Pairwise Learning with Adaptive Online Gradient Descent

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we propose an adaptive online gradient descent method with momentum for pairwise learning, in which the stepsize is determined by historical information. Due to the structure of pairwise learning, the sample pairs are dependent on the parameters, causing difficulties in the convergence analysis. To this end, we develop novel techniques for the convergence analysis of the proposed algorithm. We show that the proposed algorithm can output the desired solution in strongly convex, convex, and nonconvex cases. Furthermore, we present theoretical explanations for why our proposed algorithm can accelerate previous workhorses for online pairwise learning. All assumptions used in the theoretical analysis are mild and common, making our results applicable to various pairwise learning problems. To demonstrate the efficiency of our algorithm, we compare the proposed adaptive method with the non-adaptive counterpart on the benchmark online AUC maximization problem.

1 Introduction

Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a closed convex set (can be the full space $\mathbb{R}^d$) representing the parameter space. Given a statistical sample space Ξ with probability distribution $\mathcal{P}$; let $F(\cdot;\xi,\xi') : \mathbb{R}^d \to \mathbb{R}$ be a closed function associated with two samples $\xi, \xi' \in \Xi$. This paper considers the following pairwise learning problem

$$\min_{\boldsymbol{x}\in\mathcal{K}\subseteq\mathbb{R}^d} \Big\{ f(\boldsymbol{x}) := \mathbb{E}_{(\xi,\xi')\sim\mathcal{P}\oplus\mathcal{P}} F(\boldsymbol{x};\xi,\xi') \Big\}, \tag{1}$$

where the function $F(\boldsymbol{x};\xi,\xi')$ can be either convex or nonconvex in $\boldsymbol{x}$. The pairwise learning model equation 1 describes various classical machine learning tasks, including the metric learning Weinberger & Saul (2009); Kulis et al. (2013); Xing et al. (2002); Ying & Li (2012), the area under ROC curve (AUC) maximization Zhao et al. (2011); Gao et al. (2013); Ying et al. (2016); Liu et al. (2018), ranking Rejchel (2012); Agarwal & Niyogi (2009), two-stage multiple kernel learning Kumar et al. (2012), neural link prediction Wang et al. (2021), and the minimum error entropy principle Hu et al. (2013).

There are two major kinds of workhorses for model equation 1, i.e., *offline* and *online*. The offline one is similar to the empirical risk minimization (ERM): given an i.i.d. sample set $\{\xi^1,\ldots,\xi^n\}$, we solve

$$\min_{\boldsymbol{x}\in\mathcal{K}} \frac{1}{n(n-1)} \sum_{i,j\in[n],i\neq j}^{n} F(\boldsymbol{x};\xi^i,\xi^j),$$

where $[n] := \{1,2,\ldots,n\}$. The major difference between the offline pairwise learning model and the ERM lies in the efficiency of the samples and whether the objective functions are independent of each other. A n-samples training set outputs a finite-sum minimization with $\mathcal{O}(n)$ sub-functions in ERM, while the same training set results in $\mathcal{O}(n^2)$ in the offline pairwise learning. Furthermore, the objective functions in the ERM are independent of each other, which is broken for the offline pairwise learning[1]. The online pairwise learning assumes the i.i.d. samples $(\xi^k)_{k\in[n]}$ are continuously received by the model. In the kth iteration, the online style algorithm proceeds to sample new data ξ^k from $\mathcal{P}$ and reuses the previous samples $(\xi^{j_i})_{1\le i\le s}$ with $\{j_i\}_{1\le i\le s} \subseteq [k-1]$ to get the mini-batch stochastic gradient $\boldsymbol{g}^k = \frac{1}{s}\sum_{i=1}^{s} \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{j_i})$ Wang et al.

[1]For example, $F(\boldsymbol{x};\xi^1,\xi^2)$ and $F(\boldsymbol{x};\xi^1,\xi^3)$ are not independent because they have a shared data ξ^1.

Algorithm 1 Online Gradient Descent (OGD) for Pairwise Learning Yang et al. (2021b)

Parameters: $\eta > 0$.
Initialization: $\boldsymbol{x}^0 = \mathbf{0}$, $\xi^1 \sim \mathcal{P}$
for $k = 1, 2, 3, \ldots$
step 1: receive ξ^k and calculate $\boldsymbol{g}^k = \nabla F(\boldsymbol{x}^k; \xi^k, \xi^{k-1})$
step 2: $\boldsymbol{x}^{k+1} = \mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \eta \boldsymbol{g}^k)$
End for

(2012); Zhao et al. (2011); Ying & Zhou (2016). Thus, the online method needs to employ a $\mathcal{O}(s)$ memory to store the previously sampled data and $\mathcal{O}(s)$ computations to calculate the gradient. Nevertheless, the mini-batch version suffers two drawbacks: 1) It has been proved that its excess generalization bound can be as large as $\mathcal{O}\Big(\frac{1}{\sqrt{s}} + \frac{1}{\sqrt{n}}\Big)$ Wang et al. (2012), and we need to use a large s to improve generalization of the online method. 2) A large s causes tremendous computational and memory costs that are unacceptable for online settings. To this end, a simple yet efficient online gradient descent (OGD) is proposed (presented as Algorithm 1), in which the stochastic gradient $\boldsymbol{g}^k$ is set to be $\nabla F\big(\boldsymbol{x}^k; \xi^k, \xi^{k-1}\big)$ Yang et al. (2021b). An interesting finding is that the OGD can achieve $\mathcal{O}\big(\frac{1}{\sqrt{n}}\big)$ excess generalization bound. The favorable memory and computation costs make the OGD applicable to broader online settings.

In this paper, we focus on developing provably convergent online algorithms with adaptive stepsizes for pairwise learning. Considering the efficiency of the sampling method of OGD Yang et al. (2021b), our algorithm inherits such kind of sampling. Moreover, our algorithm employs momentum.

Algorithm 2 Adaptive Online Gradient Descent (AOGD) for Pairwise Learning

Parameters: $\eta > 0$, $0 \leq \theta < 1$.
Initialization: $\boldsymbol{x}^0 = \boldsymbol{m}^0 = \mathbf{0}$, $\xi^1 \sim \mathcal{P}$
for $k = 1, 2, 3, \ldots$
step 1: receive ξ^k and calculate $\boldsymbol{g}^k = \nabla F(\boldsymbol{x}^k; \xi^k, \xi^{k-1})$
step 2: $\boldsymbol{m}^k = \theta \boldsymbol{m}^{k-1} + (1-\theta)\boldsymbol{g}^k$
step 3: $v^k = v^{k-1} + \|\boldsymbol{g}^k\|^2$
step 4: $\boldsymbol{x}^{k+1} = \mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \eta \boldsymbol{m}^k/(v^k)^{\frac{1}{2}})$
End for

1.1 The Adaptive Online Gradient Descent for Pairwise Learning

The OGD performs an SGD-style (stochastic gradient descent) iterations but with biased stochastic gradients since ξ^{k-1} is related to $\boldsymbol{x}^k$. Motivated by the remarkable success of the adaptive variants of SGD for machine learning Duchi et al. (2011); McMahan & Streeter (2010); Tieleman & Hinton (2012); Kingma & Ba (2015); Reddi et al. (2018); Ward et al. (2019), we propose the adaptive variant of OGD (AOGD) for pairwise learning, presented as Algorithm 2. AOGD directly uses the historical sum of the moment rather than the weighted average form used in Adam Kingma & Ba (2015), and AOGD can be rewritten in the weighted average form: let $\hat{v}^k := \sum_{i=1}^k \|\boldsymbol{g}^i\|^2/k = v^k/k$, then steps 3 and 4 of AOGD can be reformulated as

$$\begin{aligned} \hat{v}^k &= \left(1 - \frac{1}{k}\right)\hat{v}^{k-1} + \frac{1}{k}\|\boldsymbol{g}^k\|^2, \\ \boldsymbol{x}^{k+1} &= \mathbf{Proj}_{\mathcal{K}}\left(\boldsymbol{x}^k - \frac{\eta}{\sqrt{k}}\frac{\boldsymbol{m}^k}{(\hat{v}^k)^{1/2}}\right). \end{aligned} \tag{2}$$

In reformulation equation 2, weights of the moment are $\big\{1/k\big\}_{k\geq 1}$ and stepsizes are $\big\{\eta/\sqrt{k}\big\}_{k\geq 1}$, satisfying the sufficient conditions for the convergence of Adam-type algorithms Zou et al. (2019); Chen et al. (2019).

Compared with OGD, AOGD employs the momentum and adaptive stepsize generated by the historical information. Thus unlike OGD, AOGD needs a memory of history, but without much computational overhead since AOGD only computes gradient once in each iteration. Another difference between OGD and AOGD

lies in the hyper-parameter η: in OGD, η shall be set as small as the desired error ϵ; while in AOGD, η can be set as a constant that is independent of ϵ.

1.2 Comparison with A Closely Related Work

In Ding et al. (2015), by adopting the AdaGrad-style adaptive gradient update Duchi et al. (2011), the authors have proposed an adaptive method for the online AUC maximization, which is a kind of pairwise learning. Although both Ding et al. (2015) and our paper consider the adaptive online method for pairwise learning, there are four major differences between Ding et al. (2015) and our paper, summarized below.

1) We follow the sampling method used by the simple yet efficient OGD algorithm in Yang et al. (2021b).

2) We consider more general cases and provide the corresponding theoretical guarantees, including the more general model (pairwise learning rather than only AUC maximization), and convergence for more general settings, including the nonconvex case, and more general schemes e.g., the use of momentum.

3) We get rid of using the regret bound because it does not directly tell us whether the algorithm converges to the desired minimizer. Another important reason for not using the regret bound for the analysis is that the regret bound has difficulties in covering the nonconvex cases. Based on the above reasons, a non-regret analysis is necessary.

4) We develop new analysis techniques to get the non-regret convergence analysis. Notice that the regret bound is not affected by the stochasticity of the data, and thus the analysis in Ding et al. (2015) does not need to consider how to deal with the biased stochastic gradients.

1.3 Challenges in the Analysis and Difference from Existing Analysis

The primary source of challenges in theoretical analysis comes from the fact that $\boldsymbol{x}^k$ *is dependent on* ξ^{k-1}, which immediately breaks the unbiased expectation of the stochastic gradient $\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})$, i.e., $\mathbb{E}\big[\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})\big] \neq \nabla f(\boldsymbol{x}^k)$. As such, we cannot directly follow the techniques from adaptive SGD Duchi et al. (2011); Reddi et al. (2018); Chen et al. (2019); Ward et al. (2019); Zou et al. (2019); Li & Orabona (2019). In paper Yang et al. (2021b), the authors consider the following decomposition

$$\eta\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) = \eta\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) + \eta\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \eta\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}).$$

Notice that $\boldsymbol{x}^{k-1}$ is independent of ξ^k,ξ^{k-1}, and $\mathbb{E}\big[\eta\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\big] = \eta\nabla f(\boldsymbol{x}^{k-1})$. The Lipschitz property of the gradient then gives us

$$\Big\|\eta\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \eta\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\Big\| = \mathcal{O}(\eta\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|),$$

and the proof is similar to the "delayed" SGD.

However, our proof cannot directly follow the technique above because AOGD involves two extra recipes, i.e., momentum and adaptive stepsize. When momentum exists, we need to deal with both $\boldsymbol{g}^k$ and $\boldsymbol{m}^k$ rather than only $\boldsymbol{g}^k$, and we have to modify the techniques from Yang et al. (2021b) to analyze the effects of momentum [2]. Another difficulty is the use of the adaptive stepsize variable $\eta/(v^k)^{\frac{1}{2}}$. Furthermore, v^k is dependent of the pair (ξ^k,ξ^{k-1}), making the analysis more challenging.

