (f_{\lambda,\sigma}(x_t, x_j) - f_t(x_t, x_j))
+ \lambda \left( \frac{1}{2}(\left\| f_{\lambda,\sigma} \right\|_K^2 + \left\| f_t \right\|_K^2) - \left\| f_t \right\|_K^2 \right)
= \frac{1}{t-1} \sum_{j=1}^{t-1} V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) (f_{\lambda,\sigma}(x_t, x_j) - f_t(x_t, x_j))
+ \lambda \left( \frac{1}{2} \left\| f_{\lambda,\sigma} \right\|_K^2 - \frac{1}{2} \left\| f_t \right\|_K^2 \right).
\]
The convexity of \( V_\sigma(r), r \in \mathbb{R} \) implies that
\[
V'_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) (f_{\lambda,\sigma}(x_t, x_j) - f_t(x_t, x_j))
\leq V_\sigma(f_{\lambda,\sigma}(x_t, x_j) - \delta(y_t, y_j)) - V_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)).
\]
Thus
\[ B \leq \frac{1}{t-1} \sum_{j=1}^{t-1} V_\sigma(f_{\lambda,\sigma}(x_t, x_j) - \delta(y_t, y_j)) - \frac{1}{t-1} \sum_{j=1}^{t-1} V_\sigma(f_t(x_t, x_j) - \delta(y_t, y_j)) \]
\[ + \frac{\lambda}{2} \|f_{\lambda,\sigma}\|_K^2 - \frac{\lambda}{2} \|f_t\|_K^2 = \tilde{E}_{\lambda,\sigma}^t(f_{\lambda,\sigma}) - \tilde{E}_{\lambda,\sigma}^t(f_t). \tag{5.9} \]

Combining (5.7), (5.8) and (5.9), we have
\[ E(\|f_{t+1} - f_{\lambda,\sigma}\|_K^2) \leq E(\|f_t - f_{\lambda,\sigma}\|_K^2) + 4\eta_t \kappa^2 \sigma^2 + 2\eta_t E(\tilde{E}_{\lambda,\sigma}^t(f_{\lambda,\sigma}) - \tilde{E}_{\lambda,\sigma}^t(f_t)). \]

This completes the proof of Lemma 5.4.

To prove the main results, we need the concept of Rademacher complexity and its important property.

**Definition 5.5** ([5]). Let \( F \) be a class of uniformly bounded functions. For any integer \( n \), we call
\[ E_{\mathbf{z}}E_{\tau}\left( \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \tau_if(z_i) \right) \]
the Rademacher average over \( F \), where \( \mathbf{z} = \{z_i\}_{i=1}^{n} \) are independent random variables distributed according to some probability measure and \( \tau = \{\tau_i\}_{i=1}^{n} \) are independent Rademacher random variables, i.e. \( P(\tau_i = 1) = P(\tau_i = -1) = \frac{1}{2} \).

We restate the following important property of Rademacher complexity (see [43] for details).

**Lemma 5.6** ([43]). Let \( \Theta \) be a domain, \( \{g_j(\theta)\} \) and \( \{h_j(\theta)\} \) be sets of functions defined for all \( \theta \in \Theta \). If for each \( j, \theta, \theta' \) that
\[ |g_j(\theta) - g_j(\theta')| \leq |h_j(\theta) - h_j(\theta')|, \]
then
\[ E_{\tau}\left( \sup_{\theta \in \Theta} \sum_{j=1}^{m} \tau_jg_j(\theta) \right) \leq E_{\tau}\left( \sup_{\theta \in \Theta} \sum_{j=1}^{m} \tau_jh_j(\theta) \right). \]

**Lemma 5.7.** If the stepsize \( \{\eta_t : t \in \mathbb{N}\} \) in algorithm (3.1) satisfies \( \eta_t \lambda \leq 1, \forall t \geq 2 \), then
\[ E(\tilde{E}_{\lambda,\sigma}^t(f_{\lambda,\sigma}) - \tilde{E}_{\lambda,\sigma}^t(f_t)) \leq E(\mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\lambda,\sigma}(f_t)) + \frac{2\kappa^2 \sigma^2}{\lambda \sqrt{t-1}}. \]

**Proof.** One has
\begin{align*}
E(\tilde{E}_{\lambda,\sigma}^t(f_{\lambda,\sigma}) - \tilde{E}_{\lambda,\sigma}^t(f_t)) &= E(\mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\lambda,\sigma}(f_t)) + \frac{\lambda}{2} \|f_{\lambda,\sigma}\|_K^2 - \mathcal{E}_{\sigma}(f_t) + \frac{\lambda}{2} \|f_t\|_K^2) \\
&= E(\mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\lambda,\sigma}(f_t)) + \frac{\lambda}{2} \|f_{\lambda,\sigma}\|_K^2 - \mathcal{E}_{\sigma}(f_t) + \frac{\lambda}{2} \|f_t\|_K^2) \\
&= E(\mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\lambda,\sigma}(f_t)) - E(\mathcal{E}_{\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\sigma}(f_t)) \\
&= E(\mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\lambda,\sigma}(f_t)) - E(\mathcal{E}_{\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\sigma}(f_t)) \tag{5.10}.
\end{align*}

It remains to estimate \( E(\mathcal{E}_{\sigma}(f_t) - \tilde{E}_{\sigma}^t(f_t)) \). Let
\[ B_{\lambda,\sigma} := \{ f \in \mathcal{H}_K : \|f\|_K \leq \frac{\kappa \sigma}{\lambda} \}. \]
and for arbitrary $f \in B_{\lambda,\sigma}$, $z' = (x', y')$, we take

$$L_\sigma(f, z') = \int_Z V_\sigma(f(x, x') - \delta(x, y')) \, d\rho(x, y).$$

Then we can write $\tilde{E}_\sigma^t(f) = \frac{1}{t-1} \sum_{j=1}^{t-1} L_\sigma(f, z_j)$, and $E_\sigma(f) = \int Z L_\sigma(f, z') \, d\rho(x', y')$.

Let $\{z_j^T\}_{j=1}^t$ be another independent sample set, and $\tau_j (j = 1, \ldots, T)$ be independent variables drawn from the Rademacher distribution. Then

$$E_{z_1, \ldots, z_{t-1}} \left( \sup_{f \in B_{\lambda,\sigma}} \left( E_\sigma(f) - \tilde{E}_\sigma^t(f) \right) \right) \leq E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j (L_\sigma(f, z_j') - L_\sigma(f, z_j)) \right) \right)$$

$$\leq 2 E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j L_\sigma(f, z_j) \right) \right) .$$

(5.11)

And by using a standard symmetry technique (see for example [4]), we know that

$$E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j (L_\sigma(f, z_j') - L_\sigma(f, z_j)) \right) \right)$$

$$\leq 2 E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j L_\sigma(f, z_j) \right) \right) .$$

(5.12)

Substituting (5.12) into (5.11) and using the definition of $L_\sigma(f, z_j)$, we get

$$E_{z_1, \ldots, z_{t-1}} \left( \sup_{f \in B_{\lambda,\sigma}} \left( E_\sigma(f) - \tilde{E}_\sigma^t(f) \right) \right)$$

$$\leq 2 E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j \int_Z V_\sigma(f(x, x_j') - \delta(y, y_j)) \, d\rho(x, y) \right) \right)$$

$$\leq 2 E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j V_\sigma(f(x, x_j') - \delta(y, y_j)) \right) \right) .$$

