Adv Comput Math
DOI 10.1007/s10444-016-9479-7

Online regularized learning with pairwise loss functions

Zheng-Chu Guo[1] · **Yiming Ying**[2] · **Ding-Xuan Zhou**[3]

Received: 13 May 2014 / Accepted: 5 August 2016

Abstract Recently, there has been considerable work on analyzing learning algorithms with pairwise loss functions in the batch setting. There is relatively little theoretical work on analyzing their online algorithms, despite of their popularity in practice due to the scalability to big data. In this paper, we consider online learning algorithms with pairwise loss functions based on regularization schemes in reproducing kernel Hilbert spaces. In particular, we establish the convergence of the last iterate of the online algorithm under a very weak assumption on the step sizes and derive satisfactory convergence rates for polynomially decaying step sizes. Our technique uses Rademacher complexities which handle function classes associated with pairwise loss functions. Since pairwise learning involves pairs of examples, which are no longer i.i.d., standard techniques do not directly apply to such pairwise learning algorithms. Hence, our results are a non-trivial extension of those in the setting of univariate loss functions to the pairwise setting.

Keywords Pairwise learning · Online learning · Regularization · RKHS

Communicated by: Karsten Urban

✉ Yiming Ying
mathying@gmail.com

[1] School of Mathematical Sciences, Zhejiang University, Hangzhou, 310027, China

[2] Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY, 12222, USA

[3] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

Published online: 16 August 2016

Mathematics Subject Classfication (2010) 68Q32 · 68T05 · 62J02 · 62L20

1 Introduction

For any $T \in \mathbb{N}$, the input space $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and the output space $\mathcal{Y} \subseteq \mathbb{R}$. In the standard framework of learning theory [11, 29], 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 $h : \mathcal{X} \to \mathcal{Y}$.

There are various important learning problems involving 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 $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For example, metric learning [12, 17, 30, 35] aims to learn a metric D such that examples with the same label stay closer while pushing apart examples with distinct labels. In this setting, a typical pairwise loss function is given, for any $(x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}$, by $\ell(D, (x, y), (x', y')) = (1 + r(y, y')D(x, x'))_+ = \max(0, 1 + r(y, y')D(x, x'))$ where $r(y, y') = 1$ if $y = y'$ and -1 otherwise. Another prominent example is the problem of bipartite ranking [1, 8, 10, 25], which aims to predict the ordering between objects from their observed features. The quality of a ranking rule $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be measured by a least-square pairwise loss function $\ell(f, (x, y), (x', y')) = (y - y' - f(x, x'))^2$. Other learning problems associated with pairwise loss functions include AUC maximization [38], gradient learning [21, 22], minimum error entropy principles [13, 15] and similarity learning [6, 9].

This paper considers learning problems associated with pairwise loss functions which, for simplicity, is referred to as *pairwise learning* problems. In this context, we assume that the hypothesis function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for pairwise learning problems 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. [11, 29]. 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 $g \in \mathcal{H}_K$, $\langle K_{(x,x')}, g\rangle_K = g(x, x')$.

A general regularization scheme in a RKHS $\mathcal{H}_K$ for pairwise learning can be written as

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

where $\lambda > 0$ is a regularization parameter. The above formulation is a common regularization formulation in the batch learning setting in the sense that the algorithm

uses the training data $\mathbf{z}$ at once. Recently, there has been considerable work on analyzing the generalization performance of the above batch learning algorithm and its related variants using the techniques of U-Statistics [6, 10, 25] or the concept of algorithmic stability [1, 17]. In contrast to well-studied pairwise learning in the batch setting, online learning algorithms only need to access the data sequentially and are more popular in practice due to their ability of analyzing big data. However, there is little work on designing and analyzing online learning with pairwise loss functions except recent work by [16, 32]. Specifically, Wang et al. [32] and Kar et al. [16] established generalization bounds for the average of the iterates of online learning with uniformly bounded pairwise loss functions. These results are established in the same spirit as the online to batch conversion bounds [7] for learning algorithms associated with univariate loss functions.

In this paper, we study the regularized online learning with pairwise loss functions and establish generalization bounds for its last iterate instead of the average of its iterates as studied in [16, 32]. Our technique uses Rademacher complexities in order to handle function classes associated with pairwise loss functions. Since pairwise learning involves pairs of examples, which are no longer i.i.d., and standard techniques in [37] do not directly apply to such pairwise learning algorithms. Hence, our results are a non-trivial extension of those in the setting of univariate loss functions [37] to the pairwise setting.

The remainder of this paper is organized as follows. In Section 2, we introduce online regularized learning algorithm associated with pairwise loss functions and state the main results. In particular, a general convergence theorem is established for the above online algorithms and their convergence rates with polynomial-decaying step sizes are established. Related work is discussed in Section 2.1. Section 3 develops some technical results which are needed to prove the main results stated in Section 2. Section 4 summarizes this paper and discuss some possible directions for future work.

2 Learning algorithm and main results

In this section, we introduce an online regularized learning algorithm associated with a pairwise loss $\ell(f(x, x'), r(y, y'))$ in a reproducing kernel Hilbert space $\mathcal{H}_K$, which is motivated by the learning algorithm [32] in the linear setting. For simplicity, we restrict our attention to the hinge loss, i.e. $\ell(f(x, x'), r(y, y')) = (1 - r(y, y')f(x, x'))_+$. Here, r is a function from $\mathcal{Y} \times \mathcal{Y}$ to a bounded interval $[-M, M]$ with some constant $M > 0$, i.e.

$$\sup_{y,y' \in \mathcal{Y}} |r(y, y')| \leq M.$$

The definition of function r can vary in different learning settings. For example, $r(y, y') = sgn(y - y')$ for the problem of ranking and, for metric learning, $r(y, y') = 1$ if x and x' are from the same class and -1 otherwise.

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 regularized pairwise learning (ORPL) is given by $f_1 = f_2 = 0$ and

$$f_{t+1} = f_t - \eta_t \left[\frac{1}{t-1} \sum_{j=1}^{t-1} \ell'(f_t(x_t, x_j), r(y_t, y_j)) K_{(x_t, x_j)}(\cdot) + \lambda f_t \right], \quad \forall t \in \mathbb{N} \text{ and } 2 \le t \le T, \tag{2}$$

where $\{\eta_t > 0 : t \in \mathbb{N}\}$ is usually called the step size, $\lambda > 0$ is the regularization parameter and $\ell'(s, r(y, y'))$ denotes the sub-gradient of the hinge loss ℓ with respect to the first argument $s \in \mathbb{R}$.

In the above definition, the sub-gradient of the hinge loss can be defined by

$$\ell'(s, r(y, y')) = \begin{cases} -r(y, y') & \text{if } sr(y, y') \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The above online learning algorithm ORPL only needs a sequential access to the training data, and, from the above definition, we know that f_t only depends on the variables $\mathbf{z}_{t-1} = \{z_1, z_2, \ldots, z_{t-1}\}$. Specifically, at each time step $t+1$, the algorithm ORPL presumes a hypothesis $f_t \in \mathcal{H}_K$ upon which a new data z_t is revealed. The quality of f_t is assessed on the local regularized empirical error:

$$\widehat{\mathcal{E}}_\lambda^t(f) = \frac{1}{t-1} \sum_{j=1}^{t-1} \ell(f(x_t, x_j), r(y_t, y_j)) + \frac{\lambda}{2} \|f\|_K^2. \tag{3}$$

Then, a gradient step is made to update f_t based on the gradient of the above local empirical error $\widehat{\mathcal{E}}_\lambda^t(f_t)$ which is exactly given by

$$\nabla \widehat{\mathcal{E}}_\lambda^t(f)|_{f=f_t} = \frac{1}{t-1} \sum_{j=1}^{t-1} \ell'(f_t(x_t, x_j), r(y_t, y_j)) K_{(x_t, x_j)} + \lambda f_t.$$

Here, $\nabla \widehat{\mathcal{E}}_\lambda^t(f)$ denotes the functional gradient of the functional $\widehat{\mathcal{E}}_\lambda^t$ in the RKHS $\mathcal{H}_K$. Denote the true risk of a hypothesis f by

$$\mathcal{E}(f) = \iint \ell(f(x, x'), r(y, y')) d\rho(x, y) d\rho(x', y'). \tag{4}$$

Our main aim is to consider the generalization performance of f_{T+1}, i.e. the last iterate of ORPL. Consider the regularization function f_λ defined by

$$f_\lambda = \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_\lambda(f) =: \mathcal{E}(f) + \frac{\lambda}{2} \|f\|_K^2 \right\}. \tag{5}$$

Now we state the following theorem which shows that the last iterate of ORPL converges to the regularization function f_λ under certain conditions on the step sizes $\{\eta_t : t \in \mathbb{N}\}$.

Theorem 1 *For every fixed $\lambda > 0$, if the step sizes $\{\eta_t : t \in \mathbb{N}\}$ in algorithm (2) satisfy*

$$\lim_{t\to\infty} \eta_t = 0 \ \text{and} \ \sum_{t=2}^{\infty} \eta_t = \infty, \tag{6}$$

then we have

$$\lim_{T\to\infty} \mathbb{E}\left[\|f_T - f_\lambda\|_K\right] = 0.$$

Let $\kappa = \sup_{x,x'\in\mathcal{X}} \sqrt{K((x,x'),(x,x'))}$. If we choose explicit step sizes, the convergence rate of the last iterate of ORPL can be stated as follows.

Theorem 2 *Let $0 < \lambda \le 1$, $\{f_t : t = 2, 3, \ldots, T+1\}$ be defined by algorithm (2). If we choose the step size as $\eta_t = \frac{1}{\lambda t^\alpha}$ with $0 < \alpha < 1$, then*

$$\mathbb{E}\left[\|f_{T+1} - f_\lambda\|_K^2\right] \le C_1 \left(\frac{1}{\lambda^2 T^\alpha}\right),$$

where C_1 is a constant independent of T (see its explicit expression in the proof).

In particular, if the step sizes are chosen as $\eta_t = \frac{1}{\lambda t}$ then we can further obtain the following result.

