Online Pairwise Learning Algorithms

Yiming Ying[†] and Ding-Xuan Zhou[‡]

[†]Department of Mathematics and Statistics
State University of New York at Albany, Albany, NY 12222, USA
[‡]Department of Mathematics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China

Abstract

Pairwise learning usually refers to a learning task which involves a loss function depending on pairs of examples, among which most notable ones include bipartite ranking, metric learning and AUC maximization. In this paper, we study an online algorithm for pairwise learning with a least-square loss function in an unconstrained setting of a reproducing kernel Hilbert space (RKHS), which we refer to as the Online Pairwise lEaRning Algorithm (OPERA). In contrast to existing works [18, 36] which require that the iterates are restricted to a bounded domain or the loss function is strongly-convex, OPERA is associated with a non-strongly convex objective function and learns the target function in an unconstrained RKHS. Specifically, we establish a general theorem which guarantees the almost surely convergence for the last iterate of OPERA without any assumptions on the underlying distribution. Explicit convergence rates are derived under the condition of polynomially decaying step sizes. We also establish an interesting property for a family of widely-used kernels in the setting of pairwise learning and illustrate the above convergence results using such kernels. Our methodology mainly depends on the characterization of RKHSs using its associated integral operators and probability inequalities for random variables with values in a Hilbert space.

1 Introduction

For any $T \in \mathbb{N}$, the input space $\mathcal{X}$ is a compact domain of $\mathbb{R}^d$ and the output space $\mathcal{Y} \subseteq \mathbb{R}$. In the standard problems of regression and classification [14, 32], one considers learning from a set of examples $\mathbf{z} = \{z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \ldots, T\}$ drawn independently and identically (i.i.d) from an unknown distribution ρ on

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Associated with a specific learning problem, typically a univariate loss function $\ell(h, x, y)$ is used to measure the quality of a hypothesis function $h : \mathcal{X} \to \mathcal{Y}$.

This paper is motivated by the recently growing interest in an important family of learning problems which, for simplicity, we refer to as *pairwise learning* problems. In contrast to classical regression and classification, such learning problems involve pairwise loss functions, i.e. the loss function depends on a pair of examples which can be expressed by $\ell(f, (x, y), (x', y'))$ for a hypothesis function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Many machine learning tasks can be formulated as pairwise learning problems. Such tasks include ranking [1, 10, 13, 17, 25], similarity and metric learning [5, 8, 11, 35, 40], AUC maximization [44], and gradient learning [21, 22]. For instance, the task of ranking is to learn a ranking function capable of predicting an ordering of objects according to some attached relevance information. It generally involves the use of a misranking loss $\ell(f, (x, y), (x', y')) = \mathbb{I}_{\{(y-y')f(x,x')<0\}}$ or its surrogate loss $\ell(f, (x, y), (x', y')) = (1-(y-y')f(x, x'))^2$, where $\mathbb{I}(\cdot)$ is the indicator function. The goal of ranking is to find a ranking rule f in a hypothesis space $\mathcal{H}$ from the available data that minimizes the expected misranking risk

$$\mathcal{R}(f) = \iint_{\mathcal{Z}\times\mathcal{Z}} \ell(f, (x, y), (x', y'))d\rho(x, y)d\rho(x', y'). \tag{1.1}$$

In this paper, we assume that the hypothesis function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for pairwise learning belongs to a *reproducing kernel Hilbert space* (RKHS) defined on the product space $\mathcal{X}^2 = \mathcal{X} \times \mathcal{X}$. Specifically, let $K : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ be a *Mercer kernel*, i.e. a continuous, symmetric and positive semi-definite kernel, see e.g. [14, 32]. According to [2], the RKHS $\mathcal{H}_K$ associated with kernel K is defined to be the completion of the linear span of the set of functions $\{K_{(x,x')}(\cdot) := K((x, x'), (\cdot, \cdot)) : (x, x') \in \mathcal{X}^2\}$ with an inner product satisfying the reproducing property, i.e., for any $x', x \in \mathcal{X}$ and $f \in \mathcal{H}_K$, $\langle K_{(x,x')}, f\rangle_K = f(x, x')$.

Recently, a large amount of work focuses on pairwise learning algorithms in the batch setting in the sense that the algorithm uses the training data $\mathbf{z}$ at once. A general regularization scheme in a RKHS $\mathcal{H}_K$ for pairwise learning can be formulated as

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{T(T-1)} \sum_{\substack{i,j=1\\ i\neq j}}^{T} \ell(f, (x_i, y_i), (x_j, y_j)) + \frac{\lambda}{2}\|f\|_K^2 \right\}. \tag{1.2}$$

where $\lambda > 0$ is a regularization parameter. The above general formulation was studied for ranking [1, 25] and metric learning [5, 8] under choices of different pairwise kernels (see further discussions in Subsection 2.1). Their generalization analysis was established using the concept of algorithmic stability [1], robustness [5] or U-statistics and U-process [8, 13, 25]. However, there is relatively little work related to online algorithms for pairwise learning, despite of its potential capability of dealing with large datasets. Until most recently, [36] established the first generalization analysis of online learning methods for pairwise learning in the linear case. In particular, they showed online to batch conversion bounds hold true which are similar to those in the univariate loss function case [9].

In this paper, we study an Online Pairwise lEaRning Algorithms (OPERA) with a least-square loss function in a reproducing kernel Hilbert space (RKHS). In particular, a general convergence theorem is established which guarantees the almost surely convergence of the last iterate of OPERA. Explicit convergence rates are derived under the condition of polynomially decaying step sizes. In contrast to existing works [18, 36, 37] which require that the iterates are restricted to a bounded domain or the loss function is strongly-convex, OPERA is associated with a non-strongly convex objective function and learns the target function in an unconstrained RKHS (see more discussions in Section 3). Our novel methodology mainly depends on the characterization of RKHSs using the associated integral operators and probability inequalities for random variables with values in the Hilbert space of Hilbert-Schmidt operators.

The paper is organized as follows. Section 2 introduces OPERA and presents main results together with particular examples of specific pairwise kernels. Section 3 discusses the related work. Section 4 presents novel error decomposition for analyzing OPERA and establishes the associated technical estimates. The main results are proved in Section 5. The paper concludes in Section 6. The proofs for technical lemmas are postponed to the Appendix.

2 Main Results

In this section, we introduce an online pairwise learning algorithm associated with the least-square loss $\ell(f,(x,y),(x',y')) = (f(x,x')-y+y')^2$ in a reproducing kernel Hilbert space $\mathcal{H}_K$, and state our main results. In particular, denote the true risk, for any function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, by

$$\mathcal{E}(f) = \iint_{\mathcal{Z}\times\mathcal{Z}} (f(x,x') - y + y')^2 d\rho(x,y)d\rho(x',y').$$

Define $\widetilde{f}_\rho$ by the difference of two standard regression functions, i.e.

$$\widetilde{f}_\rho(x,x') = \int_{\mathcal{X}} y d\rho(y|x) - \int_{\mathcal{X}} y d\rho(y|x') = f_\rho(x) - f_\rho(x'). \tag{2.1}$$

Denote by $L^2_\rho(\mathcal{X}^2)$ the space of square integrable functions on the domain $\mathcal{X}\times\mathcal{X}$, i.e.

$$L^2_\rho(\mathcal{X}^2) = \left\{ f : \mathcal{X}\times\mathcal{X} \to \mathbb{R} : \|f\|_\rho = \Big(\iint_{\mathcal{X}\times\mathcal{X}} |f(x,x')|^2 d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(x') \Big)^{1/2} < \infty \right\},$$

where $\rho_{\mathcal{X}}$ is the marginal distribution of ρ over $\mathcal{X}$. Similar to the standard least-square regression problem (see e.g. [12]), the following property holds true

$$\mathcal{E}(f) - \mathcal{E}(\widetilde{f}_\rho) = \|f - \widetilde{f}_\rho\|_\rho^2.$$

Consequently, $\widetilde{f}_\rho$ is the minimizer of the functional $\mathcal{E}(\cdot)$ among all measurable functions. Through out this paper, we refer to $\widetilde{f}_\rho$ as the *pairwise regression function.*

In this paper, we study the following online pairwise learning algorithm which aims to learn the pairwise regression function $\widetilde{f}_\rho$ from data.

Definition 1. *Given the i.i.d. generated training data* $\mathbf{z} = \{z_i = (x_i, y_i) : i = 1, 2, \ldots, T\}$*, the Online Pairwise lEaRning Algorithm (OPERA) is given by* $f_1 = f_2 = 0$ *and, for* $2 \le t \le T$*,*

$$f_{t+1} = f_t - \frac{\gamma_t}{t-1} \sum_{j=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j) K_{(x_t, x_j)}, \tag{2.2}$$

where $\{\gamma_t > 0 : t \in \mathbb{N}\}$ *is usually referred to as the sequence of step sizes.*

OPERA is similar to the online projected gradient descent algorithm in [18, 36], i.e., $f_0 = 0$ and $\eta = \frac{R^2}{T}$, and, for $1 \le t \le T$,

$$f_t = \mathrm{Proj}_{\mathcal{B}_R}\big[f_{t-1} - \frac{\eta}{t-1} \sum_{j=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j) K_{(x_t, x_j)})\big], \tag{2.3}$$

where $\mathrm{Proj}_{\mathcal{B}_R}(\cdot)$ denotes the projection to a prescribed ball $\mathcal{B}_R = \{\|f\|_K \le R : f \in \mathcal{H}_K\}$ with radius R. In contrast, OPERA does not have this additional projection step and is implemented in the unconstrained setting.

The sequence $\{f_t : t = 1, 2, \ldots, T+1\}$ is usually referred to as the *learning sequence* generated by OPERA. We call the above algorithm OPERA an online learning algorithm in the sense that it only needs a sequential access to the training data. Specifically, let $\mathbf{z}^t = \{z_1, z_2, \ldots, z_t\}$ and at each time step $t+1$, OPERA presumes a hypothesis $f_t \in \mathcal{H}_K$ upon which a new data z_t is revealed. The quality of the pairwise function f_t is estimated on the local empirical error:

$$\widehat{\mathcal{E}}^t(f_t) = \frac{1}{2(t-1)} \sum_{j=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j)^2. \tag{2.4}$$

The next iterate f_{t+1} given by equation (2.2) is exactly obtained by performing a gradient descent step from the current iterate f_t based on the gradient of the local empirical error, which is given by

$$\nabla \widehat{\mathcal{E}}^t(f)|_{f=f_t} = \frac{1}{t-1} \sum_{j=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j) K_{(x_t, x_j)}.$$

Here, $\nabla\widehat{\mathcal{E}}^t(\cdot)$ denotes the functional gradient of the functional $\widehat{\mathcal{E}}^t$ in the RKHS $\mathcal{H}_K$.

Now denote $\kappa := \sup_{x,x\in\mathcal{X}} \sqrt{K((x,x'),(x,x'))}$, and throughout the paper we assume that $|y| \leq M$ almost surely for some $M > 0$. In addition, we introduce the notion of $\mathcal{K}$-functional [6] in approximation theory as

$$\mathcal{K}(s,\widetilde{f}_\rho) := \inf_{f\in\mathcal{H}_K}\{\|f-\widetilde{f}_\rho\|_\rho + s\|f\|_K\}, \quad s>0. \tag{2.5}$$

We can establish the following general theorem about the convergence of the last iterate f_{T+1} generated by OPERA.

Theorem 1. *Let $\gamma_t = \frac{1}{\mu}t^{-\theta}$ for any $t\in\mathbb{N}$ with some $\theta\in(\frac{1}{2},1)$ and $\mu\geq\kappa^2$, and $\{f_t : t=1,\ldots,T+1\}$ be given by OPERA (2.2). For any $0<\delta<1$, we have with probability $1-\delta$*

$$\|f_{T+1}-\widetilde{f}_\rho\|_\rho \leq \mathcal{K}\big(\sqrt{6\mu}(1+\kappa)T^{-\frac{1-\theta}{2}},\widetilde{f}_\rho\big) + C_{\theta,\kappa}T^{-\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\log T\log(8T/\delta), \tag{2.6}$$

where $C_{\theta,\kappa}$ depends on κ,θ but independent of T (see its explicit form in the proof).

Recall the well-known result (e.g. [6, 41]) that

$$\lim_{s\to 0+}\mathcal{K}(s,\widetilde{f}_\rho) = \inf_{f\in\mathcal{H}_K}\|f-\widetilde{f}_\rho\|_\rho.$$

Then, assuming $\theta\in(1/2,1)$ and letting $T\to\infty$ in inequality (2.6), we can prove the following corollary.

Corollary 1. *If $\gamma_t = \frac{1}{\mu}t^{-\theta}$ for any $t\in\mathbb{N}$ with $\theta\in(\frac{1}{2},1)$ and $\mu\geq\kappa^2$, and $\{f_t : t=1,\ldots,T+1\}$ be given by OPERA (2.2). Then, $\|f_{T+1}-\widetilde{f}_\rho\|_\rho$ converges to $\inf_{f\in\mathcal{H}_K}\|f-\widetilde{f}_\rho\|_\rho$ almost surely.*

Let us discuss the implication of the above corollary. Recall that a kernel is universal if its associates RKHS is dense in the space of continuous functions on $\mathcal{X}\times\mathcal{X}$ under the uniform norm. Typical examples of universal kernels [20, 32] include the Gaussian kernel $K((x^1,x^2),(\hat{x}^1,\hat{x}^2)) = \exp(-\frac{\|(x^1,x^2)-(\hat{x}^1,\hat{x}^2)\|^2}{\sigma})$ and the Laplace kernel $K((x^1,x^2),(\hat{x}^1,\hat{x}^2)) = \exp(-\frac{\|(x^1,x^2)-(\hat{x}^1,\hat{x}^2)\|}{\sigma})$. In this case, $\inf_{f\in\mathcal{H}_K}\|f-\widetilde{f}_\rho\|_\rho = 0$, which equivalently implies that, as $T\to\infty$, $\|f_{T+1}-\widetilde{f}_\rho\|_\rho\to 0$ almost surely.

