Journal of Machine Learning Research 23 (2022) 1-47 Submitted 12/20; Published 7/22

Towards Practical Adam: Non-Convexity, Convergence Theory, and Mini-Batch Acceleration*

Congliang Chen † CONGLIANGCHEN@LINK.CUHK.EDU.CN
The Chinese University of Hong Kong, Shenzhen

Li Shen ‡† MATHSHENLI@GMAIL.COM
JD Explore Academy

Fangyu Zou † FANGYUZOU@GMAIL.COM
Meta

Wei Liu‡ WL2223@COLUMBIA.EDU
Tencent

Editor: Sanjiv Kumar

Abstract

Adam is one of the most influential adaptive stochastic algorithms for training deep neural networks, which has been pointed out to be divergent even in the simple convex setting via a few simple counterexamples. Many attempts, such as decreasing an adaptive learning rate, adopting a big batch size, incorporating a temporal decorrelation technique, seeking an analogous surrogate, *etc.*, have been tried to promote Adam-type algorithms to converge. In contrast with existing approaches, we introduce an alternative easy-to-check sufficient condition, which merely depends on the parameters of the base learning rate and combinations of historical second-order moments, to guarantee the global convergence of generic Adam for solving large-scale non-convex stochastic optimization. This observation, coupled with this sufficient condition, gives much deeper interpretations on the divergence of Adam. On the other hand, in practice, mini-Adam and distributed-Adam are widely used without any theoretical guarantee. We further give an analysis on how the batch size or the number of nodes in the distributed system affects the convergence of Adam, which theoretically shows that mini-batch and distributed Adam can be linearly accelerated by using a larger mini-batch size or a larger number of nodes. At last, we apply the generic Adam and mini-batch Adam with the sufficient condition for solving the counterexample and training several neural networks on various real-world datasets. Experimental results are exactly in accord with our theoretical analysis.

Keywords: Adam, non-convexity, convergence rate, mini-batch/distributed Adam, linear speedup.

1. Introduction

Large-scale non-convex stochastic optimization (Bottou et al., 2018), covering a slew of applications in statistics and machine learning (Jain et al., 2017; Bottou et al., 2018) such as learning a latent variable from massive data whose probability density distribution is unknown, takes the following

†. The first three authors contribute equally. ‡. Corresponding authors.

∗. This paper is an extension of (Zou et al., 2019), which was part of our CVPR 2019 paper. In this version, we renew our analysis technique that yields a cleaner convergence rate compared with (Zou et al., 2019). In addition, we also derive the linear speedup properties for mini-batch Adam and distributed Adam in the Parameter-server model. This work is done when Congliang Chen is a research intern at Tencent AI Lab, China.

generic formulation:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathbb{P}}\left[\widetilde{f}(\boldsymbol{x},\xi)\right], \tag{1}$$

where $f(\boldsymbol{x})$ is a non-convex function and ξ is a random variable with an unknown distribution $\mathbb{P}$.

Alternatively, a compromised approach to handle this difficulty is to use an unbiased stochastic estimate of $\boldsymbol{\nabla} f(\boldsymbol{x})$, denoted as $g(\boldsymbol{x},\xi)$, which leads to the stochastic gradient descent (SGD) algorithm (Robbins and Monro, 1985). Its coordinate-wise version is defined as follows:

$$\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \eta_{t,k}\boldsymbol{g}_{t,k}(\boldsymbol{x}_t,\xi_t), \tag{2}$$

for $k = 1, 2, \ldots, d$, where $\eta_{t,k} \geq 0$ is the learning rate of the k-th component of stochastic gradient $\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)$ at the t-th iteration. Under some mild assumptions (e.g., the optimal solution exists), a sufficient condition (Robbins and Monro, 1985) to ensure the global convergence of vanilla SGD in Eq. (2) is to require η_t to meet the following diminishing condition:

$$\sum_{t=1}^{\infty} \|\eta_t\| = \infty \ \text{ and } \ \sum_{t=1}^{\infty} \|\eta_t\|^2 < \infty. \tag{3}$$

Although the vanilla SGD algorithm with learning rate η_t satisfying condition (3) does converge, its empirical performance could be still stagnating, since it is difficult to tune an effective learning rate η_t via condition (3).

To further improve the empirical performance of SGD, a large variety of adaptive SGD algorithms, including AdaGrad (Duchi et al., 2011), RMSProp (Hinton et al., 2012), Adam (Kingma and Ba, 2014), Nadam (Dozat, 2016), AdaBound (Luo et al., 2019), *etc.*, have been proposed to automatically tune the learning rate η_t by using second-order moments of historical stochastic gradients $\{\boldsymbol{g}_t\}$. Let $\boldsymbol{v}_{t,k}$ and $\boldsymbol{m}_{t,k}$ be the exponential moving average of the historical second-order moments $(\boldsymbol{g}_{1,k}^2, \boldsymbol{g}_{2,k}^2, \cdots, \boldsymbol{g}_{t,k}^2)$ and stochastic gradient estimates $(\boldsymbol{g}_{1,k}, \boldsymbol{g}_{2,k}, \cdots, \boldsymbol{g}_{t,k})$, respectively. More specifically, two groups of hyperparameters (β_t, θ_t) will be involved into the calculation of $m_{t,k} = \beta_t m_{t-1,k} + (1-\beta_t)g_{t,k}$ and $v_{t,k} = \theta_t v_{t-1,k} + (1-\theta_t)g_{t,k}^2$. Then, the generic iteration scheme of these adaptive SGD algorithms (Reddi et al., 2018; Chen et al., 2018a) is summarized as

$$\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \eta_{t,k}\boldsymbol{m}_{t,k}, \ \text{ with } \eta_{t,k} = \alpha_t/\sqrt{\boldsymbol{v}_{t,k}}, \tag{4}$$

for $k = 1, 2, \ldots, d$, where $\alpha_t > 0$ is called base learning rate and it is independent of stochastic gradient estimates $(\boldsymbol{g}_{1,k}, \boldsymbol{g}_{2,k}, \cdots, \boldsymbol{g}_{t,k})$ for all $t \geq 1$. Although Adam works well for solving large scale convex and non-convex optimization problems such as training deep neural networks, it has been disclosed to be divergent in some scenarios via counterexamples (Reddi et al., 2018). Thus, without any further assumptions for corrections, Adam should not be directly used. Recently, developing sufficient conditions to guarantee global convergences of Adam -type algorithms has attracted much attention from both machine learning and optimization communities. The existing successful attempts can be divided into four categories: decreasing a learning rate, adopting a big batch size, incorporating a temporal decorrelation, and seeking an analogous surrogate. However, some of them are either hard to check or impractical. In this work, we will first introduce an alternative easy-to-check sufficient condition to guarantee the global convergences of the original Adam.

Meanwhile, in practice, stochastic Adam, where a single sample is used to estimate gradient, converges slowly to the optimal point. People usually use mini-batch Adam instead to get faster convergence performance. In SGD, although how the sample size will affect the convergence has been well studied (Li et al., 2014), few works give analysis on mini-batch adaptive gradient methods especially on Adam, since mini-batch size largely affects adaptive learning rate $\eta_{t,k}$ in Eq .(4), which makes the analysis difficult. In this work, we give the first complexity analysis for mini-batch Adam, which shows that mini-batch Adam can also be theoretically accelerated by using a larger mini-batch size.

On the other hand, as the data size goes larger in machine learning problems, it is hard to collect, store and process data in a single machine. Several machines are involved in the optimization process. Hence, distributed optimization methods are proposed, where distributed Adam is also popularly used. Different from mini-batch Adam, where only one machine is used for optimization, several machines are involved. In the distributed setting, machines are connected via a network graph. More specifically, there are two kinds of structures used in distributed Adam: parameter-server structure and decentralized structure. In the parameter-server structure, there is one special machine called as parameter server and the rest called workers. The parameter server connects to all workers, but workers don't connect to each other. Therefore, workers can share information with the parameter server in each communication round but cannot share information with the other workers. However, in the decentralized structure, there is not a server involved in the structure. A pre-defined graph connects all machines. A machine can only share information with its direct neighbors in each communication round. Still, few works answer how the local batch size and number of machines will affect the convergence of distributed Adam. In this work, because the analysis of distributed Adam under the parameter-server model is similar to Mini-batch Adam, we answer this question and show that distributed Adam under a parameter-server model can also achieve a linear speedup property as distributed SGD (Yu et al., 2019).

In summary, the contributions of this work are five-fold:

(1) We introduce an easy-to-check sufficient condition to ensure the global convergences (i.e., averaged expected gradient norm converges to 0) of generic Adam in the common smooth non-convex stochastic setting with mild assumptions. Moreover, this sufficient condition is distinctive from the existing conditions and is easier to verify.

(2) We provide a new explanation on the divergences of original Adam and RMSProp, which are possibly due to an incorrect parameter setting of the combinations of historical second-order moments.

(3) We find that the sufficient condition extends the restrictions of RMSProp (Mukkamala and Hein, 2017) and covers many convergent variants of Adam, e.g., AdamNC, AdaGrad with momentum, *etc.* Thus, their convergences in the non-convex stochastic setting naturally hold.

(4) We theoretically show that mini-batch Adam can be further accelerated by adopting a larger mini-batch size, and that distributed Adam can achieve a linear speed up property in the parameter-server distributed system by using commonly used sufficient condition parameters.

(5) We conduct experiments to validate the sufficient condition for the convergences of Adam and mini-batch Adam. The experimental results match our theoretical results.

The paper is organized as follows. In Section 2, we first give the formulation of generic Adam and then discuss several works related to Adam including several existing sufficient convergence conditions, analysis of mini-batch, and distributed stochastic gradient methods. In Section 3, we derive the sufficient condition for convergence of Adam and provide several insights for the divergence of vanilla Adam. In Section 4, we give the complexity analysis on practical Adam with a commonly used sufficient condition parameter, including mini-batch Adam and distributed Adam. At last, in Section 5, we conduct some experiments under both theoretical settings and practical settings to verify the established theory. In addition, by practical Adam, we mean that we give a thorough analysis for Adam, mini-batch Adam, and distributed Adam, which have been commonly used for training deep neural networks without theoretical guarantees.

2. Related work

2.1 Generic Adam

For readers' convenience, we first clarify a few necessary notations used in the forthcoming Generic Adam. First, we denote $\boldsymbol{x}_{t,k}$ as the k-th component of $\boldsymbol{x}_t \in \mathbb{R}^d$, and $\boldsymbol{g}_{t,k}$ as the k-th component of the stochastic gradient at the t-th iteration respectively, and call $\alpha_t > 0$ base learning rate and β_t momentum parameter, respectively. Let $\epsilon > 0$ be a sufficiently small constant. Denote $\mathbf{0} = (0, \cdots, 0)^\top \in \mathbb{R}^d$, and $\boldsymbol{\epsilon} = (\epsilon, \cdots, \epsilon)^\top \in \mathbb{R}^d$. All operations, such as multiplying, dividing, and taking the square root, are executed in the coordinate-wise mode.

Algorithm 1 Generic Adam

1: **Parameters:** Set suitable base learning rate $\{\alpha_t\}$, momentum parameter $\{\beta_t\}$, and exponential moving average parameter $\{\theta_t\}$, respectively. Choose $\boldsymbol{x}_1 \in \mathbb{R}^d$ and set initial values $\boldsymbol{m}_0 = \mathbf{0} \in \mathbb{R}^d$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon} \in \mathbb{R}^d$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Sample a stochastic gradient $\boldsymbol{g}_t$;
4: **for** $k = 1, 2, \ldots, d$ **do**
5: $\boldsymbol{v}_{t,k} = \theta_t \boldsymbol{v}_{t-1,k} + (1-\theta_t)\boldsymbol{g}_{t,k}^2$;
6: $\boldsymbol{m}_{t,k} = \beta_t \boldsymbol{m}_{t-1,k} + (1-\beta_t)\boldsymbol{g}_{t,k}$;
7: $\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \alpha_t \boldsymbol{m}_{t,k}/\sqrt{\boldsymbol{v}_{t,k}}$;
8: **end for**
9: **end for**

It is not hard to check that Generic Adam covers RMSProp by setting $\beta_t = 0$ directly. Moreover, it covers Adam with a bias correction (Kingma and Ba, 2014) as follows:

Remark 1. *The vanilla Adam with the bias correction (Kingma and Ba, 2014) takes constant parameters $\beta_t = \beta$ and $\theta_t = \theta$. The iteration scheme is written as $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \widehat{\alpha}_t \frac{\widehat{\boldsymbol{m}}_t}{\sqrt{\widehat{\boldsymbol{v}}_t}}$, with $\widehat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-\beta^t}$ and $\widehat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1-\theta^t}$. Let $\alpha_t = \widehat{\alpha}_t \frac{\sqrt{1-\theta^t}}{1-\beta^t}$. Then, the above can be rewritten as $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \boldsymbol{m}_t/\sqrt{\boldsymbol{v}_t}$. Thus, it is equivalent to taking constant β_t, constant θ_t, and new base learning rate α_t in Generic Adam.*

2.2 Convergence Conditions for Adam

First, because Reddi et al. (2018) gave counterexamples on divergence of origin Adam, several sufficient conditions have been proposed to guarantee global convergences of Adam that can be summarized into the following four categories:

(C1) Decreasing a learning rate. Reddi et al. (2018) have declared that the core cause of divergences of Adam and RMSProp is largely controlled by the difference between the two adjacent learning rates, *i.e.*,

$$\Gamma_t = 1/\boldsymbol{\eta}_t - 1/\boldsymbol{\eta}_{t-1} = \sqrt{\boldsymbol{v}_t}/\alpha_t - \sqrt{\boldsymbol{v}_{t-1}}/\alpha_{t-1}. \tag{5}$$

Once positive definiteness of Γ_t is violated, Adam and RMSProp may suffer from divergence (Reddi et al., 2018). Based on this observation, two variants of Adam called AMSGrad and AdamNC have been proposed with convergence guarantees in both the convex (Reddi et al., 2018) and non-convex (Chen et al., 2018a) stochastic settings by requiring $\Gamma_t \succ 0$. In addition, Padam (Zhou et al., 2018a) extended from AMSGrad has been proposed to contract the generalization gap in training deep neural networks, whose convergence has been ensured by requiring $\Gamma_t \succ 0$. As a relaxation of $\Gamma_t \succ 0$, Barakat and Bianchi (2020) showed that when $\alpha_t/\sqrt{v_t} \leq \alpha_{t-1}/(c\sqrt{v_{t-1}})$ holds for all t and some positive c, the algorithm Adam can converge. In the strongly convex stochastic setting, by using the long-term memory technique developed in (Reddi et al., 2018), Huang et al. (2018) have proposed NosAdam by attaching more weights on historical second-order moments to ensure its convergence. Prior to that, the convergence rate of RMSProp (Mukkamala and Hein, 2017) has already been established in the convex stochastic setting by employing similar parameters to those of AdamNC (Reddi et al., 2018).

(C2) Adopting a big batch size. Basu et al. (2018), for the first time, showed that deterministic Adam and RMSProp with original iteration schemes are convergent by using a full-batch gradient. On the other hand, both Adam and RMSProp can be reshaped as specific signSGD-type algorithms (Balles and Hennig, 2018; Bernstein et al., 2018) whose $\mathcal{O}(1/\sqrt{T})$ convergence rates have been provided in the non-convex stochastic setting by setting batch size as large as the number of maximum iterations (Bernstein et al., 2018). Recently, Zaheer et al. (2018) have established $\mathcal{O}(1/\sqrt{T})$ convergence rate of original Adam directly in the non-convex stochastic setting by requiring the batch size to be the same order as the number of maximum iterations. We comment that this type of requirement is impractical when Adam and RMSProp are applied to tackle large-scale problems (1), since these approaches cost a huge number of computations to estimate big-batch stochastic gradients in each iteration.

(C3) Incorporating a temporal decorrelation. By exploring the structure of the convex counterexample in (Reddi et al., 2018), Zhou et al. (2018b) have pointed out that the divergence of RMSProp is fundamentally caused by the imbalanced learning rate rather than the absence of $\Gamma_t \succ 0$. Based on this viewpoint, Zhou et al. (2018b) have proposed AdaShift by incorporating a temporal decorrelation technique to eliminate the inappropriate correlation between $\boldsymbol{v}_{t,k}$ and the current second-order moment $\boldsymbol{g}_{t,k}^2$, in which the adaptive learning rate $\eta_{t,k}$ is required to be independent of $\boldsymbol{g}_{t,k}^2$. However, the convergence of AdaShift in (Zhou et al., 2018b) was merely restricted to RMSProp for solving the convex counterexample in (Reddi et al., 2018).

(C4) Seeking an analogous surrogate. Due to the divergences of Adam and RMSProp (Reddi et al., 2018), Zou et al. (2018) proposed a class of new surrogates called AdaUSM to approximate Adam and RMSProp by integrating weighted AdaGrad with a unified heavy ball and Nesterov accelerated gradient momentums. Its $\mathcal{O}(\log(T)/\sqrt{T})$ convergence rate has also been provided in the non-convex stochastic setting by requiring a non-decreasing weighted sequence. Besides, many other adaptive stochastic algorithms without combining momentums, such as AdaGrad (Ward et al., 2018; Li and Orabona, 2019) and stagewise AdaGrad (Chen et al., 2018b), have been guaranteed to be convergent and work well in the non-convex stochastic setting.

In contrast with the above four types of modifications and restrictions, we introduce an alternative easy-to-check sufficient condition (abbreviated as **(SC)**) to guarantee the global convergences of original Adam. The proposed **(SC)** merely depends on the parameters in estimating $\boldsymbol{v}_{t,k}$ and base learning rate α_t. **(SC)** neither requires the positive definiteness of Γ_t like **(C1)** nor needs the batch size as large as the same order as the number of maximum iterations like **(C2)** in both the convex and non-convex stochastic settings. Thus, it is easier to verify and more practical compared with **(C1)**-**(C3)**. On the other hand, **(SC)** is partially overlapped with **(C1)** since the proposed **(SC)** can cover AdamNC (Reddi et al., 2018), AdaGrad with exponential moving average (AdaEMA) momentum (Chen et al., 2018a), and RMSProp (Mukkamala and Hein, 2017) as instances whose convergences are all originally motivated by requiring the positive definiteness of Γ_t. While, based on **(SC)**, we can directly derive their global convergences in the non-convex stochastic setting as byproducts without checking the positive definiteness of Γ_t step by step. Besides, **(SC)** can serve as an alternative explanation on divergences of original Adam and RMSProp, which are possibly due to incorrect parameter settings for accumulating the historical second-order moments rather than the imbalanced learning rate caused by the inappropriate correlation between $\boldsymbol{v}_{t,k}$ and $\boldsymbol{g}_{t,k}^2$ like **(C3)**. In addition, AdamNC and AdaEMA are convergent under **(SC)**, but violate **(C3)** in each iteration. Meanwhile, there are lots of work improving upper bounds for the above algorithms, e.g., Défossez et al. (2020) improved the constants related to β by introducing a novel average scheme in the analysis.

