Online Pairwise Learning Algorithms with Convex Loss Functions

Junhong Lin, Yunwen Lei, Bo Zhang, and Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China
jhlin5@hotmail.com, yunwen.lei@hotmail.com, bozhang37-c@my.cityu.edu.hk, mazhou@cityu.edu.hk

Abstract

Online pairwise learning algorithms with general convex loss functions without regularization in a Reproducing Kernel Hilbert Space (RKHS) are investigated. Under mild conditions on loss functions and the RKHS, upper bounds for the expected excess generalization error are derived in terms of the approximation error when the stepsize sequence decays polynomially. In particular, for Lipschitz loss functions such as the hinge loss, the logistic loss and the absolute-value loss, the bounds can be of order $O(T^{-\frac{1}{3}} \log T)$ after T iterations, while for the least squares loss, the bounds can be of order $O(T^{-\frac{1}{4}} \log T)$. In comparison with previous works for these algorithms, a broader family of convex loss functions is studied here, and refined upper bounds are obtained.

Keywords: Learning theory; Online Learning; Reproducing kernel Hilbert space; Pairwise learning

1 Introduction

Many classical learning tasks can be modeled as learning a good estimator or predictor $f : X \to Y$ based on an observed dataset $\{(x_t, y_t)\}_{t=1}^T$ of input-output samples from $X \times Y$, where X is an input space and $Y \subseteq \mathbb{R}$ an output space. Learning algorithms are often implemented by minimizing $\frac{1}{T}\sum_{t=1}^T V(y_t, f(x_t))$ over a hypothesis space of functions in various ways including regularization schemes [26]. Here $V : \mathbb{R}^2 \to \mathbb{R}_+$ is a *loss function* used for measuring the performance of a predictor f. It induces a *local error* $V(y, f(x))$ over an input-output sample $(x, y) \in X \times Y$. For non-parametric regression with $Y = \mathbb{R}$, the least squares loss function $V(y, a) = (y - a)^2$ is often used and, for an input $x \in X$ and an estimator f, the induced local error $V(y, f(x)) = (y - f(x))^2$ measures how well the predicted value $f(x)$ approximates the output value $y \in \mathbb{R}$. For binary classification with $Y = \{1, -1\}$ consisting of the two labels corresponding to the two classes, the misclassification loss function $V(y, a) = \chi_{(-\infty,0)}(ya)$ generated by the characteristic function of the interval $(-\infty, 0)$

† The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 104113]. The corresponding author is Yunwen Lei. Junhong Lin is now within the LCSL, MIT & Istituto Italiano di Tecnologia, Cambridge, MA 02139, USA

is a natural choice, and the induced local error $V(y, f(x)) = \chi_{(-\infty,0)}(yf(x))$ over a sample $(x, y) \in X \times Y$ equals 1 when the sign of $f(x)$ and y correspond to the two different labels in Y (that is, $yf(x) < 0$), while $V(y, f(x)) = 0$ when they correspond to a same label with $yf(x) \geq 0$. But the characteristic function $\chi_{(-\infty,0)}$ is not convex, and the optimization problems involved in the related learning algorithms are not convex. For designing efficient learning algorithms, $\chi_{(-\infty,0)}$ may be replaced by a convex function $\phi : \mathbb{R} \to \mathbb{R}_+$, leading to convex optimization problems involving the local error $V(y, f(x)) = \phi(yf(x))$. One choice of ϕ is the hinge loss $\phi_h(v) = \max\{1 - v, 0\}$ used in the classical support vector machines for solving binary classification problems [26]. The above learning framework has been well developed within the last two decades [26, 9]. It might be categorized as "pointwise learning", as the local error $V(y, f(x))$ takes only one sample point $(x, y) \in X \times Y$ into account.

In this paper, we study another important family of learning problems categorized as "pairwise learning" in which the local error takes a pair $\{(x, y), (x', y')\}$ of two samples from $X \times Y$ into account. Its learning tasks include ranking [1, 8], similarity and metric learning [5, 28], AUC maximization [34], and gradient learning [20, 30, 19]. The goal of *pairwise learning* is to learn a good predictor $f : X^2 \to \mathbb{R}$ predicting a value $f(x, x') \in \mathbb{R}$ for each input pair $(x, x') \in X^2$ according to various tasks. To measure the learning performance of a predictor f, we use a loss function $V : \mathbb{R}^2 \to \mathbb{R}_+$ to induce the local error $V(r(y, y'), f(x, x'))$ over two input-output samples $(x, y), (x', y') \in X \times Y$, where $r : Y \times Y \to \mathbb{R}$ is a function, called *reducing function*, chosen according to the learning task. The reducing function r is an essential concept making pairwise learning different from pointwise learning. We demonstrate how to choose the reducing function r by the following examples.

1. For the least squares regression with $Y = \mathbb{R}$ and $V(y, a) = (y - a)^2$, a sample (x, y) is drawn from a probability measure and the expected value of $y \in \mathbb{R}$ given $x \in X$ equals $f^*(x)$, the value of the conditional mean (regression) function f^* at x. So $y - y' = f^*(x) - f^*(x')$ in expectation and we choose the reducing function $r : Y \times Y \to \mathbb{R}$ as the output value difference $r(y, y') = y - y'$. Then the local error $V(r(y, y'), f(x, x')) = (y - y' - f(x, x'))^2$ measures how well the predicted value $f(x, x')$ for an input pair (x, x') approximates $f^*(x) - f^*(x')$ via the output value difference $y - y'$.

2. For metric learning in binary classification with $Y = \{1, -1\}$, we aim to learn a metric f such that a pair (x, x') of inputs (objects) from the same class $(y = y')$ are close to each other while a pair from different classes $(y \neq y')$ have a large distance $f(x, x')$. A typical choice of the reducing function $r : Y \times Y \to \mathbb{R}$ is given by $r(y, y') = 1$ if $y = y'$ and -1 otherwise [5]. The local error induced by the convex loss function $V(y, a) = \max\{0, 1 + ya\}$ is $V(r(y, y'), f(x, x')) = \max\{0, 1 + r(y, y')f(x, x')\}$. It gives a large local error $1 + f(x, x')$ if the distance $f(x, x')$ between the input pair (x, x') from the same class $(y = y')$ is large.

3. For ranking in a regression framework with $Y = \mathbb{R}$, we aim to learn a good ordering f between objects (inputs) based on their observed features such that $f(x, x') < 0$ if x is preferred over x' meaning that the ranking labels satisfy $y < y'$. A typical choice [21] of the reducing function $r : Y \times Y \to \mathbb{R}$ is given by $r(y, y') = \mathrm{sign}(y - y')$, the sign

of $y - y'$. Then the local error induced by the hinge loss ϕ_h is $V(r(y, y'), f(x, x')) = \phi(\text{sign}(y - y')f(x, x'))$.

Batch learning and online learning are two kinds of learning algorithms. The former uses an entire dataset to perform learning tasks, while the latter uses the dataset in a stream way. For batch learning algorithms in the pairwise learning framework, theoretical error and robustness analysis has been carried out in [1, 8, 21, 5, 7]. One challenge in conducting analysis in pairwise learning is that pairs of training samples are not independent. For example, given the independently and identically distributed (i.i.d.) samples $\{z_t = (x_t, y_t)\}_{t=1}^T$, a batch algorithm for pairwise learning possibly involves a target function

$$\frac{T(T-1)}{2} \sum_{1 \le i < j \le T} V(r(y_i, y_j), f(x_i, x_j)) + \text{pen}(f, \lambda), \tag{1.1}$$

where $\text{pen}(f, \lambda) \ge 0$ is some regularization term used to avoid overfitting. In this case, local errors $V(r(y_i, y_j), f(x_i, x_j))$ and $V(r(y_i, y_{j'}), f(x_i, x_{j'}))$ are indeed dependent. Thus, standard techniques for classification and regression cannot be directly applied, and new tools such as U-statistics [8] or algorithmic stability [1] are necessary for the analysis.

In spite of their good theoretical guarantees, batch algorithms for pairwise learning may be difficult to implement for large-scale learning problems in practice. Indeed, even for the simpler case of univariate learning, the computational complexity of batch algorithms with many loss functions is of order $O(T^3)$. Moreover, batch algorithms for pairwise learning suffer from extra computational burden of optimizing an objective defined over $O(T^2)$ possible sample pairs.