We can derive explicit error rates under some regularity assumptions on the pairwise regression function. The regularity of $\widetilde{f}_\rho$ can be typically measured by the integral operator $L_K : L^2_\rho(\mathcal{X}^2)\to L^2_\rho(\mathcal{X}^2)$ defined by

$$L_K f = \iint_{\mathcal{X}\times\mathcal{X}} f(x,x')K_{(x,x')}d\rho_\mathcal{X}(x)d\rho_\mathcal{X}(x').$$

Since K is a Mercer kernel, L_K is compact and positive. Therefore, the fractional power operator L_K^β is well-defined for any $\beta>0$. In particular, we know from [12, 14] that $L_K^{1/2}(L^2_\rho(\mathcal{X}^2)) = \mathcal{H}_K$.

Theorem 2. *Let $\{f_t : t = 1, \ldots, T+1\}$ be given by OPERA (2.2). Suppose $\widetilde{f}_\rho \in L_K^\beta(L_\rho^2)$ with some $\beta > 0$ and choose $\gamma_t = \frac{1}{\mu} t^{-\min\left\{\frac{2\beta+1}{2\beta+2}, \frac{2}{3}\right\}}$ with some $\mu \geq \kappa^2$. Then, for any $0 < \delta < 1$ we have, with probability $1 - \delta$, that*

$$\|f_{T+1} - \widetilde{f}_\rho\|_\rho \leq C_{\beta,\kappa} T^{-\min\left(\frac{\beta}{2\beta+2}, \frac{1}{6}\right)} \log T \log(8T/\delta), \tag{2.7}$$

where $C_{\beta,\kappa}$ depends on β, κ and μ but independent of T (see the explicit form in the proof).

The algorithm OPERA depends on selecting an appropriate pairwise kernel for a given learning task. In the next subsection, we consider a specific class of pairwise kernels and their associated RKHSs which are induced by a kernel $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

2.1 Examples with specific pairwise kernels

Observe that the pairwise regression function $\widetilde{f}_\rho(x, x') = f_\rho(x) - f_\rho(x')$, and hence a natural motivation is to use a pairwise function $f(x, x') = g(x) - g(x')$ to approximate the desired function $\widetilde{f}_\rho$, where $g \in \mathcal{H}_G$ with $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ being a kernel.

Indeed, we can introduce a specific pairwise kernel K such that any function $f \in \mathcal{H}_K$ can be represented by as $f(x, x') = g(x) - g(x')$ with $g \in \mathcal{H}_G$. Specifically, given the univariate kernel G, let the pairwise function $K : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ defined, for any $x^1, x^2, \hat{x}^1, \hat{x}^2 \in \mathcal{X}$, by

$$\begin{aligned} K((x^1, x^2), (\hat{x}^1, \hat{x}^2)) &= G(x^1, \hat{x}^1) + G(x^2, \hat{x}^2) - G(x^1, \hat{x}^2) - G(x^2, \hat{x}^1) \\ &= \langle G_{x^1} - G_{x^2}, G_{\hat{x}^1} - G_{\hat{x}^2} \rangle_G. \end{aligned} \tag{2.8}$$

It can be easily verified that K defined by (2.8) is positive semi-definite on $\mathcal{X}^2 \times \mathcal{X}^2$, and thus K is a (pairwise) Mercer kernel on $\mathcal{X} \times \mathcal{X}$ if G is a Mercer kernel on $\mathcal{X}$. The following proposition characterizes the relationship between $\mathcal{H}_K$ and the original RKHS $\mathcal{H}_G$.

Proposition 1. *Let $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel and its associated pairwise kernel be induced by (2.8). Then, the following statements hold true:*

(a) Assume the constant function $1_\mathcal{X} \in \mathcal{H}_G$ and let $\mathcal{I}_G = span\{1_\mathcal{X} \in \mathcal{H}_G\}$ containing all constant functions and $\mathcal{I}_G^\perp = \{g \in \mathcal{H}_G : \langle g, 1_\mathcal{X} \rangle_G = 0\}$ be the subspace orthogonal to $\mathcal{I}_G$. Then, the mapping $\Im : \mathcal{I}_G^\perp \to \mathcal{H}_K$ defined by $\Im(g)(x^1, x^2) = g(x^1) - g(x^2)$ is a bijection with property $\|\Im(g)\|_K = \|g\|_G$.

(b) If the constant function $1_\mathcal{X} \notin \mathcal{H}_G$, then the mapping $\Im : \mathcal{H}_G \to \mathcal{H}_K$ defined by $\Im(g)(x^1, x^2) = g(x^1) - g(x^2)$ is a bijection with property $\|\Im(g)\|_K = \|g\|_G$.

Part (b) used the assumption $1_{\mathcal{X}} \notin \mathcal{H}_G$. Various kernels induce RKHSs satisfying this assumption. For instance, the homogeneous linear kernel $G(x,x') = x^\top x'$ and the Gaussian kernel $G(x,x') = \exp(-\frac{\|x-x'\|^2}{\sigma})$ [33] are such kernels. However, in general the assumption in part (b) is not true, and thus only part (a) holds true.

From the above proposition, we can rewrite OPERA (2.2) as $g_1 = g_2 = 0$ and, for $2 \le t \le T$,

$$g_{t+1} = g_t - \gamma_t \Big[\frac{1}{t-1} \sum_{j=1}^{t-1} (g_t(x_t) - g_t(x_j) - y_t + y_j)(G_{x_t} - G_{x_j}) \Big]. \tag{2.9}$$

The learning sequence $\{f_t : t = 1, 2, \ldots, T+1\}$ of OPERA can be recovered by

$$f_t(x^1, x^2) = \Im(g_t)(x^1, x^2) := g_t(x^1) - g_t(x^2), \qquad \forall x^1, x^2 \in \mathcal{X}. \tag{2.10}$$

Denote

$$L^2_\rho(\mathcal{X}) = \Big\{ f : \mathcal{X} \to \mathbb{R} : \|f\|_\rho = \big(\int_{\mathcal{X}} |f(x)|^2 d\rho_{\mathcal{X}}(x) \big)^{1/2} < \infty \Big\},$$

and, by applying Proposition 1, we can see that the K-functional $\mathcal{K}$ defined by (2.5) is reduced to

$$\mathcal{K}_G(s, \widetilde{f}_\rho) := \begin{cases} \inf_{g \in \mathcal{I}_G^\perp} \{ \|\Im(g) - \widetilde{f}_\rho\|_\rho + s\|g\|_G \}, & \text{if } 1_{\mathcal{X}} \in \mathcal{H}_G \\ \inf_{g \in \mathcal{H}_G} \{ \|\Im(g) - \widetilde{f}_\rho\|_\rho + s\|g\|_G \}, & \text{otherwise.} \end{cases} \tag{2.11}$$

Equipped with the above notations, we can obtain the following theorem.

Theorem 3. *Let $\gamma_t = \frac{t^{-\theta}}{\kappa^2}$ for any $t \in \mathbb{N}$ with $\theta \in (\frac{1}{2}, 1)$ and $\{g_t : t = 1, 2, \ldots, T+1\}$ be given by algorithm (2.9). Then, the following statements hold true.*

(a) Let the $\mathcal{K}$-functional associated with $\mathcal{H}_G$ be defined by (2.11). Then, for any $1 < \delta < 1$ we have, with probability $1 - \delta$, that

$$\|\Im(g_{T+1}) - \widetilde{f}_\rho\|_\rho \le \mathcal{K}_G\big(\sqrt{6}\kappa(1+\kappa)T^{-\frac{1-\theta}{2}}, \widetilde{f}_\rho\big) + C_{\theta,\kappa}\, T^{-\min(\theta - \frac{1}{2}, \frac{1-\theta}{2})} \log T \log(8T/\delta).$$

(b) Suppose $1_{\mathcal{X}} \notin \mathcal{H}_G$ and $f_\rho \in L_G^\beta(L^2_\rho(\mathcal{X}))$ with some $0 < \beta \le 1/2$ and choose $\gamma_t = \frac{1}{\kappa^2} t^{-\frac{2\beta+1}{2\beta+2}}$. Then, for any $0 < \delta < 1$ we have, with probability $1 - \delta$, that

$$\|\Im(g_{T+1}) - \widetilde{f}_\rho\|_\rho] \le \widetilde{C}_{\beta,\kappa} T^{-\frac{\beta}{2\beta+2}} \log T \log(8T/\delta). \tag{2.12}$$

The above theorem implies the following result. Suppose that the original univariate kernel G is a Gaussian kernel in (2.8). Choosing $\gamma_t = \frac{t^{-\theta}}{\kappa^2}$ with $\theta \in (1/2, 1)$ in (2.9), by a similar argument to the proof for Corollary 1 we can have $\|\Im(g_{T+1}) - \widetilde{f}_\rho\|_\rho \to 0$ almost surely as $T \to \infty$. It remains a question to us whether the assumption $1_{\mathcal{X}} \notin \mathcal{H}_G$ in part (b) of the above theorem can be removed .

3 Related work and Discussions

In this section, we discuss the related work on pairwise learning in the batch setting and stochastic online learning algorithms in the univariate case.

Firstly, we briefly review existing work on pairwise learning, among which most of them addressed the batch setting. In [25], the generalization analysis for the general formulation (1.2) was conducted using empirical process and U-statistics (see discussions in Example 3 there). Specifically, the author proved nice generalization bounds for the excess risk of such estimators with rates faster than $\mathcal{O}(1/\sqrt{T})$, where T is the sample number. In Section 5.2 of [1], the following regularization formulation was studied for ranking:

$$\min_{g\in\mathcal{H}_G}\Big\{\frac{2}{T(T-1)}\sum_{\substack{i,j=1\\ i<j}}^{T}\psi(g(x_i)-g(x_j),y_i-y_j)+\frac{\lambda}{2}\|g\|_G^2\Big\}, \tag{3.1}$$

where $\mathcal{H}_G$ denotes the RKHS on $\mathcal{X}$ with inner product $\|\cdot\|_G$ and ψ is a ranking loss function (see Definition 1 there). This formulation can be regarded as a special formulation of the general framework (1.2) since, by Proposition 1, one can choose $K((x^1,x^2),(\hat{x}^1,\hat{x}^2))=\langle G_{x^1}-G_{x^2},G_{\hat{x}^1}-G_{\hat{x}^2}\rangle_G$, and then, for any $f\in\mathcal{H}_K$, there exists a $g\in\mathcal{H}_G$ such that $f(x_i,x_j)=g(x_i)-g(x_j)$ with property $\|f\|_K=\|g\|_G$. In contrast to the batch setting, there is relatively little work on online algorithms for pairwise learning. Most recently, in [36] and [18] online to batch conversion bounds were nicely established for pairwise learning, which shares the same spirit of [9] in the univariate case. Specifically, Kar et al. [18] proved the following result. [1]

Theorem A[18] *Let $f_1,f_2,\ldots,f_{T-1}$ be an ensemble of hypotheses from the space $\mathcal{H}$ generated by an online learning algorithm with a B-bounded loss function $\ell:\mathcal{H}\times\mathcal{Z}\times\mathcal{Z}\to[0,B]$ that guarantees a regret bound of $\Re_T$, i.e.*

$$\sum_{t=2}^{T}\frac{1}{t-1}\sum_{\tau=1}^{t-1}\ell(f_{t-1},z_t,z_\tau)\le\inf_{f\in\mathcal{H}}\sum_{t=2}^{T}\frac{1}{t-1}\sum_{\tau=1}^{t-1}\ell(f,z_t,z_\tau)+\Re_T. \tag{3.2}$$

Then, for any $0<\delta<1$, we have with probability $1-\delta$,

$$\frac{1}{T-1}\sum_{t=2}^{T}\mathcal{E}_\ell(f_t)\le\inf_{f\in\mathcal{H}}\mathcal{E}_\ell(f)+\frac{4}{T-1}\sum_{t=2}^{T}\mathcal{R}_{t-1}(\ell\circ\mathcal{H})+\frac{\Re_T}{T-1}+6B\sqrt{\frac{\log\frac{T}{\delta}}{T-1}},$$

where, for any $f\in\mathcal{H}$, $\mathcal{E}_\ell(f)=\iint_{\mathcal{Z}\times\mathcal{Z}}\ell(f,z,z')d\rho(z)d\rho(z')$, and the Rademacher averages $\mathcal{R}_T(\ell\circ\mathcal{H})$ is defined as $\mathcal{R}_{t-1}(\ell\circ\mathcal{H})=\mathbb{E}\big[\sup_{h\in\mathcal{H}}\frac{1}{t-1}\sum_{\tau=1}^{t-1}\varepsilon_\tau\ell(h,z,z_\tau)\big]$ with the expectation being over ε_τ, z, and z_τ.

[1] The authors mainly focused on the linear case. However, the results there can be easily extended to the kernelized case.