2.3 Mini-batch Stochastic Gradient Methods

In practice, people usually use mini-batch stochastic gradient methods instead of single sample stochastic gradient methods or full gradient methods for faster convergence. For mini-batch SGD algorithms, Li et al. (2014) have shown that mini-batch SGD boosts $\mathcal{O}(\frac{1}{\sqrt[4]{T}})$ convergence rate of SGD to $\mathcal{O}(\frac{1}{\sqrt[4]{sT}})$ where s is the mini-batch size. However, as it is much harder to show the convergence of adaptive gradient methods, few works analyze how sample size will affect the convergence of the adaptive gradient algorithms. Li and Orabona (2019) gave an analysis on Adagrad and showed the convergence rate is linear in the sample size. Zaheer et al. (2018) gave the analysis on Adam, showing that large batch size can help convergence, but the batch size should increase with iteration increasing, which may not be practical. In this work, we theoretically show that mini-batch Adam can be accelerated by adopting a larger mini-batch size as mini-batch SGD (Li et al., 2014) in the same order.

2.4 Distributed Stochastic Gradient Methods

Distributed stochastic gradient descent was first introduced in Agarwal and Duchi (2011) in the parameter-server setting. Further, in the decentralized setting, Lian et al. (2017) gave the analysis on the stochastic gradient descent. The analysis shows that the convergence speed will be linear in the number of workers in the parameter-server setting or will be linear to some constant related to the decentralized graph structure. For the adaptive gradient methods, in the parameter-server setting, Reddi et al. (2020) gave algorithms in the federated scenario called FedAdam, FedAdagrad, and FedYogi. Moreover, they showed that the convergence speed will be linear in the number of workers. However, instead divided by $\sqrt{\boldsymbol{v}_t}$, they divide the gradient with $\sqrt{\boldsymbol{v_t} + \boldsymbol{\epsilon}}$. Meanwhile, in their assumptions, ϵ in the algorithm should be in the order of $O(\frac{G}{L})$, where G is the upper bound of gradient norm, and L is the Lipschitz constant of the objective function. However, in practice, ϵ is always set to be a small value, much smaller than G/L. On the other hand, the large ϵ may dominate the adaptive term in their algorithms. Hence, their methods may degrade to stochastic gradient descent. Carnevale et al. (2020) shoed that Adam with gradient tracking method can be linearly accelerated with an increasing number of nodes in the decentralized and strongly convex setting. Still, it is unclear whether, in the nonconvex setting, this linear speedup will hold when Adam is used. Moreover, Chen et al. (2020) gave an analysis of Adagrad and showed the convergence speed will be linear in the number of workers. Meanwhile, Nazari et al. (2019) gave the analysis of Adagrad in the decentralized setting. Xie et al. (2019) also gave a variant on Adagrad algorithm called AdaAlter in the centralized setting and showed the convergence will linearly speed up by increasing the number of workers. Recently, Chen et al. (2021, 2022) extend Adam to the distributed quantized Adam with error compensation technique Stich et al. (2018). However, the linear speedup property in (Chen et al., 2021, 2022) does not hold. To the best of our knowledge, whether the distributed Adam can achieve a linear speedup is still open. This paper theoretically demonstrates that the distributed Adam in the parameter-server model can achieve a linear speedup concerning the number of workers.

3. Novel Sufficient Condition for Convergence of Adam

In this section, we characterize the upper-bound of gradient residual of problem (1) as a function of parameters (θ_t, α_t). Then the convergence rate of Generic Adam is derived directly by specifying appropriate parameters (θ_t, α_t). Below, we state the necessary assumptions that are commonly used for analyzing the convergence of a stochastic algorithm for non-convex problems:

(A1) The minimum value of problem (1) is lower-bounded, *i.e.*, $f^* = \min_{\boldsymbol{x}\in\mathbb{R}^d} \ f(\boldsymbol{x}) > -\infty$;

(A2) The gradient of f is L-Lipschitz continuous, *i.e.*, $\|\boldsymbol{\nabla} f(\boldsymbol{x}) - \boldsymbol{\nabla} f(\boldsymbol{y})\| \leq L\|\boldsymbol{x}-\boldsymbol{y}\|$, $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$;

(A3) The stochastic gradient $\boldsymbol{g}_t$ is an unbiased estimate, *i.e.*, $\mathbb{E}\left[\boldsymbol{g}_t\right] = \boldsymbol{\nabla} f_t(\boldsymbol{x}_t)$;

(A4) The second-order moment of stochastic gradient $\boldsymbol{g}_t$ is uniformly upper-bounded, *i.e.*, $\mathbb{E}\left\|\boldsymbol{g}_t\right\|^2 \leq G$.

In addition, we also suppose that the parameters $\{\beta_t\}$, $\{\theta_t\}$, and $\{\alpha_t\}$ satisfy the following restrictions:

(R1) The parameters $\{\beta_t\}$ satisfy $0 \leq \beta_t \leq \beta < 1$ for all t for some constant β;

(R2) The parameters $\{\theta_t\}$ satisfy $0 < \theta_t < 1$ and θ_t is non-decreasing in t with $\theta := \lim\limits_{t\to\infty} \theta_t > \beta^2$;

(R3) The parameters $\{\alpha_t\}$ satisfy that $\chi_t := \frac{\alpha_t}{\sqrt{1-\theta_t}}$ is "almost" non-increasing in t, by which we mean that there exist a non-increasing sequence $\{a_t\}$ and a positive constant C_0 independent of t such that $a_t \leq \chi_t \leq C_0 a_t$.

The restriction (R3) indeed says that χ_t is the product between some non-increasing sequence $\{a_t\}$ and some bounded sequence. This is a slight generalization of χ_t itself being non-decreasing. If χ_t itself is non-increasing, we can then take $a_t = \chi_t$ and $C_0 = 1$. For most of the well-known Adam-type methods, χ_t is indeed non-decreasing. For instance, for AdaGrad with EMA momentum we have $\alpha_t = \eta/\sqrt{t}$ and $\theta_t = 1 - 1/t$, so $\chi_t = \eta$ is constant; for Adam with constant $\theta_t = \theta$ and non-increasing α_t (say $\alpha_t = \eta/\sqrt{t}$ or $\alpha_t = \eta$), $\chi_t = \alpha_t/\sqrt{1-\theta}$ is non-increasing. The motivation, instead of χ_t being decreasing, is that it allows us to deal with the bias correction steps in Adam (Kingma and Ba, 2014).

We fix a positive constant $\theta' > 0$[1] such that $\beta^2 < \theta' < \theta$. Let $\gamma := \beta^2/\theta' < 1$ and

$$C_1 := \prod_{j=1}^{N} \big(\frac{\theta_j}{\theta'}\big), \tag{6}$$

where N is the maximum of the indices j with $\theta_j < \theta'$. The finiteness of N is guaranteed by the fact that $\lim_{t\to\infty} \theta_t = \theta > \theta'$. When there are no such indices, *i.e.*, $\theta_1 \geq \theta'$, we take $C_1 = 1$ by convention. In general, $C_1 \leq 1$. Our main results on estimating gradient residual state as follows:

Theorem 2. *Let $\{\boldsymbol{x}_t\}$ be a sequence generated by Generic Adam for initial values $\boldsymbol{x}_1$, $\boldsymbol{m}_0 = \boldsymbol{0}$, and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$. Assume that f and stochastic gradients $\boldsymbol{g}_t$ satisfy assumptions (A1)-(A4). Let τ be randomly chosen from $\{1, 2, \ldots, T\}$ with equal probabilities $p_\tau = 1/T$. Then, we have*

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|\right] \leq \sqrt{\frac{C + C' \sum_{t=1}^{T} \alpha_t \sqrt{1-\theta_t}}{T\alpha_T}},$$

where $C' = 2C_0^2 C_3 d\sqrt{G^2+\epsilon d}/[(1-\beta)\theta_1]$ and

$$C = \frac{2C_0\sqrt{G^2+\epsilon d}}{1-\beta}\big[(C_4 + C_3 C_0 d\chi_1 \log\big(1+\frac{G^2}{\epsilon d}\big)\big],$$

in which C_4 and C_3 are defined as $C_4 = f(x_1) - f^*$ and $C_3 = \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\big[\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2} + 2\big(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\big)^2 G\big]$, respectively.

Remark 3. *Below, we give two comments on the above Theorem:*
(i) *The constants C and C' depend on priory known constants $C_0, C_1, \beta, \theta', G, L, \epsilon, d, f^*, \theta_1, \alpha_1, \boldsymbol{x}_1$.*
(ii) *Convergence in expectation in the above theorem is on the term $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|]$ which is slightly weaker than $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^2]$. The latter form holds for most SGD variants with global learning rates, namely, the learning rate for each coordinate is the same, because $\frac{1}{\sum_{t=1}^{T}\alpha_t}\mathbb{E}\sum_{t=1}^{T}\alpha_t \|\nabla f(\boldsymbol{x}_t)\|^2$ is exactly $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^2]$ if τ is randomly selected via distribution $\mathbb{P}(\tau = k) = \frac{\alpha_k}{\sum_{t=1}^{T}\alpha_t}$. This does not apply to coordinate-wise adaptive stochastic methods because the learning rate for each coordinate is different, and hence unable to randomly select an index according to some distribution uniform for each coordinate. On the other hand, the convergence rates of AMSGrad and AdaEMA (Chen et al., 2018a) are able to achieve the bound for $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|^2]$. This is due to the strong assumption $\|g_t\| \leq G$ which results in a uniform lower bound for each coordinate of the adaptive learning rate $\eta_{t,k} \geq \alpha_t/G$. Thus, the proof of AMSGrad (Chen et al., 2018a) can be dealt with in a way similar to the case of global learning rate. In our paper we use a coordinate-wise adaptive learning rate and assume a weaker assumption $\mathbb{E}[\|g_t\|^2] \leq G$ instead of $\|g_t\|^2 \leq G$.*

1. In the special case that $\theta_t = \theta$ is constant, we can directly set $\theta' = \theta$.

3.1 Proof Sketch of Theorem 2

In this section, we will give some important lemmas that will be used to prove Theorem 2. First, for simplicity, we give some notations used in the lemmas and proof. Denote $\hat{\boldsymbol{v}}_t = \theta_t \boldsymbol{v}_{t-1} + (1 - \theta_t)\mathbb{E}\left[\boldsymbol{g}_t^2 | \boldsymbol{x}_t\right]$, $\hat{\boldsymbol{\eta}}_t = \alpha_t / \sqrt{\hat{\boldsymbol{v}}_t}$, $\boldsymbol{\Delta}_t = \alpha_t \boldsymbol{m}_t / \sqrt{\boldsymbol{v}_t}$, and $\|\boldsymbol{x}\|_{\hat{\boldsymbol{\eta}}}^2 = \sum_{i=1}^d \hat{\boldsymbol{\eta}}_i \boldsymbol{x}_i^2$.

Lemma 4. *Let $\{\boldsymbol{x}_t\}$ be the sequence generated by Algorithm 1. For $T \geq 1$, it holds that*

$$f^* \leq f(\boldsymbol{x}_1) + \sum_{t=1}^{T} \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + L \left\|\boldsymbol{\Delta}_t\right\|^2\right].$$

Lemma 4 is widely used to prove convergence of the stochastic gradient algorithms. To further prove the convergence related to gradient norm, we introduce the following lemma to bound term $\sum_{t=1}^T \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + L \left\|\boldsymbol{\Delta}_t\right\|^2\right]$.

Lemma 5. *(Lemma 33 in Appendix) Let $\{\boldsymbol{x}_t\}$ be the sequence generated by Algorithm 1. For $T \geq 1$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + L \left\|\boldsymbol{\Delta}_t\right\|^2\right] \leq -\frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right] + \frac{\zeta_0}{\theta_1}\sum_{t=1}^{T} \alpha_t \sqrt{1-\theta_t} + \zeta_1,$$

for some constants ζ_0 and ζ_1.

With Lemma 5, we obtain the upper bound for $\mathbb{E}\left[\sum_{t=1}^T \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]$ with respect to $\alpha_t,\ \theta_t,$ and β. To prove Lemma 5, the most important step is to remove dependence between the adaptive learning rate and stochastic gradient. Hence, $\hat{\eta}_t$ is introduced, as it is independent of the stochastic gradient. The rest of the proof is just bounding the error of introducing $\hat{\boldsymbol{\eta}}_t$ instead of using $\alpha_t/\sqrt{\boldsymbol{v}_t}$. Then, the last step is to build connection between $\frac{1}{T}\mathbb{E}[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|]$ and $\mathbb{E}\left[\sum_{t=1}^T \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\boldsymbol{\eta}_t}^2\right]$.

Lemma 6. *(Lemma 36 in Appendix) Let τ be an integer that is randomly chosen from $\{1, 2, \cdots, T\}$ with equal probabilities. We have the following estimate*

$$\mathbb{E}[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|] \leq \sqrt{\frac{\zeta_2}{T\alpha_T}\mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]},$$

for some constant ζ_2.

Thus, using the above three lemmas, we can prove Theorem 2.

3.2 Discussion of Theorem 2

Corollary 7. *Take $\alpha_t = \eta/t^s$ with $0 \leq s < 1$. Suppose $\lim_{t\to\infty} \theta_t = \theta < 1$. Define $Bound(T) := \frac{C + C'\sum_{t=1}^T \alpha_t \sqrt{1-\theta_t}}{T\alpha_T}$. Then the $Bound(T)$ in Theorem 2 is bounded from below by constants*

$$Bound(T) \geq C'\sqrt{1-\theta}. \tag{7}$$

In particular, when $\theta_t = \theta < 1$, we have the following more subtle estimate of lower and upper-bounds for $Bound(T)$

$$\frac{C}{\eta T^{1-s}} + C'\sqrt{1-\theta} \leq Bound(T) \leq \frac{C}{\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{1-s}.$$

Remark 8. **(i)** *Corollary 7 shows that if $\lim_{t\to\infty}\theta_t = \theta < 1$, the bound in Theorem 2 is only $\mathcal{O}(1)$, hence not guaranteeing convergence. This result is not surprising as Adam with constant θ_t has already shown to be divergent (Reddi et al., 2018). Hence, $\mathcal{O}(1)$ is its best convergence rate we can expect. We will discuss this case in more details in Section 3.4.*
(ii) *Corollary 7 also indicates that in order to guarantee convergence, the parameter has to satisfy $\lim_{t\to\infty}\theta_t = 1$. Although we do not assume this in our restrictions (R1)-(R3), it turns out to be the consequence from our analysis. Note that if $\beta < 1$ in (R1) and $\lim_{t\to\infty}\theta_t = 1$, then the restriction $\lim_{t\to\infty}\theta_t > \beta^2$ is automatically satisfied in (R2).*

We are now ready to give the Sufficient Condition for convergence of Generic Adam.

Corollary 9 (Sufficient Condition(**SC**))**.** *Generic Adam is convergent if the parameters $\{\alpha_t\}$, $\{\beta_t\}$, and $\{\theta_t\}$ satisfy the following four conditions:*

1. $0 \leq \beta_t \leq \beta < 1$*;*
2. $0 < \theta_t < 1$ *and θ_t is non-decreasing in t;*
3. $\chi_t := \alpha_t/\sqrt{1-\theta_t}$ *is "almost" non-increasing;*
4. $\left(\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}\right)/\left(T\alpha_T\right) = o(1)$.

3.3 Convergence Rate of Generic Adam

We now provide the convergence rate of Generic Adam with specific parameters $\{(\theta_t, \alpha_t)\}$, *i.e.*,

$$\alpha_t = \eta/t^s \ \text{ and } \ \theta_t = \begin{cases} 1-\alpha/K^r, & t < K, \\ 1-\alpha/t^r, & t \geq K, \end{cases} \tag{8}$$

for positive constants α, η, K, where K is taken such that $\alpha/K^r < 1$. Note that α can be taken bigger than 1. When $\alpha < 1$, we can take $K = 1$ and then $\theta_t = 1 - \alpha/t^r, t \geq 1$. To guarantee (R3), we require $r \leq 2s$. For such a family of parameters we have the following corollary.

Corollary 10. *Generic Adam with the above family of parameters (i.e. (8)) converges as long as $0 < r \leq 2s < 2$, and its non-asymptotic convergence rate is given by*

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|] \leq \begin{cases} \mathcal{O}(T^{-r/4}), & r/2+s<1 \\ \mathcal{O}(\sqrt{\log(T)/T^{1-s}}), & r/2+s=1 \\ \mathcal{O}(1/T^{(1-s)/2}), & r/2+s>1 \end{cases}.$$

Remark 11. *Corollary 10 recovers and extends the results of some well-known algorithms below:*

- ***AdaGrad with exponential moving average (EMA)***. *When $\theta_t = 1 - 1/t$, $\alpha_t = \eta/\sqrt{t}$, and $\beta_t = \beta < 1$, Generic Adam is exactly AdaGrad with EMA momentum (AdaEMA) (Chen et al., 2018a). In particular, if $\beta = 0$, this is the vanilla coordinate-wise AdaGrad. It corresponds to taking $r = 1$ and $s = 1/2$ in Corollary 10. Hence, AdaEMA has convergence rate $\mathcal{O}\left(\sqrt{\log(T)/\sqrt{T}}\right)$.*

- ***AdamNC***. *Taking $\theta_t = 1 - 1/t$, $\alpha_t = \eta/\sqrt{t}$, and $\beta_t = \beta\lambda^t$ in Generic Adam, where $\lambda < 1$ is the decay factor for the momentum factors β_t, we recover AdamNC (Reddi et al., 2018). Its $\mathcal{O}\left(\sqrt{\log(T)/\sqrt{T}}\right)$ convergence rate can be directly derived via Corollary 10.*

- ***RMSProp***. *Mukkamala and Hein (2017) have reached the same $\mathcal{O}\left(\sqrt{\log(T)/\sqrt{T}}\right)$ convergence rate for RMSprop with $\theta_t = 1 - \alpha/t$, when $0 < \alpha \leq 1$ and $\alpha_t = \eta/\sqrt{t}$ under the convex assumption. Since RMSProp is essentially Generic Adam with all momentum factors $\beta_t = 0$, we recover Mukkamala and Hein's results by taking $r = 1$ and $s = 1/2$ in Corollary 10. Moreover, our result generalizes to the non-convex stochastic setting, and it holds for all $\alpha > 0$ rather than only $0 < \alpha \leq 1$.*

The summarization of the above algorithms is provided in Table 1.