Theorem 3 *Let $0 < \lambda \le 1$ and the iterates $\{f_t : t = 2, 3, \ldots, T+1\}$ be generated by algorithm (2). If we take the step size as $\eta_t = \frac{1}{\lambda t}$, then we have, for any $T \ge 2$, that*

$$\mathbb{E}\left[\|f_{T+1} - f_\lambda\|_K^2\right] \le C_2 \left(\frac{\log T}{\lambda^2 T}\right),$$

where C_2 is independent of T (see its explicit expression in the proof).

The proofs of Theorems 1, 2 and 3 will be given in Section 3.2. For fixed λ, Theorem 3 states that $\mathbb{E}\left[\|f_{T+1} - f_\lambda\|_K^2\right] = \mathcal{O}(\log T/T)$. This rate is consistent with the rate of standard stochastic gradient descent algorithms in the setting of classification and regression [24, 26].

In the literature of learning theory [11, 29], we are often interested in the *excess generalization error* $\mathcal{E}(f_{T+1}) - \inf_f \mathcal{E}(f)$, where the minimization is taken over all measurable pairwise functions. The above theorems mainly describe the convergence of $\|f_{T+1} - f_\lambda\|_K$ which is usually referred to as the *sample error*. By combining the *approximation error* which describes the difference between f_λ and $f_c = \arg\inf_f \mathcal{E}(f)$, we can drive the overall convergence rate of the excess generalization error. To this end, we prove the following lemma which may be interesting in its own right.

Lemma 2.1 Consider $\mathcal{Y} = \{\pm 1\}$ and define, for any $y, y' \in \mathcal{Y}$, $r(y, y') = yy'$. Let, for any $x, x' \in \mathcal{X}$, $\eta(x) = P(Y = 1|x)$ and $f_\rho(x, x') = 2\big[\eta(x)\eta(x') + (1 - \eta(x))(1 - \eta(x'))\big] - 1$. Then, $f_c = \text{sign}(f_\rho)$, i.e., for any $x, x' \in \mathcal{X}$,

$$f_c(x, x') = \begin{cases} 1, & f_\rho(x, x') \geq 0, \\ -1, & \text{otherwise.} \end{cases} \tag{7}$$

In binary classification (e.g. [11, 29]), one is often interested in finding an estimator from data to approximate the minimizer $g_c = \arg\inf_g \iint_{\mathcal{X}\times\mathcal{Y}}(1 - yg(x))_+ d\rho(x, y)$. It is well-known [34, 39] that g_c is identical to the *Bayes rule*, i.e. $g_c = \text{sign}(g_\rho)$ where g_ρ is defined by $g_\rho(x) = 2\eta(x) - 1$. Lemma 2.1 can be considered as extension of this classical result to the scenario of pairwise learning. We are not aware of any results similar to Lemma 2.1. Therefore, we outline its proof in Section 3.2 for completeness.

Now we state the overall rate for the excess generalization error.

Corollary 4 *Consider $\mathcal{Y} = \{\pm 1\}$ and define, for any $y, y' \in \mathcal{Y}$, $r(y, y') = yy'$. Assume, for some $0 < \beta \leq 1$, that $\mathcal{D}(\lambda) := \inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) - \mathcal{E}(f_c) + \frac{\lambda}{2}\|f\|_K^2\right\} = \mathcal{O}(\lambda^\beta)$. Let the iterates $\{f_t : t = 2, 3, \ldots, T + 1\}$ be generated by algorithm (2) with the choice $\eta_t = \frac{1}{\lambda t}$, then, by choosing $\lambda = \left(\frac{\log^2 T}{T}\right)^{\frac{1}{2(1+\beta)}}$, we have*

$$\mathbb{E}\big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_c)\big] = \mathcal{O}\left(T^{-\frac{\beta}{2(1+\beta)}}\sqrt{\log T}\right). \tag{8}$$

In the above corollary, the decay assumption on the approximation error $\mathcal{D}(\lambda)$ is standard, see e.g. [11, 29]. In the particular case of $f_c \in \mathcal{H}_K$, we have $\mathcal{D}(\lambda) \leq \lambda\|f_c\|_K^2$. This means that $\mathbb{E}\big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_c)\big] = \mathcal{O}\left(T^{-\frac{1}{4}}(\log T)^{1/2}\right)$. This rate is apparently sub-optimal when compared with the rates in the batch learning setting [1, 25]. Improving the learning rates of the online pairwise learning is one of the future research directions.

2.1 Related work and discussion

Recently, pairwise learning has attracting increasing attention. A key characteristic of pairwise learning is that pairs of examples are not i.i.d., and hence, standard techniques of generalization analysis for learning algorithms with a univariate loss function do not directly apply to pairwise learning. The generalization bounds of pairwise learning in the batch setting can be established by using U-statistics and algorithmic stability [5, 6, 10, 13, 14]. Below, we discuss some existing work which is closely related to ours.

Zhao et al. [38] proposed an online learning algorithm for maximizing area under ROC curve (AUC) for imbalanced classification. The main challenge for online AUC maximization (OAM) is that it requires to optimize the pairwise loss between two examples from distinct classes. They presented an effective online algorithm based on the reservoir sampling [31].

By employing the covering number approach, Wang et al. [32] investigated the regret bounds and generalization performance of online learning algorithms with pairwise loss functions in the linear setting where the algorithm is similar to ours. The authors showed the data-dependent bounds for the average risk of the sequences of hypotheses generated from an arbitrary online learner. Such results can be regarded as an extension of online to batch conversion bounds [7] for learning algorithms associated with univariate loss functions. More recently, Kar et al. [16] considered the generalization bounds for the online learning algorithms with pairwise loss functions. The authors improved the results in [32] by using the Rademacher complexity techniques.

Now we compare our result with that of the average iterates in [16] where the generalization ability of the online learning algorithms with pairwise loss functions is investigated. By careful checking the proof of Theorem 5 in [16], we can restate the result there as $\frac{1}{T}\sum_{t=1}^{T} L(h_t) \leq L(h^*) + \frac{\mathcal{R}_T}{T} + C_d\mathcal{O}\big(\frac{\sqrt{\mathcal{B}_T \log T \log(T/\delta)}}{T\lambda}\big)$, where $\mathcal{R}_T$ denotes the regret bound and $\mathcal{B}_T = \max\{\mathcal{R}_T, 2C_d \log T \log(T/\delta)\}$. Note that, in the original result stated in [16, Theorem 5], the strongly convex parameter λ is absorbed in the $\mathcal{O}(\cdot)$ notation. By the properties of the strong convexity, the above result means that $\|\frac{h_1+\cdots+h_T}{T} - h^*\|^2 \leq \frac{\mathcal{R}_T}{T\lambda} + \mathcal{O}\big(\frac{\sqrt{\mathcal{B}_T \log T \log(T/\delta)}}{T\lambda^2}\big)$. We can see our result stated in Theorem 3 is comparable to the ones appeared in the literature which means, in theory, the performance of the last iterate of online pairwise learning algorithm is competitive to that of the average of iterates.

Our work is mainly motivated by [16, 32]. The main novelty of this paper is that we established, for the first time, the convergence rate for the individual iterate of online pairwise learning algorithms. The previous literature [14,30] focused on the average of its iterates, i.e. $\frac{1}{t+1}\sum_{j=1}^{t+1} f_j$. One can directly derive the convergence rate for the average of the iterates from those of the individual iterates. Indeed, if we have, for some $\vartheta \in (0, 1]$, $\mathbb{E}\big[\|f_{t+1} - f_\lambda\|_K^2\big] = \mathcal{O}(t^{-\vartheta})$ for any t, then $\mathbb{E}\big[\|\bar{f}_{t+1} - f_\lambda\|_K^2\big] \leq \frac{1}{t+1}\sum_{j=1}^{t+1}\mathbb{E}\big[\|f_j - f_\lambda\|_K^2\big] = \mathcal{O}(t^{-\vartheta})$, for $\vartheta \in (0, 1)$ and $\mathbb{E}\big[\|\bar{f}_{t+1} - f_\lambda\|_K^2\big] = \mathcal{O}(t^{-1}\log t)$, for $\vartheta = 1$. In this sense, the previous results can be regarded as corollaries of our new results.

We end this section with some remarks on the applicability of algorithm (2). The implementation of ORPL algorithm (2) requires, at iteration t, to store the previous examples $\{z_1, z_2, \ldots, z_{t-1}\}$. Consequently, the memory (space) complexity is very high. In order to improve the applicability of ORPL, one intriguing approach is to develop a memory-efficient implementation which, instead of keeping all previous $t-1$ examples at iteration t, stores only a buffer set of a limited size as employed in [16, 33] using, e.g., reservoir sampling techniques [31]. In this case, ORPL would work with finite buffers associated with the local error $\mathcal{L}_t(f) = \frac{1}{|B_t|}\sum_{j\in B_t}\ell(f(x_t, x_j), r(y_t, y_j)) + \frac{\lambda}{2}\|f\|_K^2$, where B_t is the state of buffer at iteration t. Notice that each iterate function f_{t+1} can be represented as $f_{t+1} = \sum_{j=1}^{t}\alpha_j^t K_{(x_t,x_j)}$. Therefore, another direction to improve the applicability of algorithm (2) is to design online learning algorithms for pairwise learning, which would encourage the predictor f_{t+1} to have sparse support vectors (i.e. with a large

proportion of zero coefficients α_j^t). Such algorithms should be able to reduce the computational time at both training and testing stages.

3 Technical results and proofs

This section proves our main results.

3.1 General technical results

For any $\lambda > 0$, we can establish the uniform bound for the learning sequence $\{f_t : t \in \mathbb{N}\}$ as follows. Our main aim is to consider the generalization performance of f_t which is expected to converge to the true regularization function f_λ.

We can establish the following lemma.