For a fair comparison with our results, let the loss function $\ell(f,(x,y),(x',y')) = (f(x,x')-y+y')^2$ and the hypothesis space $\mathcal{H}$ be a bounded ball in an RKHS $\mathcal{H}_K$, i.e. $\mathcal{H} = \mathcal{B}_R := \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$ with some $R > 0$. In this case, the constant B in Theorem A is given by $B = (2M+\kappa R)^2$, and $\ell(\cdot, z, z')$ is Lipschitz continuous with constant $L = 2M + \kappa R$. By standard techniques to estimate the Rademacher averages [4], we can have $\mathcal{R}_t(\ell \circ \mathcal{H}) \leq \mathcal{O}(\frac{R^2}{\sqrt{t}})$ with R sufficiently large. Then, using an argument similar to the Section 5.3 of [37] we know that the online projected gradient descent algorithm (2.3) enjoys the regret bound $\Re_T \leq (2M+\kappa R)R\sqrt{T}$. Putting this regret bound with the above estimation for the Rademacher averages together, from Theorem A we get, with probability $1-\delta$, that

$$\begin{aligned}\tfrac{1}{T-1}\textstyle\sum_{t=2}^T \mathcal{E}(f_t) - \inf_{\|f\|_K \leq R} \mathcal{E}(f) &\leq \mathcal{O}\Big(\tfrac{R^2}{T}\textstyle\sum_{t=2}^T \tfrac{1}{\sqrt{t}} + \tfrac{R^2}{\sqrt{T}} + R^2\sqrt{\log\tfrac{T}{\delta}/(T-1)}\Big) \\ &\leq \mathcal{O}\Big(R^2\sqrt{\tfrac{\log\frac{T}{\delta}}{T}}\Big).\end{aligned}$$

Let $\overline{f}_T = \frac{1}{T-1}\sum_{t=2}^T f_t$ and then we have $\mathcal{E}(\overline{f}_T) \leq \frac{1}{T-1}\sum_{t=2}^T \mathcal{E}(f_t)$. Consequently, $\mathcal{E}(\overline{f}_T) - \inf_{f\in\mathcal{H}}\mathcal{E}(f) \leq \mathcal{O}\Big(R^2\sqrt{\frac{\log\frac{T}{\delta}}{T}}\Big)$. This estimation combined with the fact, for any f, that $\mathcal{E}(f) - \mathcal{E}(\widetilde{f}_\rho) = \|f - \widetilde{f}_\rho\|_\rho^2$ implies that

$$\|\overline{f}_T - \widetilde{f}_\rho\|_\rho^2 \leq \inf_{\|f\|_K\leq R}\|f - \widetilde{f}_\rho\|_\rho^2 + \mathcal{O}\Big(R^2\sqrt{\log\Big(\frac{T}{\delta}\Big)/T}\Big). \tag{3.3}$$

The first term on the righthand side of the above inequality is known as approximation error. Suppose the pairwise regression function $\widetilde{f}_\rho \in L_K^\beta(L_\rho^2)$ with some $0 < \beta < 1/2$. Then, we know from [14, 29] that $\inf_{\|f\|_K\leq R}\|f - \widetilde{f}_\rho\|_\rho^2 \leq R^{-\frac{4\beta}{1-2\beta}}\|L_K^{-\beta}\widetilde{f}_\rho\|_\rho^{\frac{2}{1-2\beta}}$, which implies that $\|\overline{f}_T - \widetilde{f}_\rho\|_\rho^2 \leq \mathcal{O}\Big(R^{-\frac{4\beta}{1-2\beta}}\|L_K^{-\beta}\widetilde{f}_\rho\|_\rho^{\frac{2}{1-2\beta}} + R^2\sqrt{\log\Big(\frac{T}{\delta}\Big)/T}\Big)$. Choosing $R = T^{\frac{1-2\beta}{4}}$ implies, with probability $1-\delta$, that

$$\|\overline{f}_T - \widetilde{f}_\rho\|_\rho^2 \leq \mathcal{O}\Big(T^{-\beta}\big(\log\sqrt{T/\delta} + \|L_K^{-\beta}\widetilde{f}_\rho\|_\rho^{\frac{2}{1-2\beta}}\big)\Big). \tag{3.4}$$

From Theorem 2, for $0 < \beta < 1/2$ the last iterate of OPERA has the convergence rate:

$$\|f_T - \widetilde{f}_\rho\|_\rho^2 \leq \mathcal{O}\Big(T^{-\frac{\beta}{1+\beta}}(\log T\log(8T/\delta))^2\Big). \tag{3.5}$$

Comparing the rates in (3.4) and (3.5), we can see that our rate (3.5) for the last iterate of OPERA is suboptimal to that of the average of iterates generated by algorithm (2.3). However, the online projected gradient descent algorithm (2.3) requires that all iterates are restricted to a prescribed ball with radius R, which leads to a challenging question on how to tune R appropriately according to the real-data at hand. In addition, the analysis techniques [18, 36, 37] critically depend on the bounded-domain assumption and do not directly apply to the unconstrained setting here. OPERA is performed in the unconstrained setting and hence is parameter-free expect the choice of step sizes. Indeed, theorems in Section 2 show that choosing $\gamma_t = \mathcal{O}(t^{-\theta})$ with $1/2 < \theta < 1$ always guarantees that the last iterate of OPERA converges almost surely without additional assumptions on the underlying distribution ρ. It should be mentioned the above comparison assumes that the number of examples T is fixed and is known in advance. In the general online learning setting, the number of examples is not known. In this sense, the above comparison is only for the theoretical purpose.

Secondly, we discuss the related work on (stochastic) online learning algorithms in the univariate case. There is a large amount work on (stochastic) online learning algorithms in the univariate case [7, 9, 27, 28, 41, 42] or under a more general name called stochastic approximation [3, 23, 26]. The main idea is to use a randomized gradient to replace the gradient of the empirical loss, where the original idea dates back to the work [26] in the 1950s. Most of approaches in stochastic approximation assume the hypothesis space is of finite dimensional and the gradient is bounded. In fact, when the hypothesis space is of finite dimensional, a simple averaging scheme for stochastic gradient descent [3] can achieve the optimal rate $\mathcal{O}(\frac{1}{T})$ under the assumption that the covariance operator $\int_{\mathcal{X}} xx^\top d\rho_{\mathcal{X}}(x)$ is invertible. Stochastic online learning with a least square loss in an infinite-dimensional RKHS has been pioneered by [28] and the results were established for general loss functions by [42], in which the objective functions are all strongly convex.

OPERA (2.2) shares a similar idea with the above algorithms in the univariate case in the sense that, at each iteration, it uses a computationally-cheap gradient estimator to replace the true gradient. However, the objective function of OPERA is not strongly convex and the hypothesis space $\mathcal{H}_K$ is not bounded. In particular, OPERA is more close to the online algorithm in [41], where the authors studied the following stochastic gradient descent in a RKHS $\mathcal{H}_G$:

$$\begin{cases} g_1 = 0 \quad \text{and} \ , \forall t \in 1, 2, \ldots, T \\ g_{t+1} = g_t - \gamma_t (g_t(x_t) - y_t) G_{x_t}. \end{cases}$$

The analysis in [41] heavily depends on the fact that the randomized gradient $(g_t(x_t) - y_t)G_{x_t}$ is, conditionally on $\{z_1, z_2, \ldots, z_{t-1}\}$, an unbiased estimator of the true gradient $\iint_{\mathcal{X}} (g_t(x) - y) G_x d\rho(x, y)$. However, the randomized gradient $\frac{1}{t-1} \sum_{j=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j) K_{(x_t, x_j)}$ in OPERA (2.2) is not an unbiased estimator of the true gradient $\iint_{\mathcal{X}\times\mathcal{X}} f_t(x, x') - y + y') K_{(x,x')} d\rho(x, y) d\rho(x', y')$, even conditionally on $\{z_1, z_2, \ldots, z_{t-1}\}$. This introduces the main difficulty in analyzing its convergence. Our new methodology relies on the novel error decomposition presented in the next section. This enable us to overcome this analysis difficulty by further employing the characterization of RKHSs using the associated integral operators and probability inequalities for random variables with values in the Hilbert space of Hilbert-Schmidt operators.

4 Error Decomposition and Technical Estimates

This section mainly presents an error decomposition for OPERA which is critical to prove the main results in Section 2.

To this end, we introduce some necessary notations. For any $1 \leq j < t$, denote the linear operator

$$L_{(x_t,x_j)} = \langle \cdot, K_{(x_t,x_j)} \rangle_K K_{(x_t,x_j)} : \mathcal{H}_K \to \mathcal{H}_K$$

by $L_{(x_t,x_j)}(g) = g(x_t, x_j) K_{(x_t,x_j)}$ for any $g \in \mathcal{H}_K$, and let $\widehat{L}_t = \frac{1}{t-1} \sum_{j=1}^{t-1} L_{(x_t,x_j)}$. In addition, define

$$S_{(z_t,z_j)} = (y_t - y_j) K_{(x_t,x_j)}, \text{ and } \hat{S}_t = \frac{1}{t-1} \sum_{j=1}^{t-1} S_{(z_t,z_j)}.$$

We also define an auxiliary operator $\widetilde{L}_t = \displaystyle\int_{\mathcal{X}} \hat{L}_t d\rho(z_t)$, i.e., for any $f \in \mathcal{H}_K$

$$\widetilde{L}_t(f) = \frac{1}{t-1} \sum_{\ell=1}^{t-1} \int_{\mathcal{X}} f(x, x_\ell) K_{(x,x_\ell)} d\rho_{\mathcal{X}}(x).$$

Similarly, define

$$\widetilde{S}_t = \int_{\mathcal{X}} \hat{S}_t d\rho(z_t) = \frac{1}{t-1} \sum_{\ell=1}^{t-1} \int_{\mathcal{X}} (f_\rho(x) - y_\ell) K_{(x,x_\ell)} d\rho_{\mathcal{X}}(x).$$

In addition, let

$$\hat{\mathcal{A}}^t = (\widetilde{L}_t - L_K) f_t - (\widetilde{S}_t - L_K \widetilde{f}_\rho), \quad \hat{\mathcal{B}}^t = (\hat{L}_t - \widetilde{L}_t) f_t - (\hat{S}_t - \widetilde{S}_t).$$

With these notations, for any $t \geq 2$ we can rewrite equality (2.2) as

$$f_{t+1} = f_t - \gamma_t (\hat{L}_t(f_t) - \hat{S}_t) = (I - \gamma_t L_K) f_t - \gamma_t (\hat{L}_t - L_K)(f_t) + \gamma_t \hat{S}_t,$$

and

$$\begin{aligned} f_{t+1} - \widetilde{f}_\rho &= (I - \gamma_t L_K)(f_t - \widetilde{f}_\rho) - \gamma_t (\hat{L}_t - L_K) f_t + \gamma_t (\hat{S}_t - L_K \widetilde{f}_\rho) \\ &= (I - \gamma_t L_K)(f_t - \widetilde{f}_\rho) - \gamma_t \hat{\mathcal{A}}^t - \gamma_t \hat{\mathcal{B}}^t. \end{aligned} \tag{4.1}$$

For any $t, j \in \mathbb{N}$ denote $\omega_j^t(L_k) = \prod_{\ell=j}^t (I - \gamma_\ell L_K)$ for any $j \leq t$ and we use the conventional notation, for any $t \in \mathbb{N}$, $\omega_{t+1}^t(L_k) = I$ and $\sum_{\ell=t+1}^t \gamma_\ell = 0$.

Consequently, from the above equality we can derive, for any $t \geq 2$, that

$$f_{t+1} - \widetilde{f}_\rho = -\omega_2^t(L_K)(\widetilde{f}_\rho) - \sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K) \hat{\mathcal{A}}^j - \sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K) \hat{\mathcal{B}}^j \tag{4.2}$$

The above error decomposition is similar to the well-known ones in learning theory in order to perform the error analysis for learning algorithms with univariate loss functions, see e.g. [15, 28, 38, 39]. The term $\omega_2^t(L_K)(\widetilde{f_\rho})$ is deterministic which is usually referred to as *approximation error* and the other term, i.e. $\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}^j + \sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{B}}^j$, depends on the random samples which is often called the *sample error.* Consequently, from the error decomposition (4.2) we have

$$\begin{aligned} \|f_{t+1} - \widetilde{f_\rho}\|_\rho &\le \|\omega_2^t(L_K)(\widetilde{f_\rho})\|_\rho + \|\textstyle\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}^j\|_\rho \\ &+ \|\textstyle\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{B}}^j\|_\rho. \end{aligned} \tag{4.3}$$

In the following subsections we estimate the terms on the right-hand side of inequality (4.3).

4.1 Estimation of the sample error

We now turn our attention to estimating the sample error, i.e. the last two terms on the right-hand side of inequality (4.3). To this end, we first establish some useful lemmas. The following lemma gives an upper-bound of the learning sequence $\{f_t : t \in \mathbb{N}\}$ under the $\mathcal{H}_K$ norm, which is mainly inspired by a similar estimation in [19] for bounding the iterates of online gradient descent algorithm in the univariate case.