Algorithm	θ_t	α_t	β_t	Convergence Result
Adagrad with EMA (Chen et al. (2018a))	$1 - 1/t$	$\eta/\sqrt{t}$	β	$\mathcal{O}(\sqrt{\log T/\sqrt{T}})$
AdamNC (Reddi et al. (2018))	$1 - 1/t$	$\eta/\sqrt{t}$	$\beta\lambda^t$	$\mathcal{O}(\sqrt{\log T/\sqrt{T}})$
RMSProp (Mukkamala and Hein, 2017)	$1 - \alpha/t$	$\eta/\sqrt{t}$	0	$\mathcal{O}(\sqrt{\log T/\sqrt{T}})$

Table 1: Convergence results for some well-known algorithms

Comparison with Reddi et al. (2018) . Most of the convergent modifications of original Adam, such as AMSGrad, AdamNC, and NosAdam, all require $\Gamma_t \succ 0$ in Eq. (5), which is equivalent to decreasing the adaptive learning rate η_t step by step. Since the term Γ_t (or adaptive learning rate η_t) involves the past stochastic gradients (hence not deterministic), the modification to guarantee $\Gamma_t \succ 0$ either needs to change the iteration scheme of Adam (like AMSGrad) or needs to impose some strong restrictions on the base learning rates α_t and θ_t (like AdamNC). Our sufficient condition provides an easy-to-check criterion for the convergence of Generic Adam in Corollary 9. It is not necessary to require $\Gamma_t \succ 0$. Moreover, we use exactly the same iteration scheme as the original Adam without any modifications. Our work shows that the positive definiteness of Γ_t may not be an essential issue for the divergence of the original Adam. The divergence may be due to the incorrect setting of moving average parameters instead of non-positive definiteness of Γ_t.

3.4 Constant θ_t case: insights for divergence

The currently most popular RMSProp and Adam's parameter setting takes constant θ_t, *i.e.*, $\theta_t = \theta < 1$. The motivation behind is to use the exponential moving average of squares of past stochastic gradients. In practice, parameter θ is recommended to be set very close to 1. For instance, a commonly adopted θ is taken as 0.999.

Although great performance in practice has been observed, such a constant parameter setting has the serious flaw that there is no convergence guarantee even for convex optimization, as proved by the counterexamples in (Reddi et al., 2018). Ever since much work has been done to analyze the divergence issue of Adam and to propose modifications with convergence guarantees, as summarized in the introduction section. However, there is still not a satisfactory explanation that touches the fundamental reason for the divergence. In this section, we try to provide more insights for the divergence issue of Adam/RMSProp with constant parameter θ_t, based on our analysis of the sufficient condition for convergence.

From the sufficient condition perspective. Let $\alpha_t = \eta/t^s$ for $0 \le s < 1$ and $\theta_t = \theta < 1$. According to Corollary 7, $Bound(T)$ in Theorem 2 has the following estimate:

$$\frac{C}{\eta T^{1-s}} + C'\sqrt{1-\theta} \le Bound(T) \le \frac{C}{\eta T^{1-s}} + \frac{C'\sqrt{1-\theta}}{(1-s)}.$$

The bounds tell us some points on Adam with constant θ_t:

1. $Bound(T) = \mathcal{O}(1)$, so the convergence is not guaranteed. This result coincides with the divergence issue demonstrated in (Reddi et al., 2018). Indeed, since in this case Adam is not convergent, this is the best bound we can have.

2. Consider the dependence on parameter s. The bound is decreasing in s. The best bound in this case is when $s = 0$, *i.e.*, the base learning rate is taken constant. This explains why in practice taking a more aggressive constant base learning rate often leads to even better performance, comparing with taking a decaying one.

3. Consider the dependence on parameter θ. Note that the constants C and C' depend on θ_1 instead of the whole sequence θ_t. We can always set $\theta_t = \theta$ for $t \ge 2$ while fix $\theta_1 < \theta$, by which we can take C and C' independent of constant θ. Then the principal term of $Bound(T)$ is linear in $\sqrt{1-\theta}$, so decreases to zero as $\theta \to 1$. This explains why setting θ close to 1 often results in better performance in practice.

Moreover, Corollary 10 shows us how the convergence rate continuously changes when we continuously vary parameters θ_t. Let us fix $\alpha_t = 1/\sqrt{t}$ and consider the following continuous family of parameters $\{\theta_t^{(r)}\}$ with $r \in [0, 1]$:

$$\theta_t^{(r)} = 1 - \alpha^{(r)}/t^r, \ \text{ where } \ \alpha^{(r)} = r\bar{\theta} + (1 - \bar{\theta}), \ 0 < \bar{\theta} < 1.$$

Note that when $r = 1$, then $\theta_t = 1 - 1/t$, this is the AdaEMA, which has the convergence rate $\mathcal{O}\left(\sqrt{\log T/\sqrt{T}}\right)$; when $r = 0$, then $\theta_t = \bar{\theta} < 1$, this is the original Adam with constant θ_t, which only has the $\mathcal{O}(1)$ bound; when $0 < r < 1$, by Corollary 10, the algorithm has the $\mathcal{O}(T^{-r/4})$ convergence rate. Along with this continuous family of parameters, we observe that the theoretical convergence rate continuously deteriorates as the real parameter r decreases from 1 to 0, namely, as we gradually shift from AdaEMA to Adam with constant θ_t. In the limiting case, the latter is not guaranteed with convergence anymore.

4. Complexity Analysis for Practical Adam: Mini-batch/Distributed Adam

Due to the limited time, limited computational resources, and noise in data collection and processing, it is almost impossible to achieve the accurate stationary point. Thus, instead of achieving the accurate stationary point, people get more attention to achieving some approximated stationary point in practice. The crucial question under this situation will become how much time is needed to achieve some approximated solution. This section will answer this question by answering how many iterations are needed to obtain an ε-stationary point. First, we define ε-stationary point as follows:

Definition 12. *We define a random variable $\boldsymbol{x}$ as an ε-stationary point of problem* (1)*, if*

$$\mathbb{E}_{\boldsymbol{x}}\left[\left\|\mathbb{E}_{\xi\sim\mathbb{P}}\left[\nabla\tilde{f}(\boldsymbol{x},\xi)\right]\right\|\right] \le \varepsilon.$$

According to Definition 12 and Theorem 2, for generic Adam, we can directly give the following corollary:

Corollary 13. *For any $\varepsilon > 0$, if we take $T \ge C_5^2\varepsilon^{-4}$, $\alpha_t = \frac{\alpha}{\sqrt{T}}$, $\beta_t = \beta, \theta_t = 1-\frac{\theta}{T}$, which satisfy $\gamma = \frac{\beta}{1-\frac{\theta}{T}} < 1$ and $\theta_t \ge \frac{1}{4}$, then by taking τ uniformly from $\{1, 2, \cdots, T\}$, it holds that*

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|\,,\right] \le \varepsilon$$

where

$$C_5 = \frac{2\sqrt{G^2+\epsilon d}}{\alpha(1-\beta)}\left[f(x_1) - f^* + C_6 d\frac{\alpha}{\sqrt{\theta}}\log\big(1+\frac{G^2}{\epsilon d}\big) + \frac{4C_6 d\alpha}{\sqrt{\theta}}\right],$$

$$C_6 = \frac{1}{1-\sqrt{\gamma}}\left[\frac{\alpha L}{\sqrt{\theta}(1-\sqrt{\gamma})^2} + 2\big(\frac{2\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)}}+1\big)^2 G\right].$$

Remark 14. *Difference from Corollary 7. Corollary 7 shows when $lim_{t\to\infty}\theta_t < 1$ the algorithm cannot converge to a stationary point. However, because the goal of the algorithm switches to an ε-stationary point, the choice of α_t and θ_t is not contradicting Corollary 7. We will use the same choice in the following section.*

Remark 15. *With the parameter setting in Corollary 10, the number of iterations T should satisfy $\frac{T}{\log^2 T} = \Omega(\varepsilon^{-4})$, instead of $T = \Omega(\varepsilon^{-4})$, which gives a larger iteration number T. However, we can use the same parameter setting in Corollary 10 when T changes, while we need to change parameters in Corollary 13 for different T.*

Remark 16. *From Theorem 13, to achieve an ε-stationary point, only $\Omega(\varepsilon^{-4})$ iterations are needed. Comparing the result with SGD (Li et al., 2014), we have the same order of iterations to achieve an ε-stationary point.*

In the following two sections, we will analyze two practical Adam variations, i.e., mini-batch and distributed Adam. Although they can use the same technique for analysis, we list two algorithms for readers in different communities (single machine learning algorithm (mini-batch Adam) v.s. multi-machine learning algorithm (distributed Adam)).

4.1 Convergence Analysis for Mini-Batch Adam

It has been shown that when s samples are used to estimate the gradient in the stochastic gradient descent algorithm, the convergence speed can be accelerated s times than the single sample algorithms Li et al. (2014). Meanwhile, in practice, the mini-batch technique is widely used to optimize problem (1) such as training a neural network with the Adam algorithm. In this section, we will give the analysis on mini-batch Adam. The Mini-batch Adam algorithm is defined in the following Algorithm 2. Different from Algorithm 1, in Algorithm 2 s samples which are identically distributed and independent when the iterate $\boldsymbol{x}_t$ is used to estimate the gradient $\boldsymbol{\nabla} f(\boldsymbol{x}_t)$. We average the s estimates and use the averaged stochastic gradient, which should be a more accurate estimation to update $\boldsymbol{x}_t$.

Algorithm 2 Mini-batch Adam

1: **Parameters:** Choose $\{\alpha_t\}$, $\{\beta_t\}$, and $\{\theta_t\}$. Choose $\boldsymbol{x}_1 \in \mathbb{R}^d$ and set initial values $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Sample s stochastic gradients: $\boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)}, \cdots, \boldsymbol{g}_t^{(s)}$;
4: Average s stochastic gradients: $\bar{\boldsymbol{g}}_t = \frac{1}{s}\sum_{i=1}^{s} \boldsymbol{g}_t^{(i)}$;
5: **for** $k = 1, 2, \ldots, d$ **do**
6: $\boldsymbol{v}_{t,k} = \theta_t \boldsymbol{v}_{t-1,k} + (1-\theta_t)\bar{\boldsymbol{g}}_{t,k}^2$;
7: $\boldsymbol{m}_{t,k} = \beta_t \boldsymbol{m}_{t-1,k} + (1-\beta_t)\bar{\boldsymbol{g}}_{t,k}$;
8: $\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \alpha_t \boldsymbol{m}_{t,k}/\sqrt{\boldsymbol{v}_{t,k}}$;
9: **end for**
10: **end for**

To link sample size and convergence rate, we give a new assumption on the stochastic gradient and state it as follows:

(A5) The variance of stochastic gradient $\boldsymbol{g}_t$ is uniformly upper-bounded, *i.e.*, $\mathbb{E}\left[\|\boldsymbol{g}_t - \nabla f_t(\boldsymbol{x}_t)\|^2\right] \le \sigma^2$.

We add **(A5)** to establish the relation between sample size and the convergence rate. Intuitively, with an increasing size of samples, the variance of the gradient estimator should reduce. Utilizing this reduction, we can obtain an ε-stationary point, with fewer iterations but a larger sample size. Also, this assumption is widely used in analysis such as Yan et al. (2018). The following results are given under assumptions **(A1)** to **(A5)**, and the result of mini-batch Adam is given as follows:

Theorem 17. *For any $\varepsilon > 0$, if we take $\alpha_t = \frac{\alpha}{\sqrt{T}}$, $\beta_t = \beta$ and $\theta_t = 1 - \frac{\theta}{T}$, which satisfy $\gamma = \frac{\beta_t^2}{\theta_t} < 1$, $\theta_t \ge \frac{1}{4}$ and $F_T(T, s) \le \varepsilon$, then there exists $t \in \{1, 2, \cdots, T\}$ such that*

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] \le \varepsilon,$$

where by taking $\epsilon = \frac{1}{sd}$, it holds that

$$F_T(T, s) = \mathcal{O}(T^{-1}s^{1/2}d^{1/2} + T^{-1/2}d + T^{-1/3}s^{-1/4} + T^{-1/4}s^{-1/4}d^{1/2}).$$

Thus, to achieving an ε-stationary point, $\Omega(\varepsilon^{-4}s^{-1})$ iterations are needed.

More specifically, it holds that

$$F_T(\varepsilon, s) = \frac{1}{\sqrt{T}}\left((4C_8C_{10})^{1/2} + (4C_8C_{11})^{2/3} + (4C_9C_{10}) + (4C_9C_{11})^2)\right),$$
$$C_8 = \frac{\sqrt{T}\sqrt{2\sigma^2\theta + \epsilon sd}}{\sqrt{s}\alpha},\ C_9 = \sqrt{\frac{2\theta}{\alpha^2}},\ C_{11} = \frac{2C_4}{1-\beta}\sqrt[4]{\frac{2\theta}{d\epsilon T}},$$
$$C_{10} = \frac{2}{1-\beta}\left(f(x_1) - f^* + C_7 d\theta + 2C_7 d\sqrt[4]{1 + \frac{2\sigma^2}{d\epsilon s}}\right),$$
$$C_7 = \frac{1}{1-\sqrt{\gamma}}\left(\frac{\alpha^2 L}{\theta\left(1-\sqrt{\gamma}\right)^2} + \frac{2\left(\frac{2\beta/(1-\beta)}{\sqrt{1-\gamma}} + 1\right)^2 G\alpha}{\sqrt{\theta}}\right).$$

Remark 18. *Below, we give three comments on the above results:*
(i) *From Theorem 17, to achieve an ε-stationary point, when we only consider the order with respect to ε, $\Omega(\varepsilon^{-4})$ iterations are needed. Besides, by jointly considering ε and batch size s, we can accelerate the algorithm to achieve an ε-stationary point, where $\Omega(\varepsilon^{-4}s^{-1})$ iterations are needed, which indicates that Mini-batch Adam can be linearly accelerated with respect to the mini-batch size. The result is in the same order of mini-batch SGD in (Li et al., 2014).*
(ii) *Deriving the linear speedup property of mini-batch Adam with respect to mini-batch size is much more difficult than the analysis techniques for mini-batch SGD (Li et al., 2014) since the adaptive learning rate in Algorithm 2 is highly coupled with mini-batch stochastic gradient estimates. In fact, the adaptive learning rate implicitly adjusts the magnitude of the learning rate with respect to mini-batch size, while the hand-crafted learning rate in mini-batch SGD has to be tuned carefully via a linear LR scaling technique (Krizhevsky, 2014; You et al., 2017) for a large mini-batch training.*
(iii) *The dimension dependence of the above analysis is $O(\sqrt{d})$. Meanwhile, some analyses on (variants of) Adam (Chen et al. (2018a); Défossez et al. (2020); Zou et al. (2018)) achieve the same dimension dependence, while Zhou et al. (2018a) showed that in AMSGrad, RMSProp and Adagrad the dependence can be $O(d^{-1/4})$.*

4.2 Proof Sketch of Theorem 17

Similar to the proof of Theorem 2, we first try to bound $\sum_{t=1}^{T} \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t\rangle + L\left\|\boldsymbol{\Delta}_t\right\|^2\right]$. Because we introduce a new assumption that stochastic gradient has bounded variance σ^2, it is easy to verify that the averaged stochastic gradient $\bar{\boldsymbol{g}}_t$ with variance σ^2/s, which is a key property to establish the relation between sample size and convergence rate. To take advantage of variance, together with the constant value assigned to α_t and θ_t, we derive the mini-batch/distributed variants of Lemmas 5 and 6.

Lemma 19. *(Lemma 40 in Appendix) Let $\{\boldsymbol{x}_t\}$ be the sequence generated by Algorithm 2 or 3. For $T \geq 1$, when $\alpha_t = \frac{\alpha}{\sqrt{T}}$ and $\theta_t = 1 - \frac{\theta}{T}$, we have*

$$\sum_{t=1}^{T} \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + L \|\boldsymbol{\Delta}_t\|^2\right] \leq \zeta_3 + \zeta_4 \sqrt[4]{\frac{\sigma^2}{s}} + \frac{\zeta_5}{\sqrt[4]{T}} \sqrt{\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right]} - \frac{1-\beta}{2} \mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\boldsymbol{\eta}_t}\right],$$

for some constants ζ_3, ζ_4 and ζ_5.

Different from Lemma 5 in which ℓ_2 norm of $\boldsymbol{\nabla} f(\boldsymbol{x}_t)$ doesn't occur, in Lemma 19 when taking the benefit of small variance, both $\hat{\boldsymbol{\eta}}_t$ norm and ℓ_2 norm of term $\boldsymbol{\nabla} f(\boldsymbol{x}_t)$ occur in the right hand side. The key step for this lemma is in how we bound $\mathbb{E}(\|g_t\|^2)$, where in Lemma 5, $\mathbb{E}\|g_t\|^2$ is directly bounded by G^2. However, as we want to explore the benefit of having mini-batch, we bound $\mathbb{E}(\|g_t\|^2)$ by $\mathbb{E}(\|\nabla f(x_t)\|^2)$ and σ^2. Hence, because the formulation becomes much more complicated, some further calculation on $\boldsymbol{\nabla} f(\boldsymbol{x}_t)$ is needed and we will introduce the calculation in Lemma 21.

Lemma 20. *(Lemma 41 in Appendix) Let $\{\boldsymbol{x}_t\}$ be the sequence generated by Algorithm 2 or 3. For $T \geq 1$, it holds that*

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right]^2 \leq \left(\sqrt{T}(\zeta_6 \frac{\sigma}{\sqrt{s}} + \zeta_7 \sqrt{\epsilon d}) + \zeta_8 \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right]\right) \mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right],$$

for some constants ζ_6, ζ_7 and ζ_8.

Different from Lemma 6, the left hand side is not the summation of ℓ_2 norm but square root of the summation of squared ℓ_2 norm which is smaller than the summation of ℓ_2. Therefore, Theorem 17 gives a weaker result than Theorem 2, which only shows the existence of $t \in \{1, 2, \cdots, T\}$ such that $\mathbb{E}(\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|) = \mathcal{O}(1/\sqrt[4]{sT})$ rather than $\mathbb{E}(\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|) = \mathcal{O}(1/\sqrt[4]{T})$. After combining Lemma 4, Lemma 19 and Lemma 20, we can get an inequality as follows:

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right]^2 \leq \left(\zeta_9 + \zeta_{10} \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right]\right) \left(\zeta_{11} + \zeta_{12} \sqrt{\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right]}\right). \tag{9}$$

Lemma 21. *When inequality* (9) *holds, it holds that*

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right] \leq (4\zeta_9\zeta_{11})^{1/2} + (4\zeta_9\zeta_{12})^{2/3} + (4\zeta_{10}\zeta_{11}) + (4\zeta_{10}\zeta_{12})^2.$$

Hence, combining Lemma 4, Lemma 19, Lemma 20 and Lemma 21, we are able to prove $\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}\right] = \mathcal{O}(T^{1/4}s^{-1/4})$. By dividing $\sqrt{T}$ on both sides we can prove Theorem 17.