Lemma 3.1 For any $\lambda > 0$ and $t \in \mathbb{N}$, if the step size satisfies $\eta_t \lambda \leq 1$ for $t \geq 2$, then we have

$$\|f_t\|_K \leq \frac{\kappa M}{\lambda}, \quad \forall t \in \mathbb{N}. \tag{9}$$

Proof We prove the claim (9) by induction. The initial functions f_1 and f_2 certainly satisfy inequality (9). Observe that $|\ell'(f_t(x_t, x_j), r(y_t, y_j))| \leq M$, consequently,

$$\begin{aligned}\|f_{t+1}\|_K &= \|\big(1-\eta_t\lambda\big)f_t - \tfrac{\eta_t}{t-1}\textstyle\sum_{j=1}^{t-1}\ell'(f_t(x_t,x_j),r(y_t,y_j))K_{(x_t,x_j)}\|_K \\ &\leq (1-\eta_t\lambda)\|f_t\|_K + \eta_t M\kappa.\end{aligned}$$

Putting induction assumption $\|f_t\|_K \leq \frac{\kappa M}{\lambda}$ into the above inequality yields the desired estimation for $\|f_{t+1}\|_K \leq (1-\eta_t\lambda)\frac{\kappa M}{\lambda} + \eta_t M\kappa = \frac{\kappa M}{\lambda}$. This completes the proof of the lemma. □

To prove the main convergence results, we need to introduce the concept of Rademacher complexity [4] which is defined as follows.

Definition 2 Let F be a class of uniformly bounded functions. For every integer n, we call

$$R_n(F) := \mathbb{E}_{\mathbf{z}}\mathbb{E}_\varepsilon\left[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(z_i)\right],$$

the Rademacher average over F, where $\mathbf{z} = \{z_i\}_{i=1}^n$ are independent random variables distributed according to some probability measure and $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ are independent Rademacher random variables, i.e. $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = \frac{1}{2}$.

The Rademacher complexity has the following useful property (see e.g. [20]) which will be used in the later proof. This property is a refined version of the well-known contraction inequality due to Ledoux and Talagrand [18, Corollary 3.17].

Lemma 3.2 Let $\{g_j(\theta)\}$ and $\{h_j(\theta)\}$ be sets of functions on Θ. If for each j, θ, θ' that $|g_j(\theta) - g_j(\theta')| \le |h_j(\theta) - h_j(\theta')|$, then

$$\mathbb{E}_\epsilon \left[\sup_{\theta\in\Theta} \sum_{j=1}^{m} \epsilon_j g_j(\theta) \right] \le \mathbb{E}_\epsilon \left[\sup_{\theta\in\Theta} \sum_{j=1}^{m} \epsilon_j h_j(\theta) \right]. \tag{10}$$

Now we define

$$\widehat{\mathcal{E}}^t(f) = \frac{1}{t-1} \sum_{j=1}^{t-1} \ell(f(x_t, x_j), r(y_t, y_j)),$$

and

$$\widetilde{\mathcal{E}}^t(f) = \frac{1}{t-1} \sum_{j=1}^{t-1} \int_{\mathcal{Z}} \ell(f(x, x_j), r(y, y_j)) d\rho(x, y).$$

Furthermore, let $\widetilde{\mathcal{E}}^t_\lambda(f) = \widetilde{\mathcal{E}}^t(f) + \frac{\lambda}{2}\|f\|_K^2$. With the above notations, we can get the following recursive inequality which is very critical to prove the convergence of the ORPL algorithms.

Theorem 5 *Assume $\eta_t\lambda \le 1$ for any $t \ge 2$, then we have*

$$\mathbb{E}\left[\|f_{t+1} - f_\lambda\|_K^2\right] \le (1 - \eta_t\lambda)\, \mathbb{E}\left[\|f_t - f_\lambda\|_K^2\right] + 4M^2\eta_t^2\kappa^2 + 2\eta_t \left(\frac{2560e(\kappa M)^2}{\lambda t}\right). \tag{11}$$

To prove Theorem 5, we need the following technical lemma which is mainly motivated by the peeling and re-weighting techniques [3, 28] for Rademacher averages.

Lemma 3.3 Let $t \ge 2$, then for any $0 < \tau < 1$ there holds

$$\mathbb{E}\left[\widetilde{\mathcal{E}}^t_\lambda(f_\lambda) - \widetilde{\mathcal{E}}^t_\lambda(f_t)\right] \le \tau\, \mathbb{E}\left[\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f_t)\right] + \frac{1280e\kappa^2M^2}{(1-\tau)t\lambda}. \tag{12}$$

Proof Let $\mathcal{F}_\lambda = \{f \in \mathcal{H}_K : \|f\| \le \kappa M/\lambda\}$ which can be further written as $\mathcal{F}_\lambda = \bigcup_{i=1}^{\infty} \mathcal{F}(4^i b)$, where, for any $i \ge 1$,

$$\mathcal{F}(4^i b) = \left\{f \in \mathcal{F}_\lambda : 4^{i-1}b \le b + \mathcal{E}_\lambda(f) - \mathcal{E}_\lambda(f_\lambda) < 4^i b\right\}.$$

Here, $b > 0$ is a constant to be determined later. Define, for any $i \ge 0$,

$$\mathcal{R}_i = \sup_{f\in\mathcal{F}(4^i b)} \left[\mathcal{E}(f) - \mathcal{E}(f_\lambda) - \widetilde{\mathcal{E}}^t(f) + \widetilde{\mathcal{E}}^t(f_\lambda)\right].$$

To prove the desired result in the lemma, we start to estimate the term $\frac{\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f_t)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)}{b+\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)}$. Indeed, we know from (9) that $f_t \in \mathcal{F}_\lambda$. Therefore,

$$\begin{aligned}
\frac{\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f_t)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)}{b+\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)} &\le \sup_{f\in\mathcal{F}_\lambda}\left[\frac{\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)}{b+\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)}\right] \\
&= \sup_i \sup_{f\in\mathcal{F}(4^i b)}\left[\frac{\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)}{b+\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)}\right] \\
&\le \sum_{i=1}^{\infty}\sup_{f\in\mathcal{F}(4^i b)}\left[\frac{\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)}{b+\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)}\right] \\
&\le b^{-1}\sum_{i=1}^{\infty}4^{-(i-1)}\sup_{f\in\mathcal{F}(4^i b)}\left[\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)\right] \\
&= b^{-1}\sum_{i=1}^{\infty}4^{-(i-1)}\mathcal{R}_i. \qquad (13)
\end{aligned}$$

It remains to estimate $\mathcal{R}_i$. To this end, note that $\mathcal{R}_i$ is a function of $\{z_1, z_2, \ldots, z_{t-1}\}$, i.e. $\mathcal{R}_i = \mathcal{R}_i(z_1, z_2, \ldots, z_{t-1})$. In addition, observe from the strong convexity of $\mathcal{E}_\lambda[\cdot]$ and the definition of f_λ that $\lambda\|f - f_\lambda\|_K^2 \le \mathcal{E}_\lambda(f) - \mathcal{E}_\lambda(f_\lambda)$. Consequently,

$$\|f - f_\lambda\|_K \le 2^i\sqrt{\frac{b}{\lambda}}, \qquad \forall f \in \mathcal{F}(4^i b). \qquad (14)$$

Now, for any z_j being replaced by z'_j, observe, for $f \in \mathcal{F}(4^i b)$, that

$$|\mathcal{R}_i(z_1, \ldots, z'_j, \ldots, z_{t-1}) - \mathcal{R}_i(z_1, \ldots, z_j, \ldots, z_{t-1})| \le \frac{2\kappa M}{t-1}\|f - f_\lambda\|_K \le \frac{2^{i+1}\kappa M}{t-1}\sqrt{\frac{b}{\lambda}},$$

where the last inequality used the estimation (14). Therefore, we know from the McDiarmid inequality that, with probability $1 - 2^{-i}\delta$, there holds

$$\mathcal{R}_i - \mathbb{E}[\mathcal{R}_i] \le 2^{i+1}\kappa M\sqrt{\frac{2b\log(\frac{2^i}{\delta})}{t\lambda}}.$$

Now we move on to the estimation of $\mathbb{E}[\mathcal{R}_i]$ using the standard symmetrization trick (e.g. [3]). To this end, for any $z = (x, y)$, let $L_f(z) = \mathbb{E}_{\tilde{z}}\ell(f(\tilde{x}, x), r(\tilde{y}, y)) - \mathbb{E}_{\tilde{z}}\ell(f_\lambda(\tilde{x}, x), r(\tilde{y}, y))$.

Let $\mathbf{z}'_t = \{z'_j = (x'_j, y'_j) : j = 1, 2, \ldots, t\}$ be the i.i.d. copy of $\mathbf{z}_t = \{z_1, \ldots, z_t\}$, then

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_i] &= \mathbb{E}_{\mathbf{z}_t}\left[\sup_{f\in\mathcal{F}(4^i b)}\left[\mathcal{E}(f) - \mathcal{E}(f_\lambda) - \frac{1}{t-1}\sum_{j=1}^{t-1}L_f(x_j, y_j)\right]\right] \\
&= \mathbb{E}_{\mathbf{z}_t}\left[\sup_{f\in\mathcal{F}(4^i b)}\left[\mathbb{E}_{\mathbf{z}'_t}\left[\frac{1}{t-1}\sum_{j=1}^{t-1}L_f(x'_j, y'_j)\right] - \frac{1}{t-1}\sum_{j=1}^{t-1}L_f(x_j, y_j)\right]\right] \\
&\le \mathbb{E}_{\mathbf{z}_t}\mathbb{E}_{\mathbf{z}'_t}\left[\sup_{f\in\mathcal{F}(4^i b)}\left[\frac{1}{t-1}\sum_{j=1}^{t-1}L_f(x'_j, y'_j) - \frac{1}{t-1}\sum_{j=1}^{t-1}L_f(x_j, y_j)\right]\right].
\end{aligned}$$

Hence, for any $\varepsilon_j \in \{1, -1\}$ with $j = 1, \cdots, t-1$, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}_t}\mathbb{E}_{\mathbf{z}'_t}\left[\sup_{f\in\mathcal{F}(4^i b)}\Big[\frac{1}{t-1}\sum_{j=1}^{t-1}L_f(x'_j, y'_j) - \frac{1}{t-1}\sum_{j=1}^{t-1}L_f(x_j, y_j)\Big]\right] \\
&= \mathbb{E}_{\mathbf{z}_t}\mathbb{E}_{\mathbf{z}'_t}\left[\sup_{f\in\mathcal{F}(4^i b)}\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j\big[L_f(x'_j, y'_j) - L_f(x_j, y_j)\big]\right] \\
&\le 2\mathbb{E}_{\mathbf{z}_t}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}(4^i b)}\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j L_f(x_j, y_j)\right] \\
&\le 2\mathbb{E}_{\mathbf{z}_t}\mathbb{E}_{\tilde{z}}\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}(4^i b)}\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j\big[\ell(f(\tilde{x}, x_j), r(\tilde{y}, y_j)) - \ell(f_\lambda(\tilde{x}, x_j), r(\tilde{y}, y_j))\big].
\end{aligned}$$