In contrast to the existing analysis, our approach is not directly establishing the Lyapunov descent starting from bounding $\mathbb{E}[f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k)]$. Instead, we recruit some intermediate variables. Taking the nonconvex case as an example, we first establish a *Lyapunov-like* descent property as follows

$$\mathbb{E}\langle -\nabla f(\boldsymbol{x}^k), \boldsymbol{m}^k/(v^k)^{\frac{1}{2}}\rangle \leq \theta\mathbb{E}\langle -\nabla f(\boldsymbol{x}^{k-1}), \boldsymbol{m}^{k-1}/(v^{k-1})^{\frac{1}{2}}\rangle + \varphi(\boldsymbol{x}^k,\boldsymbol{x}^{k-1},\xi^k,\xi^{k-1},\xi^{k-2}),$$

[2] Indeed, in the conclusion part of Yang et al. (2021b), the authors have listed the momentum variant as future work.

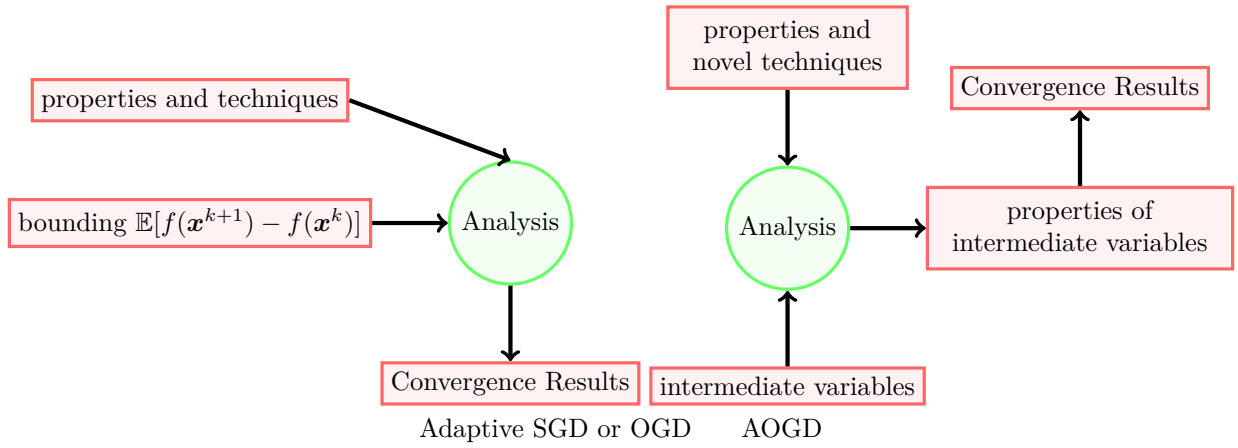

Figure 1: Contrasting the analysis of AOGD against adaptive SGD/OGD. Our analysis of AOGD studies the properties of some intermediate variables. To this end, we develop novel techniques. In the last step, we use the properties of the intermediate variables to derive the convergence guarantee for AOGD.

where $\varphi(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}, \xi^k, \xi^{k-1}, \xi^{k-2})$ is a function of variables $\boldsymbol{x}^k$, $\boldsymbol{x}^{k-1}$, ξ^k, ξ^{k-1}, and ξ^{k-2}. We stress that the descent is not Lyapunov because $\theta \neq 1$ and $\varphi(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}, \xi^k, \xi^{k-1}, \xi^{k-2})$ is not always negative. Then, we build the correspondence between the mathematical convergence measurement and $\mathbb{E}\big[\langle -\nabla f(\boldsymbol{x}^k), \boldsymbol{m}^k/(v^k)^{\frac{1}{2}}\rangle\big]$. A big picture of the difference in analyzing adaptive SGD/OGD and AOGD is presented in Figure 1.

1.4 Contributions

Our major contributions are threefold, which are summarized below.

- We propose an adaptive online gradient descent algorithm for pairwise learning with a simple sampling strategy. The proposed algorithm uses adaptive stepsize and momentum, requiring only a small overhead in memory and computational costs.
- We present the convergence results for the proposed algorithm under different settings, including strongly convex, general convex, and nonconvex cases. The use of adaptive stepsize and momentum requires non-trivial techniques for the convergence analysis. We also provide theoretical explanations for why our proposed algorithm can accelerate OGD.
- We verify the efficiency of the proposed AOGD on the benchmark online AUC maximization task, showing that AOGD outperforms OGD.

1.5 Notation

Throughout this paper, we use bold face letters to denote vectors, e.g., $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. The jth coordinate of the vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}_j$. The L_2 norm of the vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|$. We denote $\mathbb{E}[\cdot]$ as the expectation with respect to the underlying probability space. We denote the minimum value of the function f over $\mathcal{K}$ as $\min_{\mathcal{K}} f$. For two positive sequences $(a_k, b_k)_{k\geq 0}$, $a_k = \mathcal{O}(b_k)$ means that there exists $C > 0$ such that $a_k \leq C b_k$. The notation $a_k = \Theta(b_k)$ indicates that $a_k = \mathcal{O}(b_k)$ and $b_k = \mathcal{O}(a_k)$. We use $a_k = \tilde{\mathcal{O}}(b_k)$ and $a_k = \tilde{\Theta}(b_k)$ to hide the logarithmic factor but still with the same order. We use $a_k \geq \Theta(b_k)$ to present the relation $a_k \geq C b_k$ with $C > 0$.

1.6 Organization

We organize this paper as follows: In Section 2, we present assumptions and theoretical convergence results for AOGD under general convex, strongly convex, and nonconvex settings. We numerically verify the efficiency

of AOGD and compare it to the benchmark OGD in Section 3. More related works are discussed in Section 4, followed by concluding remarks. The detailed proofs are provided in the supplementary materials.

2 Convergence Analysis

2.1 Assumptions

We first collect several necessary assumptions for the convergence analysis of AOGD, which are presented below.

- **Assumption 1**: *The function $F(\cdot;\xi,\xi')$ is differentiable and Lipschitz with constant $L>0$, i.e.,*

$$\|\nabla F(\boldsymbol{x};\xi,\xi')-\nabla F(\boldsymbol{y};\xi,\xi')\|\leq L\|\boldsymbol{x}-\boldsymbol{y}\|, \tag{3}$$

  $\forall \boldsymbol{x},\boldsymbol{y}\in\mathcal{K},\xi,\xi'\in\Xi$.

- **Assumption 2**: *The gradient of $F(\boldsymbol{x};\xi,\xi')$ is uniformly bounded, i.e., $\|\nabla F(\boldsymbol{x};\xi,\xi')\|\leq B$ for some constant $B>0$, $\forall \boldsymbol{x}\in\mathcal{K}$, and $\xi,\xi'\in\Xi$.*

The Lipschitz smooth gradient assumption is widely used in the (non)convex optimization and pairwise learning communities. While Assumption 2 is frequently used in the adaptive SGD community, see, e.g., Duchi et al. (2011); Reddi et al. (2018); Chen et al. (2019); Ward et al. (2019); Zou et al. (2019); Li & Orabona (2019) [3]. Note that when $\mathcal{K}$ is bounded — Duchi et al. (2011); Reddi et al. (2018) have assumed the boundedness of the constrained set — Assumption 2 directly holds for the continuity of the gradient [4], but not vice versa. Moreover, Assumption 2 indicates the following estimate of $\nabla f(\boldsymbol{x})$

$$\|\nabla f(\boldsymbol{x})\|=\|\mathbb{E}_{\xi,\xi'\sim\mathcal{P}\oplus\mathcal{P}}\nabla F(\boldsymbol{x};\xi,\xi')\|\leq\mathbb{E}_{\xi,\xi'\sim\mathcal{P}\oplus\mathcal{P}}\|\nabla F(\boldsymbol{x};\xi,\xi')\|\leq B.$$

Using mathematical induction, we can see that the momentum $\boldsymbol{m}^k$ also enjoys the uniform bound according to Assumption 2. Furthermore, we stress that we do not need to assume the variance is bounded, which is indicated by Assumption 2 since we have

$$\mathbb{E}\|\nabla F(\boldsymbol{x};\xi,\xi')-\nabla f(\boldsymbol{x})\|^2\leq\mathbb{E}\|\nabla F(\boldsymbol{x};\xi,\xi')\|^2\leq B^2.$$

Assumptions 1 and 2 will be used in the analysis of AOGD for all different scenarios in the subsequent analysis.

2.2 General Convex Cases

In this subsection, we present the convergence result of AOGD for the general convex case, i.e., $F(\boldsymbol{x};\xi,\xi')$ is convex with respect to $\boldsymbol{x}$ and any fixed ξ,ξ'. Note that the convexity of $F(\boldsymbol{x};\xi,\xi')$ indicates the convexity of $f(\boldsymbol{x})$ in equation 1, but not vice versa.

Theorem 1 *(General Convexity) Let Assumption 1 hold, and $\|\boldsymbol{g}^0\|^2\geq\delta>0$ for some constant δ, and $F(\boldsymbol{x};\xi,\xi')$ be convex. Assume $\{\boldsymbol{x}^k\}_{k\geq1}$ is generated by AOGD for pairwise learning, $\boldsymbol{x}^*:=\arg\min_{\boldsymbol{x}\in\mathcal{K}}f(\boldsymbol{x})$, and $\mathcal{K}$ is additionally bounded, i.e., $\max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{K}}\|\boldsymbol{x}-\boldsymbol{y}\|\leq D$. Then we have*

$$\mathbb{E}\left[f\left(\frac{\sum_{k=1}^K\boldsymbol{x}^k}{K}\right)-\min_{\mathcal{K}}f\right]\leq\frac{c_1+c_2(K)}{K}, \tag{4}$$

where $c_1:=\frac{\mathbb{E}\sqrt{v^1}\|\boldsymbol{x}^*-\boldsymbol{x}^1\|^2}{2(1-\theta)\eta}+\frac{3B^2}{\sqrt{\delta}}$, *and* $c_2(K):=\left[\frac{\eta+2D^2}{2(1-\theta)}+(\eta\theta+1)\right]\cdot\mathbb{E}\sqrt{v^{K+1}}+\left[4L+\frac{2L^2}{\sqrt{\delta}}\right]\ln\frac{\mathbb{E}\sqrt{v^{K+1}}}{\sqrt{\delta}}$.

[3] The uniform assumption presented in Li & Orabona (2019) enjoys another presentation, i.e., (Assumption (H2)) presented as $|f(\boldsymbol{x})-f(\boldsymbol{y})|\leq G\|\boldsymbol{x}-\boldsymbol{y}\|$. Note that when f is differentiable, it is equivalent to $\sup_{\boldsymbol{x}}\|\nabla f(\boldsymbol{x})\|\leq G$. With bounded assumption for $F(\boldsymbol{x};\xi)-\nabla f(\boldsymbol{x})$ in Li & Orabona (2019), which is indeed the uniform bound assumption.

[4] Any continuous function is uniformly bounded over a closed bounded subset of $\mathbb{R}^d$.

The Assumption 2 is implicitly used in Theorem 1 because we have assumed the boundedness of the constrained set $\mathcal{K}$. The bounded constrained set is indeed stronger than the uniform bounded gradient assumption. We leave how to relax the bounded set assumption as future work. From Theorem 1, we can see that the convergence rate is dependent on $\mathbb{E}[\sqrt{v^K}]$. The boundedness of the stochastic gradients indicates that $\mathbb{E}[\sqrt{v^K}] = \mathcal{O}(\sqrt{K})$, which means the worst convergence rate of AOGD is $\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{K}}\Big)$. We notice that the convergence rate $\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{K}}\Big)$ coincides with the rate of OGD for pairwise learning in the general convex case. However, in some cases $\mathbb{E}[\sqrt{v^K}]$ can decay faster than $\mathcal{O}(\sqrt{K})$, based on which we can establish an accelerated rate of AOGD.