(5.13)

On the other hand, for any $f, g \in B_{\lambda,\sigma}$, we know by Proposition 2 that

$$| V_\sigma(\delta(y, y') - f(x, x')) - V_\sigma(\delta(y, y') - g(x, x')) | \leq \sigma \, | f(x, x') - g(x, x') | ,$$

which together with Lemma 5.6 implies

$$E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j V_\sigma(\delta(y, y_j) - f(x, x_j)) \right) \right)$$

$$\leq \sigma \, E_{z_1, \ldots, z_{t-1}} E_{z_1', \ldots, z_{t-1}'} \left( \sup_{f \in B_{\lambda,\sigma}} \left( \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j f(x, x_j) \right) \right) .$$

(5.14)
Substituting (5.14) into (5.13), we have

\[
E_{z_1,\ldots,z_{t-1}} \left( \sup_{f \in B_{\lambda,\sigma}} (\mathcal{E}_\sigma(f) - \tilde{\mathcal{E}}_\sigma^t(f)) \right)
\leq 2\sigma E_{z} E_{z_1,\ldots,z_{t-1}} \mathbb{E}_T \left( \sup_{f \in B_{\lambda,\sigma}} \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j f(x, x_j) \right).
\]

The reproducing property and the Schwarz inequality admit

\[
E_{z_1,\ldots,z_{t-1}} \left( \sup_{f \in B_{\lambda,\sigma}} (\mathcal{E}_\sigma(f) - \tilde{\mathcal{E}}_\sigma^t(f)) \right)
= 2\sigma E_{z} E_{z_1,\ldots,z_{t-1}} \mathbb{E}_T \left( \sup_{f \in B_{\lambda,\sigma}} \|f\|_K \cdot \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j K(x, x_j) \right)
\leq 2\sigma E_{z} E_{z_1,\ldots,z_{t-1}} \mathbb{E}_T \left( \sup_{f \in B_{\lambda,\sigma}} \|f\|_K \cdot \frac{1}{t-1} \sum_{j=1}^{t-1} \tau_j K(x, x_j) \right).
\]

By inequality \(E[\|f\|_K] \leq (E[\|f\|_K^2])^{\frac{1}{2}}\) and the definition of \(B_{\lambda,\sigma}\), one can see that

\[
E_{z_1,\ldots,z_{t-1}} \left( \sup_{f \in B_{\lambda,\sigma}} (\mathcal{E}_\sigma(f) - \tilde{\mathcal{E}}_\sigma^t(f)) \right) \leq \frac{2\sigma^2}{\lambda(t-1)} \left( E_{z} E_{z_1,\ldots,z_{t-1}} \mathbb{E}_T \left\| \sum_{j=1}^{t-1} \tau_j K(x, x_j) \right\|_K^2 \right)^{\frac{1}{2}}
\]

(5.15)

And since \(\tau_j (j = 1, \ldots, T)\) are independent variables with \(E(\tau_j) = 0, E(\tau_j^2) = 1\). This implies that

\[
E(\tau_i \tau_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}
\]

So

\[
E_{z} E_{z_1,\ldots,z_{t-1}} \mathbb{E}_T \left[ \left\| \sum_{j=1}^{t-1} \tau_j K(x, x_j) \right\|_K^2 \right]
= E_{z} E_{z_1,\ldots,z_{t-1}} \mathbb{E}_T \left[ \sum_{i=1}^{t-1} \sum_{j=1}^{t-1} \tau_i \tau_j \langle K(x, x_i), K(x, x_j) \rangle_K \right]
= E_{z} E_{z_1,\ldots,z_{t-1}} \mathbb{E}_T \left[ \sum_{j=1}^{t-1} \tau_j^2 \left\| K(x, x_j) \right\|_K^2 \right] = \sum_{j=1}^{t-1} E_{x, x_j} \left\| K(x, x_j) \right\|_K^2.
\]

(5.16)

Combined (5.15) with (5.16), it can be seen that

\[
E_{z_1,\ldots,z_{t-1}} \left( \sup_{f \in B_{\lambda,\sigma}} (\mathcal{E}_\sigma(f) - \tilde{\mathcal{E}}_\sigma^t(f)) \right) \leq \frac{2\kappa^2\sigma^2}{\lambda(t-1)} \left( \sum_{j=1}^{t-1} E_{x, x_j} \left\| K(x, x_j) \right\|_K^2 \right)^{\frac{1}{2}}
\leq \frac{2\kappa^2\sigma^2}{\lambda(t-1)} \left( (t-1)\kappa^2 \right)^{\frac{1}{2}} = \frac{2\kappa^2\sigma^2}{\lambda\sqrt{t-1}}.
\]

(5.17)

Substituting (5.17) into (5.10) together with Lemma 5.3, we have

\[
E(\tilde{\mathcal{E}}_{\lambda,\sigma}^t(f_{\lambda,\sigma}) - \tilde{\mathcal{E}}_{\lambda,\sigma}^t(f_t)) - E(\mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \mathcal{E}_{\lambda,\sigma}(f_t))
= E(\mathcal{E}_\sigma(f_t) - \tilde{\mathcal{E}}_\sigma^t(f_t)) \leq E \left( \sup_{f \in B_{\lambda,\sigma}} (\mathcal{E}_\sigma(f) - \tilde{\mathcal{E}}_\sigma^t(f)) \right) \leq \frac{2\kappa^2\sigma^2}{\lambda\sqrt{t-1}}.
\]
This implies that
\[
\mathbb{E}(\tilde{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) - \tilde{E}_{\lambda,\sigma}(f_t)) \leq \mathbb{E}(E_{\lambda,\sigma}(f_{\lambda,\sigma}) - E_{\lambda,\sigma}(f_t)) + \frac{2\kappa^2\sigma^2}{\lambda \sqrt{t-1}}. \quad \square
\]

**Lemma 5.8.** For any \( f \in \mathcal{H}_K \) there holds
\[
\frac{\lambda}{2} \|f - f_{\lambda,\sigma}\|^2_K \leq E_{\lambda,\sigma}(f) - E_{\lambda,\sigma}(f_{\lambda,\sigma}).
\]

**Proof.** From Lemma 5.2, \( E_{\sigma} \) is convex and differentiable on \( \mathcal{H}_K \). So by the definition of \( f_{\lambda,\sigma} \) we have
\[
\nabla E_{\sigma}(f_{\lambda,\sigma}) + \lambda f_{\lambda,\sigma} = 0.
\]

For any \( f \in \mathcal{H}_K \), (i) of Lemma 5.1 implies that
\[
E_{\lambda,\sigma}(f) - E_{\lambda,\sigma}(f_{\lambda,\sigma}) = E_{\sigma}(f) - E_{\sigma}(f_{\lambda,\sigma}) + \frac{\lambda}{2} \|f\|^2_K - \frac{\lambda}{2} \|f_{\lambda,\sigma}\|^2_K
\]
\[
\geq \langle f - f_{\lambda,\sigma}, \nabla E_{\sigma}(f_{\lambda,\sigma}) \rangle_K + \frac{\lambda}{2} (\|f\|^2_K - \|f_{\lambda,\sigma}\|^2_K)
\]
\[
= \langle f - f_{\lambda,\sigma}, \lambda f_{\lambda,\sigma} \rangle_K + \frac{\lambda}{2} (\|f\|^2_K - \|f_{\lambda,\sigma}\|^2_K) = \frac{\lambda}{2} \|f - f_{\lambda,\sigma}\|^2_K. \quad \square
\]