Lemma 1. *Let the learning sequence $\{f_t : t \in \mathbb{N}\}$ be given by OPERA (2.2) and assume, for any $t \in \mathbb{N}$, that $\gamma_t \kappa^2 \le 1$. Then we have*

$$\|f_t\|_K \le 2M \sqrt{\sum_{j=2}^{t-1} \gamma_j}, \qquad \forall t \in \mathbb{N}. \tag{4.4}$$

Proof. For $t = 1$ or $t = 2$, by definition $f_1 = f_2 = 0$ which certainly satisfy (4.4). It suffices to prove the case of $t \ge 2$ by induction. Recalling equality (2.2), we have

$$\begin{aligned} \|f_{t+1}\|_K^2 &= \|f_t\|_K^2 - \tfrac{2\gamma_t}{t-1} \sum_{j=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j) f_t(x_t, x_j) \\ &+ \tfrac{\gamma_t^2}{(t-1)^2} \sum_{j,j'=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j)(f_t(x_t, x_{j'}) - y_t + y_{j'}) K((x_t, x_j), (x_t, x_{j'})) \\ &\le \|f_t\|_K^2 + \tfrac{\gamma_t^2 \kappa^2}{t-1} \sum_{j}^{t-1} (f_t(x_t, x_j) - y_t + y_j)^2 \\ &- \tfrac{2\gamma_t}{t-1} \sum_{j=1}^{t-1} (f_t(x_t, x_j) - y_t + y_j) f_t(x_t, x_j). \end{aligned}$$

Define a univariate function F_j by $F_j(s) = \kappa^2\gamma_t(s - y_t + y_j)^2 - 2(s - y_t + y_j)s$. It is easy to see that $\sup_{s\in\mathbb{R}} F_j(s) = \frac{(y_t-y_j)^2}{2-\kappa^2\gamma_t} \leq (2M)^2$ since $\gamma_t\kappa^2 \leq 1$ and $|y_j|+|y_t| \leq 2M$. Therefore, from the above estimation we can get, for $t \geq 2$, that

$$\|f_{t+1}\|_K^2 \quad \leq \|f_t\|_K^2 + \tfrac{\gamma_t}{t-1}\sum_{j=1}^{t-1}\sup_j F_j(s) \leq \|f_t\|_K^2 + (2M)^2\gamma_t.$$

Combining the above inequality with the induction assumption that $\|f_t\|_K \leq 2M\sqrt{\sum_{j=2}^{t-1}\gamma_j}$ implies the desired result. This completes the proof of the lemma. □

Denote the operator norm $\|\omega_j^t(L_K)L_K^\beta\|_{\mathcal{L}(L_\rho^2)} = \sup_{\|f\|_\rho\leq 1}\|\omega_j^t(L_K)L_K^\beta(f)\|_\rho$. The following technical lemma estimates the operator norm, which is simply implied in the proof of Lemma 3 in [41].

Lemma 2. *Let $\beta > 0$ and $\gamma_\ell\kappa^2 \leq 1$ for any integer $\ell \in [j,t]$. Then there holds*

$$\|\omega_j^t(L_K)L_K^\beta\|_{\mathcal{L}(L_\rho^2)} \leq \left(\left(\frac{\beta}{e}\right)^\beta + \kappa^{2\beta}\right)\min\left\{1, \left(\sum_{\ell=j}^t \gamma_\ell\right)^{-\beta}\right\}.$$

The estimation of the sample error also relies on an important characterization of $\mathcal{H}_K$ by the fractional operator $L_K^{1/2}$ (see Theorem 4 and Remark 3 in [12]). Specifically, for any $f \in \mathcal{H}_K$ there exists $g \in L_\rho^2(\mathcal{X}^2)$ such that $L_K^{1/2}g = f$ with property $\|f\|_K = \|L_K^{1/2}g\|_K = \|g\|_\rho$. With this characterization of $\mathcal{H}_K$, it is easy to see, for any $j < t$ and $f \in \mathcal{H}_K$, that

$$\begin{aligned}\|\omega_{j+1}^t(L_K)f\|_\rho \quad &= \|\omega_{j+1}^t(L_K)L_K^{1/2}g\|_\rho \leq \|\omega_{j+1}^t(L_K)L_K^{1/2}\|_{\mathcal{L}(L_\rho^2)}\|g\|_\rho \\ &= \|\omega_{j+1}^t(L_K)L_K^{1/2}\|_{\mathcal{L}(L_\rho^2)}\|f\|_K.\end{aligned} \tag{4.5}$$

We also need the following probabilistic inequalities in a Hilbert space. The first one is the Bennett's inequality for random variables in Hilbert spaces, which can be easily derived from [28, Theorem B 4].

Lemma 3. *Let $\{\xi_i : i = 1, 2, \ldots, t\}$ be independent random variables in a Hilbert space $\mathcal{H}$ with norm $\|\cdot\|$. Suppose that almost surely $\|\xi_i\| \leq B$ and $\mathbb{E}\|\xi_i\|^2 \leq \sigma^2 < \infty$. Then, for any $0 < \delta < 1$, the following holds with probability at least $1 - \delta$,*

$$\left\|\frac{1}{t}\sum_{i=1}^t[\xi_i - \mathbb{E}\xi_i]\right\| \leq \frac{2B\log\frac{2}{\delta}}{t} + \sigma\sqrt{\frac{\log\frac{2}{\delta}}{t}}$$

The second probabilistic inequality is the Pinelis-Bernstein inequality [34, Proposition A.3] for martingale difference sequence in a Hilbert space, which is derived from [24, Theorem 3.4].

Lemma 4. *Let $\{S_k : k \in \mathbb{N}\}$ be a martingale difference sequence in a Hilbert space. Suppose that almost surely $\|S_k\| \leq B$ and $\sum_{k=1}^{t} \mathbb{E}[\|S_k\|^2 | S_1, \ldots, S_{k-1}] \leq \sigma_t^2$. Then, for any $0 < \delta < 1$, the following holds with probability at least $1 - \delta$,*

$$\sup_{1 \leq j \leq t} \left\| \sum_{k=1}^{j} S_k \right\| \leq 2 \left(\frac{B}{3} + \sigma_t \right) \log \frac{2}{\delta}.$$

We also need some facts on Hilbert-Schmidt operators on $\mathcal{H}_K$, see [15, 29]. Specifically, let $HS(\mathcal{H}_K)$ be the Hilbert space of Hilbert-Schmidt operators on $\mathcal{H}_K$ with inner product $\langle A, B \rangle_{HS} = \mathrm{Tr}(B^T A)$ for any $A, B \in HS(\mathcal{H}_K)$. Here Tr denotes the trace of a linear operator. Indeed, the space $HS(\mathcal{H}_K)$ is a subspace of the space of bounded linear operators on $\mathcal{H}_K$, which is usually denoted by $(\mathcal{L}(\mathcal{H}_K), \|\cdot\|_{\mathcal{L}(\mathcal{H}_K)})$ with the property, for any $A \in HS(\mathcal{H}_K)$, that

$$\|A\|_{\mathcal{L}(\mathcal{H}_K)} \leq \|A\|_{HS}. \tag{4.6}$$

With the above preparations, we are ready to estimate the sample error for algorithm (2.2) which, according to the error decomposition (4.3), consists of terms $\| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}^j \|_\rho$ and $\| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{B}}^j \|_\rho$. Let us start with the estimation of $\| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}^j \|_\rho$.

Theorem 4. *Assume $\gamma_t \kappa^2 \leq 1$ for any $t \in \mathbb{N}$ and let $\{f_t : t \in \mathbb{N}\}$ be given by equation (2.2). For any $t \geq 2$ and $0 < \delta < 1$, with probability $1 - \delta$ there holds*

$$\| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}^j \|_\rho \leq \left[12\kappa(1+\kappa)^2 M \log \frac{4t}{\delta} \right] \sum_{j=2}^{t} \frac{\gamma_j (1 + (\sum_{\ell=2}^{j-1} \gamma_\ell)^{1/2})}{\sqrt{j} \left(1 + \sum_{\ell=j+1}^{t} \gamma_\ell\right)^{1/2}}.$$

Proof. Write

$$\sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}^j := \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}_1^j + \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}_2^j,$$

where $\hat{\mathcal{A}}_1^j = (\widetilde{L}_j - L_K) f_j$ and $\hat{\mathcal{A}}_2^j = -(\widetilde{S}_j - L_K \widetilde{f}_\rho)$. Hence,

$$\begin{aligned} \| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}^j \|_\rho & \leq \| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}_1^j \|_\rho \\ & + \| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t (L_K) \hat{\mathcal{A}}_2^j \|_\rho. \end{aligned} \tag{4.7}$$

For the first term on the right-hand side of equation (4.7), we have

$$\begin{aligned}
\| \sum_{j=3}^{t} \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}_1^j \|_\rho &= \sum_{j=3}^{t} \gamma_j \| \omega_{j+1}^t(L_K)\hat{\mathcal{A}}_1^j \|_\rho \\
&\leq \sum_{j=3}^{t} \gamma_j \| \omega_{j+1}^t(L_K) L_K^{1/2} \|_{\mathcal{L}(L_\rho^2)} \ \| \hat{\mathcal{A}}_1^j \|_K \\
&\leq \sum_{j=3}^{t} \gamma_j \| \omega_{j+1}^t(L_K) L_K^{1/2} \|_{\mathcal{L}(L_\rho^2)} \ \| \widetilde{L}_j - L_K \|_{\mathcal{L}(\mathcal{H}_K)} \| f_j \|_K \\
&\leq \sum_{j=3}^{t} \gamma_j \| \omega_{j+1}^t(L_K) L_K^{1/2} \|_{\mathcal{L}(L_\rho^2)} \ \| \widetilde{L}_j - L_K \|_{HS} \| f_j \|_K,
\end{aligned} \tag{4.8}$$

where the second inequality used (4.5) and the last inequality used (4.6).

Let the vector-valued random variable $\xi(x) = \int_{\mathcal{X}} \langle \cdot, K_{(x',x)} \rangle_K K_{(x',x)} d\rho_{\mathcal{X}}(x')$. By following the proof of Lemma 2 in [15], we have that $\| \langle \cdot, K_{(x',x)} \rangle_K K_{(x',x)} \|_{HS} \leq \kappa^2$. Hence, $\| \xi \|_{HS} \leq \int_{\mathcal{X}} \| \langle \cdot, K_{(x',x)} \rangle_K K_{(x',x)} \|_{HS} d\rho_{\mathcal{X}}(x') \leq \kappa^2$. Applying Lemma 3 with $B = \sigma = \kappa^2$ and $\mathcal{H} = HS(\mathcal{H}_K)$, we have, with probability $1 - \frac{\delta}{t}$, that

$$\begin{aligned}
\| \widetilde{L}_j - L_K \|_{HS} &= \big\| \tfrac{1}{j-1} \sum_{\ell=1}^{j-1} \xi(x_\ell) - \mathbb{E}(\xi) \big\|_{HS} \\
&\leq \tfrac{2\kappa^2 \log \frac{2t}{\delta}}{j-1} + \kappa^2 \sqrt{\tfrac{\log \frac{2t}{\delta}}{j-1}} \leq \tfrac{3\sqrt{2}\kappa^2 \log \frac{2t}{\delta}}{\sqrt{j}}.
\end{aligned} \tag{4.9}$$

Applying Lemma 2 with $\beta = 1/2$ implies, for any $2 \leq j \leq t$, that

$$\begin{aligned}
\| \omega_{j+1}^t(L_K) L_K^{1/2} \|_{\mathcal{L}(L_\rho^2)} &\leq \big(\big(\tfrac{1}{2e}\big)^{1/2} + \kappa \big) \min \Big\{ 1, \Big(\sum_{\ell=j+1}^{t} \gamma_\ell \Big)^{-1/2} \Big\} \\
&\leq \sqrt{2}(1+\kappa) / \big(1 + \textstyle\sum_{\ell=j+1}^{t} \gamma_\ell \big)^{1/2},
\end{aligned} \tag{4.10}$$

where we used the conventional notation $\sum_{\ell=t+1}^{t} \gamma_\ell = 0$. Putting estimations (4.9), (4.10) and inequality (4.4) in Lemma 1 back into (4.8), with probability $1-\delta$ there holds

$$\| \sum_{j=3}^{t} \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}_1^j \|_\rho \leq \big[12\kappa^2(1+\kappa) M \log \frac{2t}{\delta} \big] \sum_{j=3}^{t} \frac{\gamma_j (\sum_{\ell=2}^{j-1} \gamma_\ell)^{1/2}}{\sqrt{j} \big(1 + \sum_{\ell=j+1}^{t} \gamma_\ell \big)^{1/2}}. \tag{4.11}$$

For the term $\| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}_2^j \|_\rho$, we observe from (4.5) again that

$$\begin{aligned}
\| \sum_{j=2}^{t} \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}_2^j \|_\rho &\leq \sum_{j=2}^{t} \gamma_j \| \omega_{j+1}^t(L_K) L_K^{1/2} \|_{\mathcal{L}(L_\rho^2)} \ \| \hat{\mathcal{A}}_2^j \|_K \\
&\leq \sum_{j=2}^{t} \gamma_j \| \omega_{j+1}^t(L_K) L_K^{1/2} \|_{\mathcal{L}(L_\rho^2)} \ \| \widetilde{S}_j - L_K \widetilde{f}_\rho \|_K.
\end{aligned} \tag{4.12}$$

Let the vector-valued random variable $\xi(z) = \int_{\mathcal{X}}(f_\rho(x') - y)K_{(x',x)}d\rho_{\mathcal{X}}(x') \in \mathcal{H}_K$. Observe that $\|\xi\|_K \leq \int_{\mathcal{X}} |f_\rho(x') - y| \|K_{(x',x)}\|_K d\rho_{\mathcal{X}}(x') \leq 2\kappa M$. Applying Lemma 3 with $B = \sigma = 2\kappa M$ and $\mathcal{H} = \mathcal{H}_K$, we have, with probability $1 - \frac{\delta}{t}$, that

$$\begin{aligned} \|\widetilde{S}_j - L_K \widetilde{f}_\rho\|_K &= \|\tfrac{1}{j-1}\sum_{\ell=1}^{j-1} \xi(z_\ell) - \mathbb{E}(\xi)\|_K \\ &\leq \tfrac{4\kappa M \log \frac{2t}{\delta}}{j-1} + 2\kappa M \sqrt{\tfrac{\log \frac{2t}{\delta}}{j-1}} \\ &\leq \tfrac{6\sqrt{2}\kappa M \log \frac{2t}{\delta}}{\sqrt{j}}. \end{aligned}$$

Putting the above estimation and inequality (4.10) into (4.12) implies, with probability $1 - \delta$, that

$$\|\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}_2^j\|_\rho \leq \big[12\kappa(1+\kappa)M \log \frac{2t}{\delta}\big] \sum_{j=2}^t \frac{\gamma_j}{\sqrt{j}\big(1 + \sum_{\ell=j+1}^t \gamma_\ell\big)^{1/2}}. \quad (4.13)$$

Combining inequalities (4.11) and (4.13), we have, with probability $1 - \delta$, that

$$\|\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{A}}^j\|_\rho \leq \big[12\kappa(1+\kappa)^2 M \log \frac{4t}{\delta}\big] \sum_{j=2}^t \frac{\gamma_j(1 + (\sum_{\ell=2}^{j-1}\gamma_\ell)^{1/2})}{\sqrt{j}\big(1 + \sum_{\ell=j+1}^t \gamma_\ell\big)^{1/2}}.$$

This completes the proof of the theorem. □

We move on to the estimation of the term $\|\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{B}}^j\|_\rho$.