Proposition 1 *Assume the conditions of Theorem 1 hold, and assume* $\mathbb{E}[\sqrt{v^K}] = \mathcal{O}(K^\alpha)$ *with* $0 < \alpha \leq \frac{1}{2}$. *Then we have*

$$\mathbb{E}\left[f\left(\frac{\sum_{k=1}^K \boldsymbol{x}^k}{K}\right) - \min_{\mathcal{K}} f\right] = \widetilde{\mathcal{O}}\left(\frac{1}{K^{1-\alpha}}\right). \tag{5}$$

If $\alpha < \frac{1}{2}$, the convergence rate of AOGD is thus better than $\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{K}}\Big)$. Proposition 1 then provides a theoretical interpretation of why AOGD is possible to be faster than OGD.

Remark 1 *Note that* $\alpha = 1/2$ *directly holds for the boundedness of the stochastic gradients. Indeed, this fast decaying condition is standard for analyzing adaptive stochastic optimization algorithms, see, e.g., Reddi et al. (2018); Chen et al. (2018; 2019); Liu et al. (2019).*

2.3 Strongly Convex Cases

In this subsection, we consider the case that the function $F(\boldsymbol{x};\xi,\xi')$ is ν-strongly convex for some constant $\nu > 0$, i.e., $F(\boldsymbol{x};\xi,\xi') - F(\boldsymbol{y};\xi,\xi') - \langle \nabla F(\boldsymbol{y};\xi,\xi'), \boldsymbol{x}-\boldsymbol{y}\rangle \geq \frac{\nu}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2$. In particular, if $\nu = 0$, $F(\boldsymbol{x};\xi,\xi')$ then reduces to general convex.

Stepsize rule for the strongly convex case. Before developing a convergence guarantee of AOGD for strongly convex cases, we need to select the appropriate stepsize rule for AOGD. We first need to explain the stepsize rule of AOGD for the general convex case, i.e., $\eta/\sqrt{v^k}$ for some constant $\eta > 0$, is inappropriate for solving strongly convex problems. The boundedness of gradient directly gives us $\frac{1}{\sqrt{v^k}} \geq \Theta\Big(\frac{1}{\sqrt{k}}\Big)$. However, such a stepsize rule causes the accumulation of stochastic noise; it will not improve the convergence rate of AOGD under strong convexity compared to the rate of AOGD under general convexity. A proper stepsize choice for the strongly convex case is $\Theta\Big(\frac{1}{k}\Big)$ Bach & Moulines (2013). Indeed, in papers Duchi et al. (2011); Sun et al. (2020), the authors use the $\frac{1}{\sqrt{k}\sqrt{v^k}}$ stepsize rule for AdaGrad when the problem is strongly convex, and we follow this stepsize rule for the strongly convex online pairwise learning. Note that $\frac{1}{\sqrt{k}\sqrt{v^k}} \geq \Theta\Big(\frac{1}{k}\Big)$, which coincides with the stepsize used for SGD when the underlying problem is strongly convex. In summary, in the strongly convex case, we replace **step 4** of Algorithm 2 with **step 4'**, which is given below

$$\textbf{step 4'}: \ \boldsymbol{x}^{k+1} = \mathbf{Proj}_{\mathcal{K}}\left(\boldsymbol{x}^k - \frac{\eta}{\sqrt{k}}\boldsymbol{m}^k/\sqrt{v^k}\right). \tag{6}$$

Next, we present the convergence rate of AOGD for pairwise learning with the strong convexity assumption.

Theorem 2 *Let Assumption 1 hold, and* $\|\boldsymbol{g}^0\|^2 \geq \delta > 0$, *and* $F(\boldsymbol{x};\xi,\xi')$ *be strongly convex. Assume* $\{\boldsymbol{x}^k\}_{k\geq 1}$ *is generated by the AOGD for pairwise learning with* $\theta = 0$ *using stepsize rule* **step 4'**, *and* $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathcal{K}} f(\boldsymbol{x})$. *By setting* $\eta = \frac{B}{2\nu}$, *then we have*

$$\mathbb{E}\big[\|\boldsymbol{x}^K - \boldsymbol{x}^*\|^2\big] = \mathcal{O}\left(\frac{\ln K}{K}\right). \tag{7}$$

In the strongly convex case, Assumption 2 is equivalent to the boundedness assumption of the constrained set $\mathcal{K}$. This is because the function $f(\boldsymbol{x})$ is also ν-strongly convex [5], yielding $\|\nabla f(\boldsymbol{x})\| = \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^\dagger)\| \geq \nu\|\boldsymbol{x} - \boldsymbol{x}^\dagger\|$ with $\boldsymbol{x}^\dagger$ being the global minimizer of f. When Assumption 2 holds, in Subsection 2.1 we have shown that $\|\nabla f(\boldsymbol{x})\| \leq B$ over $\mathcal{K}$. Thus, we get $\|\boldsymbol{x}\| \leq \|\boldsymbol{x} - \boldsymbol{x}^\dagger\| + \|\boldsymbol{x}^\dagger\| \leq \frac{B}{\nu} + \|\boldsymbol{x}^\dagger\|$ when $\boldsymbol{x} \in \mathcal{K}$. Notice that $\mathbf{x}^\dagger$ is fixed, the set $\mathcal{K}$ is then uniform bounded.

Theorem 2 above shows that AOGD can achieve a faster convergence rate under the strong convexity assumption than that for general convex cases. Theorem 2 also shows that in the strongly convex case, AOGD achieves the optimal convergence rate of SGD Rakhlin et al. (2012), i.e., $\widetilde{\mathcal{O}}\Big(\frac{1}{K}\Big)$. The result in Theorem 2 does not generalize to the general convex case since we set $\eta = \frac{B}{2\nu}$, which is infinity when $\nu = 0$. For technical reasons, we set $\theta = 0$ in Theorem 2, i.e., we only consider the momentum-free case. We will consider how to build the convergence rate $\widetilde{\mathcal{O}}\Big(\frac{1}{K}\Big)$ for AOGD with momentum in our future work.

2.4 Nonconvex Cases

In this part, we consider the case when $F(\boldsymbol{x}; \xi, \xi')$ is nonconvex. The assumptions for the nonconvex case are much milder than the convex and strongly convex cases. We can even get rid of using the projection operator $\mathbf{Proj}_{\mathcal{K}}(\cdot)$ for AOGD. The convergence result of AOGD for the nonconvex case is presented as follows.

Theorem 3 *Let Assumptions 1, 2 hold, and let $\{\boldsymbol{x}^k\}_{k\geq 1}$ be generated by AOGD, and $\|\boldsymbol{g}^0\|^2 \geq \delta$ for some constant $\delta > 0$. Suppose $\sqrt{v^k} \leq C \cdot k^\alpha$ for two constants $C > 0$ and $0 < \alpha \leq \frac{1}{2}$, and $\mathcal{K}$ is the full space. Then, we have*

$$\min_{1\leq k\leq K} \Big\{\mathbb{E}\|\nabla f(\boldsymbol{x}^k)\|^2\Big\} \leq \frac{c_3 + c_4(K)}{K^{1-\alpha}}, \tag{8}$$

where $c_3 := \frac{2(7-6\theta)B^2C}{(1-\theta)\sqrt{\delta}} + \frac{4Cf(\boldsymbol{x}^1)}{\eta}$, *and* $c_4(K) := \Big(\frac{(1+\theta)\eta}{1-\theta} + 2L^2/\sqrt{\delta} + L^2/\delta + 5/2\Big)\ln\Big(\frac{CK}{\delta}\Big)$.

From Theorem 3, we can see that the convergence rate of AOGD is $\widetilde{\mathcal{O}}\Big(\frac{1}{K^{1-\alpha}}\Big)$, and $\alpha = \frac{1}{2}$ is the worst case due to the boundedness of the stochastic gradient. When $\alpha < \frac{1}{2}$, we get a faster convergence of AOGD compared with OGD or SGD in the general nonconvex case.

The conditions of Theorem 3 can also be satisfied by the general convex case without projection. Therefore, equation 8 also holds when $F(\boldsymbol{x}; \xi, \xi')$ is convex. However, equation 8 is weaker than equation 4 since the convergence rate in equation 8 is not established with respect to the function values.

3 Numerical Results

Table 1: Statistics of the dataset used for contrasting the performance of AOGD and OGD, where n is the number of samples in each dataset, and d is the number of features of each instance in a given dataset. All datasets come from the LIBSVM website Chang & Lin (2011), and they are used in Yang et al. (2021b).

	diabtes	german	ijcnn1	letter	mnist	usps
n	768	1,000	49,990	15,000	60,000	7,291
d	8	24	22	161	780	256

In this section, we numerically validate our theoretical findings. To this end, we compare our proposed AOGD against the baseline OGD proposed in Yang et al. (2021b) for pairwise learning in terms of generalization and rate of convergence with respect to the number of iteration [6]. Following Yang et al. (2021b), we consider

[5] Taking expectation of both sides of the inequality $F(\boldsymbol{x}; \xi, \xi') - F(\boldsymbol{y}; \xi, \xi') - \langle \nabla F(\boldsymbol{y}; \xi, \xi'), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \frac{\nu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ gives us the strong convexity of f.

[6] In Yang et al. (2021b), the authors have shown that OGD can remarkably outperform existing algorithms, including OLP Kar et al. (2013), OAM_{gra} Zhao et al. (2011), SGD_{pair} Lei et al. (2020), and SPAUC Lei & Ying (2021).

six benchmark datasets, summarized in Table 1. Also, following the data split strategy used in Yang et al. (2021b), for the dataset with multiple classes, we convert the first half of classes to be the positive class and the second half of classes to be the negative class.

In particular, we run experiments on AUC maximization using the following loss function

$$f(\boldsymbol{w}; (\boldsymbol{x}, y), (\boldsymbol{x}', y')) = \ell(\boldsymbol{w}^\top(\boldsymbol{x} - \boldsymbol{x}'))\mathbb{I}_{[y=1 \wedge y'=-1]}$$

where

$$\mathbb{I}_{[y=1 \wedge y'=-1]} = \begin{cases} 1 & \text{if } y = 1 \text{ and } y' = -1, \\ 0 & \text{otherwise.} \end{cases}$$

And ℓ is a surrogate loss function, e.g., the hinge loss

$$\ell(t) = (1 - t)_+ = \begin{cases} 0 & \text{if } 1 - t < 0, \\ 1 - t & \text{otherwise.} \end{cases}$$

For each dataset, we use 80% of the data for training and the remaining 20% for testing. All the reported results are based on 25 runs with random shuffling. The generalization performance is reported using the average AUC score and standard deviation on the test data. To determine proper hyper parameters for OGD and AOGD, we conduct 5-fold cross-validation on the training sets: 1) for OGD, we select stepsizes $\eta_t = \eta \in 10^{[-5:5]}$[7] and the parameter space $\mathcal{K}$ is set to be the L^2-ball centered at the origin with radius $R \in 10^{[-3,3]}$; 2) for AOGD, we let $\theta = 0.9$ and we select stepsizes $\eta_t = \eta \in 10^{[-5:5]}$ and the parameter space $\mathcal{K}$ is also the L^2-ball centered at the origin with radius $R \in 10^{[-3,3]}$.

Figure 2 plots the AUC scores of AOGD and OGD against the number of iterations on the six benchmark datasets listed in Table 1. The numerical results on the six benchmark datasets show that AOGD converges faster than OGD in general, confirming our theoretical results.

Table 2: Average AUC scores ± standard deviation across the six benchmark datasets listed in Table 1. The best results are highlighted in boldface.

Algorithm	`diabetes`	`german`	`ijcnn1`	`letter`	`mnist`	`usps`
`AOGD`	**.831 ± .027**	**.795 ± .026**	**.934 ± .002**	**.814 ± .006**	.931 ± .002	.925 ± .004
`OGD`	.831 ± .030	.793 ± .021	**.934 ± .002**	.810 ± .007	**.932 ± .001**	**.926 ± .006**

Table 2 summarizes the generalization performance between AOGD and OGD. The results for AOGD are obtained from our above experiment and the results for OGD are adapted from Yang et al. (2021b). Overall, AOGD generalizes as well as OGD, and the two methods perform almost identically. Establishing the generalization of OAGD is an interesting future direction.