In practical applications, online learning may be more favorable, due to its scalability to large datasets and applicability to situations where the samples are collected sequentially. Theoretical results for online learning in classification and regression have been well developed (see for example [6, 24, 31, 2, 22, 18] and references therein), but there is relatively little work for online learning in pairwise learning. Recent research of this direction can be found in [15, 27, 32]. In particular, online pairwise learning in a linear space was investigated in [15, 27], and convergence results were established for the average of the iterates under the assumption of uniform boundedness of the loss function, with a rate $O(1/\sqrt{T})$ in the general convex case, or a rate $O(1/T)$ in the strongly convex case. Online pairwise learning in a RKHS with the least squares loss was studied in [32] where bounds in probability were derived for the excess generalization error.

In this paper, we improve the analysis of online pairwise learning (see Algorithm 1 in the next section) in a RKHS with general convex loss functions. Our main purpose is to develop convergence results for such learning algorithms using polynomially decaying stepsize sequences. Unlike [15, 27], we do not assume that the iterates are restricted to a bounded domain or the loss function is strongly convex. In particular, we will provide bounds for the expected excess generalization error, under a mild condition on approximation errors and an increment condition on the loss. For Lipschitz loss functions such as the hinge loss and the logistic loss, our bounds can be of order $O(T^{-\frac{1}{3}} \log T)$, while for the least squares loss, our bounds can be of order $O(T^{-\frac{1}{4}} \log T)$. For general convex loss functions, previous error analysis techniques dealing with the least squares loss in [32], which rely on integral operators, do not apply and are replaced by tools from convex analysis and Rademacher complexity. The key to our proof is an error decomposition, which enables us to study the weighted excess generalization error in terms of the weighted average and the moving weighted average. The novelty lies in an estimate of the differences between partial and generalization errors of the learning sequence. We shall establish bounds for the learning sequence using tools from convex analysis, and give uniform bounds for the differences between partial and full generalization errors over any given ball using Rademacher complexity. Our methods may be applied to pairwise learning with non-convex loss functions. In particular, it would be interesting to extend our methods to online learning or gradient descent methods for a minimum error entropy principle [10, 14].

2 Main Results with Discussions

In this section, after stating our pairwise learning problems and basic assumptions, we present our main results with some simulations and discussions. Proofs are postponed till the next section.

Let the input space X be a separable metric space and ρ be a Borel probability measure on $Z := X \times Y$.

For a predictor $f : X^2 \to \mathbb{R}$, we use a loss function $V : \mathbb{R}^2 \to \mathbb{R}_+$ and a reducing function $r : Y \times Y \to \mathbb{R}$ to give the local error $V(r(y,y'), f(x,x'))$ for $z = (x,y), z' = (x',y') \in Z$. The *generalization error* or risk $\mathcal{E} = \mathcal{E}^V$ associated with the loss function V is defined as

$$\mathcal{E}(f) = \int_Z \int_Z V(r(y,y'), f(x,x')) d\rho(z) d\rho(z').$$

We assume that there exists at least one minimizer f_ρ^V of the generalization error $\mathcal{E}(f)$, among all measurable functions $f : X^2 \to \mathbb{R}$. The goal of pairwise learning is to learn f_ρ^V from the sample set $S = \{z_t = (x_t, y_t)\}_{t=1}^T$ of size $T \in \mathbb{N}$. Throughout this paper, we assume that the samples are independently drawn according to ρ.

Our learning algorithm is a kernel method, where a RKHS is our hypothesis space. Let $K\colon X^2 \times X^2 \to \mathbb{R}$ be a Mercer Kernel, i.e., a continuous, symmetric and positive semi-definite kernel. The kernel K defines the RKHS $(\mathcal{H}_K, \|\cdot\|_K)$ as the completion of the linear span of the set $\{K_{(x,x')}(\cdot) := K((x,x'),(\cdot,\cdot)) : (x,x') \in X^2\}$ with respect to an inner product $\langle , \rangle_K$ satisfying the reproducing property: i.e., $\langle K_{(x,x')}, g\rangle_K = g(x,x')$ for any $(x,x') \in X^2$ and $g \in \mathcal{H}_K$.

We assume in this paper that V is convex with respect to the second variable. That is, for any fixed $y \in \mathbb{R}$, the univariate function $V(y,\cdot)$ on $\mathbb{R}$ is convex, hence its left-hand derivative $V'_-(y,f)$ exists at every $f \in \mathbb{R}$ and is non-decreasing.

The online pairwise learning algorithm considered in this paper is as follows.

Algorithm 1. *The* online pairwise learning algorithm *associated with the loss function V* *and the kernel K is defined by $f_1 = f_2 = 0$ and*

$$f_{t+1} = f_t - \frac{\eta_t}{t-1} \sum_{j=1}^{t-1} V'_-(r(y_t, y_j), f_t(x_t, x_j)) K_{(x_t, x_j)}, \qquad t = 2, \ldots, T, \tag{2.1}$$

where $\{\eta_t > 0\}_t$ is a step size sequence.

The main purpose of this paper is to estimate the expected excess generalization error $\mathbb{E}[\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)]$. To this end, we shall make the following assumptions.

Assumption 2.1. *We assume*

$$|V|_0 := \sup_{y,y' \in Y} V(r(y, y'), 0) < \infty \tag{2.2}$$

and an increment condition for the left-hand derivative $V'_-(y, \cdot)$ that for some $q \geq 0$ and constant $c_q > 0$, there holds

$$\left| V'_-(r(y, y'), f) \right| \leq c_q(1 + |f|^q), \qquad \forall f \in \mathbb{R}, y, y' \in Y. \tag{2.3}$$

We assume the kernel to be bounded with

$$\kappa = \max \left(\sup_{x,x' \in X} \sqrt{K((x, x'), (x, x'))}, 1 \right) < \infty. \tag{2.4}$$

Assumption (2.2) automatically holds for loss functions widely used for classification, where V takes the form $V(y, f) = \phi(-yf)$ with $\phi : \mathbb{R} \to \mathbb{R}_+$ being a convex function, including the hinge loss ϕ_h, the exponential loss $\phi(v) = \exp(-v)$ and the logistic loss $\phi(v) = \log(1 + \exp(-v))$. Assumption (2.2) is equivalent to the boundedness assumption on the output space Y for $r(y, y') = y - y'$ and loss functions of the form $V(y, f) = \phi(y - f)$ for regression with $\lim_{|y| \to \infty} \phi(y) = \infty$, including the p-norm absolute distance loss $\phi(y) = |y|^p$ with $p \geq 1$. Note that (2.2) may also hold for the case that Y is not bounded, e.g., the ranking problems with $r(y, y') = \text{sign}(y - y')$. The increment condition on loss functions (2.3) and the boundness assumption on the kernel are quite common in learning theory. For specific loss functions, one can easily compute the constants q and c_q in (2.3). For example, if the loss function is the hinge loss $V(y, f) = \phi_h(yf)$, we know [25] that (2.3) is satisfied with $q = 0$ and $c_q = \sup_{y,y' \in Y} |r(y, y')|$, and in this case $|V|_0 = 1$.

We also need a notion of approximation error to state our main results.

Definition 2.2. *The approximation error associated with the tripe (ρ, V, K) is defined by*

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho^V) + \lambda \|f\|_K^2 \right\}, \qquad \forall \lambda > 0. \tag{2.5}$$

Our main result of this paper is stated as follows.

Theorem 2.3. *Under Assumption 2.1, let* $\{\eta_{t+1} = \eta_1 t^{-\theta}\}_{t\in\mathbb{N}}$ *with* $\frac{q}{q+1} \le \theta < 1$ *and* η_1 *satisfying*

$$0 < \eta_1 \le \min\left\{ \frac{\sqrt{1-\theta}}{2\sqrt{2}c_q\kappa(\kappa+1)^q}, \frac{1-\theta}{4|V|_0} \right\}. \tag{2.6}$$

Then the sequence $\{f_t\}_t$ *generated by Algorithm 1 satisfies*

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} \le \widetilde{C}_0\mathcal{D}((T-1)^{\theta-1}) + \widetilde{C}_1\Lambda_{T-1},$$

where Λ_{T-1} *is the quantity defined by*

$$\Lambda_{T-1} = \begin{cases} (T-1)^{-(1-\theta)}, & \text{when } \theta > \frac{q+2}{q+3}, \\ (T-1)^{-\frac{q\theta+\theta-q}{2}}\log(eT), & \text{when } \theta \le \frac{q+2}{q+3}, \end{cases} \tag{2.7}$$

and $\widetilde{C}_0$ *and* $\widetilde{C}_1$ *are constants independent of* T *(given explicitly in the proof).*

To state explicit convergence rates, we need the following assumption for the decay of the approximation error.