Theorem 5. *Assume $\gamma_t \kappa^2 \leq 1$ for any $t \in \mathbb{N}$ and let $\{f_t : t \in \mathbb{N}\}$ be given by equation (2.2). For any $t \geq 2$ and $0 < \delta < 1$, with probability $1 - \delta$ there holds*

$$\|\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{B}}^j\|_\rho \leq \frac{64}{3}\big(\kappa(1+\kappa)^2 M \log \frac{2}{\delta}\big)\Big(\sum_{j=2}^t \frac{\gamma_j^2(1 + \sum_{\ell=2}^{j-1}\gamma_\ell)}{1 + \sum_{\ell=j+1}^t \gamma_\ell}\Big)^{\frac{1}{2}}.$$

Proof. Notice, from the recursive equality (2.2), that f_j only depends on samples $\{z_1, \ldots, z_{j-1}\}$ and $f_1 = f_2 = 0$. Therefore, for any $j \geq 2$, there holds

$$\mathbb{E}(\hat{\mathcal{B}}^j | z_1, \ldots, z_{j-1}) = 0, \quad (4.14)$$

which means that $\{\xi_j := \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{B}}^j : j = 2, \ldots, t\}$ is a martingale difference sequence. In the following, we will apply Lemma 4 to estimate $\|\sum_{j=2}^t \gamma_j \omega_{j+1}^t(L_K)\hat{\mathcal{B}}^j\|_\rho$. To this end, it remains to estimate B and σ_t^2.

Recall that $\hat{\mathcal{B}}^j = (\hat{L}_j - \widetilde{L}_j)f_j - (\hat{S}_j - \widetilde{S}_j)$. By (4.6) and Lemma 1, we have

$$\begin{aligned} \|\hat{\mathcal{B}}_j\|_K &\leq \|\hat{L}_j - \widetilde{L}_j\|_{\mathcal{L}(\mathcal{H}_K)}\|f_j\|_K + \|\hat{S}_j - \widetilde{S}_j\|_K \\ &\leq \|\hat{L}_j - \widetilde{L}_j\|_{HS}\|f_j\|_K + \|\hat{S}_j - \widetilde{S}_j\|_K \\ &\leq 2\kappa^2\|f_j\|_K + 2\kappa M \leq 4\kappa^2 M\big(\sum_{\ell=2}^{j-1}\gamma_\ell\big)^{\frac{1}{2}} + 2\kappa M. \end{aligned}$$

Consequently,

$$\begin{aligned}
\|\omega^t_{j+1}(L_K)\hat{\mathcal{B}}_j\|_\rho &\leq \|\omega^t_{j+1}(L_K)L_K^{1/2}\|_{\mathcal{L}(L^2_\rho)}\|\hat{\mathcal{B}}^j_1\|_K \\
&\leq \frac{\sqrt{2}(1+\kappa)}{\big(1+\sum\limits_{\ell=j+1}^{t}\gamma_\ell\big)^{1/2}}\big(4\kappa^2 M\big(\textstyle\sum_{\ell=2}^{j-1}\gamma_\ell\big)^{\frac{1}{2}}+2\kappa M\big) \qquad (4.15)\\
&\leq 8\kappa(1+\kappa)^2 M\Big(\tfrac{1+\sum_{\ell=2}^{j-1}\gamma_\ell}{1+\sum_{\ell=j+1}^{t}\gamma_\ell}\Big)^{\frac{1}{2}}.
\end{aligned}$$

where the second inequality used Lemma (4.10). From the above estimation, we have

$$\begin{aligned}
&\sum_{j=2}^{t}\gamma_j^2\mathbb{E}(\|\omega^t_{j+1}(L_K)\hat{\mathcal{B}}_j\|^2_\rho|z_1,\ldots,z_{j-1})\\
&\leq \sigma_t^2 := 64\kappa^2(1+\kappa)^4M^2\sum_{j=2}^{t}\frac{\gamma_j^2(1+\sum_{\ell=2}^{j-1}\gamma_\ell)}{1+\sum_{\ell=j+1}^{t}\gamma_\ell},
\end{aligned}$$

and

$$\begin{aligned}
B &= \sup\nolimits_{2\leq j\leq t}\gamma_j\|\omega^t_{j+1}(L_K)\hat{\mathcal{B}}^j\|_\rho \leq 8\kappa(1+\kappa)^2M\Big(\sup\nolimits_{2\leq j\leq t}\tfrac{\gamma_j^2(1+\sum_{\ell=2}^{j-1}\gamma_\ell)}{1+\sum_{\ell=j+1}^{t}\gamma_\ell}\Big)^{\frac{1}{2}}\\
&\leq 8\kappa(1+\kappa)^2M\Big(\sum_{j=2}^{t}\frac{\gamma_j^2(1+\sum_{\ell=2}^{j-1}\gamma_\ell)}{1+\sum_{\ell=j+1}^{t}\gamma_\ell}\Big)^{\frac{1}{2}}.
\end{aligned}$$

Applying Lemma 4 yields that, with probability $1-\delta$,

$$\|\sum_{j=2}^{t}\gamma_j\omega^t_{j+1}(L_K)\hat{\mathcal{B}}^j\|_\rho \leq \frac{64}{3}\big(\kappa(1+\kappa)^2M\log\frac{2}{\delta}\big)\Big(\sum_{j=2}^{t}\frac{\gamma_j^2(1+\sum_{\ell=2}^{j-1}\gamma_\ell)}{1+\sum_{\ell=j+1}^{t}\gamma_\ell}\Big)^{\frac{1}{2}}.$$

This completes the proof of the theorem. □

4.2 Estimates of the approximation error

Here, we establish some basic estimates for the deterministic approximation error involving $\|\omega^t_2(L_K)\widetilde{f_\rho}\|_\rho$. To this end, we recall the notion of $\mathcal{K}$-functional [6] in approximation theory, namely

$$\mathcal{K}(s,f_\rho) := \inf_{f\in\mathcal{H}_K}\{\|f-f_\rho\|_\rho + s\|f\|_K\}, \quad s>0. \qquad (4.16)$$

We can estimate the quantity $\|\omega^t_2(L_K)\widetilde{f_\rho}\|_\rho$ as follows.

Lemma 5. *Assume $\gamma_t\kappa^2 \leq 1$ for each $t\in\mathbb{N}$. Then the following statements hold true.*

(a) Let the $\mathcal{K}$-functional defined by (2.5). Then, we have

$$\|\omega_2^t(L_K)\widetilde{f_\rho}\|_\rho \leq \mathcal{K}\big(\sqrt{2}(1+\kappa)\big(\sum_{j=2}^{t}\gamma_j\big)^{-\frac{1}{2}}, \widetilde{f_\rho}\big). \tag{4.17}$$

(b) If $\widetilde{f_\rho} \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with some $\beta > 0$ then

$$\|\omega_2^t(L_K)\widetilde{f_\rho}\|_\rho \leq 2\bigg(\Big(\frac{\beta}{e}\Big)^\beta + \kappa^{2\beta}\bigg)\|L_K^{-\beta}\widetilde{f_\rho}\|_\rho\big(\sum_{j=2}^{t}\gamma_j\big)^{-\beta}. \tag{4.18}$$

Proof. Part (a) is proved as follows. For any $f \in \mathcal{H}_K$, from (4.5) we have

$$\begin{aligned}\|\omega_2^t(L_K)\widetilde{f_\rho}\|_\rho &\leq \|f - \widetilde{f_\rho}\|_\rho + \|\omega_2^t(L_K)f\|_\rho. \\ &= \|f - \widetilde{f_\rho}\|_\rho + \|\omega_2^t(L_K)L_K^{\frac{1}{2}}\|_{\mathcal{L}(L^2_\rho)}\|f\|_K.\end{aligned} \tag{4.19}$$

Applying Lemma 2 with $\beta = \frac{1}{2}, j = 2$, implies that $\|\omega_2^t(L_K)L_K^{\frac{1}{2}}\|_{\mathcal{L}(L^2_\rho)} \leq \sqrt{2}(1+\kappa)\big(\sum_{j=2}^{t}\gamma_j\big)^{-\frac{1}{2}}$. Then, substituting this into the right-hand side of (4.19) yields that

$$\|\omega_2^t(L_K)\widetilde{f_\rho}\|_\rho \leq \inf_{f\in\mathcal{H}_K}\Big\{\|f - \widetilde{f_\rho}\|_\rho + \sqrt{2}(1+\kappa)\big(\sum_{j=2}^{t}\gamma_j\big)^{-\frac{1}{2}}\|f\|_K\Big\}. \tag{4.20}$$

Part (b) can be directly proved by applying Lemma 2 and the following observation

$$\|\omega_2^t(L_K)\widetilde{f_\rho}\|_\rho \leq \|\omega_2^t(L_K)L_K^\beta\|_{\mathcal{L}(L^2_\rho)}\ \|L_K^{-\beta}\widetilde{f_\rho}\|_\rho.$$

□

5 Proof of Main Results

In this section, we prove the results presented in Section 2. Let us start with the proofs for Theorems 1 and 2. To this end, we need some technical lemmas.

Lemma 6. *Let $\gamma_j = \frac{j^{-\theta}}{\mu}$ for any $j \in \mathbb{N}$ with $\theta \in (\frac{1}{2}, 1)$ and $\mu > 0$. Then we have, for any $t \geq 4$, that*

$$\sum_{j=2}^{t}\frac{\gamma_j(1+\sum_{\ell=2}^{j-1}\gamma_\ell)}{\sqrt{j}\big(1+\sum_{\ell=j+1}^{t}\gamma_\ell\big)^{1/2}} \leq C_\theta t^{-\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\log t, \tag{5.1}$$

where

$$C_\theta = \begin{cases}\frac{26\max\big(\sqrt{\mu(1-\theta)})^{-1},\sqrt{\mu(1-\theta)}\big)}{\mu(1-\theta)|3\theta-2|} + \sqrt{\frac{5}{2\mu}}, & \text{if } \theta \neq 2/3 \\ \frac{20\max\big(\sqrt{\mu(1-\theta)})^{-1},\sqrt{\mu(1-\theta)}\big)}{\mu(1-\theta)} + \sqrt{\frac{5}{2\mu}}, & \text{if } \theta = 2/3.\end{cases}$$

Lemma 7. *Let* $\gamma_j = \frac{j^{-\theta}}{\mu}$ *for any* $j \in \mathbb{N}$ *with* $\theta \in (0,1)$*. Then we have, for any* $t \geq 4$*, that*

$$\Big(\sum_{j=2}^{t} \frac{\gamma_j^2 (1 + (\sum_{\ell=2}^{j-1} \gamma_\ell)^2)}{1 + \sum_{\ell=j+1}^{t} \gamma_\ell}\Big)^{1/2} \leq \widetilde{C}_\theta t^{-\min(\theta - \frac{1}{2}, \frac{1-\theta}{2})} \log t, \tag{5.2}$$

where $\widetilde{C}_\theta = \begin{cases} \Big(\frac{5}{8\mu} + \frac{16 \max((\mu(1-\theta))^{-1}, \mu(1-\theta))}{\mu^2 (1-\theta)|3\theta - 2|}\Big)^{1/2}, & \textit{if } \theta \neq 2/3 \\ \Big(\frac{5}{8\mu} + \frac{16 \max((\mu(1-\theta))^{-1}, \mu(1-\theta))}{\mu^2 (1-\theta)}\Big)^{1/2}, & \textit{if } \theta = 2/3. \end{cases}$

The proofs for Lemma 6 and Lemma 7 are given in the Appendix. With the above lemmas, we are ready to establish the main results stated in Section 2.