4 More Related Works

Because offline methods employ the ERM-like policy, the core problem of most offline methods is establishing the generalization bound of the finite-sum model with statistical learning theory or algorithmic stability Agarwal & Niyogi (2009); Jin et al. (2009); Wang et al. (2019); Gao & Zhou (2013); Lei et al. (2020). It is worth mentioning that the difficulty in the generalization analysis for pairwise learning lies in that the objective functions fail to be i.i.d. with each other, which breaks the fundamental assumption in statistical learning theory and the algorithmic stability communities.

The online methods for pairwise learning assume the model accesses a data stream of i.i.d. samples, including online AUC maximization algorithms Zhao et al. (2011); Ying et al. (2016); Liu et al. (2018); Natole et al. (2018); Lei & Ying (2021); Guo et al. (2020), online metric learning algorithms Shalev-Shwartz et al. (2004); Davis et al. (2007); Jain et al. (2008); Jin et al. (2009), online learning to rank Rejchel (2012); Schuth et al.

[7] $[-5:5]$ stands for integers in the interval $[-5, 5]$.

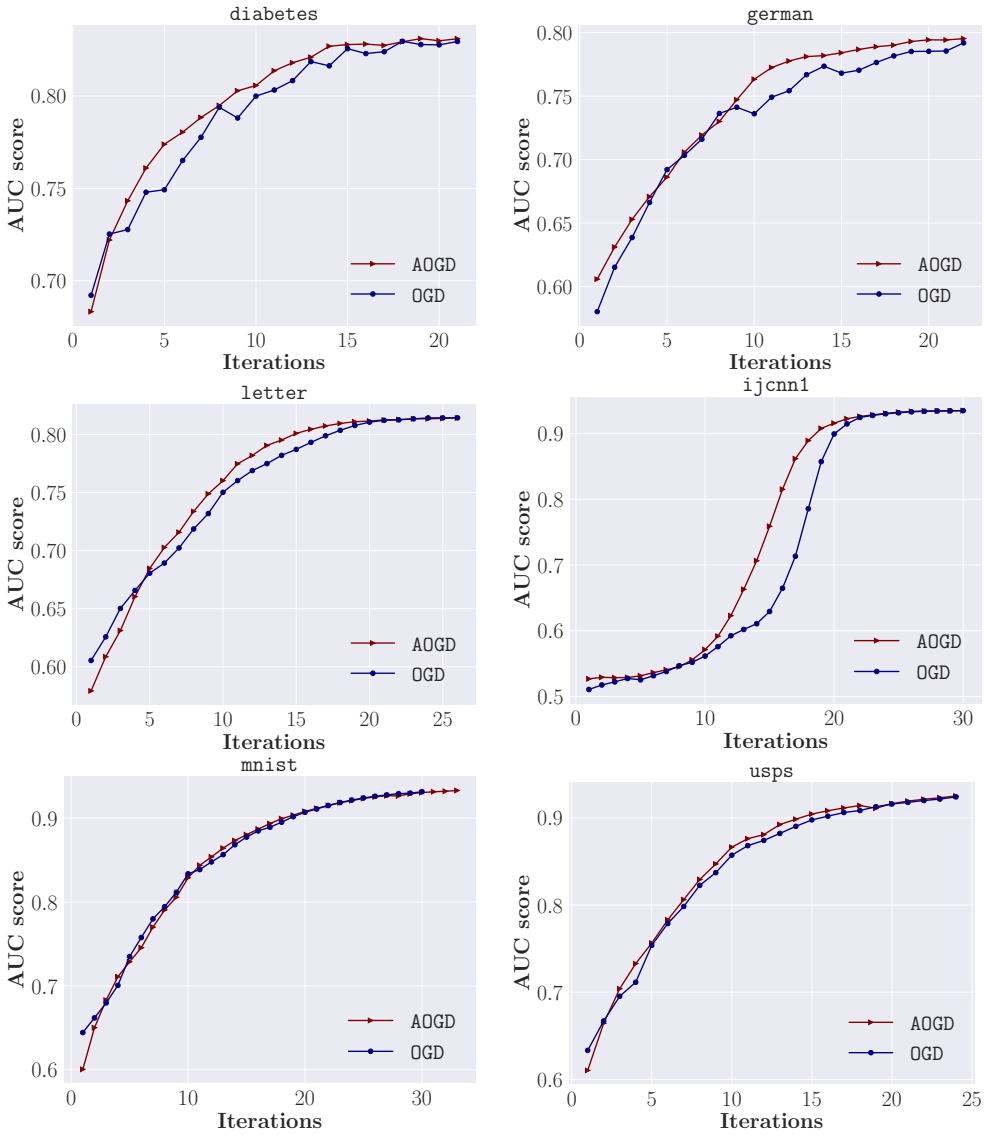

Figure 2: The AUC score of AOGD and OGD against the number of iterations for AUC maximization with hinge loss. It is evident that AOGD converges faster than OGD in general, confirming our established theoretical results for AOGD.

(2013); Zoghi et al. (2017); Li et al. (2019), neural link prediction Wang et al. (2021), etc. The online AUC maximization is proposed by the authors of Zhao et al. (2011) with theoretical guarantees. In paper Ying et al. (2016), a stochastic online AUC maximization algorithm is proposed from the perspective of a saddle representation. The main advantage of the algorithm in Ying et al. (2016) is to avoid storing all previous examples and second-order covariance matrices. Leveraging saddle representation, the authors of Liu et al. (2018) propose a faster online AUC maximization algorithm with provably improved statistical convergence rates. The stochastic proximal algorithms for AUC maximization with non-differentible regularization are proposed and studied in Natole et al. (2018). To make the algorithm scalable to large-scale streaming data, the authors of Lei & Ying (2021) propose a new stochastic proximal algorithm. In paper Guo et al. (2020), the authors consider the distributed setting and propose a communication-efficient stochastic AUC maximization with deep neural networks. The authors of Shalev-Shwartz et al. (2004) propose an online algorithm for supervised learning of pseudo-metrics. In paper Davis et al. (2007), the authors present an information-theoretic approach for online metric learning. In Jain et al. (2008), leveraging the LogDet regularization, the authors propose a fast online metric learning for the similarity search. The generalization

bound of regularized distance metric learning is established in Jin et al. (2009). The authors of Rejchel (2012) consider ranking estimators that minimize the convex empirical risks and prove their generalization bounds. A framework of online learning to rank is proposed by the authors of Schuth et al. (2013). In paper Zoghi et al. (2017), the authors investigate the online learning to rank in stochastic click models. Paper Li et al. (2019) introduces a new model for online ranking with features. The differentially private pairwise learning has been recently studied, and representative works include Huai et al. (2020); Yang et al. (2021a); Xue et al. (2021); Yang et al. (2021b).

5 Conclusions

In this paper, we propose the adaptive online gradient descent algorithms to solve pairwise learning problems and establish their theoretical performance bounds in strongly convex, convex, and nonconvex settings. Our theoretical results explain why the convergence speed of adaptive online gradient descent can outperform the one without adaptive stepsize for pairwise learning. We also provide numerical experiments to demonstrate the efficiency of the proposed algorithm.

Limitation and future work. There are two major limitations in our analysis: 1) we assume the set $\mathcal{K}$ is bounded in establishing Theorem 1, and 2) the convergence rate in Theorem 2 is analyzed for the adaptive online gradient descent without momentum. We leave how to overcome the above two limitations as future work. There are numerous other avenues for future work, including 1) Can we establish the lower bound of the convergence rates for the adaptive online gradient descent applied to pairwise learning? 2) Can we extend the online adaptive gradient descent to the proximal settings to solve nonsmooth pairwise learning problems?

Broader Impact Statement

N/A

References

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(2), 2009.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathrm{O}(1/n)$. *Advances in neural information processing systems*, 26, 2013.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1x-x309tm`.

Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, and Tianbao Yang. Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *International Conference on Learning Representations*, 2018.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 209–216, 2007.

Yi Ding, Peilin Zhao, Steven Hoi, and Yew-Soon Ong. An adaptive gradient method for online auc maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Wei Gao and Zhi-Hua Zhou. Uniform convergence, stability and learnability for ranking problems. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer, 2013.

Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass AUC optimization. In *International conference on machine learning*, pp. 906–914. PMLR, 2013.

Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, and Tianbao Yang. Communication-efficient distributed stochastic AUC maximization with deep neural networks. In *International Conference on Machine Learning*, pp. 3864–3874. PMLR, 2020.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14(2), 2013.

Mengdi Huai, Di Wang, Chenglin Miao, Jinhui Xu, and Aidong Zhang. Pairwise learning with differential privacy guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 694–701, 2020.

Prateek Jain, Brian Kulis, Inderjit Dhillon, and Kristen Grauman. Online metric learning and fast similarity search. *Advances in neural information processing systems*, 21, 2008.

Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: Theory and algorithm. *Advances in neural information processing systems*, 22, 2009.

Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 441–449, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/kar13.html`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Brian Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

Abhishek Kumar, Alexandru Niculescu-Mizil, Koray Kavukcoglu, and Hal Daumé. A binary classification framework for two-stage multiple kernel learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1331–1338, 2012.

Yunwen Lei and Yiming Ying. Stochastic proximal AUC maximization. *Journal of Machine Learning Research*, 22(61):1–45, 2021.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33:21236–21246, 2020.

Shuai Li, Tor Lattimore, and Csaba Szepesvári. Online learning to rank with features. In *International Conference on Machine Learning*, pp. 3856–3865. PMLR, 2019.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 983–992, 2019.

Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic AUC maximization with $O(1/n)$-convergence rate. In *International Conference on Machine Learning*, pp. 3189–3197. PMLR, 2018.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations*, 2019.

H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *COLT 2010*, pp. 244, 2010.

Michael Natole, Yiming Ying, and Siwei Lyu. Stochastic proximal algorithms for AUC maximization. In *International Conference on Machine Learning*, pp. 3710–3719. PMLR, 2018.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1571–1578, 2012.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=ryQu7f-RZ`.

Wojciech Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(5), 2012.

Anne Schuth, Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. Lerot: An online learning to rank framework. In *Proceedings of the 2013 workshop on Living labs for information retrieval evaluation*, pp. 23–26, 2013.

Shai Shalev-Shwartz, Yoram Singer, and Andrew Y Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 94, 2004.

Tao Sun, Linbo Qiao, Qing Liao, and Dongsheng Li. Novel convergence results of adaptive stochastic gradient descents. *IEEE Transactions on Image Processing*, 30:1044–1056, 2020.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 2012.

Boyu Wang, Hejia Zhang, Peng Liu, Zebang Shen, and Joelle Pineau. Multitask metric learning: Theory and algorithm. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3362–3371. PMLR, 2019.

Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, pp. 13–1. JMLR Workshop and Conference Proceedings, 2012.

Zhitao Wang, Yong Zhou, Litao Hong, Yuanhang Zou, and Hanjing Su. Pairwise learning for neural link prediction. *arXiv preprint arXiv:2112.02936*, 2021.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686. PMLR, 2019.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15, 2002.

Zhiyu Xue, Shaoyang Yang, Mengdi Huai, and Di Wang. Differentially private pairwise learning revisited. IJCAI, 2021.

Zhenhuan Yang, Yunwen Lei, Siwei Lyu, and Yiming Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In *International Conference on Artificial Intelligence and Statistics*, pp. 2026–2034. PMLR, 2021a.

Zhenhuan Yang, Yunwen Lei, Puyu Wang, Tianbao Yang, and Yiming Ying. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34, 2021b.

Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *The Journal of Machine Learning Research*, 13(1):1–26, 2012.