Assumption 2.4. *Assume that for some* $\beta \in (0,1]$ *and* $c_\beta > 0$*, the approximation error satisfies*

$$\mathcal{D}(\lambda) \le c_\beta \lambda^\beta, \qquad \forall \lambda > 0. \tag{2.8}$$

The assumption (2.8) on the approximation error is independent of the samples, and measures the approximation ability of the space $\mathcal{H}_K$ to f_ρ^V with respect to (ρ, V). It is standard in learning theory both in pairwise [32] and pointwise learning [25, 29, 11]. Note that in the ideal case with $f_\rho^V \in \mathcal{H}_K$, condition (2.8) always holds with $\beta = 1$ and $c_\beta \le \|f_\rho^V\|_K^2$.

We now have the following corollary, which follows directly from Theorem 2.3.

Corollary 2.5. *Under the assumptions and notations of Theorem 2.3, and Assumption 2.4, we have*

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = O(T^{(\theta-1)\beta} + \Lambda_T). \tag{2.9}$$

In particular, we have

(I) for $\theta = \frac{q+2}{q+3}$,

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = O(T^{-\frac{\beta}{q+3}}\log T).$$

(II) for $\theta = \frac{q+2\beta}{q+1+2\beta}$,

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = O(T^{-\frac{\beta}{q+1+2\beta}}\log T).$$

The above result gives bounds on the expected excess generalization error, where the general convergence rate in (2.9) depends on three parameters: q, β, and θ. In general, it is easy to compute the increment parameter q for a given loss function, whereas the parameter β is unknown. Given only the growth parameter q, Part (I) of Corollary 2.5 suggests that the optimal convergence rate is achieved by setting $\theta = \frac{q+2}{q+3}$. If furthermore, the parameter β is provided, the optimal convergence rate is achieved by setting $\theta = \frac{q+2\beta}{q+1+2\beta}$.

Specifying the loss function in the above results, we have the following convergence rates with the hinge loss and the least squares loss.

Corollary 2.6 (Hinge loss)**.** *Let the loss function $V(y,a)$ be given with the hinge loss as $V(y,a) = \phi_h(ya)$. Assume (2.4), (2.8) and $M := \sup_{y,y' \in Y} |r(y,y')| < \infty$. Choose $\{\eta_{t+1} = \eta_1 t^{-\theta}\}_{t\in\mathbb{N}}$ with η_1 satisfying (2.6), where $q = 0, c_q = M$ and $|V|_0 = 1$. Then for the sequence $\{f_t\}_t$ generated by Algorithm 1, we have the following convergence rates.*

(I) If $\theta = \frac{2}{3}$, then

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = O\left(T^{-\frac{\beta}{3}} \log T\right).$$

Specially, if $\beta = 1$, i.e., $f_\rho^V \in \mathcal{H}_K$, then the upper bound is of order $O\left(T^{-\frac{1}{3}} \log T\right)$.

(II) If $\theta = \frac{2\beta}{2\beta+1}$, then

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = O\left(T^{-\frac{\beta}{2\beta+1}} \log T\right).$$

Corollary 2.7 (Least squares loss)**.** *Let V be given by the least squares loss as $V(y,a) = (y-a)^2$. Assume (2.4), (2.8) and $M := 2\max\left(\sup_{y,y'\in Y} |r(y,y')|, 1\right) < \infty$. Choose $\{\eta_{t+1} = \eta_1 t^{-\theta}\}_{t\in\mathbb{N}}$ with η_1 satisfying (2.6), where $q = 1, c_q = M$ and $|V|_0 = \sup_{y,y'\in Y}(r(y,y'))^2$. Then for the sequence $\{f_t\}_t$ generated by Algorithm 1, we have the following convergence rates.*

(I) If $\theta = \frac{3}{4}$, then

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = O\left(T^{-\frac{\beta}{4}} \log T\right).$$

Specially, if $\beta = 1$, i.e., $f_\rho^V \in \mathcal{H}_K$, then the upper bound is of order $O\left(T^{-\frac{1}{4}} \log T\right)$.

(II) If $\theta = \frac{2\beta+1}{2\beta+2}$, then

$$\mathbb{E}_{z_1,\cdots,z_T}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = O\left(T^{-\frac{\beta}{2\beta+2}} \log T\right).$$

Simulations. We perform simulation experiments here to illustrate the derived convergence rates with polynomial decaying stepsizes. We consider the ranking problem with the loss function $V(y,a)$ given by the hinge loss as $V(y,a) = \phi_h(ya)$ and the reducing function $r(y,y') = \text{sign}(y-y')$. We consider the Boston housing dataset [13], which has 506 examples and 13 features, including *per capita crime rate by town, weighted distances to five Boston employment centres and average number of rooms per dwelling.* We wish to predict the ordering based on values of houses and consider linear ranking rules with $K((x,x'),(u,u')) = (x-x')^\top(u-u')$ for $x, x', u, u' \in \mathbb{R}^{13}$. Here $x^\top$ denotes the transpose of x. Let ρ be the uniform distribution on the 506 examples in the Boston housing dataset. We define the ranking error of a predictor $f : X \times X \to \mathbb{R}$ by $L(f) = \mathbb{E}[\text{sign}(y-y')f(x,x') < 0]$. We apply

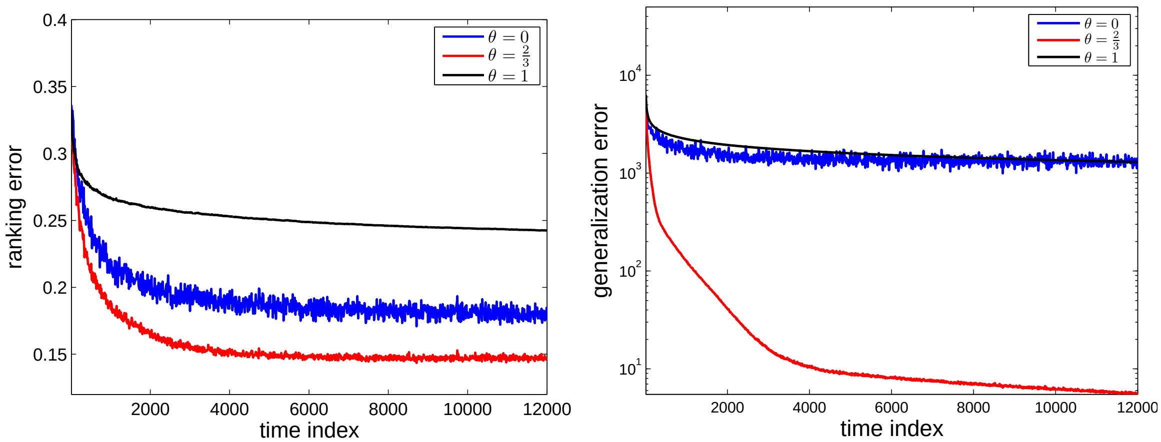

Figure 1: The behavior of Algorithm 1 on the Boston housing dataset. Left: ranking errors versus different stepsize sequences, right: generalization errors versus different stepsize sequences.

Algorithm 1 with $\eta_t = (t-1)^{-\theta}$ and $\theta \in \{0, 1, \frac{2}{3}\}$. We repeat the experiments 400 times and report the average ranking errors and average generalization errors. Figure 1 illustrates the behavior of Algorithm 1 with three different stepsize sequences. It shows that the algorithm with polynomial decaying stepsize sequence with $\theta = \frac{2}{3}$ performs better than that with the constant stepsize sequence $\eta_t \equiv 1$ and the sequence with $\theta = 1$. This is consistent with our theoretical results in Corollary 2.6.

Discussions. As mentioned before, online pairwise learning involves non-i.i.d. sample pairs. Thus, the analysis for pairwise learning is more challenging, in contrast with that for the online pointwise learning [6, 24, 31, 2, 22, 18]. With the step size $\eta_t = \eta_1 t^{-\frac{\beta}{\beta+1}}$, the convergence rate $O(T^{-\frac{\beta}{\beta+1}} \log T)$ was established in [18] for the online pointwise learning, which is comparable to the convergence rate for batch learning in the pointwise setting. The convergence rate we derived in Corollary 2.5 for the online pairwise learning is of order $O(T^{-\frac{\beta}{2\beta+1+q}} \log T)$. This is due to an essential statistical difference between these two families of learning algorithms: while the online pointwise learning uses unbiased estimators of the true gradients in the learning process, the randomized gradient $\frac{1}{t-1}\sum_{j=1}^{t-1} V'_-(r(y_t, y_j), f_t(x_t, x_j))K_{(x_t,x_j)}$ used in the online pairwise learning is a biased estimator of the true gradient $\int_Z \int_Z V'_-(y - y', f_t(x, x'))K_{(x,x')} d\rho(z) d\rho(z')$. We overcome this obstacle by applying the tool of Rademacher complexity to control the difference between partial generalization errors and generalization errors, resulting in, however, an additional term that dominates the upper bound in Proposition 3.6.