Using the property of Rademacher averages stated in Lemma 3.2 with $\theta = f$, $g_j(\theta) = \ell(f(\tilde{x}, x_j), r(\tilde{y}, y_j)) - \ell(f_\lambda(\tilde{x}, x_j), r(\tilde{y}, y_j))$, and $h_j(\theta) = M[f(\tilde{x}, x_j) - f_\lambda(\tilde{x}, x_j)]$ implies that

$$\begin{aligned}
&\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}(4^i b)}\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j\big[\ell(f(\tilde{x}, x_j), r(\tilde{y}, y_j)) - \ell(f_\lambda(\tilde{x}, x_j), r(\tilde{y}, y_j))\big] \\
&\le M\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}(4^i b)}\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j\big[f(\tilde{x}, x_j) - f_\lambda(\tilde{x}, x_j)\big] \\
&= M\mathbb{E}_{\varepsilon}\Big[\sup_{f\in\mathcal{F}(4^i b)}\langle\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j K_{(\tilde{x}, x_j)}, f - f_\lambda\rangle_K\Big]. \qquad (15)
\end{aligned}$$

Combining the above inequalities with (14) implies that

$$\begin{aligned}
&\mathbb{E}_{\varepsilon}\Big[\sup_{f\in\mathcal{F}(4^i b)}\langle\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j K_{(\tilde{x}, x_j)}, f - f_\lambda\rangle_K\Big] \le \mathbb{E}_{\varepsilon}\Big[\sup_{\|f-f_\lambda\|_K\le 2^i\sqrt{\frac{b}{\lambda}}}\langle\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j K_{(\tilde{x}, x_j)}, f - f_\lambda\rangle_K\Big] \\
&= 2^i\sqrt{\frac{b}{\lambda}}\mathbb{E}_{\varepsilon}\left\|\frac{1}{t-1}\sum_{j=1}^{t-1}\varepsilon_j K_{(\tilde{x}, x_j)}\right\|_K = 2^i\sqrt{\frac{b}{(t-1)^2\lambda}}\mathbb{E}_{\varepsilon}\Big(\sum_{k,j=1}^{t-1}\varepsilon_j\varepsilon_k K((\tilde{x}, x_j), (\tilde{x}, x_k))\Big)^{1/2} \\
&\le 2^i\sqrt{\frac{b}{(t-1)^2\lambda}}\Big(\mathbb{E}_{\varepsilon}\sum_{k,j=1}^{t-1}\varepsilon_j\varepsilon_k K((\tilde{x}, x_j), (\tilde{x}, x_k))\Big)^{1/2} = 2^i\sqrt{\frac{b}{(t-1)^2\lambda}}\Big(\sum_{j=1}^{t-1}K((\tilde{x}, x_j), (\tilde{x}, x_j))\Big)^{1/2} \\
&\le 2^i\kappa\sqrt{\frac{b}{(t-1)\lambda}}.
\end{aligned}$$

Therefore, for any $i \ge 1$, with probability $1 - 2^{-i}\delta$ we have that

$$\mathcal{R}_i \le 2^{i+1}\kappa M\sqrt{\frac{2b}{\lambda t}} + 2^{i+1}\kappa M\sqrt{\frac{2b\log(\frac{2^i}{\delta})}{t\lambda}} \le 2^{i+1}\kappa M\sqrt{\frac{2b}{t\lambda}}\left(1 + \sqrt{\log(\frac{2^i}{\delta})}\right). \qquad (16)$$

Combining the estimation (13) with (16) implies that, with probability $1-\delta$, there holds

$$\begin{aligned}\frac{\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f_t)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)}{b+\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)} &\le \sup_{f\in\mathcal{F}_\lambda}\left[\frac{\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)}{b+\mathcal{E}_\lambda(f)-\mathcal{E}_\lambda(f_\lambda)}\right]\\ &\le 8b^{-1}\sum_{i=1}^{\infty}2^{-i}\kappa M\sqrt{\frac{2b}{t\lambda}}\left(1+\sqrt{\log(\frac{2^i}{\delta})}\right)\\ &\le 8\kappa M\sqrt{\frac{2}{\lambda bt}}\left(1+\sum_{i=1}^{\infty}2^{-i}\sqrt{i+\log\frac{1}{\delta}}\right)\\ &\le 8\kappa M\sqrt{\frac{2}{\lambda bt}}\left(1+\sqrt{\log\frac{1}{\delta}}+\sum_{i=1}^{\infty}2^{-i}\sqrt{i}\right)\\ &\le 8\kappa M\sqrt{\frac{2}{\lambda bt}}\left(3+\sqrt{\log\frac{1}{\delta}}\right), \qquad (17)\end{aligned}$$

where, in the last inequality, we have used the fact that

$$\begin{aligned}&\sum_{i=1}^{\infty}2^{-i}\sqrt{i}\le\frac{1}{2}+\frac{\sqrt{2}}{4}+\frac{1}{2}\sum_{i=3}^{\infty}2^{-i}(i+1)=\frac{1}{2}+\frac{\sqrt{2}}{4}+\frac{1}{8}+\frac{1}{2}\sum_{i=3}^{\infty}2^{-i}i\\ &\le\frac{\sqrt{2}}{4}+\frac{5}{8}+\frac{1}{2}\int_2^{\infty}s2^{-s}ds=\frac{\sqrt{2}}{4}+\frac{5}{8}+\frac{1}{8}\left(\frac{1}{(\log 2)^2}+\frac{2}{\log 2}\right)\le 2.\end{aligned}$$

This implies that

$$\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f_t)+\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)\le 8\kappa M\sqrt{\frac{2}{\lambda bt}}\left(3+\sqrt{\log\frac{1}{\delta}}\right)[b+\mathcal{E}_\lambda(f_t)-\mathcal{E}_\lambda(f_\lambda)]. \qquad (18)$$

For any $0<\tau<1$, selecting $b=\left[\frac{8\kappa M}{(1-\tau)}\sqrt{\frac{2}{\lambda t}}\right]^2\left(3+\sqrt{\log\frac{1}{\delta}}\right)^2$, substituting it back into (18) and arranging terms, we obtain, with probability $1-\delta$, that

$$\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f_t)\le\tau(\mathcal{E}_\lambda(f_\lambda)-\mathcal{E}_\lambda(f_t))+\left(\frac{128\kappa^2M^2}{1-\tau}\right)\frac{\left(3+\sqrt{\log\frac{1}{\delta}}\right)^2}{t\lambda}. \qquad (19)$$

Denote the random variable $\xi=\widetilde{\mathcal{E}}^t_\lambda(f_\lambda)-\widetilde{\mathcal{E}}^t_\lambda(f_t)-\tau(\mathcal{E}_\lambda(f_\lambda)-\mathcal{E}_\lambda(f_t))$. The above estimation implies, with probability at least $1-\delta$, that

$$\xi\le\left(\frac{128\kappa^2M^2}{1-\tau}\right)\frac{\left(3+\sqrt{\log\frac{1}{\delta}}\right)^2}{t\lambda}\le\left(\frac{128\kappa^2M^2}{1-\tau}\right)\frac{10}{t\lambda}\log\frac{e}{\delta}.$$

The above inequality means $\text{Prob}[\xi > u] \le \exp\left\{1 - \frac{u}{\frac{1280\kappa^2M^2}{(1-\tau)t\lambda}}\right\}$ for any $u > 0$. Therefore,

$$\mathbb{E}[\xi] = \int_0^\infty \text{Prob}[\xi > u]du - \int_{-\infty}^0 \text{Prob}[\xi < u]du \le \int_0^\infty \text{Prob}[\xi > u]du$$
$$\le \int_0^\infty \exp\left\{1 - \frac{(1-\tau)t\lambda u}{1280\kappa^2M^2}\right\}du \le \frac{1280e\kappa^2M^2}{(1-\tau)t\lambda}.$$

This completes the proof of the desired result. □

Remark 3.1 Indeed, from (17) we can have, with probability $1-\delta$, that

$$\sup_{f\in\mathcal{F}_\lambda}\left[\frac{\mathcal{E}_\lambda(f) - \mathcal{E}_\lambda(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f) + \widetilde{\mathcal{E}}_\lambda^t(f_\lambda)}{b + \mathcal{E}_\lambda(f) - \mathcal{E}_\lambda(f_\lambda)}\right] \le 8\kappa M\sqrt{\frac{2}{\lambda bt}}\Big(3 + \sqrt{\log\frac{1}{\delta}}\Big).$$

Selecting $b = \left[\frac{8\kappa M}{(1-\tau)}\sqrt{\frac{2}{\lambda t}}\right]^2\Big(3 + \sqrt{\log\frac{1}{\delta}}\Big)^2$ and arranging terms in the above estimation, we obtain, with probability $1-\delta$, that

$$\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f) \le \tau(\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f)) + \left(\frac{128\kappa^2M^2}{1-\tau}\right)\frac{\Big(3 + \sqrt{\log\frac{1}{\delta}}\Big)^2}{t\lambda}, \qquad \forall f \in \mathcal{F}_\lambda. \tag{20}$$

Now we are ready to prove Theorem 5 using Lemma 3.3. In the following, we use the notation $\mathbb{E}[\xi|z_1, \ldots, z_t]$ to denote the conditional expectation of the random variable ξ conditioned on $\{z_1, \ldots, z_t\}$.