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

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online AUC maximization. *Advances in neural information processing systems*, 29, 2016.

Peilin Zhao, Steven CH Hoi, Rong Jin, and Tianbo Yang. Online AUC maximization. In *Proceedings of the 28th International Conference on Machine Learning ICML*, pp. 233–240, 2011.

Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. Online learning to rank in stochastic click models. In *International Conference on Machine Learning*, pp. 4199–4208. PMLR, 2017.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11127–11135, 2019.

A Proofs of Results in the Convex Scenario

A.1 Technical Lemmas

Given any $\boldsymbol{x}^* \in \arg_{\mathcal{K}} \min f$, in the (strongly) convex setting, we introduce the following notation

$$\begin{cases} \phi_k := \mathbb{E}\left[\frac{\|\boldsymbol{g}^k\|^2}{\sqrt{v^k}}\right] + 2B^2\mathbb{E}\Big[\frac{1}{\sqrt{v^{k-2}}} - \frac{1}{\sqrt{v^{k-1}}}\Big] + \frac{L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2}{\sqrt{\delta}} + B^2\mathbb{E}\Big[\frac{1}{\sqrt{v^{k-2}}} - \frac{1}{\sqrt{v^k}}\Big], \\ A_k := \mathbb{E}(\|\boldsymbol{m}^k\|^2/(v^k)^{\frac{1}{2}}), \\ B_k := \mathbb{E}\left(\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{m}^k\rangle\right), \\ C_k := (\eta\theta + \frac{1-\theta}{2})A_{k-1} + (1-\theta)\phi_k \end{cases} \tag{9}$$

We have the following lemmas.

Lemma 1 *[Lemma 9 in the appendix, Li & Orabona (2019)] Let $a_1, a_2, \ldots, a_K$ be non-negative, and h be a non-increasing function. Then we have*

$$\sum_{k=1}^{K} a_k h(a_0 + \sum_{i=1}^{k} a_i) \leq \int_{a_0}^{\sum_{k=0}^{K} a_k} h(t)dt.$$

Lemma 2 *Assume $\{\boldsymbol{x}^k\}_{k\geq 1}$ is generated by Algorithm 2, then we have*

$$\sum_{k=1}^{K} A_k \leq \sum_{k=1}^{K} \mathbb{E}[\|\boldsymbol{g}^k\|^2/(v^k)^{\frac{1}{2}}] \leq \mathbb{E}(v^k)^{\frac{1}{2}}.$$

Lemma 3 *Assume $\{\boldsymbol{x}^k\}_{k\geq 1}$ is generated by Algorithm 2 and $\|\boldsymbol{g}^0\| \geq \sqrt{\delta} > 0$, then we have*

$$\sum_{k=1}^{K} \mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 \leq \sum_{k=1}^{K} \mathbb{E}[\|\boldsymbol{g}^k\|^2/v^k] \leq \mathbb{E}\ln\frac{v^k}{\delta}.$$

Lemma 4 *Assume $\{\boldsymbol{x}^k\}_{k\geq 1}$ is generated by Algorithm 2, then we have*

$$\mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\|^2 \leq 2L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + 2\mathbb{E}\|\boldsymbol{g}^k\|^2.$$

Lemma 5 *Assume $\{\boldsymbol{x}^k\}_{k\geq 1}$ is generated by Algorithm 2, and Assumption 1, 2 hold, then the following result holds*

$$B_k \leq \theta B_{k-1} + C_k + (1-\theta)\mathbb{E}\Big[f(\boldsymbol{x}^*) - f(\boldsymbol{x}^k)\Big].$$

A.2 Proof of Theorem 1

Given $k \in \mathbb{Z}^+$, with Lemma 5 and mathematical induction, we have the following inequality

$$B_k \leq \theta^k B_1 + \sum_{i=1}^{k-1} \theta^{k-1-i} C_i + \sum_{i=1}^{k-1} (1-\theta)\theta^{k-1-i} \mathbb{E}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^i)).$$

Notice that $\boldsymbol{m}^1 = \mathbf{0}$ and $B_1 = 0$, we get

$$B_k \leq \sum_{i=1}^{k-1} \theta^{k-1-i} C_i + \sum_{i=1}^{k-1} (1-\theta)\theta^{k-1-i} \mathbb{E}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^i)). \tag{10}$$

Summing the inequality equation 10 from $k = 1$ to K gives us

$$\begin{aligned}
\sum_{k=1}^{K} B_k &\leq \sum_{k=1}^{K} \sum_{i=1}^{k-1} \theta^{k-1-i} C_i + \sum_{k=1}^{K} \sum_{i=1}^{k-1} (1-\theta)\theta^{k-1-i} \mathbb{E}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^i)) \\
&\leq \frac{\sum_{k=1}^{K} C_k}{1-\theta} + (1-\theta) \sum_{k=1}^{K} \mathbb{E}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^k)).
\end{aligned}$$

Therefore, we have

$$\sum_{k=1}^{K} \mathbb{E}(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)) \leq -\sum_{k=1}^{K} \frac{B_k}{(1-\theta)} + \frac{\sum_{k=1}^{K} C_k}{(1-\theta)^2}. \tag{11}$$

The scheme of the algorithm indicates that

$$\boldsymbol{x}^{k+1} = \mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \eta \boldsymbol{m}^k / (v^k)^{\frac{1}{2}}).$$

We then get

$$\begin{aligned}
&\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 = \|\mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \eta \boldsymbol{m}^k / (v^k)^{\frac{1}{2}}) - \boldsymbol{x}^*\|^2 \\
&= \|\mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \eta \boldsymbol{m}^k / (v^k)^{\frac{1}{2}}) - \mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^*)\|^2 \\
&\leq \|\boldsymbol{x}^k - \eta \boldsymbol{m}^k / (v^k)^{\frac{1}{2}} - \boldsymbol{x}^*\|^2.
\end{aligned}$$

Multiplying both sides with $(v^k)^{\frac{1}{2}}$, we are then led to

$$(v^k)^{\frac{1}{2}} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 \leq (v^k)^{\frac{1}{2}} \|\boldsymbol{x}^* - \boldsymbol{x}^k\|^2 + 2\eta \langle \boldsymbol{m}^k, \boldsymbol{x}^* - \boldsymbol{x}^k \rangle + \eta^2 \|\boldsymbol{m}^k\|^2 / (v^k)^{\frac{1}{2}},$$

which is equivalent to

$$\begin{aligned}
&(v^{k+1})^{\frac{1}{2}} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 \leq (v^k)^{\frac{1}{2}} \|\boldsymbol{x}^* - \boldsymbol{x}^k\|^2 + 2\eta \langle \boldsymbol{m}^k, \boldsymbol{x}^* - \boldsymbol{x}^k \rangle \\
&+ \eta^2 \|\boldsymbol{m}^k\|^2 / (v^k)^{\frac{1}{2}} + ((v^{k+1})^{\frac{1}{2}} - (v^k)^{\frac{1}{2}}) \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2.
\end{aligned}$$

Taking the total expectation of both sides of the above equation gives us

$$\begin{aligned}
-2\eta B_k \leq\ & \mathbb{E}(v^k)^{\frac{1}{2}} \|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - \mathbb{E}(v^{k+1})^{\frac{1}{2}} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 \\
& + \eta^2 A_k + (\mathbb{E}(v^{k+1})^{\frac{1}{2}} - \mathbb{E}(v^k)^{\frac{1}{2}}) D^2.
\end{aligned} \tag{12}$$

Summing equation 12 from $k = 1$ to K gives us

$$\sum_{k=1}^{K} (-B_k)/(1-\theta) \leq \frac{\mathbb{E}\sqrt{v^1} \|\boldsymbol{x}^* - \boldsymbol{x}^1\|^2}{2\eta(1-\theta)} + \frac{\eta}{2(1-\theta)} \sum_{k=1}^{K} A_k + \frac{D^2}{(1-\theta)} \mathbb{E}\sqrt{v^{K+1}}.$$

Together with equation 11, we then have

$$\sum_{k=1}^{K}\mathbb{E}(f(\boldsymbol{x}^k)-f(\boldsymbol{x}^*)) \leq \frac{\mathbb{E}\sqrt{v^1}\|\boldsymbol{x}^*-\boldsymbol{x}^1\|^2}{2\eta(1-\theta)} + \frac{\eta}{2(1-\theta)}\sum_{k=1}^{K}A_k + \frac{D^2}{(1-\theta)}\mathbb{E}\sqrt{v^{K+1}} + \frac{\sum_{k=1}^{K}C_k}{(1-\theta)^2}. \tag{13}$$

We turn to bound the right-hand side of equation 13, and get the following bound

$$\frac{\eta}{2(1-\theta)}\sum_{k=1}^{K}A_k \leq \frac{\eta}{2(1-\theta)}\mathbb{E}(v^{K+1})^{\frac{1}{2}}. \tag{14}$$

On the other hand, we can get

$$\begin{aligned}\frac{\sum_{k=1}^{K}C_k}{(1-\theta)^2} &\leq \frac{\eta\theta}{(1-\theta)^2}\mathbb{E}(v^{K+1})^{\frac{1}{2}} + \frac{2L}{(1-\theta)^2}\mathbb{E}\ln\frac{v^{K+1}}{\delta} + \frac{1}{(1-\theta)^2}\sum_{k=1}^{K}\mathbb{E}\phi_k \\ &\leq \frac{\eta\theta}{(1-\theta)^2}\mathbb{E}(v^{K+1})^{\frac{1}{2}} + \frac{2L}{(1-\theta)^2}\mathbb{E}\ln\frac{v^{K+1}}{\delta} + \frac{\frac{3B^2}{\sqrt{\delta}} + \frac{L^2}{\sqrt{\delta}}\mathbb{E}\ln\frac{v^{K+1}}{\delta} + \mathbb{E}(v^{K+1})^{\frac{1}{2}}}{(1-\theta)^2}\end{aligned} \tag{15}$$

Substituting the bounds equation 14 and equation 15 into equation 13, we are then led to

$$\begin{aligned}&\Big[\sum_{k=1}^{K}\mathbb{E}(f(\boldsymbol{x}^k)-f(\boldsymbol{x}^*))\Big] \\ &\leq \frac{\mathbb{E}\sqrt{v^1}\|\boldsymbol{x}^*-\boldsymbol{x}^1\|^2}{2\eta(1-\theta)} + [\frac{\eta+2D^2}{2(1-\theta)} + \frac{\eta\theta+1}{(1-\theta)^2}]\cdot\mathbb{E}\sqrt{v^{K+1}} \\ &\quad + [\frac{4L}{(1-\theta)^2} + \frac{2L^2}{(1-\theta)^2\sqrt{\delta}}]\mathbb{E}\ln\frac{\sqrt{v^{K+1}}}{\sqrt{\delta}} + \frac{3B^2}{(1-\theta)^2\sqrt{\delta}}.\end{aligned} \tag{16}$$

Notice that $-\ln(\cdot)$ and $-\frac{1}{\cdot}$ are both convex when $\cdot$ is positive, using Jensen's inequality gives us

$$\begin{aligned}-\ln\mathbb{E}\sqrt{v^{K+1}} \leq -\mathbb{E}\ln\sqrt{v^{K+1}},\ -\frac{1}{\mathbb{E}(\sqrt{v^k})} \leq -\mathbb{E}\frac{1}{\sqrt{v^k}} \\ \Rightarrow \mathbb{E}\ln\sqrt{v^{K+1}} \leq \ln\mathbb{E}\sqrt{v^{K+1}},\ \mathbb{E}\frac{1}{\sqrt{v^k}} \leq \frac{1}{\mathbb{E}\sqrt{v^k}}.\end{aligned}$$