Proof of Theorem 1. Applying (4.3) with $t = T$, we have

$$\begin{aligned} \|f_{T+1} - \widetilde{f}_\rho\|_\rho & \leq \|\omega_2^T(L_K)(\widetilde{f}_\rho)\|_\rho + \|\textstyle\sum_{j=2}^T \gamma_j \omega_{j+1}^T (L_K) \hat{\mathcal{A}}^j\|_\rho \\ & + \|\textstyle\sum_{j=2}^T \gamma_j \omega_{j+1}^T (L_K) \hat{\mathcal{B}}^j\|_\rho. \end{aligned} \tag{5.3}$$

By Theorem 4 and (5.1), with probability $1 - \delta$, there holds

$$\begin{aligned} \|\textstyle\sum_{j=2}^T \gamma_j \omega_{j+1}^T (L_K) \hat{\mathcal{A}}^j\|_\rho & \leq 12\kappa(1+\kappa)^2 M \log \tfrac{4T}{\delta} \displaystyle\sum_{j=2}^{T} \frac{\gamma_j (1 + \sum_{\ell=2}^{j-1} \gamma_\ell)}{\sqrt{j}\big(1 + \sum_{\ell=j+1}^{t} \gamma_\ell\big)^{1/2}} \\ & \leq 12 C_\theta \, \kappa (1+\kappa)^2 M \, T^{-\min(\theta - \frac{1}{2}, \frac{1-\theta}{2})} \log T \log \tfrac{4T}{\delta} \end{aligned} \tag{5.4}$$

From Theorem 5 and (5.2) we have, with probability $1 - \delta$, that

$$\begin{aligned} \|\textstyle\sum_{j=2}^T \gamma_j \omega_{j+1}^T (L_K) \hat{\mathcal{B}}^j\|_\rho & \leq \tfrac{64}{3} \kappa (1+\kappa)^2 M \Big(\textstyle\sum_{j=2}^T \frac{\gamma_j^2 (1 + (\sum_{\ell=2}^{j-1} \gamma_\ell)^2)}{1 + \sum_{\ell=j+1}^{T} \gamma_\ell}\Big)^{1/2} \\ & \leq \tfrac{64 \widetilde{C}_\theta}{3} \kappa (1+\kappa)^2 M \, T^{-\min(\theta - \frac{1}{2}, \frac{1-\theta}{2})} \log T \log \tfrac{2}{\delta}. \end{aligned} \tag{5.5}$$

Putting estimates (5.3), (5.4), and (5.5), with probability $1 - 2\delta$ there holds

$$\|f_{T+1} - \widetilde{f}_\rho\|_\rho \leq \|\omega_2^T(L_K)(\widetilde{f}_\rho)\|_\rho + C_{\theta,\kappa} \, T^{-\min(\theta - \frac{1}{2}, \frac{1-\theta}{2})} \log T \log \frac{4T}{\delta}, \tag{5.6}$$

where $C_{\theta,\kappa} = 4(3C_\theta + \frac{16\widetilde{C}_\theta}{3})\kappa(1+\kappa)^2 M$.

In addition, by (4.17), we have

$$\|\omega_2^T(L_K)(\widetilde{f}_\rho)\|_\rho \leq \mathcal{K}\big(\sqrt{2}(1+\kappa)\big(\sum_{j=2}^{T} \gamma_j\big)^{-\frac{1}{2}}, \widetilde{f}_\rho\big).$$

Notice that $\sum_{j=2}^T \gamma_j = \frac{1}{\mu} \sum_{j=2}^T j^{-\theta} \geq \frac{(T+1)^{1-\theta} - 2^{1-\theta}}{\mu(1-\theta)} \geq \frac{(1 - (\frac{2}{3})^{1-\theta})(T+1)^{1-\theta}}{\mu(1-\theta)} \geq \frac{T^{1-\theta}}{3\mu(1-\theta)} \geq \frac{T^{1-\theta}}{3\mu}$. Consequently,

$$\|\omega_2^T(L_K)(\widetilde{f}_\rho)\|_\rho \leq \mathcal{K}\big(\sqrt{6\mu}(1+\kappa) T^{-\frac{1-\theta}{2}}, \widetilde{f}_\rho\big). \tag{5.7}$$

Putting this back into (5.6) implies the desired result. This completes the proof of the theorem. □

Proof of Corollary 1. By the definition of the almost surely convergence, it suffices to prove, for any $\varepsilon > 0$, that

$$\lim_{t_0\to\infty} \mathbb{P}\big(\sup_{t\geq t_0}[\|f_{t+1} - \widetilde{f}_\rho\|_\rho - \inf_{f\in\mathcal{H}_K} \|f - \widetilde{f}_\rho\|_\rho] \geq 2\varepsilon\big) = 0.$$

However, it is well-known that $\lim_{s\to+0} \mathcal{K}(s, \widetilde{f}_\rho) = \inf_{f\in\mathcal{H}_K} \|f - \widetilde{f}_\rho\|_\rho$ (see e.g. Lemma 9 in [41]). This means that there exists $t_1 \in \mathbb{N}$ such that, for any $t \geq t_1$, there holds

$$\mathcal{K}\big(\sqrt{6}\kappa(1+\kappa)t^{-\frac{1-\theta}{2}}, \widetilde{f}_\rho\big) - \inf_{f\in\mathcal{H}_K} \|f - \widetilde{f}_\rho\|_\rho \leq \varepsilon.$$

Let $\mathcal{R}_t = \|f_{t+1} - \widetilde{f}_\rho\|_\rho - \mathcal{K}\big(\sqrt{6}\kappa(1+\kappa)t^{-\frac{1-\theta}{2}}, \widetilde{f}_\rho\big)$. The above estimation implies, for any $t_0 \geq t_1$, that

$$\begin{aligned}&\mathbb{P}\big(\sup_{t\geq t_0}[\|f_{t+1} - \widetilde{f}_\rho\|_\rho - \inf_{f\in\mathcal{H}_K} \|f - \widetilde{f}_\rho\|_\rho] \geq 2\varepsilon\big)\\ &\leq \mathbb{P}\big(\sup_{t\geq t_0} \mathcal{R}_t \geq \varepsilon\big) \leq \sum_{t=t_0}^{\infty} \mathbb{P}\big(\mathcal{R}_t \geq \varepsilon\big).\end{aligned} \tag{5.8}$$

From Theorem 1, we have, for any $1-\delta$, that

$$\mathbb{P}\Big(\mathcal{R}_t \geq C_{\theta,\kappa}\, t^{-\min(\theta-\frac{1}{2},\frac{1-\theta}{2})} \log t \log(4t/\delta)\Big) \leq \delta.$$

which is equivalent to

$$\mathbb{P}\Big(\mathcal{R}_t \geq \varepsilon\Big) \leq 4t \exp\Big(-\frac{t^{\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\varepsilon}{C_{\theta,\kappa}\log t}\Big).$$

Putting this back into (5.8) implies that

$$\mathbb{P}\big(\sup_{t\geq t_0}[\|f_{t+1} - \widetilde{f}_\rho\|_\rho - \inf_{f\in\mathcal{H}_K} \|f - \widetilde{f}_\rho\|_\rho] \geq 2\varepsilon\big) \leq \sum_{t=t_0}^{\infty} 4t \exp\Big(-\frac{t^{\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\varepsilon}{C_{\theta,\kappa}\log t}\Big). \tag{5.9}$$

For any $1/2 < \theta < 1$ and $\varepsilon > 0$, it is easy to see that

$$\sum_{t=2}^{\infty} 4t \exp\big(-\frac{t^{\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\varepsilon}{C_{\theta,\kappa}\log t}\big) < \infty.$$

Consequently,

$$\lim_{t_0\to\infty} \sum_{t=t_0}^{\infty} 4t \exp\Big(-\frac{t^{\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\varepsilon}{C_{\theta,\kappa}\log t}\Big) = 0.$$

Combining this with (5.9) implies, for any $\varepsilon > 0$, that

$$\lim_{t_0\to\infty} \mathbb{P}\big(\sup_{t\geq t_0}[\|f_{t+1} - \widetilde{f}_\rho\|_\rho - \inf_{f\in\mathcal{H}_K} \|f - \widetilde{f}_\rho\|_\rho] \geq 2\varepsilon\big) = 0.$$

This completes the proof of the corollary. □

From Theorem 1 and the estimation (4.18) for the approximation error, we can derive the explicit error rates for OPERA stated in Theorem 2.

Proof of Theorem 2. Applying (4.18) with $\beta > 0$ and (6.1) with $\gamma_\ell = \frac{1}{\mu}\ell^{-\theta}$, $j = 2$ and $k = T$, we have that

$$\begin{aligned}
&\|\omega_2^T(L_K)L_K^\beta\|_{\mathcal{L}(L^2_\rho)} \leq \big(\big(\tfrac{\beta}{e}\big)^\beta + \kappa^{2\beta}\big)\Big(\textstyle\sum_{\ell=2}^T \gamma_\ell\Big)^{-\beta} \leq \big(\big(\tfrac{\beta}{e}\big)^\beta + \kappa^{2\beta}\big)\Big(\textstyle\sum_{\ell=2}^T \frac{1}{\mu}\ell^{-\theta}\Big)^{-\beta} \\
&\leq \big(\big(\tfrac{\beta}{e}\big)^\beta + \kappa^{2\beta}\big)\kappa^{2\beta}\mu^\beta\Big(\textstyle\sum_{\ell=2}^T \ell^{-\theta}\Big)^{-\beta} \\
&\leq \big(\big(\tfrac{\beta}{e}\big)^\beta + \kappa^{2\beta}\big)\kappa^{2\beta}(\mu(1-\beta))^\beta\Big(T^{1-\theta} - 1\Big)^{-\beta} \\
&\leq \big[\big(\big(\tfrac{\beta}{e}\big)^\beta + \kappa^{2\beta}\big)\kappa^{2\beta}(\mu(1-\beta))^\beta(1 - (\tfrac{1}{2})^{1-\theta})^{-\beta}\big]T^{-\beta(1-\theta)} := D_{\kappa,\beta}T^{-\beta(1-\theta)}
\end{aligned}$$

Putting this estimation into Theorem 1 yields, with probability $1 - \delta$, that

$$\|f_{T+1} - \widetilde{f}_\rho\|_\rho \leq D_{\kappa,\beta}T^{-\beta(1-\theta)} + C_{\theta,\kappa}T^{-\min\{\theta - \frac{1}{2}, \frac{1-\theta}{2}\}}\log T \log(8T/\delta)). \qquad (5.10)$$

Selecting $\theta = \min\big\{\frac{2\beta+1}{2\beta+2}, \frac{2}{3}\big\}$ implies, for probability $1 - \delta$, that

$$\|f_{T+1} - \widetilde{f}_\rho\|_\rho \leq (D_{\kappa,\beta} + C_{\theta,\kappa})T^{-\min\left(\frac{\beta}{2\beta+2}, \frac{1}{6}\right)}\log T \log(8T/\delta).$$

This completes the proof of the theorem. □

We now turn our attention to the special pairwise kernel (2.8) induced by a univariate kernel G. Let us first prove Proposition 1 which describes the relationship between the space $\mathcal{H}_K$ with the pairwise kernel K and $\mathcal{H}_G$ with the univariate kernel G.

Proof of Proposition 1. To prove (a), for any $n \in \mathbb{N}$, $\{\alpha_i : i = 1, \ldots, n\}$ and $\{(x_i^1, x_i^2) \in \mathcal{X} \times \mathcal{X} : i = 1, \ldots, n\}$, let $g = \sum_{i=1}^n \alpha_i(G_{x_i^1} - G_{x_i^2}) \in \mathcal{H}_G$. Indeed, it can be further be verified that $g \in \mathcal{I}_G^\perp$ since $\langle g, 1_\mathcal{X}\rangle_G = \langle\sum_{i=1}^n \alpha_i(G_{x_i^1} - G_{x_i^2}), 1_\mathcal{X}\rangle_G = \sum_{i=1}^n \alpha_i(1_\mathcal{X}(x_i^1) - 1_\mathcal{X}(x_i^2)) = 0$. Then, for any $x^1, x^2 \in \mathcal{X}$,

$$\begin{aligned}
\textstyle\sum_{i=1}^n \alpha_i K_{(x_i^1, x_i^2)}(x^1, x^2) \quad &= \textstyle\sum_{i=1}^n \alpha_i(G_{x_i^1}(x^1) - G_{x_i^2}(x^1)) - \sum_{i=1}^n \alpha_i(G_{x_i^1}(x^2) - G_{x_i^2}(x^2)) \\
&:= g(x^1) - g(x^2),
\end{aligned}$$

From the observation that $K((x^1, x^2), (\hat{x}^1, \hat{x}^2)) = \langle G_{x^1} - G_{x^2}, G_{\hat{x}^1} - G_{\hat{x}^2}\rangle_G$, we also see that

$$\|\Im(g)\|_K = \|\sum_{i=1}^n \alpha_i K_{(x_i^1, x_i^2)}\|_K = \|\sum_{i=1}^n \alpha_i(G_{x_i^1} - G_{x_i^2})\|_G = \|g\|_G.$$

According to [2], the RKHS $\mathcal{H}_K$ is the completion of the above linear span of kernel sections $\{K_{(x_i^1, x_i^2)} : x_i^1, x_i^2 \in \mathcal{X}, i = 1, \ldots, n\}$ and likewise, $\mathcal{H}_G$ is the completion of the linear span of kernel sections $\{\{G_{x_1^i}, G_{x_i^2}\} : x_i^1, x_i^2 \in \mathcal{X}, i = 1, \ldots, n\}$ which implies that, for any $f \in \mathcal{H}_K$, there exists $g \in \mathcal{I}_G$ such that $f(x^1, x^2) = g(x^1) - g(x^2)$, and $\|f\|_K = \|g\|_G$. It remains to prove that $\Im(g) = 0$ then $g \in \mathcal{I}_G$. Indeed, $\Im(g)(x^1, x^2) = 0$ implies that $g(x^1) = g(x^2)$ for any $x^1, x^2 \in \mathcal{X}$. This means that g is a constant function which means $g \in \mathcal{I}_G$. This completes part (a) of the proposition.

Part (b) follows from part (a) since, in this case, $\mathcal{I}_G = \{0\}$ which implies $\mathcal{H}_G = \mathcal{I}_G^{\perp}$. This completes the proof of the proposition. □

Secondly, for the special pairwise kernel given by (2.8), we can establish the convergence of online pairwise learning algorithm (2.9) as stated in Theorem 3.

Proof of Theorem 3: Part (a) directly follows from Theorem 1, Proposition 1 and the definition of $\mathcal{K}_G$ given by (2.11).