Proof of Theorem 5: Let $\widehat{\mathcal{A}}_\lambda^t(f_t) = \frac{1}{t-1}\sum_{j=1}^{t-1}\ell'(f_t(x_t, x_j), r(y_t, y_j))K_{(x_t,x_j)} + \lambda f_t$. By the definition of $f_{t+1} = f_t - \eta_t\widehat{\mathcal{A}}_\lambda^t(f_t)$, we have

$$\mathbb{E}[\|f_{t+1} - f_\lambda\|_K^2] = \mathbb{E}[\|f_t - f_\lambda\|_K^2] + \eta_t^2\mathbb{E}[\|\widehat{\mathcal{A}}_\lambda^t(f_t)\|_K^2] + 2\eta_t\mathbb{E}[\langle f_\lambda - f_t, \widehat{\mathcal{A}}_\lambda^t(f_t)\rangle_K]. \tag{21}$$

By Lemma 3.1 and the definition of M, we have

$$\|\widehat{\mathcal{A}}_\lambda^t(f_t)\|_K \le \kappa M + \lambda\|f_t\|_K \le 2\kappa M. \tag{22}$$

Now we estimate the third term on the righthand side of equation (21). The reproducing property and convexity of $\ell(\cdot, r(y, y'))$ imply that the term $\langle f_\lambda - f_t, \widehat{\mathcal{A}}_\lambda^t(f_t)\rangle_K$ can be bounded by

$$\frac{1}{t-1}\sum_{j=1}^{t-1}\ell'(f_t(x_t, x_j), r(y_t, y_j))(f_{\lambda(x_t,x_j)} - f_t(x_t, x_j)) + \lambda\langle f_\lambda, f_t\rangle_K - \lambda\|f_t\|_K^2$$
$$\le \frac{1}{t-1}\sum_{j=1}^{t-1}\ell'(f_t(x_t, x_j), r(y_t, y_j))(f_{\lambda(x_t,x_j)} - f_t(x_t, x_j)) + \frac{\lambda}{2}(\|f_\lambda\|_K^2 + \|f_t\|_K^2) - \lambda\|f_t\|_K^2$$
$$\le \widehat{\mathcal{E}}_\lambda^t(f_\lambda) - \widehat{\mathcal{E}}_\lambda^t(f_t).$$

Since f_t only depends on the examples $\mathbf{z}_{t-1} = \{z_1, \ldots, z_{t-1}\}$, we have $\mathbb{E}\big[\widehat{\mathcal{E}}_\lambda^t(f_t)|z_1, \ldots, z_{t-1}\big] = \widetilde{\mathcal{E}}_\lambda^t(f_t)$. Hence,

$$\mathbb{E}\left[\widehat{\mathcal{E}}_\lambda^t(f_\lambda) - \widehat{\mathcal{E}}_\lambda^t(f_t)\right] = \mathbb{E}\Big[\mathbb{E}\big[\widehat{\mathcal{E}}_\lambda^t(f_\lambda) - \widehat{\mathcal{E}}_\lambda^t(f_t)|z_1, \ldots, z_{t-1}\big]\Big] = \mathbb{E}\big[\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t)\big] \quad (23)$$

Applying Lemma 3.3 with $\tau = 1/2$ implies that

$$\mathbb{E}\left[\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t)\right] \le \frac{1}{2}\mathbb{E}\left[\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f_t)\right] + \frac{2560e(\kappa M)^2}{\lambda t}.$$

By $\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f_t) \le -\lambda\|f_t - f_\lambda\|_K^2$, the above inequality implies that

$$\mathbb{E}\left[\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t)\right] \le -\frac{\lambda}{2}\mathbb{E}\big[\|f_t - f_\lambda\|_K^2\big] + \frac{2560e(\kappa M)^2}{\lambda t}.$$

Substituting the above estimation and inequality (22) back into (21) yields the desired result. This completes the proof of the theorem. □

The proof of the recursive inequality (11) in Theorem 5 critically depends on the estimation of the term $\mathbb{E}\big[\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t)\big] - \mathbb{E}\big[\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f_t)\big]$. This term is very similar to the well-known *sample error* in the standard framework of learning theory [11, 29]. One could use the standard Rademacher complexity approach [4] to directly estimate this critical term. Indeed, observe that $\mathbb{E}\big[\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t)\big] - \mathbb{E}\big[\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f_t)\big] = \mathbb{E}\big[\widetilde{\mathcal{E}}^t(f_\lambda) - \widetilde{\mathcal{E}}^t(f_t)\big] - \mathbb{E}\big[\mathcal{E}(f_\lambda) - \mathcal{E}(f_t)\big] = \mathbb{E}\big[\mathcal{E}(f_t) - \widetilde{\mathcal{E}}^t(f_t)\big]$. Then, one can apply the standard symmetrization technique [3] and get the following estimation:

$$\begin{aligned}
\mathbb{E}\big[\mathcal{E}(f_t) - \widetilde{\mathcal{E}}^t(f_t)\big] &\le \mathbb{E}\sup_{\|f\|_K \le \frac{\kappa}{\lambda}} \mathbb{E}\big[\mathcal{E}(f) - \widetilde{\mathcal{E}}^t(f)\big] \\
&\le 2\mathbb{E}\mathbb{E}_\varepsilon \sup_{\|f\|_K \le \frac{\kappa}{\lambda}} \frac{1}{t-1}\sum_{j=1}^{t-1} \varepsilon_j\big[\ell(f(\tilde{x}, x_j), r(\tilde{y}, y_j))\big] \\
&\le 2M\mathbb{E}\mathbb{E}_\varepsilon \sup_{\|f\|_K \le \frac{\kappa}{\lambda}} \frac{1}{t-1}\sum_{j=1}^{t-1} \varepsilon_j f(\tilde{x}, x_j) \\
&= 2M\mathbb{E}\mathbb{E}_\varepsilon\Big[\sup_{\|f\|_K \le \frac{\kappa}{\lambda}} \langle \frac{1}{t-1}\sum_{j=1}^{t-1} \varepsilon_j K_{(\tilde{x}, x_j)}, f\rangle_K\Big] \\
&\le \frac{2M^2\kappa}{t-1}\sqrt{\frac{1}{\lambda}}\,\mathbb{E}\left(\sum_{j=1}^{t-1} K((\tilde{x}, x_j), (\tilde{x}, x_j))\right)^{1/2} \\
&\le 2M^2\kappa^2\sqrt{\frac{1}{(t-1)\lambda}} \le M^2\kappa^2\sqrt{\frac{2}{t\lambda}}.
\end{aligned}$$

This estimation together the above observation means that

$$\mathbb{E}\big[\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t)\big] \le \mathbb{E}\big[\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f_t)\big] + M^2\kappa^2\sqrt{\frac{2}{t\lambda}}.$$

Using the above inequality, instead of the more refined estimation (12) in the proof of Theorem 5, we can have

$$\mathbb{E}\big[\|f_{t+1} - f_\lambda\|_K^2\big] \le \big(1-\eta_t\lambda\big)\mathbb{E}\big[\|f_t - f_\lambda\|_K^2\big] + 4M^2\eta_t^2\kappa^2 + \frac{2\sqrt{2}M^2\kappa^2\eta_t}{\sqrt{t\lambda}}. \quad (24)$$

We can see the inequality (11) is much better than inequality (24), since the last term on the righthand side of (24) depends on $\frac{1}{\sqrt{t}}$ while the counterpart in (11) only depends on $\frac{1}{t}$. This improvement comes from the Lemma 3.3 which is obtained by applying peeling and re-weighting techniques [3, 28] to the strongly convex objective function $\mathcal{E}_\lambda(f) - \mathcal{E}_\lambda(f_\lambda)$.

3.2 Proofs of main results

In this subsection, we prove the main results, i.e. Theorems 1, 2 and 3. First of all, by induction we can easily get from the recursive inequality (11) that, for any t_0, $T \in \mathbb{N}$ and $T > t_0$, there holds

$$\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big] \le \prod_{t=t_0}^{T}\big(1-\eta_t\lambda\big)\mathbb{E}\big[\|f_{t_0} - f_\lambda\|_K^2\big] + 4M^2\kappa^2\sum_{t=t_0}^{T}\eta_t^2\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big)$$
$$+ 5120eM^2\kappa^2\sum_{t=t_0}^{T}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big), \quad (25)$$

where the conventional notation $\prod_{j=T+1}^{T}\big(1-\frac{\eta_j\lambda}{2}\big) = 1$ is used.

Proof of Theorem 1 : Our proof mainly follows from [37]. By the assumption (6), there exists some $t_0 \in \mathbb{N}$ such that $\eta_t\lambda \le \frac{1}{2}$ holds for each $t \ge t_0$. Fixing such t_0, we estimate three terms on the the right hand side of (25) step by step.

Estimation of term 1: We first estimate the first term on the righthand side of (25), i.e. $\prod_{t=t_0}^{T}\big(1-\eta_t\lambda\big)\mathbb{E}\big[\|f_{t_0} - f_\lambda\|_K^2\big]$. Since $\sum_{t=t_0}^{\infty}\eta_t = \infty$ and $0 < \eta_t\lambda < 1$ for any $t \ge t_0$, we have that $\prod_{t=t_0}^{T}(1-\eta_t\lambda) \le \exp\big(-\sum_{t=t_0}^{T}\eta_t\lambda\big) \to 0$ as $T \to \infty$. Hence, for any $\epsilon > 0$, there exists some $T_1 \in \mathbb{N}$ such that, for any $T > T_1$,

$$\prod_{t=t_0}^{T}\big(1-\eta_t\lambda\big)\mathbb{E}\big[\|f_{t_0} - f_\lambda\|_K^2\big] \le \varepsilon.$$

Estimation of term 2: Now we are in a position to estimate the second term on the righthand side of (25). By the assumption $\lim_{t\to\infty}\eta_t = 0$, there exists some integer $t_\epsilon > t_0$ such that $\eta_t \le \frac{\lambda\epsilon}{2}$ for each $t > t_\epsilon$.