4.3 Convergence Analysis for Distributed Adam

For large-scale problems such as training deep convolutional neural networks over the ImageNet dataset (Russakovsky et al., 2015), it is hard to optimize problem (1) on a single machine. In this section, we extend the mini-batch Adam to the distributed Adam like the distributed SGD method (Yu et al., 2019). The simplest structure is the parameter-server model in the distributed setting, where a parameter server and multiple workers are involved in the optimization process. As it is shown in Algorithm 3, in each iteration, a worker receives the iterate $\boldsymbol{x}_t$ from the server, samples a stochastic gradient with respect to $\boldsymbol{x}_t$, and sends the gradient to the server. Meanwhile, the parameter server receives gradients from workers in each iteration, averages the gradients, and performs an Adam update.

Algorithm 3 Distributed Adam

Parameters: Choose $\{\alpha_t\}$, $\{\beta_t\}$, and $\{\theta_t\}$. Choose $\boldsymbol{x}_1 \in \mathbb{R}^d$ and set initial values $\boldsymbol{m}_0 = \boldsymbol{0}$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$.
For Server: Send $\boldsymbol{x}_1$ to workers;
for $t = 1, 2, \ldots, T$ **do**
  For Worker i**:**
  Receive iterate $\boldsymbol{x}_t$;
  Sample a stochastic gradient $\boldsymbol{g}_t^{(i)}$;
  Send $\boldsymbol{g}_t^{(i)}$ to the server;
  For the Server:
  Receive stochastic gradient $\boldsymbol{g}_t^{(i)}$ from worker i;
  Average s stochastic gradient $\bar{\boldsymbol{g}}_t = \frac{1}{s}\sum_{i=1}^{s} \boldsymbol{g}_t^{(i)}$;
  for $k = 1, 2, \ldots, d$ **do**
    $\boldsymbol{v}_{t,k} = \theta_t \boldsymbol{v}_{t-1,k} + (1-\theta_t)\bar{\boldsymbol{g}}_{t,k}^2$;
    $\boldsymbol{m}_{t,k} = \beta_t \boldsymbol{m}_{t-1,k} + (1-\beta_t)\bar{\boldsymbol{g}}_{t,k}$;
    $\boldsymbol{x}_{t+1,k} = \boldsymbol{x}_{t,k} - \alpha_t \boldsymbol{m}_{t,k} / \sqrt{\boldsymbol{v}_{t,k}}$;
  end for
  Send $\boldsymbol{x}_{t+1}$ to workers;
end for

Proposition 22. *Algorithm 3 with s workers performs the same as Algorithm 2 with s i.i.d. samples.*

Remark 23. *Below, we give two remarks on the above distributed Adam algorithm:*
(i) *For distributed Adam, to achieve an ε-stationary point, $\Omega(\varepsilon^{-4}s^{-1})$ iterations are needed, which is a linear speedup with respect to the number of workers in the network, which is in the same order as that is in distributed SGD (Yu et al., 2019).*
(ii) *Distributed Adam has been popularly used for training deep neural networks. In addition, there also exist several variants of the distributed Adam algorithm, such as PMD-LAMB (Wang et al., 2020), LAMB (You et al., 2019), LARS (You et al., 2017), etc, for training large-scale deep neural networks. However, all these works do not establish the linear speedup property for distributed adaptive methods.*

5. Experimental Results

In this section, we experimentally validate the proposed sufficient condition by applying Generic Adam and RMSProp to solve the counterexample (Chen et al., 2018a) and to train LeNet (LeCun et al., 1998) on the MNIST dataset (LeCun et al., 2010) and ResNet (He et al., 2016) on the CIFAR-100 dataset (Krizhevsky, 2009), respectively. Moreover, we use different batch sizes to train LeNet (LeCun et al., 1998) on the MNIST dataset (LeCun et al., 2010) and ResNet (He et al., 2016) on the CIFAR-100 dataset (Krizhevsky, 2009), respectively, and validate theory of the mini-batch Adam algorithm.

5.1 Synthetic Counterexample

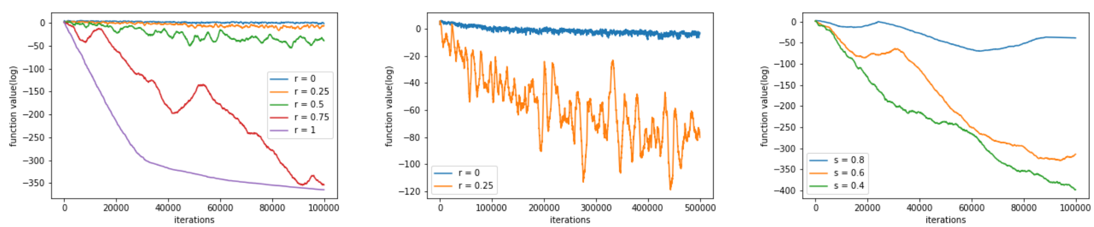

Figure 1: The above figures are for function value with different r and s values, respectively. The left figure plots the performance profiles of Generic Adam with different r values, where $\theta_t^{(r)} = 1 - (0.01 + 0.99r^2)/t^r$. The middle figure plots the performance profiles of Generic Adam with $\theta_t^{(r)} = 1 - \frac{0.01}{t^r}$ and $r = 0$ and 0.25. The right figure plots the performance profiles of Generic Adam with different s values.

In this experiment, we verify the phenomenon described in Section 3.4 how the convergence rate of Generic Adam gradually changes along a continuous path of families of parameters on the one-dimensional counterexample in Chen et al. (2018a):

$$f(x) = \mathbb{E}_\xi[\xi x^2], \tag{10}$$

where $\mathbb{P}(\xi = 5.5) = 1/11$ and $\mathbb{P}(\xi = -0.5) = 10/11$.

Sensitivity of parameter r. We set $T = 10^5$, $\alpha_t = 5/\sqrt{t}$, $\beta = 0.9$, and θ_t as $\theta_t^{(r)} = 1 - (0.01 + 0.99r^2)/t^r$ with $r \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1.0\}$, respectively. Note that when $r = 0$, Generic Adam reduces to the originally divergent Adam (Kingma and Ba, 2014) with $(\beta, \bar{\theta}) = (0.9, 0.99)$. When $r = 1$, Generic Adam reduces to AdaEMA (Chen et al., 2018a) with $\beta = 0.9$.

The experimental results are shown in the left figure of Figure 1. We can see that for $r = 1.0, 0.75$ and 0.5, Generic Adam is convergent. Moreover, the convergence becomes slower when r decreases, which exactly matches Corollary 10. On the other hand, for $r = 0$ and 0.25, Figure 1 shows that they do not converge. It seems that the divergence for $r = 0.25$ contradicts our theory. However, this is because when r is very small, the $\mathcal{O}(T^{-r/4})$ convergence rate is so slow that we may not see a convergent trend in even 10^5 iterations. Indeed, for $r = 0.25$, we actually have

$$\theta_t^{(0.25)} \leq 1 - (0.01 + 0.25 * 0.25 * 0.99)/10^{5*0.25} \approx 0.9960,$$

which is not sufficiently close to 1. As a complementary experiment, we fix the numerator and only change r when r is small. We take α_t and β_t as the same, while $\theta_t^{(r)} = 1 - \frac{0.01}{t^r}$ for $r = 0$ and 0.25, respectively. The result is shown in the middle figure of Figures 1. We can see that Generic Adam with $r = 0.25$ is indeed convergent in this situation.

Sensitivity of parameter s. Now, we show the sensitivity of s of the sufficient condition **(SC)** by fixing $r = 0.8$ and selecting s from the collection $s = \{0.4, 0.6, 0.8\}$. The right figure in Figure 1 illustrates the sensitivity of parameter s when Generic Adam is applied to solve the counterexample (10). The performance shows that when s is fixed, smaller r can lead to a faster and better convergence speed, which also coincides with the convergence results in Corollary 10.

5.2 LeNet on MNIST and ResNet-18 on CIFAR-100

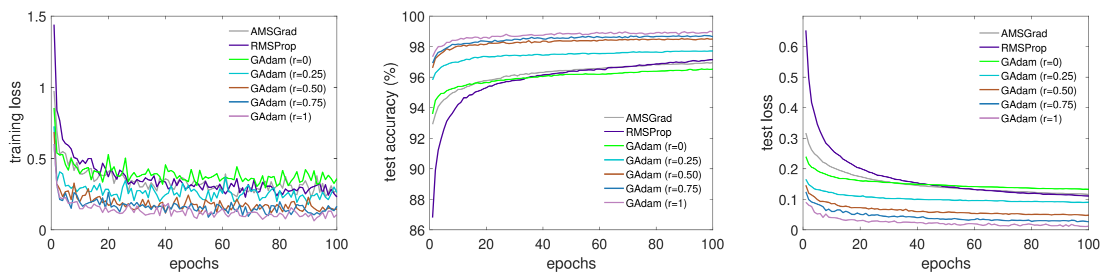

Figure 2: Performance profiles of Generic Adam with $r = \{0, 0.25, 0.5, 0.75, 1\}$, RMSProp, and AMSGrad for training LeNet on the MNIST dataset. Figures (a), (b), and (c) illustrate training loss vs. epochs, test accuracy vs. epochs, and test loss vs. epochs, respectively.

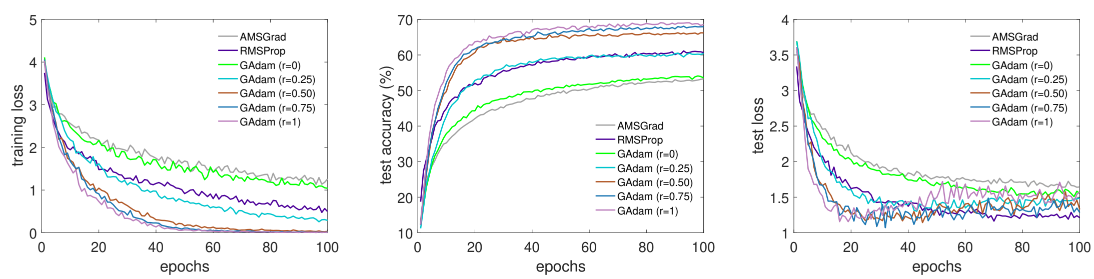

Figure 3: Performance profiles of Generic Adam with $r = \{0, 0.25, 0.5, 0.75, 1\}$, RMSProp, and AMSGrad for training ResNet on the CIFAR-100 dataset. Figures (a), (b), and (c) illustrate training loss vs. epochs, test accuracy vs. epochs, and test loss vs. epochs, respectively.

In this subsection, we apply Generic Adam to train LeNet on the MNIST dataset and ResNet-18 on the CIFAR-100 dataset, respectively, to validate the convergence rates in Corollary 10. Meanwhile, the comparisons between Generic Adam and AMSGrad (Reddi et al., 2018; Chen et al., 2018a) are also provided to distinguish their differences in training deep neural networks. We illustrate the performance profiles in three aspects: training loss vs. epochs, test loss vs. epochs, and test accuracy vs. epochs, respectively. MNIST (LeCun et al., 2010) is composed of ten classes of digits among $\{0, 1, 2, \ldots, 9\}$, which includes 60,000 training examples and 10,000 validation examples. The dimension of each example is 28×28. CIFAR-100 (LeCun et al., 2010) is composed of 100 classes of 32×32 color images. Each class includes 6,000 images. Besides, these images are divided into 50,000 training examples and 10,000 validation examples. LetNet (LeCun et al., 1998) used in the experiments is a five-layer convolutional neural network with ReLU activation function whose detailed architecture is described in (LeCun et al., 1998). The batch size is set as 64. The training stage lasts for 100 epochs in total. No ℓ_2 regularization on the weights is used. ResNet-18 (He et al., 2016) is a ResNet model containing 18 convolutional layers for CIFAR-100 classification (He et al., 2016). Input images are down-scaled to $1/8$ of their original sizes after the 18 convolutional layers and then fed into a fully-connected layer for the 100-class classification. The output channel numbers of 1-3 conv layers, 4-8 conv layers, 9-13 conv layers, and 14-18 conv layers are 64, 128, 256, and 512, respectively. The batch size is 64. The training stage lasts for 100 epochs in total. No ℓ_2 regularization on the weights is used.

In the experiments, for Generic Adam, we set $\theta_t^{(r)} = 1 - (0.001 + 0.999r)/t^r$ with $r \in \{0, 0.25, 0.5, 0.75, 1\}$ and $\beta_t = 0.9$, respectively; for RMSProp, we set $\beta_t = 0$ and $\theta_t = 1 - \frac{1}{t}$ along with the parameter settings in Mukkamala and Hein (2017). For fairness, the base learning rates α_t in Generic Adam, RMSProp, and AMSGrad are all set as $0.001/\sqrt{t}$. Figures 2 and 3 illustrate the results of Generic Adam with different r, RMSProp, and AMSGrad for training LeNet on MNIST and training ResNet-18 on CIFAR-100, respectively. We can see that AMSGrad and Adam (Generic Adam with $r = 0$) decrease the training loss slowest and show the worst test accuracy among the compared optimizers. One possible reason is due to the use of constant θ in AMSGrad and original Adam. By Figures 2 and 3, we can observe that the convergences of Generic Adam are extremely sensitive to the choice of parameter θ_t. Larger r can contribute to a faster convergence rate of Generic Adam, which corroborates the theoretical result in Corollary 10. Additionally, the test accuracies in Figures 2(b) and 3(b) indicate that a smaller training loss can contribute to a higher test accuracy for Generic Adam.

5.3 Experiments on Practical Adam

5.3.1 ABLATION STUDY BETWEEN BATCHSIZE AND OPTIMAL LEARNING RATE

In this section, we apply mini-batch Adam algorithms and mini-batch SGD algorithms to the following quadratic minimization task:

$$\min_x \mathbb{E}_{\xi \sim \mathcal{N}(0,100I)} \left[x^T A x - b^T x + \xi^T x \right],$$

where, for simplicity, $A \in \mathbb{R}^{3\times 3}$ and $b \in \mathbb{R}^3$. We optimize x with batch size 1 to 320, and grid search the best learning rate from $\{k/2000|k = 1, 2, \cdots, 2000\}$. For each pair of batch size and learning rate, we randomly sample 500 trials and take the average gradient norm as the criteria. Fig. 4 shows the result of the best learning rate and averaged gradient norm for different batch sizes after 200 optimization iterations. From the figure, we can verify that when the batch size becomes larger, the optimal learning rate for SGD and SGD-momentum increases a lot (from 0.005 to 0.04). Meanwhile, with the adaptive learning rate, the optimal learning rate does not change too much (between 0.02 to 0.04). Hence, it shows the benefit of adaptive learning rate methods compared to SGD and verifies Remark 18.

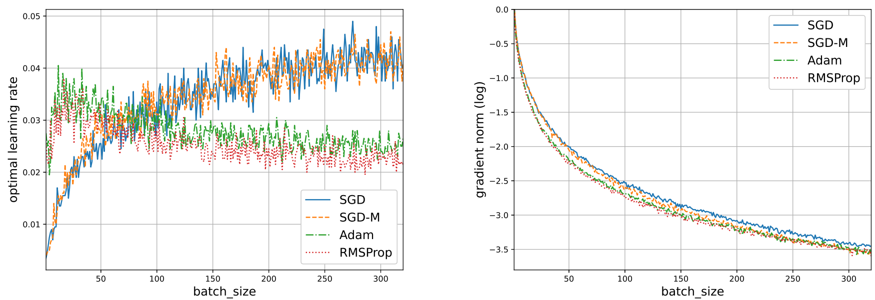

Figure 4: Performance profiles of optimal learning rate v.s. batch size and optimality gap v.s. batch size.

5.3.2 VISION TASKS

In this section we apply the mini-batch Adam algorithm to train LeNet on the MNIST dataset and ResNet-18 on the CIFAR-100 dataset. Datasets and network architecture are the same as they are described in Section 5.2. But instead of using $\alpha_t = \alpha/\sqrt{t}$, $\theta_t = 1 - \theta/t$, we set $\beta_t = 0.9$ and $\theta_t = 0.99$ for Adam and AMSGrad, and set $\beta_t = 0.9$ for RMSProp. We use different batchsizes $\{32, 64, 128\}$ to train networks. Besides, when training ResNet-18 on the CIFAR100 dataset, we use an ℓ_2 regularization on weights, the coefficient of the regularization term to $5e-4$. We use grid search in $[1e-2,\ 5e-3,\ 1e-3,\ 5e-4,\ 1e-4]$ for α_t with respect to test accuracy. In addition, when training ResNet-18 on the CIFAR100 dataset, α_t will reduce to $0.2 \times \alpha_t$ every 19550 iterations (50 epochs for the 128 batchsize setting), which (learning rate decay) is commonly used in practice . The experimental results are shown in Figure 5 and Figure 6. It can be shown that larger batchsize can give lower training loss in all experiments. However, large batchsize for training does not imply higher test accuracy or lower test loss, which needs to be further explored and examined.

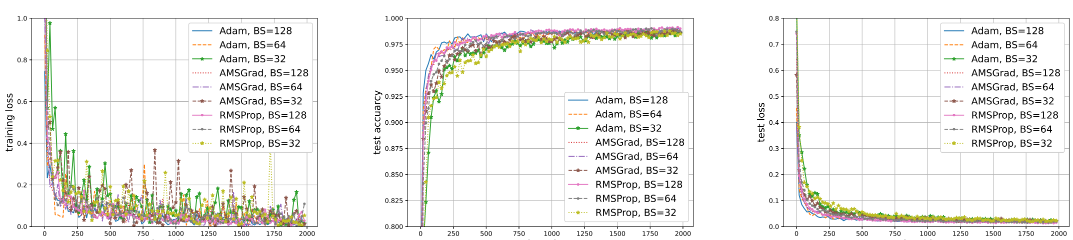

Figure 5: Performance profiles of mini-batch Adam, RMSProp and AMSGrad on MNIST, with batchsize = $\{32,\ 64,\ 128\}$.