In what follows, we compare our work with existing results on online algorithms for pairwise learning. We first compare our work with [15, 27], where the online-to-batch conversion bounds for projected online pairwise learning algorithms in Euclidean spaces were provided.

Assuming that $f_\rho^V \in \mathbb{R}^d$ is in the projected-bounded domain, upper bounds on the excess generalization error of order $O(T^{-\frac{1}{2}})$ were presented in [15] for the average iterates. In contrast, Algorithm 1 does not have any additional projection step and is implemented in the unconstrained setting on RKHSs including the Euclidean spaces. Besides, our bounds are stated in a more general setting for the last iterates, involving approximation errors. It should be mentioned that convergence rates $O(T^{-\frac{1}{2}}\log T)$ can be achieved by our analysis for the pairwise learning setting if an additional projection is performed at each iteration and $\beta = 1$. Finally, we compare our results with the existing work in [32, 33, 12]. Algorithm 1 with kernel methods was studied in [32] for the least squares loss, and in [33] for 1-activating loss V, i.e., loss function which is differentiable and satisfies

$$|V'(y, f) - V'(y, g)| \leq L|f - g|, \qquad \forall y \in \mathbb{R}, f, g \in \mathbb{R}, \tag{2.10}$$

for some $0 < L < \infty$. A convergence rate of order $O(T^{-\min\left\{\frac{\beta}{\beta+1},\frac{1}{3}\right\}}\log T)$ is achieved for the algorithm with the least squares loss in [32]. However, the analysis in [32] is based on an integral operator approach and does not apply to general convex loss functions. Note that the results in [32] are in probability while our results are stated in expectation, and it would be interesting to further develop bounds in probability for the algorithm involving convex loss functions. In comparison with the results in [33] where 1-activating loss functions are studied with an assumption on the existence of a minimizer of $\mathcal{E}(f)$ for $f \in \mathcal{H}_K$, our results hold for a broader class of loss functions and are better for 1-activating loss functions in a more general setting. First, the hinge loss and the p-absolute value loss functions with $p \neq 2$ are not covered in [33]. Second, it is easy to see that an 1-activating loss function always satisfies the growth condition (2.3) with $q = 1$. Thus, by setting $\beta = 1$ and $\eta_t = \eta_1 t^{-\frac{\alpha+2}{\alpha+3}}$ in Corollary 2.5, our optimal convergence rates are of order $O(T^{-\frac{1}{4}}\log T)$ for 1-activating loss functions, which are better than the bounds in [33] of order $O(T^{\epsilon-\frac{1}{6}})$ with an arbitrarily small $\epsilon > 0$. When the incremental exponent q satisfies $0 \leq q < 1$, the learning rates of order $O(T^{-\frac{\beta}{q+1+2\beta}}\log T)$ stated in Corollary 2.5 (II) are also better than those of order $O(T^{-\frac{\beta}{2\beta+2}}\sqrt{\log T})$ derived for online pairwise learning based on regularization schemes in RKHSs in the earlier work [12].

3 Proofs

In this section, we prove Theorem 2.3. To do so, it is necessary to prove some preliminary lemmas.

3.1 Bounding the learning sequence

For notational simplicity, we introduce the following two notations: the local empirical error of a function $f : X \times X \to \mathbb{R}$ at point z_t with respect to an ordered dataset $S = \{z_1, \cdots, z_T\}$

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

and the partial generalization error with respect to an ordered dataset $S = \{z_1, \cdots, z_T\}$

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

We first introduce the following lemma whose proof essentially makes use of the convexity and increment property of loss functions.

Lemma 3.1. *Under condition* (2.3)*, for an arbitrary fixed* $f \in \mathcal{H}_K$*, and* $t = 2, \ldots, T$*,*

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t(\widehat{\mathcal{E}}_S^t(f) - \widehat{\mathcal{E}}_S^t(f_t)), \tag{3.1}$$

where

$$G_t^2 = 4c_q^2 \kappa^2 (\kappa + 1)^{2q} \max\left\{1, \|f_t\|_K^{2q}\right\}. \tag{3.2}$$

Proof. Since f_{t+1} is given by (2.1), we have

$$\begin{aligned}\|f_{t+1} - f\|_K^2 =& \|f_t - f\|_K^2 + \eta_t^2 \left\| \frac{1}{t-1} \sum_{j=1}^{t-1} V_-'(r(y_t, y_j), f_t(x_t, x_j)) K_{(x_t, x_j)} \right\|_K^2 \\ &+ \frac{2\eta_t}{t-1} \sum_{j=1}^{t-1} V_-'(r(y_t, y_j), f_t(x_t, x_j)) \left\langle K_{(x_t, x_j)}, f - f_t \right\rangle_K.\end{aligned}$$

Observe that

$$\|K_{(x_t, x_j)}\|_K = \sqrt{K((x_t, x_j), (x_t, x_j))} \leq \kappa$$

and that

$$\|f\|_\infty \leq \kappa \|f\|_K, \qquad \forall f \in \mathcal{H}_K.$$

These together with the increment condition (2.3) yield

$$\begin{aligned}\left\| V_-'(r(y_t, y_j), f_t(x_t, x_j)) K_{(x_t, x_j)} \right\|_K &\leq \kappa \left| V_-'(r(y_t, y_j), f_t(x_t, x_j)) \right| \\ &\leq \kappa c_q (1 + |f_t(x_t, x_j)|^q) \leq \kappa c_q (1 + \kappa^q \|f_t\|_K^q).\end{aligned}$$

Therefore,

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + \frac{2\eta_t}{t-1} \sum_{j=1}^{t-1} V_-'(r(y_t, y_j), f_t(x_t, x_j)) \left\langle K_{(x_t, x_j)}, f - f_t \right\rangle_K.$$

Using the reproducing property, we get

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + \frac{2\eta_t}{t-1} \sum_{j=1}^{t-1} V_-'(r(y_t, y_j), f_t(x_t, x_j))(f(x_t, x_j) - f_t(x_t, x_j)). \tag{3.3}$$

Since $V(r(y_t, y_j), \cdot)$ is a convex function, we have

$$V_-'(r(y_t, y_j), a)(b - a) \leq V(r(y_t, y_j), b) - V(r(y_t, y_j), a), \qquad \forall a, b \in \mathbb{R}.$$

Using this relation in (3.3), we get our desired result. $\square$

Using the above lemma, we can bound the learning sequence as follows. The proof is motivated by the recent work in [16] and [17], using an induction argument.

Lemma 3.2. *Assume condition* (2.3)*. Let* $\frac{q}{q+1} \leq \theta < 1$ *and* $\eta_{t+1} = \eta_1 t^{-\theta}$ *for* $t \in \mathbb{N}$ *with* η_1 *satisfying (2.6). Then for* $t = 1, \ldots, T$,

$$\|f_{t+1}\|_K \leq (t-1)^{\frac{1-\theta}{2}}. \tag{3.4}$$

Proof. We prove our statement by induction.

Taking $f = 0$ in Lemma 3.1, we know that

$$\|f_{t+1}\|_K^2 \leq \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t(\widehat{\mathcal{E}}_S^t(0) - \widehat{\mathcal{E}}_S^t(f_t)) \leq \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t |V|_0.$$

This verifies (3.4) for the case $t = 2$ since $f_1 = f_2 = 0$ and $4\eta_1^2 c_q^2 \kappa^2 (\kappa+1)^{2q} + 2\eta_1 |V|_0 \leq 1$.