Then we divide $\sum_{t=t_0}^{T}\eta_t^2\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)$ into the following two parts, i.e.

$$\sum_{t=t_0}^{T}\eta_t^2\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)=\underbrace{\sum_{t=t_0}^{t_\epsilon}\eta_t^2\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)}_{I_1}+\underbrace{\sum_{t=t_\epsilon+1}^{T}\eta_t^2\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)}_{I_2}.$$

Fixing t_ϵ, we can find some $T_2\in\mathbb{N}$ such that for every $T\geq T_2$, there holds $\sum_{j=t_\epsilon+1}^{T}\eta_j\geq\sum_{j=t_\epsilon+1}^{T_2}\eta_j\geq\frac{1}{\lambda}\log\frac{t_\epsilon}{2\lambda^2\epsilon}$. Consequently, for $t_0\leq t\leq t_\epsilon$, we have $\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)\leq\exp\{-\sum_{j=t+1}^{T}\eta_j\lambda\}\leq\exp\{-\sum_{j=t_\epsilon+1}^{T}\eta_j\lambda\}\leq\frac{2\lambda^2\epsilon}{t_\epsilon}$. Putting this estimation and the fact that $\eta_t\lambda\leq 1/2$ for $t\geq t_0$ together yields the following bound for I_1

$$I_1=\sum_{t=t_0}^{t_\epsilon}\eta_t^2\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)\leq\frac{2\lambda^2\epsilon}{t_\epsilon}\sum_{t=t_0}^{t_\epsilon}\eta_t^2\leq\frac{2\lambda^2\epsilon}{t_\epsilon}\times t_\epsilon\times\frac{1}{4\lambda^2}=\frac{\epsilon}{2}.$$

For I_2, we can get

$$\begin{aligned}I_2&=\sum_{t=t_\epsilon+1}^{T}\eta_t^2\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)\leq\sum_{t=t_\epsilon+1}^{T}\eta_t\frac{\lambda\epsilon}{2}\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)\\&=\frac{\epsilon}{2}\sum_{t=t_\epsilon+1}^{T}\left(\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)-\prod_{j=t}^{T}\left(1-\eta_j\lambda\right)\right)\\&=\frac{\epsilon}{2}\left(1-\prod_{j=t_\epsilon+1}^{T}\left(1-\eta_j\lambda\right)\right)\leq\frac{\epsilon}{2},\end{aligned}$$

where the first inequality holds due to the fact that $\eta_t\leq\frac{\lambda\epsilon}{2}$ for each $t\geq t_\epsilon$.

Estimation of term 3: The estimation of the third term on the righthand side of (25) is similar to that of the second term. To see this, observe that there exists $t_\epsilon'\in\mathbb{N}$ such that $\frac{1}{t}\leq\frac{\lambda^2\epsilon}{2}$ when $t>t_\epsilon'$, rewrite the term $\sum_{t=t_0}^{T}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)$ as

$$\underbrace{\sum_{t=t_0}^{t_\epsilon'}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)}_{I_3}+\underbrace{\sum_{t=t_\epsilon'+1}^{T}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\left(1-\eta_j\lambda\right)}_{I_4}.$$

By the assumption that $\sum_{t=2}^{\infty}\eta_t = \infty$, there exists some $T_3 \in \mathbb{N}$ such that there holds that $\sum_{t=t'_\epsilon+1}^{T}\eta_t \geq \sum_{t=t'_\epsilon+1}^{T_3}\eta_t \geq \frac{1}{\lambda}\log\frac{t'_\epsilon}{\lambda^2 t_0\epsilon}$. That means $\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big) \leq \exp\{-\sum_{j=t+1}^{T}\eta_j\lambda\} \leq \exp\{-\sum_{j=t'_\epsilon+1}^{T}\eta_j\lambda\} \leq \frac{\lambda^2 t_0\epsilon}{t'_\epsilon}$ holds for $t_0 \leq t \leq t'_\epsilon$ and $T \geq T_3$. Consequently, I_3 can be bounded as

$$I_3 = \sum_{t=t_0}^{t'_\epsilon}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big) \leq \frac{\lambda^2 t_0\epsilon}{t'_\epsilon}\sum_{t=t_0}^{t'_\epsilon}\frac{\eta_t}{\lambda t} \leq \frac{\lambda^2 t_0\epsilon}{t'_\epsilon}\times t'_\epsilon \times \frac{1}{2\lambda^2 t_0} = \frac{\epsilon}{2}.$$

For I_4, since $\frac{1}{t} \leq \frac{\lambda^2\epsilon}{2}$ when $t > t'_\epsilon$, there holds

$$\begin{aligned} I_4 = \sum_{t=t'_\epsilon+1}^{T}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big) &\leq \frac{\epsilon}{2}\sum_{t=t'_\epsilon+1}^{T}\eta_t\lambda\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big) \\ &= \frac{\epsilon}{2}\sum_{t=t'_\epsilon+1}^{T}\left[\prod_{j=t+1}^{T}(1-\eta_j\lambda) - \prod_{j=t}^{T}(1-\eta_j\lambda)\right] \\ &= \frac{\epsilon}{2}\left[1-\textstyle\prod_{j=t'_\epsilon+1}^{T}(1-\eta_j\lambda)\right] \leq \frac{\epsilon}{2}. \end{aligned}$$

Combining all the above estimations, for $T \geq \max\{T_1, T_2, T_3\}$, we have that

$$\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big] \leq (1 + 4\kappa^2M^2 + 5120e\kappa^2M^2)\epsilon.$$

Since $\varepsilon > 0$ is arbitrary, this completes the proof of Theorem 1. □

To derive explicit rates, we need the following technical estimations from [37] and [27].

Lemma 3.4 For any $0 < \nu \leq 1$, $t < T$, $0 < \alpha \leq 1$, and $b > 0$, the following estimations hold true.

(i) $\sum_{j=t+1}^{T} j^{-\alpha} \geq \begin{cases} \frac{1}{1-\alpha}[(T+1)^{1-\alpha} - (t+1)^{1-\alpha}], & 0 < \alpha < 1 \\ \log(T+1) - \log(t+1), & \alpha = 1. \end{cases}$

(ii) $\sum_{t=1}^{T-1}\frac{1}{t^{2\alpha}}\exp\left\{-\nu\sum_{j=t+1}^{T} j^{-\alpha}\right\} le \begin{cases} \frac{18}{\nu T^\alpha} + \frac{9T^{1-\alpha}}{(1-\alpha)2^{1-\alpha}}\exp\{-\frac{\nu(1-2^{\alpha-1})}{1-\alpha}(T+1)^{1-\alpha}\}, & \alpha < 1, \\ \frac{8}{1-\nu}(T+1)^{-\nu}, & \alpha = 1. \end{cases}$

(iii) $e^{-\nu x} \leq \big(\frac{b}{\nu e}\big)^b x^{-b}$.

We are now ready to prove Theorem 2.

Proof [Proof of Theorem 2] By letting $t_0 = 2$ in inequality (25) we have that

$$\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big] \le \prod_{t=2}^{T}(1-\eta_t\lambda)\,\|f_\lambda\|_K^2 + 4M^2\kappa^2\sum_{t=2}^{T}\eta_t^2\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big) + 5120e\kappa^2M^2\sum_{t=2}^{T}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big). \tag{26}$$

By the definition of f_λ, we know that $\mathcal{E}(f_\lambda)+\frac{\lambda}{2}\|f_\lambda\|_K^2 \le \mathcal{E}(0)+\frac{\lambda}{2}\|0\|_K^2 = 1$, which implies that $\|f_\lambda\|_K \le \sqrt{\frac{2}{\lambda}}$. Taking $\eta_t = \frac{1}{\lambda t^\alpha}$ with $0<\alpha<1$ and by estimation (i) of Lemma 3.4, we can bound the first part as

$$\begin{aligned}\prod_{t=2}^{T}(1-\eta_t\lambda)\,\|f_\lambda\|_K^2 &\le \exp\left\{-\sum_{t=2}^{T}\eta_t\lambda\right\}\|f_\lambda\|_K^2 = \exp\left\{-\sum_{t=2}^{T}\frac{1}{t^\alpha}\right\}\|f_\lambda\|_K^2\\ &\le \exp\left\{-\frac{(1-(2/3)^{1-\alpha})}{(1-\alpha)}(T+1)^{1-\alpha}\right\}\frac{2}{\lambda}.\end{aligned} \tag{27}$$

Applying (iii) of Lemma 3.4 with $b = \frac{\alpha}{1-\alpha}$ and $\nu = \frac{1-(\frac{2}{3})^{1-\alpha}}{1-\alpha}$ yields that

$$\exp\left\{-\frac{(1-(2/3)^{1-\alpha})}{(1-\alpha)}(T+1)^{1-\alpha}\right\} \le \left[\frac{\alpha}{(1-(2/3)^{1-\alpha})e}\right]^{\frac{\alpha}{1-\alpha}}T^{-\alpha}.$$

Putting this estimation into (27) implies that

$$\prod_{t=2}^{T}(1-\eta_t\lambda)\,\|f_\lambda\|_K^2 \le \left[\frac{\alpha}{(1-(2/3)^{1-\alpha})e}\right]^{\frac{\alpha}{1-\alpha}}\frac{2}{\lambda T^\alpha}. \tag{28}$$

Since $\frac{1}{\lambda t} \le \eta_t = \frac{1}{\lambda t^\alpha}$, the third part can be bounded by the second part, i.e.

$$\sum_{t=2}^{T}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\big(1-\eta_j\lambda\big) \le \sum_{t=2}^{T}\eta_t^2\prod_{j=t+1}^{T}\Big(1-\eta_j\lambda\Big). \tag{29}$$

For the second part, we know from estimation (ii) of Lemma 3.4 that

$$\begin{aligned}&\sum_{t=2}^{T}\eta_t^2\prod_{j=t+1}^{T}\Big(1-\eta_j\lambda\Big) = \sum_{t=2}^{T}\frac{1}{\lambda^2t^{2\alpha}}\prod_{j=t+1}^{T}\Big(1-\frac{1}{j^\alpha}\Big)\\ &\le \frac{1}{\lambda^2}\left(\frac{18}{T^\alpha}+\frac{9T^{1-\alpha}}{(1-\alpha)2^{1-\alpha}}\exp\Big\{-\frac{1-2^{\alpha-1}}{1-\alpha}(T+1)^{1-\alpha}\Big\}+\frac{1}{T^{2\alpha}}\right)\\ &\le \frac{1}{\lambda^2}\left(\frac{19}{T^\alpha}+\frac{9T^{1-\alpha}}{(1-\alpha)2^{1-\alpha}}\exp\Big\{-\frac{1-2^{\alpha-1}}{1-\alpha}(T+1)^{1-\alpha}\Big\}\right).\end{aligned} \tag{30}$$