**Lemma 5.9.** (see [71]) Let \( 0 < \theta \leq 1 \). Then for any \( t < T \) there holds
\[
\sum_{j=t+1}^{T} j^{-\theta} \geq \begin{cases} 
\frac{1}{1 - \theta} [(T + 1)^{1-\theta} - (t + 1)^{1-\theta}], & \theta < 1, \\
\log(T + 1) - \log(t + 1), & \theta = 1.
\end{cases}
\]

**Lemma 5.10.** Let \( 0 < \theta \leq 1, 0 < \nu \leq 1, b > 0, s > 0 \). Then for any \( t < T \) there holds
\[
e^{-\nu x} \leq \left( \frac{b}{\nu e} \right)^b x^{-b}, \quad x \in (0, +\infty). \quad (5.18)
\]

\[
\sum_{j=t+1}^{T} \frac{1}{t(t+s+\theta)} \exp \left( -\nu \sum_{j=t+1}^{T} j^{-\theta} \right)
\]
\[
\leq \begin{cases} 
\frac{3}{\nu} \left( \frac{6}{T} \right)^s + \frac{3^{\nu+1} T^{1-\theta}}{(1 - \theta)^{2^{-\theta}}} \exp \left( -\nu(1 - 2^{\theta-1})^{1-\theta} (T + 1)^{1-\theta} \right), & 0 < \theta < 1, \\
\frac{(T + 1)^{\nu - s - 1}}{(T + 1)^{\nu}}, & \theta = 1.
\end{cases} \quad (5.19)
\]

**Proof.** (5.18) can be found from [58]. We only prove (5.19). When \( \theta \in (0, 1) \), it can be seen from Lemma 5.9 that
\[
\sum_{j=t+1}^{T} \frac{1}{t(t+s+\theta)} \exp \left( -\nu \sum_{j=t+1}^{T} j^{-\theta} \right)
\]
\[
\leq \exp \left( -\nu \frac{1}{1 - \theta} (T + 1)^{1-\theta} \right) \sum_{j=1}^{T} t^{-(\theta+s)} \exp \left( \frac{\nu}{1 - \theta} (t + 1)^{1-\theta} \right).
\]

For \( \forall x \in [t + 1, t + 2] \), we know that \( t^{-(\theta+s)} \leq 3^{s+1} x^{-(\theta+s)} \) and
\[
\exp \left( \frac{\nu}{1 - \theta} (t + 1)^{1-\theta} \right) \leq \exp \left( \frac{\nu}{1 - \theta} x^{1-\theta} \right).
\]
This together with (5.20), (5.21) and (5.22) implies that
\[
\sum_{t=1}^{T} \frac{1}{t^{(s+\theta)}} \exp \left( -\nu \sum_{j=t+1}^{T} j^{-\theta} \right) 
\leq 3^{s+1} \exp \left( -\frac{\nu}{1-\theta} (T+1)^{1-\theta} \right) \int_{2}^{T+1} x^{-(\theta+s)} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) dx.
\] (5.20)

Decomposing the above integral into two parts, we have
\[
\int_{2}^{T} x^{-(\theta+s)} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) dx
= \int_{2}^{T} x^{-(\theta+s)} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) dx + \int_{T}^{T+1} x^{-(\theta+s)} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) dx.
\] (5.21)

We estimate the two integrals in (5.21) respectively.

When \( x \in [2, \frac{T}{2}] \), we have \( x^{-s} \leq 1 \) and
\[
\exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) \leq \exp \left( \frac{\nu}{1-\theta} \left( \frac{T}{2} \right)^{1-\theta} \right).
\]

Also
\[
\int_{2}^{T} x^{-\theta} dx \leq \int_{1}^{T} x^{-\theta} dx \leq \frac{T^{1-\theta}}{(1-\theta)}.
\]

Thus
\[
\int_{2}^{T} x^{-(\theta+s)} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) dx \leq \frac{T^{1-\theta}}{(1-\theta)^{2^{1-\theta}}} \exp \left( \frac{\nu}{1-\theta} \left( \frac{T}{2} \right)^{1-\theta} \right).
\] (5.22)

When \( x \in [\frac{T}{2}, T+1] \), we have \( x^{-s} \leq \left( \frac{T}{2} \right)^{-s} \). Hence
\[
\int_{\frac{T}{2}}^{T+1} x^{-(\theta+s)} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) dx
\leq \left( \frac{T}{2} \right)^{-s} \int_{\frac{T}{2}}^{T+1} x^{-\theta} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right) dx = \left( \frac{T}{2} \right)^{-s} \int_{\frac{T}{2}}^{T+1} \frac{d}{\nu} \exp \left( \frac{\nu}{1-\theta} x^{1-\theta} \right)
= \frac{1}{\nu} \left( \frac{T}{2} \right)^{-s} \left( \exp \left( \frac{\nu}{1-\theta} (T+1)^{1-\theta} \right) - \exp \left( \frac{\nu}{1-\theta} \left( \frac{T}{2} \right)^{1-\theta} \right) \right)
\leq \frac{2^{s}}{\nu T^{s}} \exp \left( \frac{\nu}{1-\theta} (T+1)^{1-\theta} \right).
\] (5.23)

This together with (5.20), (5.21) and (5.22) implies that
\[
\sum_{t=1}^{T} \frac{1}{t^{(s+\theta)}} \exp \left( -\nu \sum_{j=t+1}^{T} j^{-\theta} \right) 
\leq 3^{s+1} \exp \left( -\frac{\nu}{1-\theta} (T+1)^{1-\theta} \right) \left( \frac{T^{1-\theta}}{(1-\theta)^{2^{1-\theta}}} \exp \left( \frac{\nu}{1-\theta} \left( \frac{T}{2} \right)^{1-\theta} \right) + \frac{2^{s}}{\nu T^{s}} \exp \left( \frac{\nu}{1-\theta} (T+1)^{1-\theta} \right) \right)
\leq \frac{2^{s+1}}{v T^{s}} + \frac{3^{s+1} T^{1-\theta}}{(1-\theta)^{2^{1-\theta}}} \exp \left( -\frac{\nu}{1-\theta} (1-2^{\theta-1}) (T+1)^{1-\theta} \right)
\]
Our conclusion is proved.

For all \( \eta \) satisfies Proposition 3.

Assume that the stepsize following proposition is important for the proofs in this section.

4046 SHUHUA WANG, ZHENLONG CHEN AND BAOHUAI SHENG

Applying this relation iteratively for \( j \in [t, t + 1] \) there holds

\[
\sum_{t=1}^{T-1} \frac{1}{(t+1)^\nu \prod_{j=t+1}^{T} x^\nu} = \left( \frac{t+1}{x} \right)^\nu (x)^{s+1} \leq \left( \frac{t+1}{t} \right)^\nu x^{s+1} \leq 2^{\nu s+1}.
\]

Hence

\[
t^{-(s+1)}(t+1)^\nu \leq 2^{\nu s+1} x^{-(s+1)} x^\nu = 2^{\nu s+1} x^{\nu - s - 1}.
\]

This together with (5.24) yields

\[
\sum_{t=1}^{T-1} \frac{1}{(t+1)^\nu \prod_{j=t+1}^{T} x^\nu} \leq \frac{2^{\nu s+1} x^{\nu - s - 1}}{(T + 1)^\nu} \left( \frac{T^{\nu - s}}{\nu - s} - \frac{1}{\nu - s} \right).
\]

Our conclusion is proved.