Therefore, equation 16 can be presented as

$$\mathbb{E}\left[f\Big(\frac{\sum_{k=1}^{K}\boldsymbol{x}^k}{K}\Big) - \min f\right] \leq \frac{c_1+c_2(K)}{K},$$

where

$$c_1 := \frac{\mathbb{E}\sqrt{v^1}\|\boldsymbol{x}^*-\boldsymbol{x}^1\|^2}{2(1-\theta)\eta} + \frac{3B^2}{(1-\theta)^2\sqrt{\delta}},$$

and

$$c_2(K) := [\frac{\eta+2D^2}{2(1-\theta)} + \frac{\eta\theta+1}{(1-\theta)^2}]\cdot\mathbb{E}\sqrt{v^{K+1}} + [\frac{4L}{(1-\theta)^2} + \frac{2L^2}{(1-\theta)^2\sqrt{\delta}}]\ln\frac{\mathbb{E}\sqrt{v^{K+1}}}{\sqrt{\delta}}.$$

B Proofs of Results in the Strongly Convex Scenario

B.1 Proof of Proposition 2

The boundedness of the gradient and the strong convexity indicate $\mathcal{K}$ is bounded, whose radius is assumed to be $D > 0$. Notice that the operator $\mathbf{Proj}_{\mathcal{K}}(\cdot)$ is contractive, as $\theta = 0$, we get

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|^2 = \mathbb{E}\|\mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \frac{\eta}{\sqrt{k}}\boldsymbol{m}^k/\sqrt{v^k}) - \boldsymbol{x}^*\|^2 \\
&= \mathbb{E}\|\mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \frac{\eta}{\sqrt{k}}\boldsymbol{m}^k/\sqrt{v^k}) - \mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^*)\|^2 \\
&\leq \mathbb{E}\|\boldsymbol{x}^k - \frac{\eta}{\sqrt{k}}\boldsymbol{m}^k/\sqrt{v^k} - \boldsymbol{x}^*\|^2 \\
&= \mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 - 2\frac{\eta}{\sqrt{k}}\mathbb{E}(\frac{\langle \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}), \boldsymbol{x}^k - \boldsymbol{x}^*\rangle}{\sqrt{v^k}}) + \frac{\eta^2}{k}\mathbb{E}\|\boldsymbol{m}^k\|^2/v^k \\
&\leq \mathbb{E}\Big[(1 - 2\eta\nu/\sqrt{kv^k})\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2\Big] + \frac{\eta^2}{k}\mathbb{E}\frac{\|\boldsymbol{m}^k\|^2}{v^k} \\
&\quad + 2\frac{\eta}{\sqrt{k}}\mathbb{E}([F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})]/\sqrt{v^k}),
\end{aligned}$$

where we used the strong convexity of $F(\boldsymbol{x};\xi^k,\xi^{k-1})$. Now, we turn to bound

$$\mathbb{E}([F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})]/\sqrt{v^k}).$$

With direct computations we have the decomposition

$$\begin{aligned}
&\mathbb{E}([F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})]/\sqrt{v^k}) \\
&= \mathbb{E}([F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})]/\sqrt{v^{k-2}}) \\
&+ \mathbb{E}([F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})]/\sqrt{v^{k-2}}) \\
&+ \frac{\eta}{\sqrt{k}}\mathbb{E}\Big([F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})] \times (1/\sqrt{v^k} - 1/\sqrt{v^{k-2}})\Big),
\end{aligned}$$

consequently,

$$\begin{aligned}
&\mathbb{E}([F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})]/\sqrt{v^k}) \\
&\leq \mathbb{E}(\frac{\langle \boldsymbol{m}^{k-1}, \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle}{\sqrt{k-1}\sqrt{v^{k-1}}\sqrt{v^{k-2}}}) + \frac{DB\eta}{\sqrt{k}}\mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^k}) \\
&\leq \frac{1}{2\sqrt{k-1}}\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|}{v^{k-1}} + \frac{1}{2\sqrt{k-1}}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{v^{k-2}} \\
&\quad + \frac{DB\eta}{\sqrt{k}}\mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^k}) + \frac{B}{\sqrt{k-1}\sqrt{\delta}}\mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^{k-1}}) \\
&\leq \frac{1}{\sqrt{\delta}}\mathbb{E}(\frac{\|\boldsymbol{m}^{k-1}\|^2}{\sqrt{k-1}v^{k-1}}) + \frac{DB\eta}{\sqrt{k}}\mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^k}) \\
&\quad + \frac{B}{\sqrt{k-1}\sqrt{\delta}}\mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^{k-1}}).
\end{aligned}$$

where we used $\mathbb{E}([F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})]/\sqrt{v^{k-2}}) = [f(\boldsymbol{x}^*) - f(\boldsymbol{x}^{k-1})]/\sqrt{v^{k-2}} \leq 0$. Letting $\eta = \frac{B}{2\nu}$ and $\beta_k := \frac{\eta}{\sqrt{k}\sqrt{k-1}}\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|}{v^{k-1}} + \frac{\eta}{\sqrt{k}\sqrt{k-1}}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{v^{k-2}} + \frac{2DB\eta^2}{k}\mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^k}) + \frac{2B\eta}{\sqrt{k-1}\sqrt{k}\sqrt{\delta}}\mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^{k-1}}) + \frac{\eta^2}{k}\mathbb{E}\frac{\|\boldsymbol{m}^k\|^2}{v^k}$,

$$a_{k+1} \leq (1 - \frac{1}{k})a_k + \beta_k.$$

With direct computations, we have

$$\begin{aligned}
a_3 &\leq \frac{1}{2}a_2 + \beta_2, \\
a_4 &\leq \frac{2}{3}a_3 + \beta_3 \leq \frac{1}{3}a_2 + \frac{2}{3}\beta_2 + \beta_3, \\
a_5 &\leq \frac{3}{4}a_4 + \beta_4 \leq \frac{1}{4}a_2 + \frac{2}{4}\beta_2 + \frac{3}{4}\beta_3 + \beta_4, \\
&\vdots \\
a_{K+1} &\leq \frac{a_2}{K} + \sum_{k=2}^{K} \frac{k\beta_k}{K}.
\end{aligned}$$

Notice that

$$\begin{aligned}
k\beta_k = \mathcal{O}\Big(&\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|}{v^{k-1}} + \mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{v^{k-2}} + \mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^k}) \\
&+ \mathbb{E}(1/\sqrt{v^{k-2}} - 1/\sqrt{v^{k-1}}) + \mathbb{E}\frac{\|\boldsymbol{m}^k\|^2}{v^k}\Big),
\end{aligned}$$

we just need to compute $\sum_{k=1}^{\infty}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{v^{k-2}}$. By using Lemma 4, we have

$$\mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2/v^{k-2} \leq 2L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2/\delta + 2\mathbb{E}\|\boldsymbol{g}^k\|^2/v^{k-2}.$$

With the fact

$$\mathbb{E}\frac{\|\boldsymbol{g}^k\|^2}{v^{k-2}} \leq \mathbb{E}\frac{\|\boldsymbol{g}^k\|^2}{v^k} + B^2(\mathbb{E}1/v^{k-2} - \mathbb{E}1/v^k),$$

we then get

$$\sum_{k=1}^{\infty}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{v^{k-2}} = \mathcal{O}(\ln v^K).$$

With Lemma 3, we can get

$$\sum_{k=2}^{K} k\beta_k = \mathcal{O}(\ln v^K) = \mathcal{O}(\ln K) \Rightarrow a_{K+1} = \frac{\ln K}{K}.$$

C Proofs of Results in Nonconvex Scenario

C.1 Additional Technical Lemmas

In the nonconvex case, we denote the following items to simplify the presentations of the technical lemmas

$$\begin{cases}
\hat{A}_k := \mathbb{E}\Big[\|\boldsymbol{m}^k\|^2/v^k\Big], \\
\hat{B}_k := \mathbb{E}\left(\langle -\nabla f(\boldsymbol{x}^k), \boldsymbol{m}^k/(v^k)^{\frac{1}{2}}\rangle\right), \\
\hat{C}_k := \theta\eta\hat{A}_k + 2(1-\theta)\mathbb{E}\Big[\|\boldsymbol{g}^k\|^2/v^k\Big] + 6(1-\theta)B^2\mathbb{E}[1/(v^{k-2})^{\frac{1}{2}} - 1/(v^k)^{\frac{1}{2}}] \\
\qquad +(1-\theta)(2L^2/\sqrt{\delta} + L^2/\delta + 1/2)\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2.
\end{cases} \tag{17}$$

Lemma 6 *Assume $\{\boldsymbol{x}^k\}_{k\geq 1}$ is generated by Algorithm 2, then we have*

$$\sum_{k=1}^{K}\hat{A}_k \leq \sum_{k=1}^{K}\mathbb{E}\|\boldsymbol{g}^k\|^2/v^k \leq \mathbb{E}\ln\frac{v^k}{\delta}.$$

Lemma 7 *Assume $\{\boldsymbol{x}^k\}_{k\geq 1}$ is generated by Algorithm 2 and the functions are nonconvex. Let $\mathcal{K}$ be the full space and Assumptions 1 and 2 hold, then the following result holds*

$$\hat{B}_k + \frac{(1-\theta)}{2}\mathbb{E}\Big(\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}}\Big) \leq \theta\hat{B}_{k-1} + \hat{C}_k.$$

C.2 Proof of Theorem 3

According to Lemma 7, we have

$$\frac{(1-\theta)}{2}\sum_{k=1}^{K}\mathbb{E}\Big(\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}}\Big) \leq -\hat{B}_K + (\theta-1)\sum_{k=1}^{K-1}\hat{B}_k + \sum_{k=1}^{K}\hat{C}_k$$
$$\leq (\theta-1)\sum_{k=1}^{K-1}\hat{B}_k + \sum_{k=1}^{K}\hat{C}_k + \frac{B^2}{\sqrt{\delta}}. \tag{18}$$

The Lipschitz property of the gradients gives

$$\mathbb{E}f(\boldsymbol{x}^{k+1}) - \mathbb{E}f(\boldsymbol{x}^k) \leq \mathbb{E}\langle\nabla f(\boldsymbol{x}^k), \boldsymbol{x}^{k+1}-\boldsymbol{x}^k\rangle + \frac{L\mathbb{E}\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|^2}{2} = \eta\hat{B}_k + \frac{L\eta^2}{2}\hat{A}_k. \tag{19}$$

Combining with equation 19, we get the following estimate

$$\sum_{k=1}^{K-1} -\hat{B}_k \leq L\eta\sum_{k=1}^{K-1}\hat{A}_k + \frac{2f(\boldsymbol{x}^1)}{\eta}. \tag{20}$$

On the other hand, with Lemma 3, we have the following bound

$$\frac{2}{1-\theta}\sum_{k=1}^{K}\hat{C}_k \leq \frac{2\theta\eta}{1-\theta}\sum_{k=1}^{K}\hat{A}_k + \frac{6B^2}{\sqrt{\delta}} + (2L^2/\sqrt{\delta} + L^2/\delta + 5/2)\mathbb{E}\ln(\frac{v^K}{\delta}). \tag{21}$$

Using [Lemma 2, Li & Orabona (2019)] and Lemma 6, we have

$$\sum_{k=1}^{K}\hat{A}_k \leq \mathbb{E}\ln(\frac{v^K}{\delta}). \tag{22}$$

Substituting equation 22, equation 21 and equation 20 into equation 18, then we get

$$\sum_{k=1}^{K}\mathbb{E}\Big(\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}}\Big)$$
$$\leq \Big(\frac{(1+\theta)\eta}{1-\theta} + 2L^2/\sqrt{\delta} + L^2/\delta + 5/2\Big)\mathbb{E}\ln(\frac{v^K}{\delta}) + \frac{(7-6\theta)B^2}{(1-\theta)\sqrt{\delta}} + \frac{2f(\boldsymbol{x}^1)}{\eta}. \tag{23}$$