Assume $\|f_t\|_K \leq (t-2)^{\frac{1-\theta}{2}}$ with $t \geq 3$. Then

$$G_t^2 \leq 4c_q^2 \kappa^2 (\kappa+1)^{2q} (t-2)^{(1-\theta)q}. \tag{3.5}$$

Hence

$$\begin{aligned} \|f_{t+1}\|_K^2 &\leq (t-2)^{1-\theta} + 4\eta_1^2 (t-1)^{-2\theta} c_q^2 \kappa^2 (\kappa+1)^{2q} (t-1)^{(1-\theta)q} + 2\eta_1 (t-1)^{-\theta} |V|_0 \\ &\leq (t-1)^{1-\theta} \left\{ \left(1 - \frac{1}{t-1}\right)^{1-\theta} + \frac{4\eta_1^2 c_q^2 \kappa^2 (\kappa+1)^{2q}}{(t-1)^{(q+1)\theta+1-q}} + \frac{2\eta_1 |V|_0}{t-1} \right\}. \end{aligned}$$

Since $\left(1 - \frac{1}{t-1}\right)^{1-\theta} \leq 1 - \frac{1-\theta}{t-1}$ and the condition $\theta \geq \frac{q}{q+1}$ implies $(q+1)\theta + 1 - q \geq 1$, we have

$$\|f_{t+1}\|_K^2 \leq (t-1)^{1-\theta} \left\{ 1 - \frac{1-\theta}{t-1} + \frac{4\eta_1^2 c_q^2 \kappa^2 (\kappa+1)^{2q}}{t-1} + \frac{2\eta_1 |V|_0}{t-1} \right\}.$$

Finally we use the restriction (2.6) for η_1 and find $\|f_{t+1}\|_K^2 \leq (t-1)^{1-\theta}$. This completes the induction procedure and proves our conclusion. □

With the above two lemmas, and noticing that f_t is independent of z_t, we can easily prove the following result.

Proposition 3.3. *Assume condition* (2.3)*. Let* $\frac{q}{q+1} \leq \theta < 1$ *and* $\eta_{t+1} = \eta_1 t^{-\theta}$ *for all* $t \in \mathbb{N}$ *with* η_1 *satisfying (2.6). Assume that* $t \in \{2, \ldots, T\}$ *and that* $f \in \mathcal{H}_K$ *is independent of* z_t *(but may depend on* $z_1, \cdots, z_{t-1}$*). Then we have*

$$\begin{aligned} &\mathbb{E}_{z_t} \|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 \\ &+ 4\eta_1^2 c_q^2 \kappa^2 (\kappa+1)^{2q} (t-1)^{(1-\theta)q - 2\theta} + 2\eta_t \left[\widetilde{\mathcal{E}}_S^t(f) - \widetilde{\mathcal{E}}_S^t(f_t) \right]. \end{aligned} \tag{3.6}$$

Proof. Taking expectations on both sides of (3.1) with respect to z_t, and noting that f_t is independent of z_t, we get

$$\mathbb{E}_{z_t}\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t \left[\widetilde{\mathcal{E}}_S^t(f) - \widetilde{\mathcal{E}}_S^t(f_t)\right].$$

Lemma 3.2 shows that $\|f_t\|_K \leq (t-1)^{\frac{1-\theta}{2}}$, which implies (3.5). Applying (3.5) and using $\eta_t = \eta_1(t-1)^{-\theta}$ in the above inequality yield the desired bound. □

Proposition 3.3 gives an iterated inequality related to the partial generalization error $\widetilde{\mathcal{E}}_S^t(f_t)$. Note that our goal is to derive upper bounds on the excess generalization error. It is thus necessary to develop relationships between the partial generalization error and generalization error, which will be considered in the following subsection.

3.2 From partial generalization error to generalization error

For $R > 0$, denote B_R the ball of radius R in $\mathcal{H}_K$: $B_R = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$. The following lemma gives a uniform upper bound on the differences between the partial generalization error and generalization error over any ball B_R with $R \geq 1$. Its proof uses a standard symmetry technique and some properties related to Rademacher complexity.

Lemma 3.4. *For $R \geq 1$, and all $1 \leq t \leq T$*

$$\mathbb{E}_{z_1,\cdots,z_{t-1}}\left[\sup_{f\in B_R}\{\mathcal{E}(f) - \widetilde{\mathcal{E}}_S^t(f)\}\right] \leq \frac{2c_q R\kappa(1+\kappa^q R^q)}{\sqrt{t-1}}.$$

The above inequality is also true if we replace $\{\mathcal{E}(f) - \widetilde{\mathcal{E}}_S^t(f)\}$ by $\{\widetilde{\mathcal{E}}_S^t(f) - \mathcal{E}(f)\}$.

Proof. For notational simplicity, we denote

$$\mathcal{L}(f, z_j) = \int_Z V(r(y, y_j), f(x, x_j))d\rho(z).$$

Then

$$\widetilde{\mathcal{E}}_S^t(f) = \frac{1}{t-1}\sum_{j=1}^{t-1}\mathcal{L}(f, z_j)$$

and

$$\mathcal{E}(f) = \int_Z \mathcal{L}(f, z')d\rho(z').$$

Let $S' = \{z_1', \cdots, z_T'\}$ be another independent sample set. We first note that

$$\begin{aligned}
&\mathbb{E}_S[\sup_{f\in B_R}\{\mathcal{E}(f) - \widetilde{\mathcal{E}}_S^t(f)\}] \\
&= \mathbb{E}_S[\sup_{f\in B_R}\{\mathbb{E}_{S'}[\widetilde{\mathcal{E}}_{S'}^t(f)] - \widetilde{\mathcal{E}}_S^t(f)\}] \\
&\leq \mathbb{E}_{S,S'}[\sup_{f\in B_R}\{\widetilde{\mathcal{E}}_{S'}^t(f) - \widetilde{\mathcal{E}}_S^t(f)\}].
\end{aligned}$$

Here, we abuse the notation $\mathbb{E}_S$ for $\mathbb{E}_{z_1,\cdots,z_{t-1}}$. Let $\sigma_1, \sigma_2, \ldots, \sigma_T$ be independent random variables drawn from the Rademacher distribution i.e. $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$ for $i = 1, 2, \ldots, T$. Using a standard symmetry technique, for example in [3],

$$\begin{aligned}
&\mathbb{E}_{S,S'}[\sup_{f \in B_R} \{\widetilde{\mathcal{E}}^t_{S'}(f) - \widetilde{\mathcal{E}}^t_S(f)\}] \\
&\le \mathbb{E}_{S,S',\sigma}\left[\sup_{f \in B_R}\left\{\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j(\mathcal{L}(f, z'_j) - \mathcal{L}(f, z_j))\right\}\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_S[\sup_{f \in B_R}\{\mathcal{E}(f) - \widetilde{\mathcal{E}}^t_S(f)\}] \\
&\le \mathbb{E}_{S,S',\sigma}\left[\sup_{f \in B_R}\left\{\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j(\mathcal{L}(f, z'_j) - \mathcal{L}(f, z_j))\right\}\right] \\
&\le 2\mathbb{E}_{S,\sigma}\left[\sup_{f \in B_R}\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j\mathcal{L}(f, z_j)\right] \\
&= 2\mathbb{E}_{S,\sigma}\left[\sup_{f \in B_R}\mathbb{E}_z\left[\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j V(r(y, y_j), f(x, x_j))\right]\right] \\
&\le 2\mathbb{E}_{z,S,\sigma}\left[\sup_{f \in B_R}\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j V(r(y, y_j), f(x, x_j))\right].
\end{aligned}$$

For any $z \in Z$, the term $\mathbb{E}_{S,\sigma}\left[\sup_{f \in B_R}\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j V(r(y, y_j), f(x, x_j))\right]$ is the Rademacher complexity [4] of the function class B_R with respect to ρ for sample size $t-1$. Using (2.3) and that $\|f\|_\infty \le \kappa\|f\|_K$, it is easy to see that for any $f, f' \in B_R$,

$$|V(r(y, y_j), f(x, x_j)) - V(r(y, y_j), f'(x, x_j))| \le c_q(1 + R^q\kappa^q)|f(x, x_j) - f'(x, x_j)|.$$

Applying Talagrand's contraction lemma (see e.g., [19, Theorem 7]), we have

$$\begin{aligned}
&\mathbb{E}_{S,\sigma}\left[\sup_{f \in B_R}\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j V(r(y, y_j), f(x, x_j))\right] \\
&\le c_q(1 + \kappa^q R^q)\mathbb{E}_{S,\sigma}\left[\sup_{f \in B_R}\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j f(x, x_j)\right]
\end{aligned}$$

and therefore,

$$\mathbb{E}_S[\sup_{f \in B_R}\mathbb{E}\{\mathcal{E}(f) - \widetilde{\mathcal{E}}^t(f)\}]$$

$$\leq 2c_q(1+\kappa^q R^q)\mathbb{E}_{z,S,\sigma}\left[\sup_{f\in B_R}\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j f(x,x_j)\right]$$
$$= 2c_q(1+\kappa^q R^q)\mathbb{E}_{z,S,\sigma}\left[\sup_{f\in B_R}\left\langle f,\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j K_{(x,x_j)}\right\rangle_K\right].$$