5.2. Proofs of main results. Take \( \prod_{i=m+1}^{n} a_i = 1 \) and \( \sum_{i=m+1}^{n} a_i = 0 \). Then the following proposition is important for the proofs in this section.

**Proposition 3.** Assume that the stepsize \( \{\eta_t : t \in \mathbb{N}\} \) in the algorithm (3.1) satisfies \( \eta_t \leq 1 \), then for any \( t_0, T > t_0 \), there holds

\[
\mathbb{E} \left[ \| f_{T+1} - f\|_K^2 \right] \leq \left( 1 - \lambda \eta_t \right) \mathbb{E} \left[ \| f_{t_0} - f\|_K^2 \right] + 4\kappa^2 \sigma^2 \sum_{t=t_0}^{T} \eta_t^2 \prod_{j=t+1}^{T} (1 - \lambda \eta_j) + 4\kappa^2 \sigma^2 \sum_{t=t_0}^{T} \frac{\eta_t}{\lambda \sqrt{T-t}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j).
\]

**Proof.** Combining Lemma 5.4, Lemma 5.7 and Lemma 5.8, one can see that

\[
\mathbb{E} \left[ \| f_{t+1} - f\|_K^2 \right] \\
\leq \mathbb{E} \left[ \| f_t - f\|_K^2 \right] + 4\eta_t^2 \kappa^2 \sigma^2 + 4\eta_t \mathbb{E} \left[ \tilde{E}_k(f_t) - \tilde{E}_k(f) \right] \\
\leq \mathbb{E} \left[ \| f_t - f\|_K^2 \right] + 4\eta_t^2 \kappa^2 \sigma^2 + 2\eta_t \left( \mathbb{E} \left[ E_k(f_t) - E_k(f) \right] + \frac{2\kappa^2 \sigma^2}{\lambda \sqrt{t-1}} \right) \\
\leq (1 - \lambda \eta_t) \mathbb{E} \left[ \| f_t - f\|_K^2 \right] + 4\kappa^2 \sigma^2 (\eta_t^2 + \frac{\eta_t}{\lambda \sqrt{T-t-1}}).
\]

Applying this relation iteratively for \( t = T, T-1, \ldots, t_0 \) we obtain

\[
\mathbb{E} \left[ \| f_{T+1} - f\|_K^2 \right] \leq (1 - \lambda \eta_{T}) \mathbb{E} \left[ \| f_{t_0} - f\|_K^2 \right] + 4\kappa^2 \sigma^2 \eta_{T} + \frac{4\kappa^2 \sigma^2 \eta_{T}}{\lambda \sqrt{T-1}}.
\]
\[
\leq (1 - \lambda \eta_T) \left( (1 - \lambda \eta_{T-1}) \mathbb{E}[\|f_{T-1} - f_{\lambda, \sigma}\|_K^2] + 4\kappa^2 \sigma^2 \eta^2_{T-1} + \frac{4\kappa^2 \sigma^2 \eta_T}{\lambda \sqrt{T-1}} \right) + 4\kappa^2 \sigma^2 \eta_T^2 + \frac{4\kappa^2 \sigma^2 \eta_T}{\lambda \sqrt{T-1}} \\
\leq \cdots \leq \prod_{t=t_0}^T (1 - \lambda \eta_t) \mathbb{E}[\|f_{t_0} - f_{\lambda, \sigma}\|_K^2] + 4\kappa^2 \sigma^2 \sum_{t=t_0}^T \eta_t^2 \prod_{j=t+1}^T (1 - \lambda \eta_j) + 4\kappa^2 \sigma^2 \sum_{t=t_0}^T \frac{\eta_t}{\lambda \sqrt{T-1}} \prod_{j=t+1}^T (1 - \lambda \eta_j).
\]

The proposition is proved. \(\square\)

**Proof of Theorem 4.1.** Since \(\eta_t \to 0(t \to +\infty)\), there exists \(t_0 \in \mathbb{N}\), such that for any \(t \geq t_0\), we have \(\eta_t \leq \frac{1}{2\sqrt{6\lambda \sigma}}\). For the fixed \(t_0\), Proposition 3 implies that it is sufficient to estimate the three terms on the right side of (5.25) respectively.

(i) **Bound the first term.**

Recall the assumptions \(\sum_{t=t_0}^\infty \eta_t = \infty\) and \(\lambda \eta_t \leq 1\). By \(1 - u \leq e^{-u}(u \geq 0)\), we have

\[
\prod_{t=t_0}^T (1 - \lambda \eta_t) \leq \exp \left\{ -\lambda \sum_{t=t_0}^T \eta_t \right\} \to 0, \quad T \to +\infty.
\]

This means that, for any \(\varepsilon > 0\), there exists \(T_1 \in \mathbb{N}\), such that for all \(T > T_1\), there holds

\[
\prod_{t=t_0}^T (1 - \lambda \eta_t) \mathbb{E}[\|f_{t_0} - f_{\lambda, \sigma}\|_K^2] \leq \frac{\varepsilon}{3}. \tag{5.26}
\]

(ii) **Bound the middle term.**

From the assumption \(\lim_{t \to +\infty} \eta_t = 0\), for any \(\varepsilon > 0\), there exists \(t(\varepsilon) > t_0\), such that for all \(t > t(\varepsilon)\) there holds \(\eta_t \leq \frac{\lambda^2}{24\kappa^2 \sigma^2}\).

Since

\[
\sum_{t=t_0}^T \eta_t^2 \prod_{j=t+1}^T (1 - \lambda \eta_j) = \sum_{t=t_0}^{t(\varepsilon)} \eta_t^2 \prod_{j=t+1}^T (1 - \lambda \eta_j) + \sum_{t=t(\varepsilon)+1}^T \eta_t^2 \prod_{j=t+1}^T (1 - \lambda \eta_j). \tag{5.27}
\]

For fixed \(t(\varepsilon)\) there exists \(T_2 \in \mathbb{N}\) such that for all \(T > T_2\) there holds

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

Thus for every \(t_0 \leq t \leq t(\varepsilon)\), there holds

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

And for any \(t \geq t_0\) we know that \(\lambda \eta_t \leq \frac{1}{2\sqrt{6\lambda \sigma}}\). Hence,

\[
\sum_{t=t_0}^{t(\varepsilon)} \eta_t^2 \prod_{j=t+1}^T (1 - \lambda \eta_j) \leq \frac{t(\varepsilon)}{24\lambda^2 \kappa^2 \sigma^2} \times \frac{\lambda^2 \varepsilon}{t(\varepsilon)} = \frac{\varepsilon}{24\kappa^2 \sigma^2}. \tag{5.28}
\]
On the other hand
\[
\sum_{t=t(e)+1}^{T} \eta_t^2 \prod_{j=t+1}^{T} (1 - \lambda \eta_j) \leq \sum_{t=t(e)+1}^{T} \eta_t \times \frac{\lambda \varepsilon}{24\kappa^2\sigma^2} \prod_{j=t+1}^{T} (1 - \lambda \eta_j)
\]
\[
= \frac{\varepsilon}{24\kappa^2\sigma^2} \left( 1 - \prod_{j=t(e)+1}^{T} (1 - \lambda \eta_j) \right) \leq \frac{\varepsilon}{24\kappa^2\sigma^2}. \quad (5.29)
\]