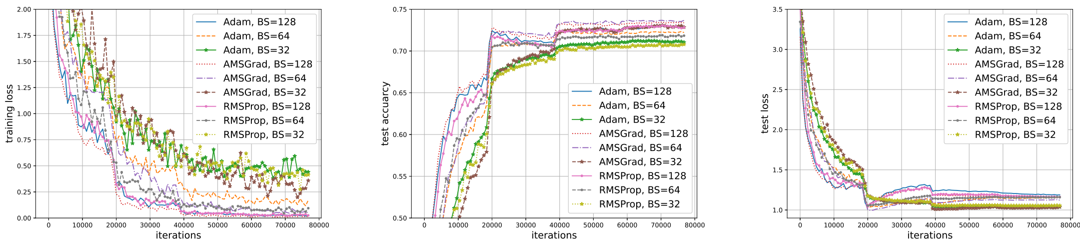

Figure 6: Performance profiles of mini-batch Adam, RMSProp and AMSGrad on CIFAR100, with batchsize = $\{32,\ 64,\ 128\}$.

5.3.3 Transformer XL on WikiText-103

Also, we applied mini-batch Adam to train a base model of Transformer XL (Dai et al., 2019) on the dataset WikiText-103 (Merity et al., 2016). The base model of Transformer XL contains 16 self-attention layers. In each self-attention layer, there are 10 heads, and the encoding dimension of each head is set to 41. The WikiText-103 dataset is a collection of over 100 million tokens extracted from the set of verified ‘Good’ and ‘Featured’ articles on Wikipedia. We adopt the same parameter settings provided by the authors but test on batch size $\{30, 60, 120\}$. The results are shown in Figure 7. Also, as it is shown in the figure, a larger batch size can give a lower training loss in all experiments. Meanwhile, in the figure, AMSGrad and Adam achieve much better performance than SGD-M, which shows the benefit of using adaptive methods instead of SGD-based methods, just like what was mentioned in Zhang et al. (2019).

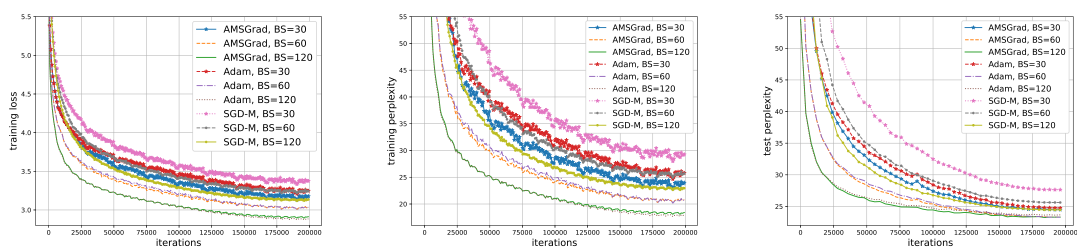

Figure 7: Performance profiles of mini-batch Adam, SGD-M and AMSGrad on WikiText-103, with batchsize = $\{30,\ 60,\ 120\}$.

6. Conclusions

In this work, we delved into the convergences of Adam, and presented an easy-to-check sufficient condition to guarantee their convergences in the non-convex stochastic setting. This sufficient condition merely depends on the base learning rate and the linear combination parameter of second-order moments. Relying on this sufficient condition, we found that the divergences of Adam are possibly due to the incorrect parameter settings. Besides, when encountering the practice Adam, we theoretically showed that the number of samples will linearly speed up the convergence in both the mini-batch setting and distributed setting, which closes the gap between theory and practice. At last, the correctness of theoretical results has also been verified via the counterexample and training deep neural networks on real-world datasets.

Acknowledgement

This work is supported by the Major Science and Technology Innovation 2030 "Brain Science and Brain-like Research" key project (No. 2021ZD0201405).

Appendix A. Key Lemma to prove Theorem 2 and Theorem 13

In this section we provide the necessary lemmas for the proofs of Theorem 2 and Theorem 13. First, we give some notations for simplifying the following proof.

Notations We use bold letters to represent vectors. The k-th component of a vector $\boldsymbol{v}_t$ is denoted as $v_{t,k}$. The inner product between two vectors $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ is denoted as $\langle \boldsymbol{v}_t, \boldsymbol{w}_t \rangle$. Other than that, all computations that involve vectors shall be understood in the component-wise way. We say a vector $\boldsymbol{v}_t \geq 0$ if every component of $\boldsymbol{v}_t$ is non-negative, and $\boldsymbol{v}_t \geq \boldsymbol{w}_t$ if $v_{t,k} \geq w_{t,k}$ for all $k = 1, 2, \ldots, d$. The ℓ_1 norm of a vector $\boldsymbol{v}_t$ is defined as $\|\boldsymbol{v}_t\|_1 = \sum_{k=1}^{d} |v_{t,k}|$. The ℓ_2 norm is defined as $\|\boldsymbol{v}_t\|^2 = \langle \boldsymbol{v}_t, \boldsymbol{v}_t \rangle = \sum_{k=1}^{d} |v_{t,k}|^2$. Given a positive vector $\hat{\boldsymbol{\eta}}_t$, it will be helpful to define the following weighted norm: $\|\boldsymbol{v}_t\|_{\boldsymbol{\eta}_t}^2 = \langle \boldsymbol{v}_t, \hat{\boldsymbol{\eta}}_t \boldsymbol{v}_t \rangle = \sum_{k=1}^{d} \hat{\eta}_{t,k} |v_{t,k}|^2$.

Lemma 24. *Given $S_0 > 0$ and a non-negative sequence $\{s_t\}$, let $S_t = S_0 + \sum_{i=1}^{t} s_i$ for $t \geq 1$. Then the following estimate holds*

$$\sum_{t=1}^{T} \frac{s_t}{S_t} \leq \log(S_T) - \log(S_0). \tag{11}$$

Proof The finite sum $\sum_{t=1}^{T} s_t/S_t$ can be interpreted as a Riemann sum $\sum_{t=1}^{T} (S_t - S_{t-1})/S_t$. Since $1/x$ is decreasing on the interval $(0, \infty)$, we have

$$\sum_{t=1}^{T} \frac{S_t - S_{t-1}}{S_t} \leq \int_{S_0}^{S_T} \frac{1}{x} dx = \log(S_T) - \log(S_0).$$

The proof is completed. ∎

Lemma 25 (Abel's Lemma - Summation by parts)**.** *Let $\{u_t\}$ and $\{s_t\}$ be two non-negative sequences. Let $S_t = \sum_{i=1}^{t} s_i$ for $t \geq 1$. Then*

$$\sum_{t=1}^{T} u_t s_t = \sum_{t=1}^{T-1} (u_t - u_{t+1}) S_t + u_T S_T. \tag{12}$$

Proof Let $S_0 = 0$. Then

$$\sum_{t=1}^{T} u_t s_t = \sum_{t=1}^{T} u_t (S_t - S_{t-1}) = \sum_{t=1}^{T-1} u_t S_t - \sum_{t=1}^{T-1} u_{t+1} S_t + u_T S_T = \sum_{t=1}^{T-1} (u_t - u_{t+1}) S_t + u_T S_T. \tag{13}$$

The proof is completed. ∎

Lemma 26. *Let $\{\theta_t\}$ and $\{\alpha_t\}$ satisfy the restrictions (R2) and (R3). For any $i \le t$, we have*

$$\chi_t \le C_0 \chi_i \;\text{ and }\; \alpha_t \le C_0 \alpha_i. \tag{14}$$

Proof For any $i \le t$, since the sequence $\{a_t\}$ is non-increasing, we have $a_t \le a_i$. Hence,

$$\chi_t = \frac{\alpha_t}{\sqrt{1-\theta_t}} \le C_0 a_t \le C_0 a_i \le C_0 \frac{\alpha_i}{\sqrt{1-\theta_i}} = C_0 \chi_i,$$

which proves the first inequality. On the other hand, since $\{\theta_t\}$ is non-decreasing, it holds

$$\alpha_t \le C_0 \frac{\sqrt{1-\theta_t}}{\sqrt{1-\theta_i}} \alpha_i \le C_0 \alpha_i = C_0 \alpha_i.$$

The proof is completed. ∎

Let $\Theta_{(t,i)} = \prod_{j=i+1}^{t} \theta_j$ for $i < t$ and $\Theta_{(t,t)} = 1$ by convention.

Lemma 27. *Fix a constant θ' with $\beta^2 < \theta' < \theta$. Let C_1 be as given as Eq. (6) in the main paper. For any $i \le t$, we have*

$$\Theta_{(t,i)} \ge C_1 (\theta')^{t-i}. \tag{15}$$

Proof For any $i \le t$, since $\theta_j \ge \theta'$ for $j \ge N$, and $\theta_j < \theta'$ for $j < N$, we have

$$\Theta_{(t,i)} = \prod_{j=i+1}^{t} \theta_j \ge \left(\prod_{j=i+1}^{N} \theta_j \right) (\theta')^{t-N} = \left(\prod_{j=i+1}^{N} (\theta_j/\theta') \right) (\theta')^{t-i} \ge \left(\prod_{j=1}^{N} (\theta_j/\theta') \right) (\theta')^{t-i}.$$

We take the constant $C_1 = \prod_{j=1}^{N} (\theta_j/\theta')$, where N is the maximum of the indices for which $\theta_j < \theta'$. The proof is completed. ∎

Remark 28. *If $\theta_t = \theta$ is a constant, we have $\Theta_{(t,i)} = \theta^{t-i}$. In this case we can take $\theta' = \theta$ and $C_1 = 1$.*

Lemma 29. *Let $\gamma := \beta^2/\theta'$. We have the following estimate*

$$\boldsymbol{m_t}^2 \le \frac{1}{C_1(1-\gamma)(1-\theta_t)} \boldsymbol{v}_t, \; \forall t. \tag{16}$$

Proof Let $B_{(t,i)} = \prod_{j=i+1}^{t} \beta_j$ for $i < t$ and $B_{(t,t)} = 1$ by convention. By the iteration formula $\boldsymbol{m}_t = \beta_t \boldsymbol{m}_{t-1} + (1-\beta_t)\boldsymbol{g}_t$ and $\boldsymbol{m}_0 = \boldsymbol{0}$, we have

$$\boldsymbol{m}_t = \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} \beta_j \right) (1-\beta_i)\boldsymbol{g}_i = \sum_{i=1}^{t} B_{(t,i)} (1-\beta_i) \boldsymbol{g}_i.$$

Similarly, by $\boldsymbol{v}_t = \theta_t \boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{g}_t^2$ and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$, we have

$$\boldsymbol{v}_t = \left(\prod_{j=1}^{t} \theta_j\right)\boldsymbol{\epsilon} + \sum_{i=1}^{t}\left(\prod_{j=i+1}^{t}\theta_j\right)(1-\theta_i)\,\boldsymbol{g}_i^2 \geq \sum_{i=1}^{t}\Theta_{(t,i)}(1-\theta_i)\boldsymbol{g}_i^2.$$

It follows by arithmetic inequality that

$$\begin{aligned}\boldsymbol{m}_t^2 &= \left(\sum_{i=1}^{t}\frac{(1-\beta_i)B_{(t,i)}}{\sqrt{(1-\theta_i)\Theta_{(t,i)}}}\sqrt{(1-\theta_i)\Theta_{(t,i)}}\boldsymbol{g}_i\right)^2 \\ &\leq \left(\sum_{i=1}^{t}\frac{(1-\beta_i)^2B_{(t,i)}^2}{(1-\theta_i)\Theta_{(t,i)}}\right)\left(\sum_{i=1}^{t}\Theta_{(t,i)}(1-\theta_i)\boldsymbol{g}_i^2\right) \leq \left(\sum_{i=1}^{t}\frac{(1-\beta_i)^2B_{(t,i)}^2}{(1-\theta_i)\Theta_{(t,i)}}\right)\boldsymbol{v}_t.\end{aligned}$$

Note that $\{\theta_t\}$ is non-decreasing by (**R**2), and $B_{(t,i)} \leq \beta^{t-i}$ by (**R**1). By Lemma 27, we have

$$\sum_{i=1}^{t}\frac{(1-\beta_i)^2B_{(t,i)}^2}{(1-\theta_i)\Theta_{(t,i)}} \leq \frac{1}{C_1(1-\theta_t)}\sum_{i=1}^{t}\left(\frac{\beta^2}{\theta'}\right)^{t-i} \leq \frac{1}{C_1(1-\theta_t)}\sum_{k=0}^{t-1}\gamma^k \leq \frac{1}{C_1(1-\gamma)(1-\theta_t)}.$$

The proof is completed. ∎

Let $\boldsymbol{\Delta}_t := \boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\alpha_t\boldsymbol{m}_t/\sqrt{\boldsymbol{v}_t}$. Let $\hat{\boldsymbol{v}}_t = \theta_t\boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{\delta}_t^2$, where $\boldsymbol{\delta}_t^2 = \mathbb{E}_t\left[\boldsymbol{g}_t^2\right]$ and let $\hat{\boldsymbol{\eta}}_t = \alpha_t/\sqrt{\hat{\boldsymbol{v}}_t}$.

Lemma 30. *With the notations above, the following equality holds*

$$\boldsymbol{\Delta}_t - \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}\boldsymbol{\Delta}_{t-1} = -(1-\beta_t)\hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t + \hat{\boldsymbol{\eta}}_t\boldsymbol{g}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\boldsymbol{A}_t + \hat{\boldsymbol{\eta}}_t\boldsymbol{\delta}_t\frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\boldsymbol{B}_t, \tag{17}$$

where

$$\begin{aligned}\boldsymbol{A}_t &= \frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}} + \frac{(1-\beta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}}, \\ \boldsymbol{B}_t &= \left(\frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{\delta}_t}{\sqrt{\hat{\boldsymbol{v}}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right) - \frac{(1-\beta_t)\boldsymbol{\delta}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}}.\end{aligned}$$

Proof We have

$$\begin{aligned}\boldsymbol{\Delta}_t - \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}\boldsymbol{\Delta}_{t-1} = & -\frac{\alpha_t\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}} + \frac{\beta_t\alpha_t\boldsymbol{m}_{t-1}}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}} = -\alpha_t\left(\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}} - \frac{\beta_t\boldsymbol{m}_{t-1}}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right) \\ = & \underbrace{-\frac{(1-\beta_t)\alpha_t\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}}_{\text{(I)}} + \underbrace{\beta_t\alpha_t\boldsymbol{m}_{t-1}\left(\frac{1}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{v}_t}}\right)}_{\text{(II)}}.\end{aligned} \tag{18}$$

For (I) we have

$$\begin{aligned}
\text{(I)} =&\ \frac{(1-\beta_t)\alpha_t \boldsymbol{g}_t}{\sqrt{\hat{\boldsymbol{v}}_t}} + (1-\beta_t)\alpha_t \boldsymbol{g}_t \left(\frac{1}{\sqrt{\boldsymbol{v}_t}} - \frac{1}{\sqrt{\hat{\boldsymbol{v}}_t}} \right) \\
=&\ (1-\beta_t)\hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t + (1-\beta_t)\alpha_t \boldsymbol{g}_t \frac{(1-\theta_t)(\boldsymbol{\delta}_t^2 - \boldsymbol{g}_t^2)}{\sqrt{\boldsymbol{v}_t}\sqrt{\hat{\boldsymbol{v}}_t}(\sqrt{\boldsymbol{v}_t} + \sqrt{\hat{\boldsymbol{v}}_t})} \\
=&\ (1-\beta_t)\hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t + \hat{\boldsymbol{\eta}}_t \boldsymbol{\delta}_t \frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \frac{(1-\beta_t)\boldsymbol{\delta}_t}{\sqrt{\boldsymbol{v}_t} + \sqrt{\hat{\boldsymbol{v}}_t}} - \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \frac{(1-\beta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t} + \sqrt{\hat{\boldsymbol{v}}_t}}.
\end{aligned} \tag{19}$$

For (II) we have

$$\begin{aligned}
\text{(II)} =&\ \beta_t \alpha_t \boldsymbol{m}_{t-1} \frac{(1-\theta_t)\boldsymbol{g}_t^2}{\sqrt{\boldsymbol{v}_t}\sqrt{\theta_t \boldsymbol{v}_{t-1}}(\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}})} \\
=&\ \beta_t \alpha_t \boldsymbol{m}_{t-1} \frac{(1-\theta_t)\boldsymbol{g}_t^2}{\sqrt{\boldsymbol{v}_t}\sqrt{\hat{\boldsymbol{v}}_t}(\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}})} + \beta_t \alpha_t \boldsymbol{m}_{t-1} \frac{(1-\theta_t)\boldsymbol{g}_t^2}{\sqrt{\boldsymbol{v}_t}(\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}})} \left(\frac{1}{\sqrt{\theta_t \boldsymbol{v}_{t-1}}} - \frac{1}{\sqrt{\hat{\boldsymbol{v}}_t}} \right) \\
=&\ \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \left(\frac{\beta_t \boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}}} \right) + \frac{\beta_t \alpha_t \boldsymbol{m}_{t-1}(1-\theta_t)^2 \boldsymbol{g}_t^2 \boldsymbol{\delta}_t^2}{\sqrt{\boldsymbol{v}_t}\sqrt{\hat{\boldsymbol{v}}_t}\sqrt{\theta_t \boldsymbol{v}_{t-1}}(\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}})(\sqrt{\hat{\boldsymbol{v}}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}})} \\
=&\ \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \left(\frac{\beta_t \boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}}} \right) + \hat{\boldsymbol{\eta}}_t \boldsymbol{\delta}_t \frac{(1-\theta_t)\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \left(\frac{\beta_t \boldsymbol{m}_{t-1}}{\sqrt{\theta_t \boldsymbol{v}_{t-1}}} \frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}}} \frac{\sqrt{1-\theta_t}\boldsymbol{\delta}_t}{\sqrt{\hat{\boldsymbol{v}}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}}} \right).
\end{aligned} \tag{20}$$

Combining Eq. (19) and Eq. (20), we obtain the desired Eq. (17). The proof is completed. ∎

Lemma 31. *Let* $M_t = \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + L \left\| \boldsymbol{\Delta}_t \right\|^2\right]$ *and* $\chi_t = \alpha_t / \sqrt{1-\theta_t}$. *Then for any* $t \geq 2$, *we have*

$$M_t \leq \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} M_{t-1} + L\,\mathbb{E}\left[\|\boldsymbol{\Delta}_t\|^2\right] + C_2 G \chi_t \mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \tag{21}$$

and

$$M_1 \leq L\,\mathbb{E}\left[\|\boldsymbol{\Delta}_1\|^2\right] + C_2 G \chi_1 \mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_1}{\sqrt{v}_1}\right\|^2\right], \tag{22}$$

where $C_2 = 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2$.