Applying the Schwarz inequality,

$$\mathbb{E}_S[\sup_{f\in B_R}\mathbb{E}\{\mathcal{E}(f)-\widetilde{\mathcal{E}}^t(f)\}]$$
$$\leq 2c_q(1+\kappa^q R^q)\mathbb{E}_{z,S,\sigma}\left[\sup_{f\in B_R}\|f\|_K\left\|\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j K_{(x,x_j)}\right\|_K\right].$$

Applying $\mathbb{E}[\|g\|_K] \leq (\mathbb{E}[\|g\|_K^2])^{\frac{1}{2}}$, and noting that $\sigma_1,\sigma_2,\ldots,\sigma_T$ are independent random variables with mean zeros,

$$\mathbb{E}_S[\sup_{f\in B_R}\mathbb{E}\{\mathcal{E}(f)-\widetilde{\mathcal{E}}^t(f)\}]$$
$$\leq \frac{2c_q(1+\kappa^q R^q)R}{t-1}\left[\mathbb{E}_{z,S,\sigma}\left\|\sum_{j=1}^{t-1}\sigma_j K_{(x,x_j)}\right\|_K^2\right]^{\frac{1}{2}}$$
$$= \frac{2c_q(1+\kappa^q R^q)R}{t-1}\left[\sum_{j=1}^{t-1}\mathbb{E}_{x,x_j}\left\|K_{(x,x_j)}\right\|_K^2\right]^{\frac{1}{2}}$$
$$\leq \frac{2c_q(1+\kappa^q R^q)R\kappa}{\sqrt{t-1}},$$

where for the last inequality we use the boundness assumption on the kernel. Thus we get the desired result. The proof is complete. □

Combining the above lemma with Lemma 3.2, we get the following corollary.

Corollary 3.5. *Under the assumptions of Lemma 3.2, we have for any* $t=3,\cdots,T$,

$$|\mathbb{E}_{z_1,\cdots,z_{t-1}}[\mathcal{E}(f_t)-\widetilde{\mathcal{E}}_S^t(f_t)]| \leq 2c_q\kappa(1+\kappa^q)(t-1)^{\frac{(1-\theta)(q+1)-1}{2}}.$$

3.3 A useful proposition

The following proposition will be used several times in our proof. Its proof follows directly from Proposition 3.3 and Corollary 3.5.

Proposition 3.6. *Under assumptions of Proposition 3.3, for any $f \in \mathcal{H}_K$ which is independent of $z_1, \cdots, z_t$, or $f = f_k$ $(3 \leq k \leq t)$, we have*

$$\begin{aligned}
&2\eta_t \mathbb{E}_{z_1,\cdots,z_{t-1}}\left[\mathcal{E}(f_t) - \mathcal{E}(f)\right] \\
\leq &\mathbb{E}_{z_1,\cdots,z_t}\left\{\|f_t - f\|_K^2 - \|f_{t+1} - f\|_K^2\right\} + C_{q,\kappa,\eta_1}(t-1)^{-q^*}.
\end{aligned} \tag{3.7}$$

Here

$$q^* = \frac{3\theta - (1-\theta)q}{2}. \tag{3.8}$$

and C_{q,κ,η_1} is a constant depending only on q, κ and η_1, given explicitly by (3.10) in the proof.

Proof. Note that for $3 \leq k \leq T$, f_k depends only on $z_1, \cdots, z_{k-1}$. By Proposition 3.3, we have

$$\begin{aligned}
&\mathbb{E}_{z_1,\cdots,z_t}\|f_{t+1} - f\|_K^2 \leq \mathbb{E}_{z_1,\cdots,z_t}\|f_t - f\|_K^2 \\
&+ 4\eta_1^2 c_q^2 \kappa^2 (\kappa+1)^{2q}(t-1)^{(1-\theta)q - 2\theta} + 2\eta_t \mathbb{E}_{z_1,\cdots,z_{t-1}}\left[\widetilde{\mathcal{E}}_S^t(f) - \widetilde{\mathcal{E}}_S^t(f_t)\right].
\end{aligned}$$

Rewrite $\mathbb{E}_{z_1,\cdots,z_{t-1}}\left[\widetilde{\mathcal{E}}_S^t(f) - \widetilde{\mathcal{E}}_S^t(f_t)\right]$ as

$$\mathbb{E}_{z_1,\cdots,z_{t-1}}\left[\mathcal{E}(f) - \mathcal{E}(f_t)\right] + \mathbb{E}_{z_1,\cdots,z_{t-1}}\left[(\widetilde{\mathcal{E}}_S^t(f) - \mathcal{E}(f)) + (\mathcal{E}(f_t) - \widetilde{\mathcal{E}}_S^t(f_t))\right]. \tag{3.9}$$

If $f = f_k$ with $1 \leq k \leq t$, by applying Corollary 3.5 to bound the last term of (3.9), and noting that $\theta \geq \frac{q}{q+1}$ implies

$$\frac{(1-\theta)(q+1) - 1}{2} - \theta = \frac{(1-\theta)q - 3\theta}{2} \geq (1-\theta)q - 2\theta,$$

we get (3.7) with

$$C_{q,\kappa,\eta_1} = 4\eta_1^2 c_q^2 \kappa^2 (\kappa+1)^{2q} + 8\eta_1 c_q \kappa (1 + \kappa^q). \tag{3.10}$$

If f is independent of $z_1, \cdots, z_t$, the last term of (3.9) is exactly

$$\mathbb{E}_{z_1,\cdots,z_{t-1}}\left[\mathcal{E}(f_t) - \widetilde{\mathcal{E}}_S^t(f_t)\right].$$

Using Corollary 3.5 to bound this term again, we get (3.7). From the above analysis, one can conclude the proof. □

3.4 Estimating excess generalization error

We now give the following general result, with which we can prove our main result, Theorem 2.3. For notational simplicity, we denote the excess generalization error of $f_* \in \mathcal{H}_K$ with respect to (ρ, V) by $\mathcal{A}(f_*)$:

$$\mathcal{A}(f_*) = \mathcal{E}(f_*) - \mathcal{E}(f_\rho^V). \tag{3.11}$$

Theorem 3.7. *Assume (2.3) with $q \geq 0$. Let $\eta_{t+1} = \eta_1 t^{-\theta}$ with $\frac{q}{q+1} \leq \theta < 1$ and η_1 satisfying (2.6). Then for every fixed $f_* \in \mathcal{H}_K$,*

$$\mathbb{E}_{z_1,\cdots,z_{T-1}} \left\{ \mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \right\} \leq \frac{\mathcal{A}(f_*)}{1-\theta} + \frac{\|f_*\|_K^2}{2\eta_1}(T-1)^{\theta-1} + \widetilde{C}_1 \Lambda_{T-1}, \tag{3.12}$$

where Λ_{T-1} is given by (2.7) and $\widetilde{C}_1$ is a positive constant depending on q, κ, θ (independent of T and f_*, and given explicitly in the proof).

The proof of this theorem is postponed to the next subsection. A novel error decomposition plays an important role in the proof. Note that the decomposition of ρ into the margin probability measure on X and the conditional probability measures allows the case with noise.

Now we are in a position to prove Theorem 2.3.

Proof of Theorem 2.3. By Theorem 3.7, we have

$$\mathbb{E}_{z_1,\cdots,z_{T-1}} \left\{ \mathcal{E}(f_T) - \mathcal{E}(f_\rho^V) \right\} \leq \widetilde{C}_0 \left(\mathcal{E}(f_*) - \mathcal{E}(f_\rho^V) + (T-1)^{\theta-1} \|f_*\|_K^2 \right) + \widetilde{C}_1 \Lambda_{T-1},$$

where

$$\widetilde{C}_0 = \frac{1}{1-\theta} + \frac{1}{2\eta_1}.$$

Since the constants $\widetilde{C}_0$ and $\widetilde{C}_1$ are independent of $f_* \in \mathcal{H}_K$, we can take infimum over $f_* \in \mathcal{H}_K$ on both sides, and conclude the desired result. □

3.5 Proof of Theorem 3.7

Before proving Theorem 3.7, we present two lemmas, whose proofs follow from Proposition 3.6 and some elementary inequalities. In the rest of this subsection, we denote $\mathbb{E}_{z_1,\cdots,z_T}$ by $\mathbb{E}$ for simplicity.