Applying (iii) of Lemma 3.4 with $b = \frac{1}{1-\alpha}$ and $\nu = \frac{1-2^{\alpha-1}}{1-\alpha}$ implies that

$$T^{1-\alpha}\exp\Big\{-\frac{(1-2^{\alpha-1})}{1-\alpha}(T+1)^{1-\alpha}\Big\} \le \Big(\frac{1}{(1-2^{\alpha-1})e}\Big)^{\frac{1}{1-\alpha}} T^{-\alpha}.$$

Putting the above estimation into (30) implies that

$$\sum_{t=2}^{T}\eta_t^2\prod_{j=t+1}^{T}\Big(1-\eta_j\lambda\Big) \le \Big(19+\frac{9}{(1-\alpha)2^{1-\alpha}}\Big(\frac{1}{(1-2^{\alpha-1})e}\Big)^{\frac{1}{1-\alpha}}\Big)\frac{1}{\lambda^2 T^\alpha}. \quad (31)$$

Putting (28), (29) and (30) into (26) implies that

$$\mathbb{E}\big[\|f_{T+1}-f_\lambda\|_K^2\big] \le C_1\frac{1}{\lambda^2 T^\alpha},$$

where $C_1 = \Big[2\Big(\frac{\alpha}{(1-(2/3)^{1-\alpha})e}\Big)^{\frac{\alpha}{1-\alpha}} + 4(1+1280e)(\kappa M)^2\Big(19+\frac{9}{(1-\alpha)2^{1-\alpha}}\Big(\frac{1}{(1-2^{\alpha-1})e}\Big)^{\frac{1}{1-\alpha}}\Big)\Big]$.

This completes the proof of the theorem. □

Proof of Theorem 3 If we take the step size as $\eta_t = \frac{1}{t\lambda}$ in inequality (26), we have

$$\begin{aligned}
\mathbb{E}\big[\|f_{T+1}-f_\lambda\|_K^2\big] &\le \prod_{t=2}^{T}\Big(1-\eta_t\lambda\Big)\mathbb{E}\big[\|f_\lambda\|_K^2\big] + 4M^2\kappa^2\sum_{t=2}^{T}\eta_t^2\prod_{j=t+1}^{T}\Big(1-\eta_j\lambda\Big)\\
&\quad + 5120eM^2\kappa^2\sum_{t=2}^{T}\frac{\eta_t}{\lambda t}\prod_{j=t+1}^{T}\Big(1-\eta_j\lambda\Big)\\
&\le \prod_{t=2}^{T}\Big(1-\frac{1}{t}\Big)\|f_\lambda\|_K^2 + 4(1+1280e)M^2\kappa^2\sum_{t=2}^{T}\frac{1}{t^2\lambda^2}\prod_{j=t+1}^{T}\Big(1-\frac{1}{j}\Big)\\
&= \frac{1}{T}\|f_\lambda\|_K^2 + \frac{4(1+1280e)M^2\kappa^2}{T\lambda^2}\sum_{t=2}^{T}\frac{1}{t}\\
&\le \frac{2}{T\lambda} + 4(1+1280e)M^2\kappa^2\frac{\log T}{T\lambda^2} \le [2+4(1+1280e)M^2\kappa^2]\frac{\log T}{T\lambda^2}.
\end{aligned}$$

This completes the proof of the theorem by taking $C_2 = [2+4(1+1280e)M^2\kappa^2]$. □

We turn our attention to the proof of Lemma 2.1 and Corollary 4.

Proof of Lemma 2.1 We rewrite the generalization error as

$$\mathcal{E}(f) = \iint (1-yy'f(x,x'))_+ d\rho(x,y)d\rho(x',y') = \iint L(f(x,x'))d\rho_X(x)d\rho_X(x').$$

Here,

$$\begin{aligned}
L(t) &= \iint (1-yy't)_+ d\rho(y|x)d\rho(y'|x')\\
&= (1-t)_+\mathrm{Prob}(yy'=1|x,x') + (1+t)_+\mathrm{Prob}(yy'=-1|x,x')\\
&= (1-t)_+[\eta(x)\eta(x') + (1-\eta(x))(1-\eta(x'))] + (1+t)_+[\eta(x)(1-\eta(x'))\\
&\quad +\eta(x')(1-\eta(x))].
\end{aligned}$$

When $t = f_c(x, x') \in \{-1, +1\}$, one can easily see that $L(f_c(x, x')) = 2\text{Prob}(yy' \neq f_c(x, x')|x, x')$. And from the definition of f_c, we have $L(f_c(x, x')) \leq 2\text{Prob}(yy' = s|x, x')$ for any $s \in \{-1, +1\}$.

Case 1: If $t \geq 1$, we have $(1 - t)_+ = 0$ and

$$\begin{aligned} L(t) &= (1 + t)[\eta(x)(1 - \eta(x')) + \eta(x')(1 - \eta(x))] \\ &\geq 2[\eta(x)(1 - \eta(x')) + \eta(x')(1 - \eta(x))] = 2\text{Prob}(yy' = -1|x, x') \geq L(f_c(x, x')). \end{aligned}$$

Case 2: If $t \leq -1$, we have $(1 + t)_+ = 0$ and

$$\begin{aligned} L(t) = (1 - t)[\eta(x)\eta(x') + (1 - \eta(x))(1 - \eta(x'))] &\geq 2[\eta(x)\eta(x') \\ &+ (1 - \eta(x))(1 - \eta(x'))] \geq L(f_c(x, x')). \end{aligned}$$

Case 3: If $-1 < t < 1$, we have

$$\begin{aligned} L(t) &= (1-t)[\eta(x)\eta(x') + (1-\eta(x))(1-\eta(x'))] + (1+t)[\eta(x)(1-\eta(x') + \eta(x')(1-\eta(x)))] \\ &\geq (1 - t)\frac{1}{2}L(f_c(x, x')) + (1 + t)\frac{1}{2}L(f_c(x, x')) = L(f_c(x, x')). \end{aligned}$$

Hence, we have $L(t) \geq L(f_c(x, x'))$ for all $t \in \mathbb{R}$, it follows that

$$\mathcal{E}(f) = \iint L(f(x, x'))d\rho_X(x)d\rho_X(x') \geq \iint L(f_c(x, x'))d\rho_X d(x)\rho_X(x') = \mathcal{E}(f_c).$$

This completes the proof of the lemma. □

We are now ready to give the proof of Corollary 4.

Proof of Corollary 4 Observe that

$$\begin{aligned} \mathcal{E}(f_{T+1}) - \mathcal{E}(f_c) &\leq \Big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\lambda)\Big] + \Big[\mathcal{E}(f_\lambda) - \mathcal{E}(f_c)\Big] \leq \Big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\lambda)\Big] + \mathcal{D}(\lambda) \\ &\leq \kappa\|f_{T+1} - f_\lambda\|_K + \mathcal{D}(\lambda). \end{aligned} \tag{32}$$

We know from Theorem 3 that, for any $0 < \lambda \leq 1$, that

$$\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K\big] \leq \big(\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big]\big)^{\frac{1}{2}} = \mathcal{O}\Big(\frac{\log T}{\lambda\sqrt{T}}\Big).$$

Putting this estimation with $\mathcal{D}(\lambda) = \mathcal{O}\big(\lambda^\beta\big)$, from (32) we obtain that

$$\mathbb{E}\big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_c)\big] = \mathcal{O}\Big(\frac{\log T}{\lambda\sqrt{T}} + \lambda^\beta\Big).$$

Choosing $\lambda = \big(\frac{\log^2 T}{T}\big)^{\frac{1}{2(1+\beta)}}$ yields the desired result. □

This paper focuses on the convergence of the individual iterate f_t in the expectation form. We end this section with a comment on describing the difficulties of

deriving the convergence of $\|f_{T+1} - f_\lambda\|_K\|_K^2$ with high confidence. To this end, let

$$\widetilde{\mathcal{A}}_\lambda^t(f_t) = \frac{1}{t-1}\sum_{j=1}^{t-1}\int_Z \ell'(f_t(x,x_j), r(y,y_j))K_{(x,x_j)}d\rho(x,y) + \lambda f_t,$$

and we know that

$$\langle f_\lambda - f_t, \widetilde{\mathcal{A}}_\lambda^t(f_t)\rangle_K \le \widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t). \tag{33}$$

Combining the above estimation, we can see, by the definition of f_{t+1} and (22), that $\|f_{t+1} - f_\lambda\|_K^2$ can be bounded by

$$\begin{aligned}
&\|f_t - f_\lambda\|_K^2 + 4\kappa^2M^2\eta_t^2 + 2\eta_t\langle f_\lambda - f_t, \widehat{\mathcal{A}}_\lambda^t(f_t) - \widetilde{\mathcal{A}}_\lambda^t(f_t)\rangle_K + 2\eta_t(\widetilde{\mathcal{E}}_\lambda^t(f_\lambda) - \widetilde{\mathcal{E}}_\lambda^t(f_t))\\
\le\ &\|f_t - f_\lambda\|_K^2 + 4\kappa^2M^2\eta_t^2 + 2\eta_t\langle f_\lambda - f_t, \widehat{\mathcal{A}}_\lambda^t(f_t) - \widetilde{\mathcal{A}}_\lambda^t(f_t)\rangle_K\\
&\qquad +\eta_t(\mathcal{E}_\lambda(f_\lambda) - \mathcal{E}_\lambda(f_t)) + 512\kappa^2M^2\left(3+\sqrt{\log\frac{1}{\delta}}\right)^2/t\lambda\\
\le\ &(1-\eta_t\lambda)\|f_t - f_\lambda\|_K^2 + 2\eta_t\langle f_\lambda - f_t, \widehat{\mathcal{A}}_\lambda^t(f_t) - \widetilde{\mathcal{A}}_\lambda^t(f_t)\rangle_K + 4\kappa^2M^2\\
&\left[\eta_t^2 + \frac{128\left(3+\sqrt{\log\frac{1}{\delta}}\right)^2}{t\lambda}\right],
\end{aligned}$$

where the second to the last inequality used (20) with $\tau = \frac{1}{2}$, the fact that $f_t \in \mathcal{F}_\lambda$ for any t, and $\mathcal{E}_\lambda(f_t) - \mathcal{E}_\lambda(f_\lambda) \ge \lambda\|f_t - f_\lambda\|_K^2$. Consequently,

$$\begin{aligned}
\|f_{T+1} - f_\lambda\|_K^2 \le\ &\prod_{t=2}^T(1-\eta_t\lambda)\|f_\lambda\|_K^2 + 4\kappa^2M^2\sum_{j=2}^T\prod_{k=j+1}^T(1-\eta_k\lambda)\left[\eta_j^2 + \frac{128\left(3+\sqrt{\log\frac{1}{\delta}}\right)^2\eta_j}{j\lambda}\right]\\
&+ 2\sum_{j=2}^T\prod_{k=j+1}^T(1-\eta_k\lambda)\eta_j\langle f_\lambda - f_j, \widehat{\mathcal{A}}_\lambda^j(f_j) - \widetilde{\mathcal{A}}_\lambda^j(f_j)\rangle_K.
\end{aligned} \tag{34}$$