Proof First, for $t \geq 2$ we have

$$\mathbb{E}\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle = \underbrace{\frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \mathbb{E}\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_{t-1} \rangle}_{\text{(I)}} + \underbrace{\mathbb{E}\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t - \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \boldsymbol{\Delta}_{t-1} \right\rangle}_{\text{(II)}}. \tag{23}$$

To estimate (I), by the Schwartz inequality and the Lipschitz continuity of the gradient of f, we have

$$\begin{aligned} \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_{t-1} \rangle \leq & \ \langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t) - \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle \\ \leq & \ \langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + L \left\| \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \right\| \left\| \boldsymbol{\Delta}_{t-1} \right\| \\ = & \ \langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + L \left\| \boldsymbol{\Delta}_{t-1} \right\|^2 . \end{aligned} \tag{24}$$

Hence, we have

$$\text{(I)} \leq \frac{\beta_t \alpha_t}{\sqrt{\theta_t} \alpha_{t-1}} \mathbb{E} \left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_{t-1}), \boldsymbol{\Delta}_{t-1} \rangle + L \left\| \boldsymbol{\Delta}_{t-1} \right\|^2 \right] = \frac{\beta_t \alpha_t}{\sqrt{\theta_t} \alpha_{t-1}} M_{t-1}. \tag{25}$$

To estimate (II), by Lemma 30, we have

$$\begin{aligned} & \mathbb{E} \left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t - \frac{\beta_t \alpha_t}{\sqrt{\theta_t} \alpha_{t-1}} \boldsymbol{\Delta}_{t-1} \right\rangle \\ = & -(1-\beta_t) \mathbb{E} \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \rangle \underbrace{- \mathbb{E} \left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \frac{(1-\theta_t) \boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \boldsymbol{A}_t \right\rangle}_{\text{(III)}} \underbrace{- \mathbb{E} \left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{\delta}_t \frac{(1-\theta_t) \boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \boldsymbol{B}_t \right\rangle}_{\text{(IV)}} . \end{aligned} \tag{26}$$

Note that $\hat{\boldsymbol{\eta}}_t$ is independent of $\boldsymbol{g}_t$ and $\mathbb{E}_t[\boldsymbol{g}_t] = \boldsymbol{\nabla} f(\boldsymbol{x}_t)$. Hence, for the first term in the right hand side of Eq. (26), we have

$$\begin{aligned} -(1-\beta_t) \mathbb{E} \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \boldsymbol{g}_t \rangle &= -(1-\beta_t) \mathbb{E} \left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \hat{\boldsymbol{\eta}}_t \mathbb{E}_t[\boldsymbol{g}_t] \right\rangle \\ &= -(1-\beta_t) \mathbb{E} \left\| \boldsymbol{\nabla} f(\boldsymbol{x}_t) \right\|^2_{\hat{\boldsymbol{\eta}}_t} \\ &\leq -(1-\beta) \mathbb{E} \left\| \boldsymbol{\nabla} f(\boldsymbol{x}_t) \right\|^2_{\hat{\boldsymbol{\eta}}_t} . \end{aligned} \tag{27}$$

To estimate (III), we have

$$\text{(III)} \ \leq \mathbb{E} \left\langle \frac{\sqrt{\hat{\boldsymbol{\eta}}_t} |\boldsymbol{\nabla} f(\boldsymbol{x}_t)| |\boldsymbol{g}_t|}{\boldsymbol{\delta}_t}, \frac{\sqrt{\hat{\boldsymbol{\eta}}_t} \boldsymbol{\delta}_t |\boldsymbol{A}_t| (1-\theta_t) |\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle . \tag{28}$$

Note that $\boldsymbol{\delta}_t \leq G$. Therefore,

$$\sqrt{\hat{\boldsymbol{\eta}}_t} \boldsymbol{\delta}_t = \sqrt{\hat{\boldsymbol{\eta}}_t \boldsymbol{\delta}_t^2} = \sqrt{\frac{\alpha_t \boldsymbol{\delta}_t^2}{\sqrt{\hat{\boldsymbol{v}}_t}}} \leq \sqrt{\frac{\alpha_t \boldsymbol{\delta}_t^2}{\sqrt{(1-\theta_t) \boldsymbol{\delta}_t^2}}} \leq \sqrt{\frac{G \alpha_t}{\sqrt{1-\theta_t}}} = \sqrt{G \chi_t}. \tag{29}$$

On the other hand,

$$|\boldsymbol{A}_t| = \left| \frac{\beta_t \boldsymbol{m}_{t-1}}{\sqrt{\boldsymbol{v}_t} + \sqrt{\theta_t \boldsymbol{v}_{t-1}}} + \frac{(1-\beta_t) \boldsymbol{g}_t}{\sqrt{\boldsymbol{v}}_t + \sqrt{\hat{\boldsymbol{v}}_t}} \right| \leq \frac{\beta_t |\boldsymbol{m}_{t-1}|}{\sqrt{\theta_t \boldsymbol{v}_{t-1}}} + \frac{(1-\beta_t) |\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}}. \tag{30}$$

By Lemma 29, we have

$$\frac{|\boldsymbol{m}_{t-1}|}{\sqrt{\boldsymbol{v}_{t-1}}} \leq \frac{1}{\sqrt{C_1 (1-\gamma)(1-\theta_t)}}. \tag{31}$$

Meanwhile,

$$\frac{|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \leq \frac{|\boldsymbol{g}_t|}{\sqrt{(1-\theta_t)\boldsymbol{g}_t^2}} = \frac{1}{\sqrt{1-\theta_t}}. \tag{32}$$

Hence, we have

$$\begin{aligned} |\boldsymbol{A}_t| &\leq \frac{\beta_t}{\sqrt{C_1(1-\gamma)(1-\theta_t)\theta_t}} + \frac{1-\beta_t}{\sqrt{1-\theta_t}} \leq \left(\frac{\beta_t/(1-\beta_t)}{\sqrt{C_1(1-\gamma)\theta_t}} + 1\right)\frac{1-\beta_t}{\sqrt{1-\theta_t}} \\ &\leq \left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)\frac{1-\beta_t}{\sqrt{1-\theta_t}} := \frac{C_2'(1-\beta_t)}{\sqrt{1-\theta_t}}, \end{aligned} \tag{33}$$

where $C_2' = \left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)$. The last inequality holds due to $\beta_t/(1-\beta_t) \leq \beta/(1-\beta)$ as $\beta_t \leq \beta$. Therefore, we have

$$\begin{aligned} &\left\langle \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)||\boldsymbol{g}_t|}{\boldsymbol{\delta}_t}, \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}\boldsymbol{\delta}_t|\boldsymbol{A}_t|(1-\theta_t)|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\left\langle \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)||\boldsymbol{g}_t|}{\boldsymbol{\delta}_t}, \sqrt{G\chi_t}C_2'(1-\beta_t)\frac{\sqrt{1-\theta_t}|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\frac{1-\beta_t}{4}\left\|\frac{\sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)||\boldsymbol{g}_t|}{\boldsymbol{\delta}_t}\right\|^2 + C_2'^2 G(1-\beta_t)\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 \\ \leq\ &\frac{1-\beta_t}{4}\left\|\frac{\hat{\boldsymbol{\eta}}_t|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|^2|\boldsymbol{g}_t|^2}{\boldsymbol{\delta}_t^2}\right\|_1 + C_2'^2 G\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \end{aligned} \tag{34}$$

Note that $\boldsymbol{\delta}_t^2 = \mathbb{E}_t[\boldsymbol{g}_t^2]$. Hence,

$$\mathbb{E}_t\left\|\frac{\hat{\boldsymbol{\eta}}_t|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|^2|\boldsymbol{g}_t|^2}{\boldsymbol{\delta}_t^2}\right\|_1 = \left\|\hat{\boldsymbol{\eta}}_t|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|^2\right\|_1 = \|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2. \tag{35}$$

Combining Eq. (28), Eq. (34), and Eq. (35), we obtain

$$\text{(III)} \leq \frac{1-\beta_t}{4}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right] + C_2'^2 G\chi_t\mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \tag{36}$$

The term (IV) is estimated similarly as term (III). First, we have

$$\begin{aligned} |\boldsymbol{B}_t| \leq\ &\left(\frac{\beta_t|\boldsymbol{m}_{t-1}|}{\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\frac{\sqrt{1-\theta_t}\boldsymbol{\delta}_t}{\sqrt{\hat{\boldsymbol{v}}_t}+\sqrt{\theta_t\boldsymbol{v}_{t-1}}}\right) + \frac{(1-\beta_t)\boldsymbol{\delta}_t}{\sqrt{\boldsymbol{v}_t}+\sqrt{\hat{\boldsymbol{v}}_t}} \\ \leq\ &\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)\frac{1-\beta_t}{\sqrt{1-\theta_t}} = \frac{C_2'(1-\beta_t)}{\sqrt{1-\theta_t}}, \end{aligned} \tag{37}$$

where C_2' is the constant defined above. We have

$$\begin{aligned} \text{(IV)} \leq\ &\mathbb{E}\left\langle \sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|, \frac{\sqrt{\hat{\boldsymbol{\eta}}_t}\boldsymbol{\delta}_t|\boldsymbol{B}_t|(1-\theta_t)|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\mathbb{E}\left\langle \sqrt{\hat{\boldsymbol{\eta}}_t}|\boldsymbol{\nabla} f(\boldsymbol{x}_t)|, \sqrt{G\chi_t}C_2'(1-\beta_t)\frac{\sqrt{1-\theta_t}|\boldsymbol{g}_t|}{\sqrt{\boldsymbol{v}_t}} \right\rangle \\ \leq\ &\frac{1-\beta_t}{4}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right] + C_2'^2 G\chi_t\mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \end{aligned} \tag{38}$$

Combining Eq. (23), Eq. (24), Eq. (26), Eq. (27), Eq. (36), and Eq. (38), we obtain

$$\begin{aligned}\mathbb{E}\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t\rangle \ &\leq \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}M_{t-1} + 2C_2'^2 G\chi_t \mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 - \frac{1-\beta_t}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \\ &\leq \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}M_{t-1} + 2C_2'^2 G\chi_t \mathbb{E}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right].\end{aligned} \tag{39}$$

Let C_2 denote the constant $2(C_2')^2$. Then $C_2 = 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}}+1\right)^2$. Thus, we obtain Eq. (21) by adding the term $L\mathbb{E}\left[\|\boldsymbol{\Delta}_t\|^2\right]$ to both sides of Eq. (39). Next, we estimate Eq. (22). When $t=1$, we have

$$M_1 = \ \mathbb{E}\left[-\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_1), \frac{\alpha_1\boldsymbol{m}_1}{\sqrt{\boldsymbol{v}_1}}\right\rangle + L\left\|\boldsymbol{\Delta}_1\right\|^2\right] = \mathbb{E}\left[-\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_1), \frac{\alpha_1(1-\beta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\right\rangle + L\left\|\boldsymbol{\Delta}_1\right\|^2\right]. \tag{40}$$

The same as what we did for term (I) in Lemma 30, we have

$$\frac{(1-\beta_1)\alpha_1\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_t}} = (1-\beta_1)\hat{\boldsymbol{\eta}}_1\boldsymbol{g}_1 + \hat{\boldsymbol{\eta}}_1\boldsymbol{\delta}_1\frac{(1-\theta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\frac{(1-\beta_1)\boldsymbol{\delta}_1}{\sqrt{\boldsymbol{v}_1}+\sqrt{\hat{\boldsymbol{v}}_1}} - \hat{\boldsymbol{\eta}}_1\boldsymbol{g}_1\frac{(1-\theta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\frac{(1-\beta_1)\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}+\sqrt{\hat{\boldsymbol{v}}_1}}. \tag{41}$$

Then the similar argument as Eq. (34) implies that

$$\begin{aligned}\mathbb{E}\left[-\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}_1), \frac{\alpha_1\boldsymbol{m}_1}{\sqrt{\boldsymbol{v}_1}}\right\rangle\right] &\leq \ C_2 G\chi_1\mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\right\|^2\right] - \frac{1-\beta_1}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_1)\|^2_{\hat{\boldsymbol{\eta}}_1}\right] \\ &\leq \ C_2 G\chi_1\mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_1}{\sqrt{\boldsymbol{v}_1}}\right\|^2\right].\end{aligned} \tag{42}$$

Combining Eq. (40) and Eq. (42), and adding both sides by $L\mathbb{E}\left[\|\boldsymbol{\Delta}\|_1^2\right]$, we obtain Eq. (22). This completes the proof. ■

Lemma 32. *The following estimate holds*

$$\sum_{t=1}^T \|\boldsymbol{\Delta}_t\|^2 \leq \frac{C_0^2\chi_1}{C_1(1-\sqrt{\gamma})^2}\sum_{t=1}^T \chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2. \tag{43}$$

Proof Note that $\boldsymbol{v}_t \geq \theta_t\boldsymbol{v}_{t-1}$, so we have $\boldsymbol{v}_t \geq \left(\prod_{j=i+1}^t \theta_j\right)\boldsymbol{v}_i = \Theta_{(t,i)}\boldsymbol{v}_i$. By Lemma 27, this follows that $\boldsymbol{v}_t \geq C_1(\theta')^{t-i}\boldsymbol{v}_i$ for all $i \leq t$. On the other hand,

$$|\boldsymbol{m}_t| \leq \sum_{i=1}^t\left(\prod_{j=i+1}^t \beta_j\right)(1-\beta_i)|\boldsymbol{g}_i| \leq \sum_{i=1}^t \beta^{t-i}|\boldsymbol{g}_i|.$$

It follows that

$$\frac{|\boldsymbol{m}_t|}{\sqrt{\boldsymbol{v}_t}} \leq \sum_{i=1}^{t} \frac{\beta^{t-i}|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_t}} \leq \frac{1}{\sqrt{C_1}} \sum_{i=1}^{t} \left(\frac{\beta}{\sqrt{\theta'}}\right)^{t-i} \frac{|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_i}} = \frac{1}{\sqrt{C_1}} \sum_{i=1}^{t} \sqrt{\gamma}^{t-i} \frac{|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_i}}. \tag{44}$$

Since $\alpha_t = \chi_t\sqrt{1-\theta_t} \leq \chi_t\sqrt{1-\theta_i}$ for $i \leq t$, it follows that

$$\begin{aligned}\|\boldsymbol{\Delta}_t\|^2 = \left\|\frac{\alpha_t \boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 \leq\ & \frac{\chi_t^2}{C_1}\left\|\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\frac{\sqrt{1-\theta_i}|\boldsymbol{g}_i|}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \leq \frac{\chi_t^2}{C_1}\left(\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\right)\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \\ \leq\ & \frac{\chi_t^2}{C_1(1-\sqrt{\gamma})}\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2.\end{aligned} \tag{45}$$

By Lemma 26, $\chi_t \leq C_0\chi_i, \forall i \leq t$. Hence,

$$\|\boldsymbol{\Delta}_t\|^2 = \left\|\frac{\alpha_t \boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 \leq \frac{C_0^2\chi_1}{C_1(1-\sqrt{\gamma})}\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\chi_i\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2. \tag{46}$$

It follows that

$$\begin{aligned}\sum_{t=1}^{T}\|\boldsymbol{\Delta}_t\|^2 \leq\ & \frac{C_0^2\chi_1}{C_1(1-\sqrt{\gamma})}\sum_{t=1}^{T}\sum_{i=1}^{t}\sqrt{\gamma}^{t-i}\chi_i\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \\ =\ & \frac{C_0^2\chi_1}{C_1(1-\sqrt{\gamma})}\sum_{i=1}^{T}\left(\sum_{t=i}^{T}\sqrt{\gamma}^{t-i}\right)\chi_i\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \\ \leq\ & \frac{C_0^2\chi_1}{C_1(1-\sqrt{\gamma})^2}\sum_{i=1}^{T}\chi_i\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2.\end{aligned} \tag{47}$$

The proof is completed. ■

Lemma 33. *(Lemma 5 in Section 3) Let $M_t = \mathbb{E}\left[\langle\boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t\rangle + L\|\boldsymbol{\Delta}_t\|^2\right]$. For $T \geq 1$ we have*

$$\sum_{t=1}^{T} M_t \leq C_3\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right]. \tag{48}$$

where the constant C_3 is given by

$$C_3 = \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2} + 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2 G\right).$$

Proof Let $N_t = L\mathbb{E}\left[\|\boldsymbol{\Delta}_t\|^2\right] + C_2 G\chi_t\mathbb{E}\left[\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right]$. By Lemma 31, we have $M_1 \leq N_1$ and

$$M_t \leq \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}M_{t-1} + N_t - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \leq \frac{\beta_t\alpha_t}{\sqrt{\theta_t}\alpha_{t-1}}M_{t-1} + N_t. \tag{49}$$

It is straightforward to acquire by induction that

$$
\begin{aligned}
M_t &\le \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} \frac{\beta_{t-1}\alpha_{t-1}}{\sqrt{\theta_{t-1}}\alpha_{t-2}} M_{t-2} + \frac{\beta_t \alpha_t}{\sqrt{\theta_t}\alpha_{t-1}} N_{t-1} + N_t - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \\
&\ \ \vdots \\
&\le \frac{\alpha_t B_{(t,1)}}{\alpha_1\sqrt{\Theta_{(t,1)}}} M_1 + \sum_{i=2}^{t} \frac{\alpha_t B_{(t,i)}}{\alpha_i\sqrt{\Theta_{(t,i)}}} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \\
&\le \sum_{i=1}^{t} \frac{\alpha_t B_{(t,i)}}{\alpha_i\sqrt{\Theta_{(t,i)}}} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right].
\end{aligned} \tag{50}
$$

By Lemma 26, it holds $\alpha_t \le C_0 \alpha_i$ for any $i \le t$. By Lemma 27, $\Theta_{(t,i)} \ge C_1(\theta')^{t-i}$. In addition, $B_{(t,i)} \le \beta^{t-i}$. Hence,

$$
\begin{aligned}
M_t \le&\ \frac{C_0}{\sqrt{C_1}} \sum_{i=1}^{t} \left(\frac{\beta}{\sqrt{\theta'}}\right)^{t-i} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \\
=&\ \frac{C_0}{\sqrt{C_1}} \sum_{i=1}^{t} \sqrt{\gamma}^{t-i} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right].
\end{aligned} \tag{51}
$$

Hence,

$$
\begin{aligned}
\sum_{t=1}^{T} M_t \le&\ \frac{C_0}{\sqrt{C_1}} \sum_{t=1}^{T}\sum_{i=1}^{t} \sqrt{\gamma}^{t-i} N_i - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \\
=&\ \frac{C_0}{\sqrt{C_1}} \sum_{i=1}^{T}\left(\sum_{t=i}^{T} \sqrt{\gamma}^{t-i}\right) N_i - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right] \\
=&\ \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})} \sum_{t=1}^{T} N_t - \frac{1-\beta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2_{\hat{\boldsymbol{\eta}}_t}\right].
\end{aligned} \tag{52}
$$