For part (b), under the assumption $1_{\mathcal{X}} \notin \mathcal{H}_G$, from Proposition 1 we have

$$\begin{aligned} \mathcal{K}_G(s, \widetilde{f}_\rho) &\leq 2 \inf_{g \in \mathcal{H}_G} \{\|g - f_\rho\|_\rho + \tfrac{s}{2}\|g\|_G\} \\ &\leq 2\sqrt{2}\Big(\inf_{g \in \mathcal{H}_G}\{\|g - f_\rho\|_\rho^2 + \tfrac{s^2}{4}\|g\|_G^2\}\Big)^{1/2}. \end{aligned} \tag{5.11}$$

According to [12, 15], $\inf_{g \in \mathcal{H}_G}\{\|g - f_\rho\|_\rho^2 + \lambda \|g\|_G^2\} \leq \lambda^{2\beta}\|L_G^{-\beta} f_\rho\|_\rho$ for any $\beta \leq 1/2$. Now applying this estimation and (5.11) with $\lambda = \frac{s^2}{4}$ and $s = \sum_{j=2}^t \gamma_j$ implies that

$$\mathcal{K}_G(\sqrt{6}\kappa(1+\kappa)T^{-\frac{1-\theta}{2}}, \widetilde{f}_\rho) \leq \mathcal{O}(T^{-(1-\theta)\beta}).$$

Putting this into (2.6) and choosing $\gamma_t = \frac{1}{\kappa^2} t^{-\frac{2\beta+1}{2\beta+2}}$ yields the desired result. This completes the proof of the theorem. □

6 Conclusion

This paper studied an online learning algorithm for pairwise learning in an unconstrained RKHS setting called OPERA. OPERA has a non-strongly convex objective function and is performed in an unconstrained setting, for which we are not aware of similar studies for such online pairwise learning algorithms. We established its almost-surely convergence and derived explicit error rates for polynomially decaying step sizes. Below we discuss some possible directions for future work.

Firstly, the rates of OPERA under the regularity assumption $\widetilde{f}_\rho \in L_K^\beta(L_\rho^2)$ are of the form $\mathbb{E}[\|f_{T+1} - \widetilde{f}_\rho\|_\rho] \leq \mathcal{O}(T^{-\frac{\beta}{2\beta+2}})$, which is suboptimal compared with the rate $\mathcal{O}(T^{-\frac{\beta}{2\beta+1}})$ in the univariate case [41]. It would be very interesting to improve the rates of OPERA.

Secondly, OPERA is not a fully online learning algorithm since it needs to save previous samples $\mathbf{z}^t = \{(x_i, y_i) : i = 1, \ldots, t\}$ at iteration t, although, in the linear case, efficient implementation may be possible. Hence, to improve the practical implementation of OPERA, the other direction would be to introduce a memory-efficient implementation which uses a finite buffer of capacity as in [18, 37]. In this case, OPERA would work with finite buffers associated with the local loss $L_t(f) = \frac{1}{|B_t|}\sum_{(x,y)\in B_t}(f(x_t,x) - y_t + y)^2$ at each iteration, where B_t is the state of the buffer at iteration t. We expect that the resultant convergence rate of this modified OPERA would be related to the capacity of the buffer set and the total number of training samples.

Finally, note that the techniques developed in this paper heavily depend on the error decomposition (i.e. equations (4.2) and (4.3)). It seems that they can not be directly applied to handle other popular loss functions such as the hinge loss and the logistic loss. We recently have developed completely different techniques in [43]. This enables us to prove the convergence of the last iterate of an online pairwise learning algorithm similar to OPERA with the least square being replaced by a smooth loss function (e.g. the logistic loss).

Acknowledgements

The work by D. X. Zhou described in this paper is supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 105011].

References

[1] S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *J. of Machine Learning Research*, **10**: 441–474, 2009.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**: 337–404, 1950.

[3] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[4] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *J. of Machine Learning Research*, **3**: 463–482, 2002.

[5] A. Bellet and A. Habrard. Robustness and generalization for metric learning. *arXiv preprint*, 2014. `http://arxiv.org/pdf/1209.1086.pdf`

[6] J. Bergh and J. Löfström. *Interpolation Spaces, An Introduction.* Springer-Verlag, New York, 1976.

[7] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[8] Q. Cao, Z. C. Guo and Y. Ying. Generalization bounds for metric and similarity learning. To appear in *Machine Leanring Journal*, 2015.

[9] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inform. Theory*, **50**: 2050–2057, 2004.

[10] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, **13**: 201–215, 2010.

[11] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *J. of Machine Learning Research*, **11**: 1109–1135, 2010.

[12] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, **39**: 1-49, 2001.

[13] S. Clémencon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, **36**: 844–874, 2008.

[14] F. Cucker and D.-X. Zhou. *Learning Theory: An Approximation Theory Viewpoint.* Cambridge Univesity Press, 2007.

[15] E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, **5**: 59–85, 2005.

[16] Z. C. Guo and Y. Ying. Guaranteed classification via regularized similarity learning. *Neural Computation*, **26**: 497-522, 2014.

[17] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM International conference on Knowledge discovery and data mining (SIGKDD)*, 2002.

[18] P. Kar, B. K Sriperumbudur, P. Jain and H. C. Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[19] J. H. Lin and D.X. Zhou. Learning theory of radomized Kaczmarz algorithm. *Submitted for publication*, 2014.

[20] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *J. of Machine Learning Research*, **7**: 2651-2667, 2006.

[21] S. Mukherjee and Q. Wu. Estimation of gradients and coordinate covariation in classification. *J. of Machine Learning Research*, **7**: 2481-2514, 2006.

[22] S. Mukherjee and D. X. Zhou. Learning coordinate covariances via gradients. *J. of Machine Learning Research*, **7**: 519-549, 2006.

[23] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, **19(4)**: 1574–609, 2008.

[24] I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *Ann. Prob.* **22**: 1679-1706, 1994.

[25] W. Rejchel. On ranking and generalization bounds. *J. of Machine Learning Research*, **13**: 1373-1392, 2012.

[26] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, **22**: 400–407, 1951.

[27] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. *Proceedings of the 30th International Conference on Machine Learning (ICML), Atlanta, GA, USA*, 2013.

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

[29] S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.* **26**: 153–172, 2007.

[30] S. Smale and D.X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, **1**: 1741, 2003.

[31] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. of Machine Learning Research*, **2**: 67–93, 2001.

[32] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer-Verlag, New York, 2008.

[33] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, **52**: 4635–4643, 2006.

[34] P. Tarrés and Y. Yao. Online learning as stochastic approximations of regularization paths. *IEEE Transactions on Information Theory*, **60**: 5716–5735, 2014.

[35] K. Q. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbour classification. *J. of Machine Learning Research*, **10**: 207–244, 2009.

[36] Y. Wang, R. Khardon, D. Pechyony and R. Jones. Generalization bounds for online learning algorithms with pairwise loss functions. *The 25th Annual Conference on Learning Theory (COLT)*, 2012.

[37] Y. Wang, R. Khardon, D. Pechyony, and R. Jones. Online learning with pairwise loss functions. *arxiv preprint, 2013.* `http://arxiv.org/pdf/1301.5332v1.pdf`

[38] Q. Wu, Y. Ying, and D. X. Zhou. Learning rates of least-square regularized regression. *Found. Comput. Math.*, **6**: 171–192, 2005.

[39] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, **26(2)**: 289-315, 2007.

[40] Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *J. of Machine Learning Research*, **13**: 1–26, 2012.

[41] Y. Ying and M. Pontil. Online gradient descent algorithms. *Found. Comput. Math.*, **5**: 561–596, 2008.

[42] Y. Ying and D. X. Zhou. Online regularized classification algorithms. *IEEE Transaction on Information Theory*, **11**: 4775-4788, 2006.

[43] Y. Ying and D.X. Zhou. Unregularized online learning algorithms with general loss functions. *ArXiv Preprint*, 2015. Available in `http://arxiv.org/pdf/1503.00623v2.pdf`

[44] P. Zhao, S. C. H. Hoi, R. Jin and T. Yang. Online AUC Maximization. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.

[45] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine learning (ICML)*, 2003.

Appendix

Here we present the proofs for Lemmas 6 and Lemma 7. To this end, we first state a technical lemma which will be used later.

Lemma 8. *Let $\gamma_j = \frac{j^{-\theta}}{\mu}$ for any $j \in \mathbb{N}$ with $\theta \in (0,1)$. Then, for any $1 \le j \le k$, there holds*

$$\frac{1}{\mu(1-\theta)}((k+1)^{1-\theta} - j^{1-\theta}) \le \sum_{\ell=j}^{k} \gamma_\ell \le \frac{1}{\mu(1-\theta)}(k^{1-\theta} - (j-1)^{1-\theta}). \tag{6.1}$$

Proof. Notice that $\ell^{-\theta} \le s^{-\theta}$ for $s \in [\ell-1, \ell]$ and $\ell^{-\theta} \ge s^{-\theta}$ for $s \in [\ell, \ell+1]$. Hence, $\frac{1}{\mu}\sum_{\ell=j}^{k} \displaystyle\int_{\ell}^{\ell+1} s^{-\theta} ds \le \sum_{\ell=j}^{k} \gamma_\ell \le \frac{1}{\mu}\sum_{\ell=j}^{k}\int_{\ell-1}^{\ell} s^{-\theta} ds$ which implies that

$$\frac{1}{\mu}\int_{j}^{k+1} s^{-\theta} ds \le \sum_{\ell=j}^{k} \gamma_\ell \le \frac{1}{\mu}\int_{j-1}^{k} s^{-\theta} ds.$$

The desired result follows directly from the above inequality. □

We are ready to establish the proof of Lemma 6.

Proof of Lemma 6. Let $\mathscr{J} := \sum_{j=2}^{t} \frac{\gamma_j (1 + (\sum_{\ell=2}^{j-1}\gamma_\ell)^{1/2})}{\sqrt{j}\big(1+\sum_{\ell=j+1}^{t}\gamma_\ell\big)^{1/2}}$. It can be written as

$$\begin{aligned}\mathscr{J} &= \Big[\tfrac{\gamma_t(1+(\sum_{\ell=2}^{t-1}\gamma_\ell)^{1/2})}{\sqrt{t}}\Big] + \Big[\tfrac{\gamma_2}{\sqrt{2}\big(1+\sum_{\ell=3}^{t}\gamma_\ell\big)^{1/2}}\Big] + \Big[\textstyle\sum_{j=3}^{t-1} \tfrac{\gamma_j(1+(\sum_{\ell=2}^{j-1}\gamma_\ell)^{1/2})}{\sqrt{j}\big(1+(\sum_{\ell=j+1}^{t}\gamma_\ell)^{1/2}\big)^{1/2}}\Big]\\ &:= \mathscr{J}_1 + \mathscr{J}_2 + \mathscr{J}_3.\end{aligned} \tag{6.2}$$

We estimate $\mathscr{J}_1, \mathscr{J}_2$, and $\mathscr{J}_3$ separately as follows.

Firstly, let us look at the term $\mathscr{J}_1$. Indeed, by (6.1) we have

$$\begin{aligned}\mathscr{J}_1 &= \tfrac{1}{\mu}t^{-\theta-1/2}\big(1 + \tfrac{1}{\sqrt{\mu}}(\textstyle\sum_{\ell=1}^{t-1}\ell^{-\theta})^{1/2}\big) \le \tfrac{1}{\mu}t^{-\theta-1/2}\big(1 + \tfrac{t^{\frac{1-\theta}{2}}}{\sqrt{\mu(1-\theta)}}\big)\\ &\le 2\max(1, (\sqrt{\mu(1-\theta)})^{-1})t^{-\frac{3\theta}{2}}.\end{aligned} \tag{6.3}$$

Secondly, for the term $\mathscr{J}_2$, we apply (6.1) again to get that

$$\mathscr{J}_2 \le \tfrac{2^{-\theta-1/2}}{\mu}\tfrac{\sqrt{\mu(1-\theta)}}{\big((t+1)^{1-\theta}-3^{1-\theta}\big)^{1/2}} \le \big(\tfrac{(1-\theta)}{(1-(3/5)^{1-\theta})\mu}\big)^{1/2} t^{-(1-\theta)/2} \le \big(\tfrac{5}{2\mu}\big)^{1/2} t^{-(1-\theta)/2}, \tag{6.4}$$

where the second to last inequality used the assumption $t \ge 4$ which implies $3^{1-\theta} \le (\frac{3}{5}(t+1))^{1-\theta}$,, and the last inequality used the property that, for any $0 < \theta < 1$ and $0 < x < 1$, that $(1-x)^{1-\theta} \ge (1-\theta)(1-x)$.