Lemma 3.8 (Weighted average)**.** *Under the assumption of Theorem 3.7, for any $T \geq 2$,*

$$\frac{1}{T-1} \sum_{t=2}^{T} 2\eta_t \mathbb{E} \left\{ \mathcal{E}(f_t) - \mathcal{E}(f_\rho^V) \right\} \leq \frac{\|f_*\|_K^2}{T-1} + \frac{2\eta_1 \mathcal{A}(f_*)}{1-\theta}(T-1)^{-\theta}$$
$$+ \begin{cases} \frac{q^* C_{q,\kappa,\eta_1}}{q^*-1}(T-1)^{-1}, & \text{when } \theta > \frac{q+2}{q+3}, \\ C_{q,\kappa,\eta_1}(T-1)^{-1}\log(eT), & \text{when } \theta = \frac{q+2}{q+3}, \\ \frac{C_{q,\kappa,\eta_1}}{1-q^*}(T-1)^{-q^*}, & \text{when } \theta < \frac{q+2}{q+3}. \end{cases}$$

Here q^* and C_{q,κ,η_1} are given by (3.8) and (3.10), respectively.

Proof. Note that by Proposition 3.6, we have (3.7). Choosing $f = f_*$ in (3.7) and adding both sides with $2\eta_t \mathcal{A}(f_*)$, we get

$$\begin{aligned} &2\eta_t \mathbb{E}\left[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right] \\ \leq &\mathbb{E}\left\{ \|f_t - f_*\|_K^2 - \|f_{t+1} - f_*\|_K^2 \right\} + C_{q,\kappa,\eta_1}(t-1)^{-q^*} + 2\eta_t \mathcal{A}(f_*), \end{aligned}$$

Taking summations over $t = 2, \dots, T$, with $f_2 = 0$, and $\eta_t = \eta_1 (t-1)^{-\theta}$,

$$\sum_{t=2}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right\} \le \|f_*\|_K^2 + C_{q,\kappa,\eta_1} \sum_{t=1}^{T-1} t^{-q^*} + 2\eta_1 \mathcal{A}(f_*) \sum_{t=1}^{T-1} t^{-\theta}.$$

Note that q^* is given by (3.8), and that from the restriction $\theta \in [\frac{q}{q+1}, 1)$, q^* satisfies $0 < q^* < 2$ and

$$q^* \begin{cases} > 1 & \text{when } \theta > \frac{q+2}{q+3}. \\ = 1 & \text{when } \theta = \frac{q+2}{q+3}, \\ < 1 & \text{when } \theta < \frac{q+2}{q+3}. \end{cases}$$

Applying

$$\sum_{t=1}^{T-1} t^{-\theta'} \le 1 + \int_1^{T-1} u^{-\theta'} du \le \begin{cases} \frac{(T-1)^{1-\theta'}}{1-\theta'}, & \text{when } \theta' < 1, \\ \log(eT), & \text{when } \theta' = 1, \\ \frac{\theta'}{\theta'-1}, & \text{when } \theta' > 1, \end{cases} \tag{3.13}$$

to bound $\sum_{t=1}^{T-1} t^{-q^*}$ and $\sum_{t=1}^{T-1} t^{-\theta}$, we get the desired result. The proof is complete. □

Lemma 3.9 (Moving weighted average)**.** *Under the assumption of Theorem 3.7, for any $T \ge 2$,*

$$\sum_{k=1}^{T-2} \frac{1}{k(k+1)} \sum_{t=T-k+1}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\right\}$$
$$\le \begin{cases} 2C_{q,\kappa,\eta_1}\left(2^{q^*} + \frac{q^*}{q^*-1}\right)(T-1)^{-1}, & \textit{when } \theta > \frac{q+2}{q+3}, \\ 4C_{q,\kappa,\eta_1}(\log T)(T-1)^{-1}, & \textit{when } \theta = \frac{q+2}{q+3}, \\ 2C_{q,\kappa,\eta_1}\left(2^{q^*} + \frac{1}{1-q^*}\right)(\log T)(T-1)^{-q^*}, & \textit{when } \theta < \frac{q+2}{q+3}. \end{cases}$$

Here q^* and C_{q,κ,η_1} are given by (3.8) and (3.10), respectively.

Proof. Let $k \in \{2, \dots, T-1\}$. Note that f_{T-k} depends only on $z_1, \cdots, z_{T-k-1}$. By Proposition 3.6, we have for $t \ge T-k$,

$$2\eta_t \mathbb{E}\left[\mathcal{E}(f_t) - \mathcal{E}(f)\right] \le \mathbb{E}\left\{\|f_t - f\|_K^2 - \|f_{t+1} - f\|_K^2\right\} + C_{q,\kappa,\eta_1}(t-1)^{-q^*}.$$

Taking summation over $t = T-k, \dots, T$ yields

$$\sum_{t=T-k+1}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\right\} = \sum_{t=T-k}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\right\}$$
$$\le C_{q,\kappa,\eta_1} \sum_{t=T-k}^{T} (t-1)^{-q^*} = C_{q,\kappa,\eta_1} \sum_{t=T-1-k}^{T-1} t^{-q^*}.$$

It thus follows that

$$\sum_{k=1}^{T-2}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\left\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\right\}\leq C_{q,\kappa,\eta_1}\sum_{k=1}^{T-2}\frac{1}{k(k+1)}\sum_{t=T-1-k}^{T-1}t^{-q^*}.$$

By applying the following elementary inequality from [16] (which will be proved in the appendix for completeness)

$$\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}t^{-q^*}\leq\begin{cases}2\left(2^{q^*}+\frac{q^*}{q^*-1}\right)T^{-1}, & \text{when } q^*\in(1,2),\\ 4(\log T)T^{-1}, & \text{when } q^*=1,\\ 2\left(2^{q^*}+\frac{1}{1-q^*}\right)(\log T)T^{-q^*}, & \text{when } q^*\in(0,1),\end{cases}\tag{3.14}$$

the desired estimate is verified. The proof is complete. □

With the above two lemmas, now we are ready to prove Theorem 3.7.

Proof of Theorem 3.7. The basic idea is to bound the weighted excess generalization error $2\eta_T\mathbb{E}_{z_1,\cdots,z_{T-1}}[\mathcal{E}(f_T)-\mathbb{E}(f_\rho^V)]$ in terms of the weighted average and the moving weighted average. To do so, we need the following fact from [22, 18] which asserts that for any sequence $\{u_j\}_{j\in\mathbb{N}}$ in $\mathbb{R}$, there holds

$$u_T=\frac{1}{T-1}\sum_{j=2}^{T}u_j+\sum_{k=1}^{T-2}\frac{1}{k(k+1)}\sum_{j=T-k+1}^{T}(u_j-u_{T-k}).\tag{3.15}$$

In fact, for $k\in\{1,\cdots,T-2\}$, we have

$$\begin{aligned}&\frac{1}{k}\sum_{j=T-k+1}^{T}u_j-\frac{1}{k+1}\sum_{j=T-k}^{T}u_j\\ =&\frac{1}{k(k+1)}\left\{(k+1)\sum_{j=T-k+1}^{T}u_j-k\sum_{j=T-k}^{T}u_j\right\}\\ =&\frac{1}{k(k+1)}\sum_{j=T-k+1}^{T}(u_j-u_{T-k}).\end{aligned}$$

Summing over $k=2,\cdots,T-1$, and rearranging terms, we get (3.15). Now, for any $k=1,\dots,T-2$, we choose $u_t=2\eta_t\mathbb{E}\left\{\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\right\}$ in (3.15) to get

$$\begin{aligned}&2\eta_T\mathbb{E}\left\{\mathcal{E}(f_T)-\mathcal{E}(f_\rho^V)\right\}=\frac{1}{T-1}\sum_{j=2}^{T}2\eta_j\mathbb{E}\left\{\mathcal{E}(f_j)-\mathcal{E}(f_\rho^V)\right\}\\ &+\sum_{k=1}^{T-2}\frac{1}{k(k+1)}\sum_{j=T-k+1}^{T}\left(2\eta_j\mathbb{E}\left\{\mathcal{E}(f_j)-\mathcal{E}(f_\rho^V)\right\}-2\eta_{T-k}\mathbb{E}\left\{\mathcal{E}(f_{T-k})-\mathcal{E}(f_\rho^V)\right\}\right),\end{aligned}$$

which can be rewritten as

$$2\eta_T \mathbb{E}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} = \frac{1}{T-1}\sum_{t=2}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right\}$$
$$+\sum_{k=1}^{T-2}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\right\}$$
$$+\sum_{k=1}^{T-2}\frac{1}{k+1}\left[\frac{1}{k}\sum_{t=T-k+1}^{T} 2\eta_t - 2\eta_{T-k}\right]\mathbb{E}\left\{\mathcal{E}(f_{T-k}) - \mathcal{E}(f_\rho^V)\right\}. \tag{3.16}$$