Finally, by Lemma 32, we have

$$
\begin{aligned}
\sum_{t=1}^{T} N_i =&\ \mathbb{E}\left[L\sum_{t=1}^{T}\|\boldsymbol{\Delta}_t\|^2 + C_2 G \sum_{t=1}^{T}\chi_t \left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \\
\le&\ \left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2} + C_2 G\right)\mathbb{E}\left[\sum_{t=1}^{T}\chi_t \left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right].
\end{aligned} \tag{53}
$$

Let

$$
\begin{aligned}
C_3 =&\ \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2} + C_2 G\right) \\
=&\ \frac{C_0}{\sqrt{C_1}(1-\sqrt{\gamma})}\left(\frac{C_0^2\chi_1 L}{C_1(1-\sqrt{\gamma})^2} + 2\left(\frac{\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)\theta_1}} + 1\right)^2 G\right).
\end{aligned}
$$

Combining Eq. (52) and Eq. (53), we then obtain the desired estimate Eq. (48). The proof is completed. ■

Lemma 34. *The following estimate holds*

$$\mathbb{E}\left[\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2\right] \leq d\left[\log\left(1+\frac{G^2}{\epsilon d}\right)+\sum_{i=1}^{t}\log(\theta_i^{-1})\right]. \tag{54}$$

Proof Let $W_0 = 1$ and $W_t = \prod_{i=1}^{T}\theta_i^{-1}$. Let $w_t = W_t - W_{t-1} = (1-\theta_t)\prod_{i=1}^{t}\theta_i^{-1} = (1-\theta_t)W_t$. We therefore have

$$\frac{w_t}{W_t} = 1-\theta_t, \quad \frac{W_{t-1}}{W_t} = \theta_t.$$

Note that $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$ and $\boldsymbol{v}_t = \theta_t\boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{g}_t$, so it holds that $W_0\boldsymbol{v}_0 = \boldsymbol{\epsilon}$ and $W_t\boldsymbol{v}_t = W_{t-1}\boldsymbol{v}_{t-1} + w_t\boldsymbol{g}_t^2$. Then, $W_t\boldsymbol{v}_t = W_0\boldsymbol{v}_0 + \sum_{i=1}^{t} w_i\boldsymbol{g}_i^2 = \boldsymbol{\epsilon} + \sum_{i=1}^{t} w_i\boldsymbol{g}_i^2$. It follows that

$$\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 = \sum_{i=1}^{t}\left\|\frac{(1-\theta_i)\boldsymbol{g}_t^2}{\boldsymbol{v}_i}\right\|_1 = \sum_{i=1}^{t}\left\|\frac{w_i\boldsymbol{g}_i^2}{W_i\boldsymbol{v}_i}\right\|_1 = \sum_{i=1}^{t}\left\|\frac{w_i\boldsymbol{g}_i^2}{\boldsymbol{\epsilon}+\sum_{\ell=1}^{i} w_\ell\boldsymbol{g}_\ell^2}\right\|_1. \tag{55}$$

Writing the norm in terms of coordinates, we obtain

$$\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 = \sum_{i=1}^{t}\sum_{k=1}^{d}\frac{w_i g_{i,k}^2}{\epsilon+\sum_{\ell=1}^{i} w_\ell g_{\ell,k}^2} = \sum_{k=1}^{d}\sum_{i=1}^{t}\frac{w_i g_{i,k}^2}{\epsilon+\sum_{\ell=1}^{i} w_\ell g_{\ell,k}^2}. \tag{56}$$

By Lemma 27, for each $k = 1, 2, \ldots, d$,

$$\sum_{i=1}^{t}\frac{w_i g_{i,k}^2}{\epsilon+\sum_{\ell=1}^{i} w_\ell g_{\ell,k}^2} \leq \log\left(\epsilon+\sum_{\ell=1}^{t} w_\ell g_{\ell,k}^2\right) - \log(\epsilon) = \log\left(1+\frac{1}{\epsilon}\sum_{\ell=1}^{t} w_\ell g_{\ell,k}^2\right). \tag{57}$$

Hence,

$$\begin{aligned}\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2 \leq& \sum_{k=1}^{d}\log\left(1+\frac{1}{\epsilon}\sum_{i=1}^{t} w_i g_{i,k}^2\right)\\ \leq& \, d\log\left(\frac{1}{d}\sum_{k=1}^{d}\left(1+\frac{1}{\epsilon}\sum_{i=1}^{t} w_i g_{i,k}^2\right)\right) = d\log\left(1+\frac{1}{\epsilon d}\sum_{i=1}^{t} w_i\left\|\boldsymbol{g}_i\right\|^2\right).\end{aligned} \tag{58}$$

The second inequality is due to the convex inequality $\frac{1}{d}\sum_{k=1}^{d}\log(z_i) \leq \log\left(\frac{1}{d}\sum_{k=1}^{d} z_i\right)$. Indeed, we have the more general convex inequality that $\mathbb{E}[\log(X)] \leq \log\mathbb{E}[X]$, for any positive random variable X. Taking X to be $1 + \frac{1}{\epsilon d}\sum_{i=1}^{t} w_i \|\boldsymbol{g}_i\|^2$ in the right hand side of Eq. (58), we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2\right] \le d\,\mathbb{E}\left[\log\left(1+\frac{1}{\epsilon d}\sum_{i=1}^{t} w_i\|\boldsymbol{g}_i\|^2\right)\right] \le d\log\left(1+\frac{1}{\epsilon d}\sum_{i=1}^{t} w_i\mathbb{E}\left[\|\boldsymbol{g}_i\|^2\right]\right)\\
\le\;& d\log\left(1+\frac{G^2}{\epsilon d}\sum_{i=1}^{t} w_i\right) = d\log\left(1+\frac{G^2}{\epsilon d}(W_t-W_0)\right) = d\log\left(1+\frac{G^2}{\epsilon d}\left(\prod_{i=1}^{t}\theta_i^{-1}-1\right)\right)\\
\le\;& d\left[\log\left(1+\frac{G^2}{\epsilon d}\right)+\log\left(\prod_{i=1}^{t}\theta_i^{-1}\right)\right].
\end{aligned} \tag{59}$$

The last inequality is due to the following trivial inequality:

$$\log(1+ab) \le \log(1+a+b+ab) = \log(1+a)+\log(1+b)$$

for any non-negative parameters a and b. It then follows that

$$\mathbb{E}\left[\sum_{i=1}^{t}\left\|\frac{\sqrt{1-\theta_i}\boldsymbol{g}_i}{\sqrt{\boldsymbol{v}_i}}\right\|^2\right] \le d\left[\log\left(1+\frac{G^2}{\epsilon d}\right)+\sum_{i=1}^{t}\log(\theta_i^{-1})\right]. \tag{60}$$

The proof is completed. ■

Lemma 35. *We have the following estimate*

$$\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \le C_0 d\left[\chi_1\log\left(1+\frac{G^2}{\epsilon d}\right)+\frac{1}{\theta_1}\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}\right]. \tag{61}$$

Proof For simplicity of notations, let $\omega_t := \left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2$, and $\Omega_t := \sum_{i=1}^{t}\omega_i$. Note that $\chi_t \le C_0 a_t$. Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \le C_0\,\mathbb{E}\left[\sum_{t=1}^{T}a_t\omega_t\right]. \tag{62}$$

By Lemma 25, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}a_t\omega_t\right] = \mathbb{E}\left[\sum_{t=1}^{T-1}(a_t-a_{t+1})\Omega_t + a_T\Omega_T\right]. \tag{63}$$

Let $S_t := \log\left(1+\frac{G^2}{\epsilon d}\right)+\sum_{i=1}^{t}\log(\theta_i^{-1})$. By Lemma 34, we have $\mathbb{E}[\Omega_t] \le dS_t$. Since $\{a_t\}$ is a non-increasing sequence, we have $a_t - a_{t+1} \ge 0$. By Eq. (63), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T-1}(a_t-a_{t+1})\Omega_t + a_T\Omega_T\right] \le d\left(\sum_{t=1}^{T-1}(a_t-a_{t+1})S_t + a_T S_T\right)\\
=&d\left(a_1S_0+\sum_{t=1}^{T}a_t(S_t-S_{t-1})\right) = d\left[a_1\log\left(1+\frac{G^2}{\epsilon d}\right)+\sum_{t=1}^{T}a_t\log(\theta_t^{-1})\right].
\end{aligned} \tag{64}$$

Note that $a_t \le \chi_t$. Combining Eq. (62), Eq. (63), and Eq. (64), we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \le\ & C_0 d\left[\chi_1\log\left(1+\frac{G^2}{\epsilon d}\right)+\sum_{t=1}^{T}\chi_t\log(\theta_t^{-1})\right]\\ =\ & C_0 d\left[\chi_1\log\left(1+\frac{G^2}{\epsilon d}\right)+\sum_{t=1}^{T}\chi_t\log(\theta_t^{-1})\right].\end{aligned}\tag{65}$$

Note that $\log(1+x)\le x$ for all $x>-1$. It follows that

$$\log(\theta_t^{-1}) = \log(1+(\theta_t^{-1}-1)) \le \theta_t^{-1}-1 \le \frac{1-\theta_t}{\theta_1}.$$

Note that $\chi_t = \alpha_t/\sqrt{1-\theta_t}$. By Eq. (62) and Eq. (64), we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\chi_t\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \le C_0 d\left[\chi_1\log\left(1+\frac{G^2}{\epsilon d}\right)-\frac{1}{\theta_1}\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}\right].\tag{66}$$

The proof is completed. ∎

Lemma 36. *(Lemma 6 in Section 3) Let τ be randomly chosen from $\{1,2,\ldots,T\}$ with equal probabilities $p_\tau = 1/T$. We have the following estimate*

$$\mathbb{E}\left[\|\nabla f(x_\tau)\|\right] \le \sqrt{\frac{C_0\sqrt{G^2+\epsilon d}}{T\alpha_T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]}.\tag{67}$$

Proof For any two random variables X and Y, by the Hölder's inequality, we have

$$\mathbb{E}[|XY|] \le \mathbb{E}\left[|X|^p\right]^{1/p}\mathbb{E}\left[|Y|^q\right]^{1/q}.\tag{68}$$

Let $X = \left(\frac{\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}{\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}}\right)^{1/2}$, $Y = \|\hat{\boldsymbol{v}}_t\|_1^{1/4}$, and let $p=2$, $q=2$. By Eq. (68), we have

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|\right]^2 \le \mathbb{E}\left[\frac{\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}{\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}}\right]\mathbb{E}\left[\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}\right].\tag{69}$$

On the one hand, we have

$$\begin{aligned}\frac{\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}{\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}} = \sum_{k=1}^{d}\frac{|\nabla_k f(\boldsymbol{x}_t)|^2}{\sqrt{\sum_{k=1}^{d}\hat{v}_{t,k}}} \le\ & \alpha_t^{-1}\sum_{k=1}^{d}\frac{\alpha_t}{\sqrt{\hat{v}_{t,k}}}|\nabla_k f(\boldsymbol{x}_t)|^2\\ =\ & \alpha_t^{-1}\sum_{k=1}^{d}\hat{\eta}_{t,k}|\nabla_k f(\boldsymbol{x}_t)|^2 = \alpha_t^{-1}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2.\end{aligned}\tag{70}$$

Since $\hat{\boldsymbol{v}}_t = \theta_t \boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{\delta}_t^2$, and all entries are non-negative, we have $\|\hat{\boldsymbol{v}}_t\|_1 = \theta_t \|\boldsymbol{v}_{t-1}\|_1 + (1-\theta_t)\|\boldsymbol{\delta}_t\|^2$. Notice that $\boldsymbol{v}_t = \theta_t \boldsymbol{v}_{t-1} + (1-\theta_t)\boldsymbol{g}_t^2$, $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$, and $\mathbb{E}_t\left[\boldsymbol{g}_t^2\right] \le G^2$. It is straightforward to prove by induction that $\mathbb{E}[\|\boldsymbol{v}_t\|_1] \le G^2 + \epsilon d$. Hence,

$$\mathbb{E}\left[\sqrt{\|\hat{\boldsymbol{v}}_t\|_1}\right] \le \sqrt{\mathbb{E}[\|\hat{\boldsymbol{v}}_t\|_1]} \le \sqrt{G^2 + \epsilon d}. \tag{71}$$

By Eq. (69), Eq. (70), and Eq. (71), we obtain

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|\right]^2 \le \left(\alpha_t^{-1}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]\right)\sqrt{G^2+\epsilon d}. \tag{72}$$

By Lemma 26, $\alpha_T \le C_0\alpha_t$ for any $t \le T$, so $\alpha_t^{-1} \le C_0\alpha_T^{-1}$. Then, we obtain

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|\right]^2 \le \frac{C_0\sqrt{G^2+\epsilon d}}{\alpha_T}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right], \ \forall t \le T. \tag{73}$$

The lemma is followed by

$$\begin{aligned}
(\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|\right]) = &\ \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|\right] \\
\le &\ \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{C_0\sqrt{G^2+\epsilon d}}{\alpha_T}\,\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]} \\
\le &\ \sqrt{\frac{C_0\sqrt{G^2+\epsilon d}}{T\alpha_T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|_{\hat{\boldsymbol{\eta}}_t}^2\right]}.
\end{aligned} \tag{74}$$

The proof is completed. ∎

Appendix B. Proof of Theorem 2

Theorem. *Let $\{\boldsymbol{x}_t\}$ be a sequence generated by Generic Adam for initial values $\boldsymbol{x}_1$, $\boldsymbol{m}_0 = \boldsymbol{0}$, and $\boldsymbol{v}_0 = \boldsymbol{\epsilon}$. Assume that f and stochastic gradients $\boldsymbol{g}_t$ satisfy assumptions (A1)-(A4). Let τ be randomly chosen from $\{1, 2, \ldots, T\}$ with equal probabilities $p_\tau = 1/T$. We have the following estimate*

$$\mathbb{E}\left[\|\boldsymbol{\nabla} f(\boldsymbol{x}_\tau)\|\right] \le \sqrt{\frac{C + C'\sum_{t=1}^{T}\alpha_t\sqrt{1-\theta_t}}{T\alpha_T}}, \tag{75}$$

where the constants C and C' are given by

$$\begin{aligned}
C &= \frac{2C_0\sqrt{G^2+\epsilon d}}{1-\beta}\left(f(x_1) - f^* + C_3C_0 d\,\chi_1\log\left(1+\frac{G^2}{\epsilon d}\right)\right), \\
C' &= \frac{2C_0^2C_3 d\sqrt{G^2+\epsilon d}}{(1-\beta)\theta_1}.
\end{aligned}$$

Proof By the L-Lipschitz continuity of the gradient of f and the descent lemma, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + \frac{L}{2} \left\| \boldsymbol{\Delta}_t \right\|^2. \tag{76}$$

Let $M_t := \mathbb{E}\left[\langle \boldsymbol{\nabla} f(\boldsymbol{x}_t), \boldsymbol{\Delta}_t \rangle + L \left\| \boldsymbol{\Delta}_t \right\|^2 \right]$. We have $\mathbb{E}[f(\boldsymbol{x}_{t+1})] \leq \mathbb{E}[f(\boldsymbol{x}_t)] + M_t$. Taking a sum for $t = 1, 2, \ldots, T$, we obtain

$$\mathbb{E}\left[f(\boldsymbol{x}_{T+1}) \right] \leq f(\boldsymbol{x}_1) + \sum_{t=1}^{T} M_t. \tag{77}$$

Note that $f(x)$ is bounded from below by f^*, so $\mathbb{E}[f(\boldsymbol{x}_{T+1})] \geq f^*$. Applying the estimate of Lemma 33, we have

$$f^* \leq f(\boldsymbol{x}_1) + C_3 \mathbb{E}\left[\sum_{t=1}^{T} \chi_t \left\| \frac{\sqrt{1-\theta_t} \boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \right\|^2 \right] - \frac{1-\beta}{2} \mathbb{E}\left[\sum_{t=1}^{T} \left\| \boldsymbol{\nabla} f(\boldsymbol{x}_t) \right\|^2_{\hat{\boldsymbol{\eta}}_t} \right], \tag{78}$$

where C_3 is the constant given in Lemma 33. By applying the estimates in Lemma 34 and Lemma 36 for the second and third terms in the right hand side of Eq. (78), and appropriately rearranging the terms, we obtain

$$\begin{aligned}
\left(\mathbb{E}\left[\left\| \boldsymbol{\nabla} f(\boldsymbol{x}_\tau^T) \right\| \right] \right) &\leq \sqrt{ \frac{C_0 \sqrt{G^2 + \epsilon d}}{T \alpha_T} \mathbb{E}\left[\sum_{t=1}^{T} \left\| \boldsymbol{\nabla} f(\boldsymbol{x}_t) \right\|^2_{\hat{\boldsymbol{\eta}}_t} \right] } \\
&\leq \sqrt{ \frac{2 C_0 \sqrt{G^2 + \epsilon d}}{(1-\beta) T \alpha_T} \left(f(\boldsymbol{x}_1) - f^* + C_3 \mathbb{E}\left[\sum_{t=1}^{T} \chi_t \left\| \frac{\sqrt{1-\theta_t} \boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}} \right\| \right] \right) } \\
&\leq \sqrt{ \frac{2 C_0 \sqrt{G^2 + \epsilon d}}{(1-\beta) T \alpha_T} \left[f(\boldsymbol{x}_1) - f^* + C_3 C_0 d \, \chi_1 \log\left(1 + \frac{G^2}{\epsilon d} \right) - \frac{C_3 C_0 d}{\theta_1} \sum_{t=1}^{T} \alpha_t \sqrt{1-\theta_t} \right] } \\
&= \sqrt{ \frac{C + C' \sum_{t=1}^{T} \alpha_t \sqrt{1-\theta_t}}{T \alpha_T} },
\end{aligned} \tag{79}$$

where

$$\begin{aligned}
C &= \frac{2 C_0 \sqrt{G^2 + \epsilon d}}{1-\beta} \left(f(x_1) - f^* + C_3 C_0 d \, \chi_1 \log\left(1 + \frac{G^2}{\epsilon d} \right) \right), \\
C' &= \frac{2 C_0^2 C_3 d \sqrt{G^2 + \epsilon d}}{(1-\beta) \theta_1}.
\end{aligned}$$

The proof is completed. ■

Appendix C. Proof of Corollary 10

Corollary. *Generic Adam with the above family of parameters converges as long as* $0 < r \le 2s < 2$*, and its non-asymptotic convergence rate is given by*

$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|] \le \begin{cases} \mathcal{O}(T^{-r/4}), & r/2+s<1 \\ \mathcal{O}(\sqrt{\log(T)/T^{1-s}}), & r/2+s=1 \\ \mathcal{O}(1/T^{(1-s)/2}), & r/2+s>1 \end{cases}.$$

Proof It is not hard to verify that the following equalities hold:

$$\begin{aligned}\sum_{t=K}^{T} \alpha_t \sqrt{1-\theta_t} &= \eta\sqrt{\alpha}\sum_{t=K}^{T} t^{-(r/2+s)} \\ &= \begin{cases} \mathcal{O}(T^{1-(r/2+s)}), & r/2+s<1 \\ \mathcal{O}(\log(T)), & r/2+s=1 \\ \mathcal{O}(1), & r/2+s>1 \end{cases}.\end{aligned}$$

In this case, $T\alpha_T = \eta T^{1-s}$. Therefore, by Theorem 2 the non-asymptotic convergence rate is given by

$$\|\nabla f(\boldsymbol{x}_\tau)\|^2 \le \begin{cases} \mathcal{O}(T^{-r/4}), & r/2+s<1 \\ \mathcal{O}(\sqrt{\log(T)/T^{1-s}}), & r/2+s=1 \\ \mathcal{O}(1/T^{(1-s)/2}), & r/2+s>1 \end{cases}.$$

To guarantee convergence, then $0 < r \le 2s < 2$. ■

Appendix D. Proof of Theorem 13

Theorem. *For any* $T > 0$*, if we take* $\alpha_t = \frac{\alpha}{\sqrt{T}},\ \beta_t = \beta, \theta_t = 1 - \frac{\theta}{T}$*, which satisfies* $\gamma = \frac{\beta}{1-\frac{\theta}{T}} < 1$ *and* $\theta_t \ge \frac{1}{4}$*, then it holds that*

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_\tau)\|\right] \le \sqrt{\frac{C_5}{\sqrt{T}}} = \mathcal{O}(T^{-1/4}),$$

where

$$\begin{aligned} C_5 &= \frac{2\sqrt{G^2+\epsilon d}}{\alpha(1-\beta)}\left[f(x_1) - f^* + C_6 d\frac{\alpha}{\sqrt{\theta}}\log\big(1+\frac{G^2}{\epsilon d}\big) + \frac{4C_6 d\alpha}{\sqrt{\theta}}\right], \\ C_6 &= \frac{1}{1-\sqrt{\gamma}}\left[\frac{\alpha L}{\sqrt{\theta}(1-\sqrt{\gamma})^2} + 2\big(\frac{2\beta/(1-\beta)}{\sqrt{C_1(1-\gamma)}}+1\big)^2 G\right]. \end{aligned}$$

Proof Based on Theorem 2, by plugging α_t, β_t and θ_t in the conclusion of Theorem 2, we can get the desired result. ■

Appendix E. Key Lemma to prove Theorem 17

In this section, we provide the additional lemmas for the proofs of Theorem 17.