Putting (5.28) and (5.29) into (5.27), one can see that for all \( T > T_2 \) there holds
\[
4\kappa^2\sigma^2 \sum_{t=t_0}^{T} \frac{\eta_t^2}{\lambda \sqrt{T-1}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j)
\]
\[
= 4\kappa^2\sigma^2 \left( \sum_{t=t_0}^{T} \frac{\eta_t}{\lambda \sqrt{T-1}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j) + \sum_{t=t'(e)+1}^{T} \frac{\eta_t}{\lambda \sqrt{T-1}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j) \right). \quad (5.31)
\]

And there exists \( T_3 \in \mathbb{N} \), such that for all \( T > T_3 \) there holds
\[
\sum_{t=t'(e)+1}^{T} \frac{\eta_t}{\lambda \sqrt{T-1}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j) \geq \sum_{t=t'(e)+1}^{T_2} \eta_j \geq \frac{1}{\lambda} \log \frac{2\sqrt{6}\kappa \sigma t'(e)}{\lambda^2 \sqrt{T_0 - 1}}.
\]

Therefore for every \( t_0 \leq t \leq t'(e) \) we have
\[
\prod_{j=t+1}^{T} (1 - \lambda \eta_j) \leq \exp \left\{ -\lambda \sum_{j=t+1}^{T} \eta_j \right\} \leq \exp \left\{ -\lambda \sum_{j=t'(e)+1}^{T} \eta_j \right\} \leq \frac{\lambda^2 \varepsilon \sqrt{T_0 - 1}}{2 \sqrt{6} \kappa \sigma t'(e)}.
\]

Thus
\[
\sum_{t=t_0}^{T} \frac{\eta_t}{\lambda \sqrt{T-1}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j) \leq \frac{\lambda^2 \varepsilon \sqrt{T_0 - 1}}{2 \sqrt{6} \kappa \sigma T'(e)} \sum_{t=t_0}^{t'(e)} \frac{\eta_t}{\lambda \sqrt{T-1}}
\]
\[
\leq \frac{\lambda^2 \varepsilon \sqrt{T_0 - 1}}{2 \sqrt{6} \kappa \sigma t'(e)} t'(e) \frac{1}{2 \sqrt{6} \kappa \sigma \lambda^2 \sqrt{T_0 - 1}} = \frac{\varepsilon}{24\kappa^2\sigma^2}. \quad (5.32)
\]

On the other hand
\[
\sum_{t=t'(e)+1}^{T} \frac{\eta_t}{\lambda \sqrt{T-1}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j) \leq \frac{\varepsilon}{24\kappa^2\sigma^2} \sum_{t=t'(e)+1}^{T} \lambda \eta_t \prod_{j=t+1}^{T} (1 - \lambda \eta_j)
\]
\[
= \frac{\varepsilon}{24\kappa^2\sigma^2} \left( 1 - \prod_{j=t'(e)+1}^{T} (1 - \lambda \eta_j) \right) \leq \frac{\varepsilon}{24\kappa^2\sigma^2}. \quad (5.33)
\]

Substituting (5.32) and (5.33) into (5.31), we see that for all \( T > T_3 \) there holds
\[
4\kappa^2\sigma^2 \sum_{t=t_0}^{T} \frac{\eta_t}{\lambda \sqrt{T-1}} \prod_{j=t+1}^{T} (1 - \lambda \eta_j) \leq 4\kappa^2\sigma^2 \left( \frac{\varepsilon}{24\kappa^2\sigma^2} + \frac{\varepsilon}{24\kappa^2\sigma^2} \right) = \frac{\varepsilon}{3}. \quad (5.34)
\]
Combined the analysis in (i),(ii) and (iii), it can be seen that for fixed $\varepsilon$, as long as $T \geq \max\{T_1, T_2, T_3\}$, then
\[
\mathbb{E}[\|f_{T+1} - f_{\lambda,\sigma}\|_K^2] \leq \varepsilon.
\]
holds. This means that $\mathbb{E}[\|f_{T+1} - f_{\lambda,\sigma}\|_K^2] \to 0, T \to +\infty$. The desired result is given. 

Proof of Theorem 4.2. It can be seen from Proposition 3 that
\[
\mathbb{E}[\|f_{T+1} - f_{\lambda,\sigma}\|_K^2] \leq \prod_{t=2}^{T} (1 - \lambda \eta_t) \mathbb{E}[\|f_{\lambda,\sigma}\|_K^2] + 4\kappa^2 \sigma^2 \sum_{t=2}^{T} \eta_t^2 \prod_{j=t+1}^{T} (1 - \lambda \eta_j)
\]
\[
+ 4\kappa^2 \sigma^2 \sum_{t=2}^{T} \eta_t \prod_{j=t+1}^{T} (1 - \lambda \eta_j)
\]
\[
= A + B + C.
\]
where
\[
A = \prod_{t=2}^{T} (1 - \lambda \eta_t) \mathbb{E}[\|f_{\lambda,\sigma}\|_K^2], \quad B = 4\kappa^2 \sigma^2 \sum_{t=2}^{T} \eta_t^2 \prod_{j=t+1}^{T} (1 - \lambda \eta_j),
\]
\[
C = 4\kappa^2 \sigma^2 \sum_{t=2}^{T} \eta_t \prod_{j=t+1}^{T} (1 - \lambda \eta_j).
\]
So we can finish the proof by bounding $A, B, C$ respectively.

From the definition of $f_{\lambda,\sigma}$, we have
\[
\mathcal{E}_{\lambda,\sigma}(f_{\lambda,\sigma}) = \mathcal{E}_{\sigma}(f_{\lambda,\sigma}) + \frac{\lambda}{2} \|f_{\lambda,\sigma}\|_K^2 \leq \mathcal{E}_{\sigma}(0) \leq |V_{\sigma}|_0.
\]
Thus $\frac{\lambda}{2} \|f_{\lambda,\sigma}\|_K^2 \leq |V_{\sigma}|_0$. This means
\[
\|f_{\lambda,\sigma}\|_K \leq \sqrt{\frac{2|V_{\sigma}|_0}{\lambda}} \leq \sqrt{\frac{2M\sigma}{\lambda}}.
\]
Using Lemma 5.9 together with inequality $1 - u \leq e^{-u}(u \geq 0)$, we get
\[
A \leq \exp\left\{- \sum_{t=2}^{T} \lambda \eta_t \right\} \|f_{\lambda,\sigma}\|_K^2 = \exp\left\{- \sum_{t=2}^{T} t^{-\theta} \right\} \|f_{\lambda,\sigma}\|_K^2 \leq \frac{2M\sigma}{\lambda} \exp\left\{- \sum_{t=2}^{T} t^{-\theta} \right\} \leq \frac{2M\sigma}{\lambda} \left( \frac{\theta}{(1 - (\frac{\theta}{2})^{1-\theta})e} \right)^{-\nu} T^{-\theta}.
\]
The last inequality holds since Lemma 5.10 (i) with $b = \frac{a}{1-\theta}, \nu = \frac{1 - (\frac{\theta}{2})^{1-\theta}}{1-\theta}$. From the assumption of $\eta_t$ together with the inequality $1 - u \leq e^{-u}(u \geq 0)$ we know
\[
B = \frac{4\kappa^2 \sigma^2}{\lambda^2} \sum_{t=2}^{T} t^{-2\theta} \prod_{j=t+1}^{T} (1 - j^{-\theta}) \leq \frac{4\kappa^2 \sigma^2}{\lambda^2} \sum_{t=1}^{T} t^{-2\theta} \exp\left\{- \sum_{j=t+1}^{T} j^{-\theta} \right\}
\]
\[
\leq \frac{4\kappa^2 \sigma^2}{\lambda^2} \left( \sum_{t=1}^{T-1} t^{-2\theta} \exp\left\{- \sum_{j=t+1}^{T} j^{-\theta} \right\} + T^{-2\theta} \right).
\]
By Lemma 5.10 (ii) with \( \nu = 1, s = \theta \), there holds
\[
\sum_{t=1}^{T-1} t^{-2\theta} \exp \left\{ - \sum_{j=t+1}^{T} j^{-\theta} \right\} + T^{-2\theta} \\
\leq \frac{19}{T^\theta} + \frac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}} \exp \left( - \frac{1 - 2^{\theta-1}}{1-\theta} (T + 1)^{1-\theta} \right). \tag{5.38}
\]