Since, $\mathcal{E}(f_{T-k}) - \mathcal{E}(f_\rho^V) \geq 0$ and that $\{\eta_t\}_{t\in\mathbb{N}}$ is a non-increasing sequence, we know that the last term of the above inequality is at most zero. Therefore, we get

$$\begin{aligned}2\eta_T \mathbb{E}\left\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\right\} &\leq \frac{1}{T-1}\sum_{t=2}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right\} \\ &+\sum_{k=1}^{T-2}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T} 2\eta_t \mathbb{E}\left\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\right\}.\end{aligned} \tag{3.17}$$

Applying lemmas 3.8 and 3.9 to bound the last two terms, we get the desired bound (3.12) with $\widetilde{C}_1$ given explicitly by

$$\widetilde{C}_1 = \begin{cases} \frac{C_{q,\kappa,\eta_1}(3q^*+2^{q^*+1}(q^*-1))}{2\eta_1(q^*-1)}, & \text{when } \theta > \frac{q+2}{q+3}, \\ \frac{3C_{q,\kappa,\eta_1}}{\eta_1}, & \text{when } \theta = \frac{q+2}{q+3}, \\ \frac{C_{q,\kappa,\eta_1}\left(2^{q^*+1}+\frac{3}{1-q^*}\right)}{2\eta_1}, & \text{when } \theta < \frac{q+2}{q+3}. \end{cases}$$

The proof of Theorem 3.7 is complete. □

4 Conclusion

This paper presents learning rates of the last iterate for online pairwise learning algorithms involving general convex loss functions which are better than the existing results under certain circumstances. Our idea is to use an error decomposition from [16, 23] to decompose the weighted excess generalization error into weighted average errors and moving weighted average errors. We apply some tools from Rademacher complexity to overcome the difficulty with the bias of the randomized gradients as estimators of the true gradients in the online pairwise learning setting. It is interesting to discuss here the connection between classification/regression tasks and pairwise learning problems. For the specific pairwise learning problem with $V(y,f) = (y-f)^2$ and $r(y,y') = y - y'$, it was proved in [32, 10] that the optimal predictor is $f_\rho^V(x,x') = \int_X y d\rho(y|x) - \int_X y d\rho(y|x')$, where $\rho(y|x)$ is the conditional measure at x. This shows that the pairwise learning based on the least squares loss is essentially a pointwise learning problem since $\tilde{f}_\rho(x) := \int_X y d\rho(y|x)$ is the regression function minimizing $\int_Z (y-f(x))^2 d\rho$. Characterizing f_ρ^V and the approximation error assumption (2.8) for a general pairwise learning loss function in terms of function space properties, such as for metric and similarity learning, is a challenging problem for further study.

References

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

[2] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[3] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

[4] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[5] Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132, 2016.

[6] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[7] Andreas Christmann and Ding-Xuan Zhou. On the robustness of regularized pairwise learning methods based on kernels, *Journal of Complexity*, 37:1–33, 2016.

[8] Stéphan Clémençon, Gabor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pages 844–874, 2008.

[9] Felipe Cucker and Ding-Xuan Zhou. Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, 2007.

[10] Jun Fan, Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and Computational Harmonic Analysis*, 41(1):164–189, 2016.

[11] Zheng-Chu Guo, Dao-Hong Xiang, Xin Guo, and D. X. Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, in press.

[12] Zheng-Chu Guo, Yiming Ying, and Ding-Xuan Zhou. Online regularized learning with pairwise loss functions. *Advances in Computational Mathematics*, pages 1–24, 2016.

[13] David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[14] Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13(04):437–455, 2015.

[15] Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *Proceedings of The 30th International Conference on Machine Learning*, pages 441–449, 2013.

[16] Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17(77):1–38, 2016.

[17] Junhong Lin and Ding-Xuan Zhou. Learning theory of randomized kaczmarz algorithm. *Journal of Machine Learning Research*, 16:3341–3365, 2015.

[18] Junhong Lin and Ding-Xuan Zhou. Online learning algorithms can converge comparably fast as batch learning. *IEEE Transactions on Neural Networks and Learning Systems*, To appear, 2017.

[19] Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

[20] Sayan Mukherjee and Qiang Wu. Estimation of gradients and coordinate covariation in classification. *The Journal of Machine Learning Research*, 7:2481–2514, 2006.

[21] Wojciech Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(1):1373–1392, 2012.

[22] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79, 2013.

[23] Lei Shi, Yunlong Feng, and Ding-Xuan Zhou. Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31:286–302, 2011.

[24] Steven Smale and Yuan Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6(2):145–170, 2006.

[25] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science Business Media, 2008.

[26] Vladimir Vapnik. Statistical Learning Theory. John Wiley & Sons, New York, 1998.

[27] Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *COLT*, volume 23, pages 13–1, 2012.

[28] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2005.

[29] Qiang Wu and Ding-Xuan Zhou. Learning with sample dependent hypothesis spaces. *Computers and Mathematics with Applications*, 56:2896-2907, 2008.

[30] Yiming Ying, Qiang Wu, and Colin Campbell. Learning the coordinate gradients. *Advances in Computational Mathematics*, 37(3):355–378, 2012.

[31] Yiming Ying and Ding Xuan Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.

[32] Yiming Ying and Ding-Xuan Zhou. Online pairwise learning algorithms. *Neural computation*, 28(4):743–777, 2016.

[33] Yiming Ying and Ding-Xuan Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[34] Peilin Zhao, Rong Jin, Tianbao Yang, and Steven C Hoi. Online auc maximization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 233–240, 2011.

A Appendix for Proving (3.14)

First note that

$$\sum_{t=T-k+1}^{T} t^{-q^*} \leq \int_{T-k}^{T} x^{-q^*} dx \leq \frac{T^{1-q^*} - (T-k)^{1-q^*}}{1-q^*}, \quad \text{when } q^* \neq 1.$$

When $0 < q^* < 1$, for $k \leq \frac{T}{2}$,

$$\sum_{t=T-k}^{T} t^{-q^*} \leq (T-k)^{-q^*}(k+1) \leq 2^{q^*} T^{-q^*}(k+1),$$

and for $k > \frac{T}{2}$

$$\sum_{t=T-k}^{T} t^{-q^*} \leq \frac{T^{1-q^*} - (T-k)^{1-q^*}}{1-q^*} + (T-k)^{-q^*} \leq \frac{T^{1-q^*}}{1-q^*}.$$

It thus follows that

$$\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*}$$

$$
\begin{aligned}
&\le \sum_{k\le T/2} \frac{1}{k(k+1)} 2^{q^*} T^{-q^*}(k+1) + \sum_{T-1\ge k>T/2} \frac{1}{k(k+1)} \frac{T^{1-q^*}}{1-q^*} \\
&\le \left(2^{q^*+1} + \frac{2}{1-q^*}\right)(\log T) T^{-q^*}.
\end{aligned}
$$

When $q^* = 1$, we have

$$
\begin{aligned}
&\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \le \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \frac{k+1}{T-k} = \frac{1}{T} \sum_{k=1}^{T-1} \left\{ \frac{1}{k} + \frac{1}{T-k} \right\} \\
&\le 4(\log T) T^{-1}.
\end{aligned}
$$

When $2 > q^* > 1$, for $k \le \frac{T}{2}$,

$$
\sum_{t=T-k}^{T} t^{-q^*} \le (T-k)^{-q^*}(k+1) \le 2^{q^*} T^{-q^*}(k+1),
$$

and for $k > \frac{T}{2}$

$$
\sum_{t=T-k}^{T} t^{-q^*} \le \frac{(T-k)^{1-q^*} - T^{1-q^*}}{q^*-1} + (T-k)^{-q^*} \le \frac{q^*}{q^*-1}.
$$

Therefore, we have

$$
\begin{aligned}
&\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \\
&\le 2^{q^*} T^{-q^*} \sum_{k\le T/2} \frac{1}{k} + \frac{q^*}{q^*-1} \sum_{T-1\ge k>T/2} \frac{1}{k(k+1)} \\
&\le 2^{q^*+1} T^{-q^*} \log T + \frac{2q^*}{q^*-1} T^{-1} \\
&\le \frac{2^{q^*+1} + 2q^*}{q^*-1} T^{-1}.
\end{aligned}
$$