Lastly, we estimate the term $\mathscr{J}_3$. To this end, by (6.1) we can estimate $\mathscr{J}_3$ as follows:

$$\begin{aligned}
\mathcal{J}_3 &\le \tfrac{1}{\mu}\textstyle\sum_{j=3}^{t-1} \frac{j^{-\theta}\left(1+\frac{1}{\sqrt{\mu(1-\theta)}}((j-1)^{1-\theta}-1)^{1/2}\right)}{\sqrt{j}\left(1+\frac{1}{\mu(1-\theta)}((t+1)^{1-\theta}-(j+1)^{1-\theta})\right)^{1/2}} \\
&\le \tfrac{2}{\mu}\max(1,(\sqrt{\mu(1-\theta)})^{-1})\textstyle\sum_{j=3}^{t-1} \frac{j^{-\frac{3\theta}{2}}}{\left(1+\frac{1}{\mu(1-\theta)}((t+1)^{1-\theta}-(j+1)^{1-\theta})\right)^{1/2}} \qquad (6.5) \\
&\le \tfrac{2}{\mu}\max(\sqrt{\mu(1-\theta)})^{-1},\sqrt{\mu(1-\theta)})\textstyle\sum_{j=3}^{t-1} \frac{j^{-\frac{3\theta}{2}}}{\left(1+((t+1)^{1-\theta}-(j+1)^{1-\theta})\right)^{1/2}}.
\end{aligned}$$

It remains to estimate $\sum_{j=3}^{t-1} \frac{j^{-\frac{3\theta}{2}}}{\left(1+((t+1)^{1-\theta}-(j+1)^{1-\theta})\right)^{1/2}}$. To this end, we further decompose it into two terms as

$$\begin{aligned}
\textstyle\sum_{j=3}^{t-1} \frac{j^{-\frac{3\theta}{2}}}{(1+((t+1)^{1-\theta}-(j+1)^{1-\theta}))^{1/2}} &= (\textstyle\sum_{j>t/2}^{t-1}+\sum_{3\le j\le t/2}) \frac{j^{-\frac{3\theta}{2}}}{(1+((t+1)^{1-\theta}-(j+1)^{1-\theta}))^{1/2}} \\
&:= \widetilde{\mathcal{J}}_{31}+\widetilde{\mathcal{J}}_{32}.
\end{aligned} \qquad (6.6)$$

For $\widetilde{\mathcal{J}}_{31}$, for any $s \in [j, j+1]$, that $j^{-\theta} \le 2^{\theta}(1+s)^{-\theta}$ and $(t+1)^{1-\theta}-(j+1)^{1-\theta} \ge (t+1)^{1-\theta}-(s+1)^{1-\theta}$. Then,

$$\begin{aligned}
\widetilde{\mathcal{J}}_{31} &\le 2^{\frac{\theta}{2}}t^{-\frac{\theta}{2}}\textstyle\sum_{j>t/2}^{t-1} \frac{j^{-\theta}}{(1+((t+1)^{1-\theta}-(j+1)^{1-\theta}))^{1/2}} \\
&\le 2^{3\theta/2}t^{-\frac{\theta}{2}}\textstyle\sum_{j>t/2}^{t-1} \displaystyle\int_j^{j+1} \frac{(1+s)^{-\theta}ds}{(1+(t+1)^{1-\theta}-(s+1)^{1-\theta})^{1/2}} \\
&\le 2^{3\theta/2}t^{-\frac{\theta}{2}} \int_{t/2}^{t} \frac{(1+s)^{-\theta}ds}{(1+(t+1)^{1-\theta}-(s+1)^{1-\theta})^{1/2}} \qquad (6.7) \\
&\le \tfrac{2^{1+3\theta/2}}{1-\theta}t^{-\frac{\theta}{2}}\left[(1+(t+1)^{1-\theta})-(t/2+1)^{1-\theta}\right]^{1/2} \\
&\le \tfrac{2^{1+3\theta/2}}{1-\theta}t^{-\frac{\theta}{2}}(t+1)^{\frac{1-\theta}{2}} \le \tfrac{4\sqrt{2}}{1-\theta}t^{\frac{1}{2}-\theta}
\end{aligned}$$

For $\widetilde{\mathcal{J}}_{32}$, the fact that $(t+1)^{1-\theta}-(j+1)^{1-\theta} \ge (1-(2/3)^{1-\theta})(t+1)^{1-\theta}$ for any $j \le t/2$ implies that

$$\widetilde{\mathcal{J}}_{32} \le \frac{1}{t^{\frac{1-\theta}{2}}(1-(2/3)^{1-\theta})}\sum_{3\le j<t/2} j^{-3\theta/2} \le \frac{3}{1-\theta}t^{-(1-\theta)/2}\sum_{3\le j<t/2} j^{-3\theta/2}. \qquad (6.8)$$

Notice that

$$\sum_{3\le j<t/2} j^{\frac{-3\theta}{2}} \le \int_2^{t/2} s^{-3\theta/2}ds \le \begin{cases} \frac{2}{|2-3\theta|}t^{-\min(0,\frac{3\theta-2}{2})}, & \text{if } \theta \ne 2/3 \\ \ln t, & \text{if } \theta = 2/3 \end{cases}$$

Putting the above inequality into (6.8) yields that

$$\widetilde{\mathcal{J}}_{32} \le A_\theta t^{-\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\ln t, \qquad (6.9)$$

where $A_\theta = \frac{6}{(1-\theta)|3\theta-2|}$ if $\theta \ne 2/3$ and $\frac{3}{1-\theta}$ otherwise. Combining (6.7) and (6.9), (6.5), and (6.6) together implies that

$$\mathcal{J}_3 \le B_\theta t^{-\min(\theta-\frac{1}{2},\frac{1-\theta}{2})}\ln t, \qquad (6.10)$$

where

$$B_\theta = \begin{cases} \frac{4\max(\sqrt{\mu(1-\theta)})^{-1},\sqrt{\mu(1-\theta)})}{\mu(1-\theta)}(2\sqrt{2}+\frac{3}{|3\theta-2|}), & \text{if } \theta \neq 2/3 \\ \frac{2(3+4\sqrt{2})\max(\sqrt{\mu(1-\theta)})^{-1},\sqrt{\mu(1-\theta)})}{\mu(1-\theta)}, & \text{if } \theta = 2/3. \end{cases}$$

Now putting estimates (6.3), (6.4), and (6.10) together yields the desired result. This completes the proof of the lemma. □

We now turn our attention to the proof for Lemma 7.

Proof of Lemma 7. Let $\mathcal{I} = \sum_{j=2}^{t} \frac{\gamma_j^2(1+\sum_{\ell=2}^{j-1}\gamma_\ell)}{1+\sum_{\ell=j+1}^{t}\gamma_\ell}$. We can write $\mathcal{I}$ as

$$\begin{aligned} \mathcal{I} &= \big[\gamma_t^2(1+\textstyle\sum_{\ell=2}^{t-1}\gamma_\ell)\big] + \big[\frac{\gamma_2^2}{(1+\sum_{\ell=3}^{t-1}\gamma_\ell)}\big] + \big[\textstyle\sum_{j=3}^{t-1}\frac{\gamma_j^2(1+\sum_{\ell=2}^{j-1}\gamma_\ell)}{1+\sum_{\ell=j+1}^{t}\gamma_\ell}\big] \\ &:= \mathcal{I}_1+\mathcal{I}_2+\mathcal{I}_3, \end{aligned} \tag{6.11}$$

where we used the conventional notation $\sum_{\ell=j+1}^{j}\gamma_\ell = 0$ for any $j \in \mathbb{N}$. We estimate $\mathcal{I}_1, \mathcal{I}_2$, and $\mathcal{I}_3$ term by term as follows.

Firstly, let us first estimate $\mathcal{I}_1$. By (6.1), we can have that

$$\begin{aligned} \mathcal{I}_1 &\leq \frac{1}{\mu^2}t^{-2\theta}(1+\frac{1}{\mu(1-\theta)}((t-1)^{1-\theta}-1)) \\ &\leq \frac{2\max(1,(\mu(1-\theta))^{-1})}{\mu^2}t^{1-3\theta}. \end{aligned} \tag{6.12}$$

Secondly, we move on to the estimation of term $\mathcal{I}_2$. By (6.1), we obtain that

$$\begin{aligned} \mathcal{I}_2 &\leq \frac{1}{4\mu^2}\frac{1}{1+\frac{1}{\mu(1-\theta)}((t+1)^{1-\theta}-3^{1-\theta})} \\ &\leq \frac{1-\theta}{4\mu\Big(1-(\frac{3}{5})^{1-\theta}\Big)}t^{-(1-\theta)} \leq \frac{5}{8\mu}t^{-(1-\theta)} \end{aligned} \tag{6.13}$$

where, in the second to last inequality, we used the assumption $t \geq 4$ which implies $3^{1-\theta} \leq (\frac{3}{5}(t+1))^{1-\theta}$, and the last inequality used the fact, for any $0<\theta<1$ and $0<x<1$, that $(1-x)^{1-\theta} \geq (1-\theta)(1-x)$.

Finally, we turn our attention to the estimation of $\mathcal{I}_3$. Applying (6.1) again to $\mathcal{I}_3$ implies that

$$\begin{aligned} \mathcal{I}_3 &\leq \frac{1}{\mu^2}\sum_{j=3}^{t-1}\frac{j^{-2\theta}\Big(1+\frac{1}{\mu(1-\theta)}((j-1)^{1-\theta}-1)\Big)}{1+\frac{1}{\mu(1-\theta)}\Big((t+1)^{1-\theta}-(j+1)^{1-\theta}\Big)} \\ &\leq \frac{2}{\mu^2}\sum_{j=3}^{t-1}\frac{j^{-2\theta}\max\Big(1,\frac{1}{\mu(1-\theta)}\Big)j^{1-\theta}}{1+\frac{1}{\mu(1-\theta)}\Big((t+1)^{1-\theta}-(j+1)^{1-\theta}\Big)} \\ &\leq \frac{2\max(1,(\mu(1-\theta)^{-1}))\max(1,\mu(1-\theta))}{\mu^2}\sum_{j=3}^{t-1}\frac{j^{1-3\theta}}{1+\Big((t+1)^{1-\theta}-(j+1)^{1-\theta}\Big)} \\ &\leq \frac{2\max(\mu(1-\theta),(\mu(1-\theta)^{-1}))}{\mu^2}\sum_{j=3}^{t-1}\frac{j^{1-3\theta}}{1+\Big((t+1)^{1-\theta}-(j+1)^{1-\theta}\Big)}. \end{aligned} \tag{6.14}$$

It now suffices to estimate the term $\mathcal{I}_3 := \sum_{j=3}^{t-1}\frac{j^{1-3\theta}}{1+\Big((t+1)^{1-\theta}-(j+1)^{1-\theta}\Big)}$, which can be written as

$$\begin{aligned} \mathcal{I}_3 &= \big(\textstyle\sum_{j>t/2}^{t}+\sum_{3\leq j\leq t/2}\big)\frac{j^{1-3\theta}}{1+\Big((t+1)^{1-\theta}-(j+1)^{1-\theta}\Big)} \\ &:= \widetilde{\mathcal{I}}_{31}+\widetilde{\mathcal{I}}_{32}. \end{aligned} \tag{6.15}$$

For the first term $\widetilde{\mathcal{I}}_{31}$, observe, for any $s \in [j, j+1]$, that $j^{-\theta} \leq 2^{\theta}(1+s)^{-\theta}$ and $(t+1)^{1-\theta} - (j+1)^{1-\theta} \geq (t+1)^{1-\theta} - (s+1)^{1-\theta}$. Therefore,

$$\begin{aligned}
\widetilde{\mathcal{I}}_{31} &:= \textstyle\sum_{j>t/2}^{t-1} \frac{j^{1-3\theta}}{1+\big((t+1)^{1-\theta}-(j+1)^{1-\theta}\big)} \\
&\leq 2^{2\theta-1} t^{1-2\theta} \textstyle\sum_{j>t/2}^{t-1} \displaystyle\int_j^{j+1} \frac{(s+1)^{-\theta}}{1+\big((t+1)^{1-\theta}-(s+1)^{1-\theta}\big)} ds \\
&\leq 2^{2\theta-1} t^{1-2\theta} \int_{t/2}^{t} \frac{(s+1)^{-\theta}}{1+\big((t+1)^{1-\theta}-(s+1)^{1-\theta}\big)} ds \\
&= \tfrac{2^{2\theta-1}t^{1-2\theta}}{1-\theta} \left[\ln(1+\big((t+1)^{1-\theta}-(t/2)^{1-\theta}\big)) - \ln(1+\big((t+1)^{1-\theta}-t^{1-\theta}\big))\right] \\
&\leq \tfrac{2^{2\theta-1}t^{1-2\theta}}{1-\theta} \ln(t+1)^{1-\theta} \leq 2^{2\theta-1} t^{1-2\theta} \ln(t+1) \leq 4t^{1-2\theta} \ln t.
\end{aligned} \tag{6.16}$$

For $\widetilde{\mathcal{I}}_{32}$, we have

$$\begin{aligned}
\widetilde{\mathcal{I}}_{32} &= \sum_{3 \leq j \leq t/2} \frac{j^{1-3\theta}}{1+\big((t+1)^{1-\theta}-(j+1)^{1-\theta}\big)} \\
&\leq \sum_{3 \leq j \leq t/2} \frac{j^{1-3\theta}}{1+(1-(2/3)^{1-\theta})(t+1)^{1-\theta}} \\
&\leq \tfrac{t^{-(1-\theta)}}{(1-(2/3)^{1-\theta})} \sum_{3 \leq j \leq t/2} j^{1-3\theta} \leq \frac{3t^{-(1-\theta)}}{(1-\theta)} \sum_{3 \leq j \leq t/2} j^{1-3\theta},
\end{aligned} \tag{6.17}$$

where we used again the fact, for any $0 < \theta < 1$ and $0 < x < 1$, that $(1-x)^{1-\theta} \geq (1-\theta)(1-x)$. Also, by a simple calculation, there holds

$$\sum_{3 \leq j \leq t/2} j^{1-3\theta} \leq \begin{cases} \frac{1}{|2-3\theta|} t^{-\min(0, 3\theta-2)}, & \text{if } \theta \neq 2/3 \\ \ln t, & \text{if } \theta = 2/3. \end{cases}$$

Putting the above estimation into (6.17) yields that

$$\widetilde{\mathcal{I}}_{32} \leq \widetilde{A}_{\theta}\, t^{-\min(2\theta-1, 1-\theta)} \ln t. \tag{6.18}$$

where $\widetilde{A}_{\theta} = \frac{3}{|3\theta-2|(1-\theta)}$ if $\theta \neq 2/3$ and $\frac{3}{1-\theta}$ otherwise. Putting (6.16) and (6.18) back into (6.14) implies that

$$\widetilde{\mathcal{I}}_{3} \leq \widetilde{B}_{\theta}\, t^{-\min(2\theta-1, 1-\theta)} \ln t, \tag{6.19}$$

where $\widetilde{B}_{\theta} = \frac{3}{(1-\theta)|3\theta-2|} + 4$ if $\theta \neq 2/3$ and $\frac{3}{1-\theta} + 4$ otherwise. Combining estimates (6.12), (6.13), and (6.19) together yields the desired result. This completes the proof of the lemma. □