And by Lemma 5.10 (i) with \( \nu = 1 - 2^{\theta-1} - 1, b = \frac{1}{1-\theta} \), we get
\[
\exp \left\{ - \frac{1 - 2^{\theta-1}}{1-\theta} (T + 1)^{1-\theta} \right\} \leq \left( \frac{1}{(1 - 2^{\theta-1})e} \right)^{\frac{1}{1-\theta}} (T + 1)^{-1}. \tag{5.39}
\]

Substituting (5.39) into (5.38) and then (5.38) into (5.37), we have
\[
B \leq \frac{4\kappa^2 \sigma^2}{\lambda^2} \left( \frac{19}{T^\theta} + \frac{9}{(1-\theta)2^{1-\theta}} \left( \frac{1}{(1 - 2^{\theta-1})e} \right)^{\frac{1}{1-\theta}} (T + 1)^{-\theta} \right)
\leq \frac{4\kappa^2 \sigma^2}{\lambda^2} \left( 19 + \frac{9}{(1-\theta)2^{1-\theta}} \left( \frac{1}{(1 - 2^{\theta-1})e} \right)^{\frac{1}{1-\theta}} \right) T^{-\theta}. \tag{5.40}
\]

We now bound the third term on the righthand of (5.35). Similarly, by the assumption of \( \eta_t \) together with the inequality \( 1 - u \leq e^{-u} (u \geq 0) \), we know that
\[
C \leq \frac{4\sqrt{2} \kappa^2 \sigma^2}{\lambda^2} \sum_{t=1}^{T} t^{-(\theta+\frac{1}{2})} \exp \left\{ - \sum_{j=t+1}^{T} j^{-\theta} \right\}
= \frac{4\sqrt{2} \kappa^2 \sigma^2}{\lambda^2} \left( \sum_{t=1}^{T-1} t^{-(\theta+\frac{1}{2})} \exp \left\{ - \sum_{j=t+1}^{T} j^{-\theta} \right\} + T^{-(\theta+\frac{1}{2})} \right). \tag{5.41}
\]

By Lemma 5.10 (ii) with \( \nu = 1, s = \frac{1}{2} \), it follows that
\[
\sum_{t=1}^{T-1} t^{-(\theta+\frac{1}{2})} \exp \left\{ - \sum_{j=t+1}^{T} j^{-\theta} \right\} + T^{-(\theta+\frac{1}{2})}
\leq \frac{3\sqrt{6} + 1}{\sqrt{T}} + \frac{3^2 T^{1-\theta}}{(1-\theta)2^{1-\theta}} \exp \left\{ - \frac{1 - 2^{\theta-1}}{1-\theta} (T + 1)^{1-\theta} \right\}
\leq \frac{3\sqrt{6} + 1}{\sqrt{T}} + \frac{3^2}{(1-\theta)2^{1-\theta}} \left( \frac{1}{(1 - 2^{\theta-1})e} \right)^{\frac{1}{1-\theta}} T^{-\theta}. \tag{5.42}
\]

Putting (5.42) back into (5.41), we have
\[
C \leq \frac{8\kappa^2 \sigma^2}{\lambda^2} \left( \frac{3\sqrt{6} + 1}{\sqrt{T}} + \frac{3^2}{(1-\theta)2^{1-\theta}} \left( \frac{1}{(1 - 2^{\theta-1})e} \right)^{\frac{1}{1-\theta}} T^{-\theta} \right). \tag{5.43}
\]

Combining (5.36), (5.40), (5.43) with (5.35) and taking \( \theta' = \min\{\theta, \frac{1}{2}\} \), we have
\[
E(\|f_{T+1} - f_{\lambda, \sigma}\|_K^2)
\leq \frac{2M\sigma}{\lambda} \left( \frac{\theta}{(1 - \frac{\theta}{2})^{1-\theta} e} \right)^{\frac{\theta}{1-\theta}} T^{-\theta} + \frac{4\kappa^2 \sigma^2}{\lambda^2} \left( 19 + \frac{9}{(1-\theta)2^{1-\theta}} \left( \frac{1}{(1 - 2^{\theta-1})e} \right)^{\frac{1}{1-\theta}} \right) T^{-\theta}
+ \frac{8\kappa^2 \sigma^2}{\lambda^2} \left( \frac{3\sqrt{6} + 1}{\sqrt{T}} + \frac{3^2}{(1-\theta)2^{1-\theta}} \left( \frac{1}{(1 - 2^{\theta-1})e} \right)^{\frac{1}{1-\theta}} T^{-\theta} \right).
\]
Since we have

\[ \xi \leq \left( E Z M E \left( E \left( \int \sigma \left| \left| \left| f \right| \right| \left| \left| \left| x,x \right| \right| \right| \right| - E \left( \int \sigma \left| \left| \left| f \right| \right| \left| \left| \left| x,x \right| \right| \right| \right| \right) \right) \right) \sigma \lambda^2 T^{\theta^*}. \]

This completes the proof of Theorem 4.2. □

**Proof of Theorem 4.3.** By Proposition 3, we can see that

\[ \mathbb{E}[\|f_{T+1} - f_{\lambda,\sigma}\|_K^2] \leq \frac{1}{T} \mathbb{E}\|f_{\lambda,\sigma}\|_K^2 + \frac{4 \kappa^2 \sigma^2}{\lambda^2 T} \sum_{t=2}^{T} \frac{1}{t} + \frac{4 \kappa^2 \sigma^2}{\lambda^2 T} \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}}. \]  

(5.44)

Since we have \( \frac{1}{\sqrt{t}} \leq \sqrt{2} \cdot \frac{1}{\sqrt{x}} \) for \( \forall x \in [t, t+1] \). Thus

\[ \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} \leq \int_{1}^{T} \sqrt{2} \cdot \frac{1}{\sqrt{x}} dx \leq 2^\frac{3}{2} \sqrt{T}. \]  

(5.45)

and

\[ \sum_{t=2}^{T} \frac{1}{t} \leq \int_{1}^{T} \frac{1}{x} dx = \log T. \]  

(5.46)