By choosing $\eta_j = \frac{1}{j\lambda}$, the first two terms on the right hand side of (34) can achieve the fast convergence rate $\mathcal{O}(\frac{\log T}{\lambda^2 T})$ since

$$\prod_{t=2}^T(1-\eta_t\lambda)\|f_\lambda\|_K^2 = \frac{1}{T}\|f_\lambda\|_K^2 \le \frac{1}{T\lambda},$$

and

$$\sum_{j=2}^T\prod_{k=j+1}^T(1-\eta_k\lambda)\Big[\eta_j^2 + \frac{128\big(3+\sqrt{\log\frac{1}{\delta}}\big)^2\eta_j}{j\lambda}\Big] \le \frac{1+128\big(3+\sqrt{\log\frac{1}{\delta}}\big)^2}{\lambda^2}\Big(\frac{\log T}{T}\Big).$$

Notice that $\{\langle f_\lambda - f_j, \widehat{\mathcal{A}}_\lambda^j(f_j) - \widetilde{\mathcal{A}}_\lambda^j(f_j)\rangle_K : j \in \mathbb{N}\}$ is a martingale difference

sequence. Choosing $\eta_j = \frac{1}{j\lambda}$ and directly applying the Pinelis-Bernstein inequality [23] to this martingale difference sequence implies that

$$\sum_{j=2}^{T}\prod_{k=j+1}^{T}(1-\eta_k\lambda)\eta_j\langle f_\lambda - f_j, \widehat{\mathcal{A}}_\lambda^j(f_j) - \widetilde{\mathcal{A}}_\lambda^j(f_j)\rangle_K$$
$$= \frac{1}{T\lambda}\sum_{j=2}^{T}\langle f_\lambda - f_j, \widehat{\mathcal{A}}_\lambda^j(f_j) - \widetilde{\mathcal{A}}_\lambda^j(f_j)\rangle_K = \mathcal{O}(\frac{1}{\sqrt{T}\lambda^2}).$$

From the above arguments, we now clearly see that the critical hurdle to get the fast convergence rate for $\|f_{T+1} - f_\lambda\|_K^2$ with high probability comes from the term $\sum_{j=2}^{T}\prod_{k=j+1}^{T}(1-\eta_k\lambda)\eta_j\langle f_\lambda - f_j, \widehat{\mathcal{A}}_\lambda^j(f_j) - \widetilde{\mathcal{A}}_\lambda^j(f_j)\rangle_K$. We can get the fast convergence rate in the form of expectation since the expectation of this term disappears.

4 Conclusion

In this paper, we first introduced an online learning algorithm (2) with pairwise loss function with focus on the hinge loss for the sake of brevity. Our analysis can be easily extended to other general pairwise loss functions. Under certain conditions on step sizes, we established general convergence results and derived the learning rate.

There are several possible directions for future work. Firstly, in this paper we consider the performance of the last iterate f_{T+1} of algorithm (2). It would be interesting to investigate the performance of the average of all iterates $(f_1 + f_2 + \cdots + f_T)/T$ or the average of the α proportion iterates with $0 < \alpha \leq 1$, it is expected to get the optimal rate by averaging scheme. For more discussion about the averaging scheme for online learning algorithms associated with univariate loss functions, see [26] and references therein. Secondly, our results only indicates a sub-linear convergence rate, it is unknown how to get an exponential convergence rate under certain conditions on the step sizes and the RKHS $\mathcal{H}_K$. Thirdly, our analysis requires the regularization parameter to be strictly positive, i.e. $\lambda > 0$. We know from [36], for the least-square loss in the univariate setting, that convergence results also can be established even when $\lambda = 0$. However, it still remains a challenging question to us on how to establish similar convergence results for algorithm (2) with $\lambda = 0$.

Acknowledgments The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 104012] and by the National Natural Science Foundation of China [Grant No. 11401524, Grant No. 11531013 and Grant No. 11471292].

References

1. Agarwal, S., Niyogi, P.: Generalization bounds for ranking algorithms via algorithmic stability. J. Mach. Learn. Res. **10**, 441–474 (2009)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)

3. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local Rademacher complexities. Ann. Stat. **33**, 1497–1537 (2005)
4. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)
5. Bellet, A., Habrard, A., Sebban, M.: Similarity learning for provably accurate sparse linear classification. ICML (2012)
6. Cao, Q., Guo, Z.C., Ying, Y.: Generalization bounds for metric and similarity learning. Machine Learning Journal **102**(1), 115–132 (2016)
7. Cesa-Bianchi, C., Gentile, C.: Improved risk tail bounds for online algorithms. IEEE Trans. Inf. Theory **54**(1), 286–390 (2008)
8. Chen, H., Pan, Z., Li, L.: Learning performance of coefficient-based regularized ranking. Neurocomputing **133**, 54–62 (2014)
9. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. J. Mach. Learn. Res. **11**, 1109–1135 (2010)
10. Clémencon, S., Lugosi, G., Vayatis, N.: Ranking and empirical minimization of U-statistics. Ann. of Stat. **36**, 844–874 (2008)
11. Cucker, F., Zhou, D.-X.: Learning Theory: An Approximation Theory Viewpoint. Cambridge Univesity Press (2007)
12. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning (ICML) (2007)
13. Fan, J., Hu, T., Wu, Q., Zhou, D.X.: Consistency analysis of an empirical minimum error entropy algorithm, Applied and Computational Harmonic Analysis **41**(1), 161–189 (2016). doi:10.1016/j.acha.2014.12.005
14. Guo, Z.C., Ying, Y.: Guaranteed classification via regularized similarity learning. Neural Comput. **26**, 497–522 (2014)
15. Hu, T., Fan, J., Wu, Q., Zhou, D.X.: Regularization schemes for minimum error entropy principle. Anal. Appl. **13**, 437–455 (2015)
16. Kar, P., Sriperumbudur, B., Jain, P., Karnick, H.: On the generalization ability of online learning algorithms for pairwise loss functions. In: Proceedings of the 30th International Conference on Machine Learning (ICML) (2013)
17. Jin, R., Wang, S., Zhou, Y.: Regularized distance metric learning: theory and algorithm. In: Advances in Neural Information Processing Systems (NIPS) (2009)
18. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: isoperimetry and processes. Springer (1991)
19. McDiarmid, C.: Surveys in Combinatorics, Chapter on the methods of bounded differences, pp. 148–188. Cambridge University Press, Cambridge (UK) (1989)
20. Meir, R., Zhang, T.: Generalization error bounds for Bayesian mixture algorithms. J. Mach. Learn. Res. **4**, 839–860 (2003)
21. Mukherjee, S., Wu, Q.: Estimation of gradients and coordinate covariation in classification. J. Mach. Learn. Res. **7**, 2481–2514 (2006)
22. Mukherjee, S., Zhou, D.X.: Learning coordinate covariances via gradients. J. Mach. Learn. Res. **7**, 519–549 (2006)
23. Pinelis, I.: Optimum bounds for the distributions of martingales in banach spaces. Ann. Prob. **22**, 1679–1706 (1994)
24. Rakhlin, A., Shamir, O., Sridharan, K.: Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Conference on Machine Learning (ICML) (2012)
25. Rejchel, W.: On ranking and generalization bounds. J. Mach. Learn. Res. **13**, 1373–1392 (2012)
26. Shamir, O., Zhang, T.: Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. In: Proceedings of the 30th International Conference on Machine Learning (ICML) (2013)
27. Smale, S., Yao, Y.: Online learning algorithms. Found. Comput. Math. **6**, 145–170 (2006)
28. Sridharan, K., Srebro, N., Shalev-Shwartz, S.: Fast rates for regularized objectives Advances in Neural Information Processing Systems (NIPS) (2008)
29. Steinwart, I., Christmann, A.: Support Vector Machines. Springer-Verlag, New York (2008)
30. Weinberger, K.Q., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbour classification. In: Advances in Neural Information Processing Systems (NIPS) (2005)

31. Vitter, J.S.: Random sampling with a reservoir. ACM Trans. Math. Softw. **11**(1), 37–57 (1985)
32. Wang, Y., Khardon, R., Pechyony, D., Jones, R.: Generalization bounds for online learning algorithms with pairwise loss functions. COLT (2012)
33. Wang, Y., Khardon, R., Pechyony, D., Jones, R.: Online learning with pairwise loss functions. ArXiv Preprint (2013). arXiv:1301.5332
34. Wu, Q., Zhou, D.X.: Analysis of support vector machine classification. J. Comput. Anal. Appl. **8**(2), 99–119 (2006)
35. Ying, Y., Li, P.: Distance metric learning with eigenvalue optimization. J. Mach. Learn. Res. **13**, 1–26 (2012)
36. Ying, Y., Pontil, M.: Online gradient descent algorithms. Found. Comput. Math. **5**, 561–596 (2008)
37. Ying, Y., Zhou, D.X.: Online regularized classification algorithms. IEEE Trans. Inf. Theory **11**, 4775–4788 (2006)
38. Zhao, P., Hoi, S.C.H., Jin, R., Yang, T.: Online AUC Maximization. In: Proceedings of the 28th International Conference on Machine Learning (ICML) (2011)
39. Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. of Stat. **32**, 56–85 (2004)