Notice that $\frac{1}{(v^k)^{\frac{1}{2}}} \geq \frac{1}{Ck^{\alpha}}$, then we have

$$\sum_{k=1}^{K}\mathbb{E}\Big(\|[\nabla f(\boldsymbol{x}^k)]^2/(v^k)^{\frac{1}{2}}\|_1\Big) \geq (\sum_{k=1}^{K}\frac{1}{k^{\alpha}C})\cdot\min_{1\leq k\leq K}\{\mathbb{E}\|\nabla f(\boldsymbol{x}^k)\|^2\}.$$

We then complete the proof by using the fact that $0 < \frac{1}{1-\alpha} \leq 2$ when $0 < \alpha \leq 1/2$. Thus, we can get the following estimate

$$\min_{1\leq k\leq K}\{\mathbb{E}\|\nabla f(\boldsymbol{x}^k)\|^2\} \leq \frac{c_3 + c_4(K)}{K^{1-\alpha}},$$

where

$$c_3 := \frac{2(7-6\theta)B^2C}{(1-\theta)\sqrt{\delta}} + \frac{4Cf(\boldsymbol{x}^1)}{\eta},$$

and

$$c_4(K) := \Big(\frac{(1+\theta)\eta}{1-\theta} + 2L^2/\sqrt{\delta} + L^2/\delta + 5/2\Big)\ln(\frac{CK}{\delta}).$$

D Proofs of the Technical Lemmas

D.1 Proof of Lemma 2

With the fact that $\boldsymbol{m}^k = (1-\theta)\sum_{j=1}^k \theta^{k-j}\boldsymbol{g}^j$ when $k \geq 1$, we have

$$
\begin{aligned}
&\|[\boldsymbol{m}^k]^2/(v^k)^{\frac{1}{2}}\| = \sum_{i=1}^d |\boldsymbol{m}_i^k/(v^k)^{\frac{1}{4}}|^2 \leq \sum_{i=1}^d (1-\theta)^2 |\sum_{j=1}^k \theta^{k-j}\boldsymbol{g}_i^j/(v^k)^{\frac{1}{4}}|^2 \\
&\overset{a)}{\leq} \sum_{i=1}^d (1-\theta)^2 (\sum_{j=1}^k \theta^{k-j}(v^k)^{\frac{1}{2}}) \times \sum_{j=1}^k \theta^{k-j}(\boldsymbol{g}_i^j)^2/(v^k) \\
&\leq \sum_{i=1}^d (1-\theta)^2 \cdot \frac{(v^k)^{\frac{1}{2}}}{1-\theta} \cdot \sum_{j=1}^k \theta^{k-j}(\boldsymbol{g}_i^j)^2/(v^k) \\
&= (1-\theta)\cdot \sum_{j=1}^k \theta^{k-j}\|\boldsymbol{g}^j\|^2/(v^k)^{\frac{1}{2}} \overset{b)}{\leq} (1-\theta)\cdot \sum_{j=1}^k \theta^{k-j}\|\boldsymbol{g}^j\|^2/(v^j)^{\frac{1}{2}}
\end{aligned}
$$

where $a)$ uses the Cauchy's inequality $(\sum_{j=1}^k a_j b_j)^2 \leq (\sum_{j=1}^k a_j^2)\cdot(\sum_{j=1}^k b_j^2)$ with $a_j = \theta^{\frac{k-j}{2}}(v^k)^{\frac{1}{4}}$ and $b_j = \theta^{\frac{k-j}{2}}\boldsymbol{g}_i^j/(v^k)^{\frac{1}{2}}$; $b)$ is due to $v^j \leq v^k$ when $j \leq k$. Thus, we are led to

$$
\begin{aligned}
\sum_{k=1}^K \sum_{j=1}^k \theta^{k-j}\|\boldsymbol{g}^j\|^2/(v^j)^{\frac{1}{2}} &= \sum_{j=1}^K \sum_{k=j}^K \theta^{k-j}\|\boldsymbol{g}^j\|^2/(v^j)^{\frac{1}{2}} \\
&= \sum_{j=1}^K \sum_{k=j}^K \theta^{k-j}\|\boldsymbol{g}^j\|^2/(v^j)^{\frac{1}{2}} \leq \frac{1}{1-\theta}\sum_{j=1}^K \|\boldsymbol{g}^j\|^2/(v^j)^{\frac{1}{2}}.
\end{aligned}
$$

Thus, we can get

$$
\sum_{k=1}^K A_k \leq \sum_{k=1}^K \mathbb{E}[\|\boldsymbol{g}^k\|^2/(v^k)^{\frac{1}{2}}].
$$

Using Lemma 1 with $a_k = \|\boldsymbol{g}^k\|^2$ and $h(\cdot) = \frac{1}{\sqrt{\cdot}}$, we get

$$
\sum_{k=1}^K \|\boldsymbol{g}^k\|^2/(v^k)^{\frac{1}{2}} \leq \sqrt{v^K} - \sqrt{v^0} \leq \sqrt{v^K},
$$

where we used the fact that $v^0 \geq 0$. The proof is then completed.

D.2 Proof of Lemma 3

Similar to the proof of Lemma 2, we have

$$
\begin{aligned}
&\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 = \|\mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \eta\boldsymbol{m}^k) - \boldsymbol{x}^k\|^2 = \|\mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k - \eta\boldsymbol{m}^k) - \mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^k)\|^2 \\
&\leq \|\boldsymbol{m}^k/(v^k)^{\frac{1}{2}}\|^2 = \sum_{i=1}^d |\boldsymbol{m}_i^k/(v^k)^{\frac{1}{2}}|^2 \leq \sum_{i=1}^d (1-\theta)^2 |\sum_{j=1}^k \theta^{k-j}\boldsymbol{g}_i^j/(v^k)^{\frac{1}{2}}|^2 \\
&\overset{a)}{\leq} \sum_{i=1}^d (1-\theta)^2 (\sum_{j=1}^k \theta^{k-j})\cdot \sum_{j=1}^k \theta^{k-j}\frac{(\boldsymbol{g}_i^j)^2}{(v^k)} \leq \sum_{i=1}^d (1-\theta)^2 \cdot \frac{1}{1-\theta}\cdot \sum_{j=1}^{k-1} \theta^{k-j}(\boldsymbol{g}_i^j)^2/(v^k) \\
&= (1-\theta)\cdot \sum_{j=1}^k \theta^{k-j}\|\boldsymbol{g}^j\|^2/(v^k) \overset{b)}{=} (1-\theta)\cdot \sum_{j=1}^k \theta^{k-j}\|\boldsymbol{g}^j\|^2/v^j
\end{aligned}
$$

where $a)$ uses the fact that $(\sum_{j=1}^{k} a_j b_j)^2 \leq \sum_{j=1}^{k} a_j^2 \sum_{j=1}^{k} b_j^2$ with $a_j = \theta^{\frac{k-j}{2}}$ and $b_j = \theta^{\frac{k-j}{2}} \boldsymbol{g}_i^j/(v^k)^{\frac{1}{2}}$, and $b)$ is due to $v^j \leq v^k$ when $j \leq k$. Thus, we can get

$$\sum_{k=1}^{K} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 \leq \sum_{k=1}^{K} \|\boldsymbol{g}^k\|^2/v^k.$$

Using Lemma 1 with $a_k = \|\boldsymbol{g}^k\|^2$ and $h(\cdot) = \frac{1}{\cdot}$, we are then led to

$$\sum_{k=1}^{K} \|\boldsymbol{g}^k\|^2/v^k \leq \ln \frac{v^k}{v^0} \leq \ln \frac{v^k}{\delta}.$$

The proof is then completed.

D.3 Proof of Lemma 4

Direct computations give us

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{g}^k\|^2 = \mathbb{E}\|\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})\|^2 \\
&= \mathbb{E}\|\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) + \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 \\
&= \mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 + \mathbb{E}\|\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 \\
&\quad + 2\mathbb{E}\langle \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}), \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \\
&\overset{a)}{\geq} \frac{1}{2}\mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 - \mathbb{E}\|\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 \\
&\geq \frac{1}{2}\mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 - L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2,
\end{aligned}$$

where $a)$ uses the inequality $2\mathbb{E}\langle \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}), \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \geq -\frac{1}{2}\mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 - 2\mathbb{E}\|\nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2$. Thus we can get

$$\mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2 \leq 2L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + 2\mathbb{E}\|\boldsymbol{g}^k\|^2.$$

D.4 Proof of Lemma 5

The convexity of $f_{i_k}(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ and $\boldsymbol{g}^k = \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})$ gives us

$$\begin{aligned}
\mathbb{E}\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{g}^k\rangle &\leq \mathbb{E}[F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})] \\
&= \mathbb{E}[F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) + F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})].
\end{aligned} \tag{24}$$

The Lipchitz gradient continuity of $F(\boldsymbol{x};\xi^k,\xi^{k-1})$ with respect to $\boldsymbol{x}$ indicates

$$\begin{aligned}
&\mathbb{E}[F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})] \\
&\leq \frac{L}{2}\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})\rangle \\
&\leq \frac{L}{2}\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \\
&\quad + \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \\
&\leq \frac{3L}{2}\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle,
\end{aligned} \tag{25}$$

where we used the Cauchy's inequality $\mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^k;\xi^k,\xi^{k-1}) - \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \leq L\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^k\|^2$. Because ξ^k, ξ^{k-1} are independent of $\boldsymbol{x}^{k-1}$,

$$\mathbb{E}(F(\boldsymbol{x}^*;\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})) = f(\boldsymbol{x}^*) - f(\boldsymbol{x}^{k-1}). \tag{26}$$

Turning back to equation 24, we get

$$\mathbb{E}\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{g}^k\rangle \leq \mathbb{E}[F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})] + f(\boldsymbol{x}^*) - f(\boldsymbol{x}^{k-1}). \tag{27}$$

Notice that $\boldsymbol{x}^{k-1}$ is independent of (ξ^k,ξ^{k-1}),

$$\begin{aligned}
&\mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \\
&= \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla f(\boldsymbol{x}^{k-1})\rangle \\
&\quad + \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - \mathbb{E}\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \\
&\leq \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla f(\boldsymbol{x}^{k-1})\rangle + \frac{1}{2}\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|^2}{\sqrt{v^{k-1}}} \\
&\quad + \frac{1}{2}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - \mathbb{E}\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{\sqrt{v^{k-1}}},
\end{aligned} \tag{28}$$

where we used $|\mathbb{E}\langle X, Y\rangle| \leq \mathbb{E}\|X\|^2 + \mathbb{E}\|Y\|^2$ with $X = \sqrt[4]{v^{k-1}}(\boldsymbol{x}^{k-1} - \boldsymbol{x}^k)$, and $Y = [\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - \mathbb{E}\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})]/\sqrt[4]{v^{k-1}}$. Now, we turn to the upper bound of $\frac{1}{2}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - \mathbb{E}\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{\sqrt{v^{k-1}}}$:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - \mathbb{E}\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{\sqrt{v^{k-1}}} \\
&= \frac{1}{2}\mathbb{E}\frac{\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - \mathbb{E}\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{\sqrt{v^{k-2}}} \\
&\quad + \frac{1}{2}\mathbb{E}\Big[(\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - \mathbb{E}\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2) \times (\frac{1}{\sqrt{v^{k-1}}} - \frac{1}{\sqrt{v^{k-2}}})\Big] \\
&\leq \frac{1}{2}\frac{\mathbb{E}\|\nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\|^2}{\sqrt{v^{k-2}}} + 2B^2\mathbb{E}(\frac{1}{\sqrt{v^{k-2}}} - \frac{1}{\sqrt{v^{k-1}}}) \\
&\overset{a)}{\leq} \frac{L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + \mathbb{E}\|\boldsymbol{g}^k\|^2}{\sqrt{v^{k-2}}} + 2B^2\mathbb{E}(\frac{1}{\sqrt{v^{k-2}}} - \frac{1}{\sqrt{v^{k-1}}}) \\
&\leq \phi_k,
\end{aligned} \tag{29}$$

where $\phi_k := \mathbb{E}\frac{\|\boldsymbol{g}^k\|^2}{\sqrt{v^k}} + 2B^2\mathbb{E}(\frac{1}{\sqrt{v^{k-2}}} - \frac{1}{\sqrt{v^{k-1}}}) + \frac{L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2}{\sqrt{\delta}} + B^2\mathbb{E}(\frac{1}{\sqrt{v^{k-2}}} - \frac{1}{\sqrt{v^k}})$, and $a)$ depends on Lemma 4. Thus, we have

$$\begin{aligned}
&\mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \\
&\leq \mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla f(\boldsymbol{x}^{k-1})\rangle + \frac{1}{2}\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|^2}{\sqrt{v^{k-1}}} + \phi_k.
\end{aligned} \tag{30}$$