Putting (5.45) and (5.46) back into (5.44), we have

\[ \mathbb{E}[\|f_{T+1} - f_{\lambda,\sigma}\|_K^2] \leq \frac{1}{T} \mathbb{E}\|f_{\lambda,\sigma}\|_K^2 + \frac{4 \kappa^2 \sigma^2}{\lambda^2 T} \log T + \frac{24 \kappa^2 \sigma^2}{\lambda^2 T} \sqrt{T} \]

\[ \leq \frac{2}{\lambda^2 T} \mathbb{E}|V_\sigma|_0 + \frac{4 \kappa^2 \sigma^2}{\lambda^2 T} \log T + \frac{24 \kappa^2 \sigma^2}{\lambda^2 T} \sqrt{T} \]

\[ \leq \frac{2}{\lambda^2 T} \log T + \frac{4 \kappa^2 \sigma^2}{\lambda^2 T} \sqrt{T} \]

\[ \leq (2M + (4 + 8 \sqrt{2}) \kappa^2) \frac{\sigma}{\lambda^2 \sqrt{T}}. \]

Theorem 4.3 is proved. □

**Proof of Corollary 1.** We have the following error decomposition

\[ \mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_0) \leq (\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_{\lambda,\sigma}))) + \mathcal{E}_\sigma(f_{\lambda,\sigma}) - \mathcal{E}_\sigma(f_0) \]  

(5.47)

There exists \( \xi \) between \( \delta(y,y') - f_{T+1}(x,x') \) and \( \delta(y,y') - f_{\lambda,\sigma}(x,x') \) such that

\[ V_\sigma(\delta(y,y') - f_{T+1}(x,x')) - V_\sigma(\delta(y,y') - f_{\lambda,\sigma}(x,x')) \]

\[ = V_\sigma'(\xi)|f_{T+1}(x,x') - f_{\lambda,\sigma}(x,x')| \]

\[ = \frac{|\xi|}{1 + (\frac{\xi}{\sigma})^2} |f_{T+1}(x,x') - f_{\lambda,\sigma}(x,x')| \]

\[ \leq \sigma |f_{T+1}(x,x') - f_{\lambda,\sigma}(x,x')|. \]

Then

\[ |\mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_{\lambda,\sigma})| \]

\[ \leq \int_Z \int_Z V_\sigma(\delta(y,y') - f_{T+1}(x,x')) - V_\sigma(\delta(y,y') - f_{\lambda,\sigma}(x,x'))|d\rho(x,y)d\rho(x',y') \]

\[ \leq \sigma \int_Z \int_Z |f_{T+1}(x,x') - f_{\lambda,\sigma}(x,x')|d\rho(x,y)d\rho(x',y') \leq \kappa \sigma \|f_{T+1} - f_{\lambda,\sigma}\|_K. \]  

(5.48)

Substituting (5.48) into (5.47), we get

\[ \mathcal{E}_\sigma(f_{T+1}) - \mathcal{E}_\sigma(f_0) \leq \kappa \sigma \|f_{T+1} - f_{\lambda,\sigma}\|_K + \mathcal{E}_\sigma(f_{\lambda,\sigma}) - \mathcal{E}_\sigma(f_0) \]

\[ \leq \kappa \sigma \|f_{T+1} - f_{\lambda,\sigma}\|_K + \mathcal{E}_\sigma(f_{\lambda,\sigma}) - \mathcal{E}_\sigma(f_0) + \frac{\lambda}{2} \|f_{\lambda,\sigma}\|_K^2. \]
\[ = \kappa \sigma \| f_{T+1} - f_{\lambda, \sigma} \|_K + \inf_{f \in \mathcal{H}_K} \left( \mathcal{E}_\sigma(f) - \mathcal{E}_\sigma(f_\sigma) + \frac{\lambda}{2} \| f_{\lambda, \sigma} \|_K^2 \right) \]

Combined which with Theorem 4.2, the desired result follows.

Acknowledgments. We would like to thank the editor and the anonymous reviewers for their valuable comments and constructive suggestions which help the authors to improve the results of the manuscript and increase the readability.

REFERENCES

[1] S. Agarwal and P. Niyogi, Generalization bounds for ranking algorithms via algorithms via algorithmic stability, J. Mach. Learn. Res., 10 (2009), 441–474.

[2] R. Aktass, A. Altin and F. Tasdelen, A note on a family of two-variable polynomials, J. Comput. Appl. Math., 235 (2011), 4825–4833.

[3] A. Altin, R. Aktas and E. Erkus-Duman, On a multivariable extension for the extended Jacobi polynomials, J. Math. Anal. Appl., 353 (2009), 121–133.

[4] P. L. Bartlett, O. Bousquet and S. Mendelson, Local rademacher complexities, Ann. Stat., 33 (2005), 1497–1537.

[5] P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, J. Mach. Learn. Res., 3 (2002), 463–482.

[6] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer-Verlag, New York, 2010.

[7] A. Bezubik, J. Hrivnák and S. Pósta, Two dimensional symmetric and antisymmetric exponential functions and orthogonal polynomials, J. Phys. Conf. Ser., 411 (2013), Art. 012007.

[8] M. Boissier, S. Lyu, Y. Ying and D. X. Zhou, Fast convergence of online pairwise learning algorithms, in Proceedings of the 19th International Conference on Artifical Intelligence and Statistics, (2016) 204–212.

[9] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge, Cambridge University Press, 2004.

[10] C. Brunner, A. Fischer, K. Luig and T. Thies, Pairwise support vector machines and their application to large scale problems, J. Mach. Learn. Res., 13 (2012), 2279–2292.

[11] Q. Cao, Z. C. Guo and Y. M. Ying, Generalization bounds for metric and similarity learning, Mach. Learn., 102 (2016), 115–132.

[12] H. Chen, Z. B. Pan and L. Q. Li, Learning performance of coefficient-based regularized learning, Neurocomputing, 133 (2014), 54–62.

[13] H. Chen and J. T. Wu, Regularized learning with convex losses and \( \ell^1 \) penalty, Abst. Appl. Anal., (2013), Art. 927827, 8 pp.

[14] R. H. Chen and Z. M. Wu, Solving partial differential equation by using multiquadratic quasi-interpolation, Appl. Math. Comput., 186 (2007), 1502–1510.

[15] X. M. Chen and Y. W. Lei, Refined bounds for online pairwise learning algorithms, Neurocomputing, 275 (2018), 2656–2665.

[16] S. O. Cheng and A. J. Smola, Hyperkernels, Conference Paper, 2002. Available from: https://www.researchgate.net/publication/221618045.

[17] S. O. Cheng, A. J. Smola and R. C. Williamson, Learning the kernel with hyperkernels, J. Mach. Learn. Res., 6 (2005), 1043–1071.

[18] A. Christmann and D. X. Zhou, On the robustness of regularized pairwise learning methods based on kernels, J. Complexity, 37 (2016), 1–33.

[19] F. Cucker and S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc., 39 (2002), 1–49.

[20] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, New York, 2007.

[21] S. R. Das, Kuhoo, D. Mishra and 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.

[22] C. F. Dunkl and Y. Xu, Orthogonal Polynomials of Several Variables, Cambridge University Press, 2014.
[23] Q. Fang, M. Xu and Y. M. Ying, Faster convergence of a randomized coordinate descent method for linearly constrained optimization problems, *Anal. Appl.*, 16 (2018), 741–755.

[24] L. Fernández, T. E. Pérez and M. A. Piñar, On Koornwinder classical orthogonal polynomials in two variables, *J. Comput. Appl. Math.*, 236 (2012), 3817–3826.

[25] W. W. Gao and Z. G. Wang, A meshless scheme for partial differential equations based on multiquadratic trigonometric b-spline quasi-interpolation, *Chin. Phys. B*, 23 (2014), Art. 110207.

[26] Z. C. Guo, Y. M. Ying and D. X. Zhou, Online regularized learning with pairwise loss functions, *Adv. Comput. Math.*, 43 (2017), 127–150.

[27] T. Hu, J. Fan and D. H. Xiang, Convergence analysis of distributed multi-penalty regularized pairwise learning, *Anal. Appl.*, 18 (2020), 109–127.

[28] S. Karlin and J. Gregor, Determinants of orthogonal polynomials, *Bull. Amer. Math. Soc.*, 68 (1962), 204–209.

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

[30] A. Klímyk and J. Patera, (Anti)Symmetric multivariable exponential functions and corresponding Fourier transforms, *J. Phys. A Math. Theor.*, 40 (2007), 10473–10489.

[31] T. Koornwinder, Two-variable analogues of the classical orthogonal polynomials, in *Theory and Application of Special Functions* (eds. R. A. Askey), Academic Press, New York, (1975), 435–495.

[32] T. Koornwinder, Yet another proof of the addition formula for Jacobi polynomials, *J. Math. Anal. Appl.*, 61 (1977), 136–141.

[33] T. Koornwinder, Orthogonal polynomials in two variables which are eigenfunctions of two algebraically independeng partial differential operators, I, II, *Indag Math.*, 36 (1974), 48–66.

[34] H. L. Krall and I. M. Sheffer, Orthogonal polynomials in two variables, *Ann. Mat. Pura Appl.*, 76 (1967), 325–376.

[35] G. Kriukova, S. V. Pereverzyev and P. Tkachenko, On the convergence rate and some applications of regularized ranking algorithms, *J. Complexity*, 33 (2016), 14–29.

[36] D. W. Lee, Partial differential equations for products of two classical orthogonal polynomials, *Bull. Korean Math. Soc.*, 42 (2005), 179–188.

[37] S. B. Lin, J. S. Zeng, J. Fang and Z. B. Xu, Learning rates of $l^q$ coefficient regularization learning with Gaussian kernel, *Neural Comput.*, 26 (2014), 2359–2378.

[38] J. H. Lin and D. X. Zhou, Online learning algorithms can converge comparably fast as batch learning, *IEEE Trans. Neural Netw. Learn. Syst.*, 29 (2017), 2367–2378.

[39] J. H. Lin, Y. W. Lei, B. Zhang and D. X. Zhou, Online pairwise learning algorithms with convex loss functions, *Inform. Sci.*, 406-407 (2017), 57–70.

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

[41] H. Rosengren, Multivariable Chiristoffel-Darboux kernels and charateristic polynomials of random Hermitian matrices, *Symmetry Integr. Geom. Meth. Appl.*, 2 (2006), Art. 085, 12 pp.

[42] B. Schölkopf, R. Herbrich and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002.
[51] B. H. Sheng, The weighted norm for some Mercer kernel matrices, *Acta. Math. Sci.*, **33A** (2013), 6–15 (in Chinese).

[52] B. H. Sheng, L. Q. Duan and P. X. Ye, Strong convex loss can increase the learning rates of online learning, *J. Computer*, **9** (2014), 1606–1611.

[53] B. H. Sheng and J. L. Wang, On the K-functional in learning theory, *Anal. Appl.*, **18** (2020), 423–446.

[54] B. H. Sheng and D. H. Xiang, The performance of semi-supervised laplacian regularized regression with least square loss, *Int. J. Wavelets Multiresolut. Inform. Process.*, **15** (2017), Art. 1750016, 31 pp.

[55] B. H. Sheng and H. Z. Zhang, Performance analysis of the LapRSSLG algorithm in learning theory, *Anal. Appl.*, **18** (2020), 79–108.

[56] B. H. Sheng and H. C. Zhu, The convergence rate of SVM for kernel-based robust regression, *Int. J. Wavelets Multiresolut. Inform. Process.*, **17** (2019), Art. 1950004.

[57] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.*, **1** (2003), 1–25.

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

[59] I. Sprinkhuizen-Kuyper, Orthogonal polynomials in two variables. A further analysis of the polynomials orthogonal over a region bounded by two lines and a parabola, *SIAM J. Math. Anal.*, **7** (1976), 501–518.

[60] I. Steinwart and C. Scovel, Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs, *Constr. Approx.*, **35** (2012), 363–417.

[61] H. Sun, Mercer theorem for RKHS on noncompact sets, *J. Complexity*, **21** (2005), 337–349.

[62] S. H. Wang, Z. L. Chen and B. H. Sheng, The convergence rate of SVM for kernel-based robust regression, *Int. J. Wavelets Multiresolut. Inform. Process.*, **17** (2019), Art. 1950004.

[63] C. Wang and T. Hu, Online regularized pairwise learning with least squares loss, *Anal. Appl.*, **18** (2020), 49–78.

[64] Y. Wang, R. Khardon, D. Pechyony and R. Jones, Generalization bounds for online learning algorithms with pairwise loss functions, in *JMLR: Workshop and Conference Proceedings, Annual Conference on Learning Theory*, **23** (2012), 13.1–13.22.

[65] Q. Wu, Y. M. Ying and D. X. Zhou, Multi-kernel regularized classifiers, *J. Complexity*, **23** (2007), 108–134.

[66] Z. M. Wu, Dynamical knot and shape parameter setting for simulating shock wave by using multi-quadratic quasi-interpolation, *Eng. Anal. Bound. Elem.*, **29** (2005), 354–358.

[67] Z. M. Wu and R. Schaback, Shape preserving properties and convergence of univariate multiquadratic quasi-interpolation, *Acta Math. Appl. Sin.*, **10** (1994), 441–446.

[68] D. H. Xiang, Conditional quantiles with varying Gaussians, *Adv. Comput. Math.*, **38** (2013), 723–735.

[69] G. B. Ye and D. X. Zhou, Fully online classification by regularization, *Appl. Comput. Harmon. Anal.*, **23** (2007), 198–214.

[70] Y. M. Ying and D. X. Zhou, Learnability of Gaussians with flexible variances, *J. Mach. Learn. Res.*, **8** (2007), 249–276.

[71] Y. M. Ying and D. X. Zhou, Online regularized classification algorithms, *IEEE. Trans. Inform. Theory*, **52** (2006), 4775–4788.

[72] Y. M. Ying and D. X. Zhou, Online pairwise learning algorithms, *Neural Comput.*, **28** (2016), 743–777.

[73] Y. M. Ying and D. X. Zhou, Unregularized online learning algorithms with general loss functions, *Appl. Comput. Harmon. Anal.*, **42** (2017), 224–244.

[74] Y. X. Zeng and D. Klabjian, Online adaptive machine learning based algorithm for implied volatility surface modeling, *Knowledge-Based Syst.*, **163** (2019), 376–391.

[75] L. L. Zhang and B. H. Sheng, Online regularized generalized gradient classification algorithms, *Anal. Theory Appl.*, **26** (2010), 278–300.

[76] Y. L. Zhao, J. Fan and L. Shi, Learning rates for regularized least squares ranking algorithm, *Anal. Appl.*, **15** (2017), 815–836.

Received July 2019; revised November 2019.