Once with the Lipchitz property,

$$\langle \nabla f(\boldsymbol{x}^{k-1}), \boldsymbol{x}^{k-1} - \boldsymbol{x}^k\rangle \leq f(\boldsymbol{x}^{k-1}) - f(\boldsymbol{x}^k) + \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2. \tag{31}$$

Combing equation 31 and equation 30, we then get

$$\begin{aligned}
&\mathbb{E}\langle \boldsymbol{x}^{k-1} - \boldsymbol{x}^k, \nabla F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1})\rangle \leq f(\boldsymbol{x}^{k-1}) - f(\boldsymbol{x}^k) \\
&\quad + \frac{L}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + \frac{1}{2}\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|^2}{\sqrt{v^{k-1}}} + \phi_k.
\end{aligned} \tag{32}$$

Substituting equation 32 into equation 25,

$$\begin{aligned}
&\mathbb{E}[F(\boldsymbol{x}^{k-1};\xi^k,\xi^{k-1}) - F(\boldsymbol{x}^k;\xi^k,\xi^{k-1})] \leq 2L\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 \\
&\quad + f(\boldsymbol{x}^{k-1}) - f(\boldsymbol{x}^k) + \frac{1}{2}\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|^2}{\sqrt{v^{k-1}}} + \phi_k.
\end{aligned} \tag{33}$$

Substituting equation 33 into equation 27, we then get

$$\mathbb{E}\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{g}^k\rangle \leq 2L\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + f(\boldsymbol{x}^*) - f(\boldsymbol{x}^k) + \frac{1}{2}\mathbb{E}\frac{\|\boldsymbol{m}^{k-1}\|^2}{\sqrt{v^{k-1}}} + \phi_k. \tag{34}$$

According to our algorithm and we denote $\Lambda := \mathbb{E}(\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{g}^k\rangle)$, then we have

$$\begin{aligned}
&\mathbb{E}\left(\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{m}^k\rangle\right) = \mathbb{E}\left(\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \theta\boldsymbol{m}^{k-1} + (1-\theta)\boldsymbol{g}^k\rangle\right)\\
&= (1-\theta)\cdot\Lambda + \theta\mathbb{E}\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{m}^{k-1}\rangle\\
&= (1-\theta)\cdot\Lambda + \theta\mathbb{E}\langle \boldsymbol{x}^* - \boldsymbol{x}^{k-1}, \boldsymbol{m}^{k-1}\rangle + \theta\mathbb{E}\langle \boldsymbol{x}^k - \boldsymbol{x}^{k-1}, \boldsymbol{m}^{k-1}\rangle\\
&\overset{b)}{\leq} (1-\theta)\cdot\Lambda + \theta\mathbb{E}\langle \boldsymbol{x}^* - \boldsymbol{x}^{k-1}, \boldsymbol{m}^{k-1}\rangle + \eta\theta\mathbb{E}\|\boldsymbol{m}^{k-1}\|^2/(v^{k-1})^{\frac{1}{2}},
\end{aligned}$$

where $b)$ depends on that $\langle \boldsymbol{x}^k - \boldsymbol{x}^{k-1}, \boldsymbol{m}^{k-1}\rangle \leq \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|\cdot\|\boldsymbol{m}^{k-1}\| = \|\mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^{k-1} - \eta\boldsymbol{m}^k) - \mathbf{Proj}_{\mathcal{K}}(\boldsymbol{x}^{k-1})\|\cdot\|\boldsymbol{m}^{k-1}\| \leq \|\boldsymbol{m}^{k-1}\|^2/(v^{k-1})^{\frac{1}{2}}$. Then, we get

$$B_k \leq (1-\theta)\mathbb{E}\langle \boldsymbol{x}^* - \boldsymbol{x}^k, \boldsymbol{g}^k\rangle + \theta B_{k-1} + \eta\theta A_{k-1}. \tag{35}$$

Substituting equation 34 into equation 35, we then proved the desired result.

D.5 Proof of Lemma 6

This proof is identical to the proof of Lemma 3 and will not be reproduced.

D.6 Proof of Lemma 7

Notice that $\mathbb{E}\nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1}) = \nabla f(\boldsymbol{x}^{k-1})$, we then get

$$\begin{aligned}
&\mathbb{E}\langle -\nabla f(\boldsymbol{x}^k)/(v^k)^{1/2}, \boldsymbol{g}^k\rangle\\
&= \mathbb{E}\langle -\nabla f(\boldsymbol{x}^k)/(v^k)^{1/2}, \nabla F(\boldsymbol{x}^k; \xi^k, \xi^{k-1})\rangle\\
&= -\mathbb{E}\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{1/2} + \mathbb{E}\langle \nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\rangle\\
&+ \underbrace{\mathbb{E}\langle \nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\rangle}_{:=(\dagger)}\\
&+ \mathbb{E}\langle \frac{\nabla f(\boldsymbol{x}^k)}{(v^k)^{\frac{1}{2}}}, \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1}) - \nabla F(\boldsymbol{x}^k; \xi^k, \xi^{k-1})\rangle.
\end{aligned} \tag{36}$$

The Cauchy's inequality together with the smooth assumption gives us

$$\begin{aligned}
&\mathbb{E}\langle \nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\rangle\\
&\leq |\mathbb{E}\langle \nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{4}}, [\nabla f(\boldsymbol{x}^{k-1}) - \nabla f(\boldsymbol{x}^k)]/(v^k)^{\frac{1}{4}}\rangle|\\
&\leq \frac{1}{4}\mathbb{E}\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}} + \mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1}) - \nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}}\\
&\leq \frac{1}{4}\mathbb{E}\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}} + L^2\mathbb{E}\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^k\|^2/(v^k)^{\frac{1}{2}}.
\end{aligned} \tag{37}$$

Similarly, the Lipschitz property of the stochastic gradient yields

$$\begin{aligned}
&\mathbb{E}\langle \nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1}) - \nabla F(\boldsymbol{x}^k; \xi^k, \xi^{k-1})\rangle\\
&\leq \mathbb{E}|\langle \frac{\nabla f(\boldsymbol{x}^k)}{(v^k)^{\frac{1}{4}}}, \frac{[\nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1}) - \nabla F(\boldsymbol{x}^k; \xi^k, \xi^{k-1})]}{(v^k)^{\frac{1}{4}}}\rangle|\\
&\leq \frac{1}{4}\mathbb{E}\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}} + L^2\mathbb{E}\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^k\|^2/(v^k)^{\frac{1}{2}}.
\end{aligned} \tag{38}$$

Substituting equation 38 and equation 37 into equation 36, we are then led to

$$\begin{aligned}\mathbb{E}\langle -\nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \boldsymbol{g}^k\rangle &\leq -\mathbb{E}\Big(\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}}\Big) + (\dagger) \\ &+ \frac{1}{2}\mathbb{E}(\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}}) + 2L^2\mathbb{E}(\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^k\|^2/(v^k)^{\frac{1}{2}}).\end{aligned} \tag{39}$$

Now, we turn to bound (†):

$$\begin{aligned}(\dagger) &= \underbrace{\mathbb{E}\langle \nabla f(\boldsymbol{x}^{k-1})/(v^{k-1})^{\frac{1}{2}}, \nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\rangle}_{=0} \\ &+ \mathbb{E}\Big\langle \nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}} - \nabla f(\boldsymbol{x}^{k-1})/(v^k)^{\frac{1}{2}} \\ &+ \nabla f(\boldsymbol{x}^{k-1})/(v^k)^{\frac{1}{2}} - \nabla f(\boldsymbol{x}^{k-1})/(v^{k-1})^{\frac{1}{2}}, \nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\Big\rangle \\ &\leq \mathbb{E}\langle [\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})]/(v^k)^{\frac{1}{2}}, \nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\rangle \\ &+ 2B^2\mathbb{E}[1/(v^{k-2})^{\frac{1}{2}} - 1/(v^k)^{\frac{1}{2}}] \\ &\leq \frac{\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2}{2} + 2B^2\mathbb{E}[1/(v^{k-2})^{\frac{1}{2}} - 1/(v^k)^{\frac{1}{2}}] \\ &+ \frac{1}{2}\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\|^2/v^k.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}&\frac{1}{2}\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\|^2/v^k \\ &= \frac{1}{2}\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\|^2/v^{k-2} \\ &+ \frac{1}{2}\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1}) - \nabla F(\boldsymbol{x}^{k-1}; \xi^k, \xi^{k-1})\|^2(1/v^k - 1/v^{k-2}) \\ &\overset{\text{Lemma 4}}{\leq} L^2\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2/v^{k-2} + 2\mathbb{E}\|\boldsymbol{g}^k\|^2/v^{k-2} + 2B^2\mathbb{E}(1/v^k - 1/v^{k-2}) \\ &\leq \frac{L^2}{\delta}\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2 + 2\mathbb{E}\|\boldsymbol{g}^k\|^2/v^k + 4B^2\mathbb{E}(1/v^k - 1/v^{k-2}).\end{aligned}$$

Thus, equation 39 can also be bounded as

$$\begin{aligned}\mathbb{E}\langle -\nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \boldsymbol{g}^k\rangle &\leq -\frac{1}{2}\mathbb{E}\Big(\|\nabla f(\boldsymbol{x}^k)\|^2/(v^k)^{\frac{1}{2}}\Big) \\ &+ 6B^2\mathbb{E}[1/(v^{k-2})^{\frac{1}{2}} - 1/(v^k)^{\frac{1}{2}}] + 2\mathbb{E}\|\boldsymbol{g}^k\|^2/v^k \\ &+ (2L^2/\sqrt{\delta} + L^2/\delta + 1/2)\mathbb{E}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|^2.\end{aligned} \tag{40}$$

We also use a shorthand notation $\Lambda := \mathbb{E}(\langle -\nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \boldsymbol{g}^k\rangle)$ and then

$$\begin{aligned}&\mathbb{E}\left(\langle -\nabla f(\boldsymbol{x}^k), \boldsymbol{m}^k/(v^k)^{\frac{1}{2}}\rangle\right) \\ &= \mathbb{E}\left(\langle -\nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \theta\boldsymbol{m}^{k-1} + (1-\theta)\boldsymbol{g}^k\rangle\right) \\ &= (1-\theta)\cdot\Lambda + \theta\mathbb{E}\langle -\nabla f(\boldsymbol{x}^k)/(v^k)^{\frac{1}{2}}, \boldsymbol{m}^{k-1}\rangle \\ &\overset{a)}{\leq} (1-\theta)\cdot\Lambda + \theta\mathbb{E}\langle -\nabla f(\boldsymbol{x}^{k-1})/(v^{k-1})^{\frac{1}{2}}, \boldsymbol{m}^{k-1}\rangle + \theta\eta\mathbb{E}\|\boldsymbol{m}^{k-1}\|^2/v^{k-1}.\end{aligned} \tag{41}$$

where $a)$ uses the Cauchy-Schwarz inequality and the Lipschitz property of f and $v^{k-1} \leq v^k$. Substituting equation 41 into equation 40, we then proved the result.