Lemma 37. *With the definitions in Algorithm 2, for any $t = 1, 2, \cdots, T$ we have the following estimation:*

$$\mathbb{E}[\|\bar{\boldsymbol{g}}_t - \nabla f(\boldsymbol{x}_t)\|^2] \leq \frac{\sigma^2}{s}.$$

Proof

$$\begin{aligned}\mathbb{E}[\|\bar{\boldsymbol{g}}_t - \nabla f(\boldsymbol{x}_t)\|^2] &= \mathbb{E}\left[\left\|\frac{1}{s}\sum_{i=1}^{s}\boldsymbol{g}_t^{(i)} - \nabla f(\boldsymbol{x}_t)\right\|^2\right] \\ &\leq \mathbb{E}\left[\frac{1}{s^2}\sum_{i=1}^{s}\|\boldsymbol{g}_t^{(i)} - \nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{\sigma^2}{s}.\end{aligned}$$

The second inequality holds, because $\boldsymbol{g}_t^{(i)}$ are independent and have the same expectation ($\mathbb{E}\left[\boldsymbol{g}_t^{(i)}\right] = \mathbb{E}[\nabla f(\boldsymbol{x}_t)]$). ∎

Lemma 38. *The following estimate holds:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\sqrt{1-\theta_t}\bar{\boldsymbol{g}}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] \leq d\theta + 2d\sqrt[4]{1+\frac{2\sigma^2}{d\epsilon s}} + \sqrt{\sqrt{\frac{2\theta}{d\epsilon T}}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right]}.$$

Proof With the similar proof in Lemma 34, it holds that

$$\begin{aligned}\sum_{t=1}^{T}\left\|\frac{\sqrt{1-\theta_t}\bar{\boldsymbol{g}}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 &\leq d\log\left(1+\frac{1}{d\epsilon}\sum_{k=1}^{T}w_k\|\bar{\boldsymbol{g}}_k\|^2\right) \\ &\leq d\log\left(1+\frac{2}{d\epsilon}\sum_{k=1}^{T}w_k\left(\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|^2 + \|\nabla f(\boldsymbol{x}_k)\|^2\right)\right) \\ &\leq 2d\log\left(\sqrt{1+\frac{2}{d\epsilon}\sum_{k=1}^{T}w_k\|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|^2} + \sqrt{\frac{2}{d\epsilon}\sum_{k=1}^{T}w_k\|\nabla f(\boldsymbol{x}_k)\|^2}\right).\end{aligned} \tag{80}$$

Thus, by taking expectation on both side, we can obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\sqrt{1-\theta_t}\bar{\boldsymbol{g}}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] &\leq \mathbb{E}\left[2d\log\left(\sqrt{1+\frac{2}{d\epsilon}\sum_{k=1}^{T}w_k\|\bar{\boldsymbol{g}}_k-\nabla f(\boldsymbol{x}_k)\|^2}+\sqrt{\frac{2}{d\epsilon}\sum_{k=1}^{T}w_k\|\nabla f(\boldsymbol{x}_k)\|^2}\right)\right]\\ &\leq 2d\log\left(\sqrt{1+\frac{2}{d\epsilon}\sum_{k=1}^{T}w_k\mathbb{E}\|\bar{\boldsymbol{g}}_k-\nabla f(\boldsymbol{x}_k)\|^2}+\sqrt{\frac{2}{d\epsilon}}\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}w_k\|\nabla f(\boldsymbol{x}_k)\|^2}\right]\right)\\ &\leq 2d\log\left(\sqrt{1+\frac{2\sigma^2W_T}{d\epsilon s}}+\sqrt{\frac{2w_T}{d\epsilon}}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right]\right),\end{aligned} \tag{81}$$

where the last inequality uses Lemma 37.

Meanwhile, we have

$$\begin{aligned}&\log W_T = T\log\left(\frac{T}{T-\theta}\right)\leq T\left(\frac{T}{T-\theta}-1\right)\leq\theta,\\ &W_T\geq 1,\\ &\frac{w_T}{W_T}=1-\theta_T=\frac{\theta}{T}.\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\sqrt{1-\theta_t}\bar{\boldsymbol{g}}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right] &\leq d\log W_T+2d\log\left(\sqrt{1+\frac{2\sigma^2}{d\epsilon s}}+\sqrt{\frac{2\theta}{d\epsilon T}}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right]\right)\\ &\leq d\theta+2d\sqrt[4]{1+\frac{2\sigma^2}{d\epsilon s}}+\sqrt{\sqrt{\frac{2\theta}{d\epsilon T}}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right]}.\end{aligned} \tag{82}$$

Hence, we obtain the desired result. ■

Lemma 39. *(Lemma 19 in Section 4.2) By the definition of M_t, it holds that*

$$\sum_{t=1}^{T}M_t\leq C_7\mathbb{E}\sum_{t=1}^{T}\left\|\frac{\sqrt{1-\theta_t}\bar{\boldsymbol{g}}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2-\frac{1-\beta}{2}\sum_{t=1}^{T}\mathbb{E}\|\nabla f\left(\boldsymbol{x}_t\right)\|_{\hat{\eta}_t}^2,$$

where

$$C_7=\frac{1}{1-\sqrt{\gamma}}\left(\frac{\alpha^2L}{\theta\left(1-\sqrt{\gamma}\right)^2}+\frac{2\left(\frac{2\beta/(1-\beta)}{\sqrt{(1-\gamma)}}+1\right)^2G\alpha}{\sqrt{\theta}}\right).$$

Proof Using Lemma 33, by plugging $C_0 = C_1 = 1$, $\chi_t = \frac{\alpha}{\sqrt{\theta}}$ and $\theta_t \geq \frac{1}{4}$, it holds that

$$\sum_{t=1}^{T} M_t \leq C_7 \mathbb{E} \sum_{t=1}^{T} \left\| \frac{\sqrt{1-\theta_t} \bar{\boldsymbol{g}}_t}{\sqrt{\boldsymbol{v}_t}} \right\|^2 - \frac{1-\beta}{2} \sum_{t=1}^{T} \mathbb{E} \|\nabla f(\boldsymbol{x}_t)\|_{\hat{\eta}_t}^2,$$

where

$$C_7 = \frac{1}{1-\sqrt{\gamma}} \left(\frac{\alpha^2 L}{\theta \left(1-\sqrt{\gamma}\right)^2} + \frac{2\left(\frac{2\beta/(1-\beta)}{\sqrt{(1-\gamma)}} + 1\right)^2 G\alpha}{\sqrt{\theta}} \right).$$

■

Lemma 40. *(Lemma 20 in Section 4) The following estimation always holds:*

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla f(\boldsymbol{x}_t)\|^2}\right]^2 \leq \left(\frac{\sqrt{T}\sqrt{2\sigma^2\theta + \epsilon s d}}{\sqrt{s}\alpha} + \sqrt{\frac{2\theta}{\alpha^2}} \mathbb{E} \sqrt{\sum_{t=1}^{T} \|\nabla f(\boldsymbol{x}_t)\|^2} \right) \mathbb{E}\left[\sum_{t=1}^{T} \|\nabla f(\boldsymbol{x}_t)\|_{\hat{\eta}_t}^2\right].$$

Proof First, we have

$$\begin{aligned}
\sqrt{\|\hat{v}_t\|_1} &= \sqrt{\sum_{k=1}^{t-1} (\theta')^{t-k} (1-\theta_k) \|\bar{\boldsymbol{g}}_k\|^2 + (1-\theta_t) \mathbb{E}_t(\|\bar{\boldsymbol{g}}_t\|^2) + (\theta')^t \epsilon d} \\
&\leq \sqrt{\sum_{k=1}^{t-1} (1-\theta_k) \|\bar{\boldsymbol{g}}_k\|^2 + (1-\theta_t) \mathbb{E}_t(\|\bar{\boldsymbol{g}}_t\|^2) + \epsilon d} \\
&\leq \sqrt{2\sum_{k=1}^{t-1} (1-\theta_k) \|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|^2 + (1-\theta_t) \sigma^2/s + 2\sum_{k=1}^{t} (1-\theta_k) \|\nabla f(\boldsymbol{x}_k)\|^2 + \epsilon d} \\
&\leq \sqrt{2\sum_{k=1}^{T-1} (1-\theta_k) \|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|^2 + (1-\theta_t) \sigma^2/s + \epsilon d} + \sqrt{2\sum_{k=1}^{T} (1-\theta_k) \|\nabla f(\boldsymbol{x}_k)\|^2}.
\end{aligned}$$

Therefore, it holds

$$\max \sqrt{\|\hat{v}_t\|_1} \leq \sqrt{2\sum_{k=1}^{T-1} (1-\theta_k) \|\bar{\boldsymbol{g}}_k - \nabla f(\boldsymbol{x}_k)\|^2 + (1-\theta_T) \sigma^2/s + \epsilon d} + \sqrt{2\sum_{k=1}^{T} (1-\theta_k) \|\nabla f(\boldsymbol{x}_k)\|^2}.$$

Then we can obtain

$$\begin{aligned}
&\mathbb{E}\left[\max\sqrt{\|\hat{v}_t\|_1}\right] \\
&\leq \mathbb{E}\sqrt{2\sum_{k=1}^{T-1}(1-\theta_k)\|\bar{\boldsymbol{g}}_k-\nabla f(\boldsymbol{x}_k)\|^2+(1-\theta_T)\,\sigma^2/s+\epsilon d}+\mathbb{E}\sqrt{2\sum_{k=1}^{T}(1-\theta_k)\|\nabla f(\boldsymbol{x}_k)\|^2} \\
&\leq \sqrt{\mathbb{E}2\sum_{k=1}^{T-1}(1-\theta_k)\|\bar{\boldsymbol{g}}_k-\nabla f(\boldsymbol{x}_k)\|^2+(1-\theta_T)\,\sigma^2/s+\epsilon d}+\mathbb{E}\sqrt{2\sum_{k=1}^{T}(1-\theta_k)\|\nabla f(\boldsymbol{x}_k)\|^2} \\
&\leq \frac{\sqrt{2\sigma^2\theta+\epsilon sd}}{\sqrt{s}}+\sqrt{\frac{2\theta}{T}}\mathbb{E}\sqrt{\sum_{k=1}^{T}\|\nabla f(\boldsymbol{x}_k)\|^2}.
\end{aligned} \tag{83}$$

Meanwhile, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\sum_{t=1}^T\|\nabla f(\boldsymbol{x}_t)\|^2}{\max\sqrt{\|\hat{v}_t\|_1}}\right] &\leq \mathbb{E}\left[\sum_{t=1}^T\frac{\|\nabla f(\boldsymbol{x}_t)\|^2}{\sqrt{\|\hat{v}_t\|_1}}\right]\leq \mathbb{E}\left[\sum_{t=1}^T\sum_{k=1}^d\frac{\nabla_k f(\boldsymbol{x}_t)^2}{\sqrt{\hat{v}_{t,k}}}\right] \\
&= \frac{\sqrt{T}}{\alpha}\mathbb{E}\left[\sum_{t=1}^T\sum_{k=1}^d\frac{\nabla_k f(\boldsymbol{x}_t)^2\alpha}{\sqrt{T}\sqrt{\hat{v}_{t,k}}}\right]=\frac{\sqrt{T}}{\alpha}\mathbb{E}\left[\sum_{t=1}^T\|\nabla f(\boldsymbol{x}_t)\|^2_{\hat{\eta}_t}\right].
\end{aligned} \tag{84}$$

With inequalities (83) and (84), we can obtain

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\sum_{t=1}^T\|\nabla f(\boldsymbol{x}_t)\|^2}\right]^2 &\leq \mathbb{E}\left[\max\sqrt{\|\hat{v}_t\|_1}\right]\mathbb{E}\left[\frac{\sum_{t=1}^T\|\nabla f(\boldsymbol{x}_t)\|^2}{\max\sqrt{\|\hat{v}_t\|_1}}\right] \\
&\leq \left(\frac{\sqrt{T}\sqrt{2\sigma^2\theta+\epsilon sd}}{\sqrt{s}\alpha}+\sqrt{\frac{2\theta}{\alpha^2}}\mathbb{E}\sqrt{\sum_{t=1}^T\|\nabla f(\boldsymbol{x}_t)\|^2}\right)\mathbb{E}\left[\sum_{t=1}^T\|\nabla f(\boldsymbol{x}_t)\|^2_{\hat{\eta}_t}\right].
\end{aligned}$$

■

Lemma 41. *(Lemma 21 in Section 4) For any $x \in \mathbb{R}$, if $x^2 \leq (A+Bx)(C+D\sqrt{x})$, then $x \leq (4BD)^2+4BC+(4AD)^{2/3}+\sqrt{4AC}$.*

Proof We discuss the solution of x in 4 different situations. First, when $Bx \geq A$ and $D\sqrt{x} \geq C$, we have

$$x^2 \leq (A+Bx)(C+D\sqrt{x}) \leq 4BDx^{3/2}.$$

Therefore, $x \leq (4BD)^2$.

Secondly, when $Bx \geq A$ and $D\sqrt{x} \leq C$, we have

$$x^2 \leq (A+Bx)(C+D\sqrt{x}) \leq 4BCx.$$

Hence, $x \leq 4BC$.

Thirdly, when $Bx \le A$ and $D\sqrt{x} \ge C$, it holds that

$$x^2 \le (A + Bx)(C + D\sqrt{x}) \le 4AD\sqrt{x}.$$

And we can obtain $x \le (4AD)^{2/3}$.

Last, when $Bx \le A$ and $D\sqrt{x} \le C$, it holds that

$$x^2 \le (A + Bx)(C + D\sqrt{x}) \le 4AC.$$

Then we have $x \le \sqrt{4AC}$.

Therefore, combining four different conditions, we have $x \le (4BD)^2 + 4BC + (4AD)^{2/3} + \sqrt{4AC}$. ■

Appendix F. Proof of Theorem 17

Theorem. *For any $T > 0$, if we take $\alpha_t = \frac{\alpha}{\sqrt{T}}$, $\beta_t = \beta$, $\theta_t = 1 - \frac{\theta}{T}$, which satisfies $\gamma = \frac{\beta_t}{\theta_t} < 1$ and $\theta_t \ge \frac{1}{4}$, then there exists $t \in \{1, 2, \cdots, T\}$ such that*

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] \le \frac{1}{\sqrt{T}}\left((4C_8C_{10})^{1/2} + (4C_8C_{11})^{2/3} + (4C_9C_{10}) + (4C_9C_{11})^2)\right),$$

where

$$C_8 = \frac{\sqrt{T}\sqrt{2\sigma^2\theta + \epsilon sd}}{\sqrt{s}\alpha},\ C_9 = \sqrt{\frac{2\theta}{\alpha^2}},\ C_{10} = \frac{2C_4}{1-\beta}\sqrt[4]{\frac{2\theta}{d\epsilon T}},$$

$$C_{11} = \frac{2}{1-\beta}\left(f(x_1) - f^* + C_4 d\theta + 2C_4 d\sqrt[4]{1 + \frac{2\sigma^2}{d\epsilon s}}\right),$$

$$C_7 = \frac{1}{1-\sqrt{\gamma}}\left(\frac{\alpha^2 L}{\theta\left(1-\sqrt{\gamma}\right)^2} + \frac{2\left(\frac{2\beta/(1-\beta)}{\sqrt{1-\gamma}} + 1\right)^2 G\alpha}{\sqrt{\theta}}\right).$$

In addition, by taking $\epsilon = \frac{1}{sd}$, it holds that

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}(T^{-1}s^{1/2}d^{1/2} + T^{-1/2}d + T^{-1/3}s^{-1/4} + T^{-1/4}s^{-1/4}d^{1/2}).$$

Proof First, according to the gradient Lipschitz condition of f, it holds

$$\begin{aligned} f(x_{t+1}) &\le f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \Delta_t\rangle + \frac{L}{2}\|\Delta_t\|^2 \\ &= f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \Delta_t\rangle + L\|\Delta_t\|^2. \end{aligned}$$

Recall that $M_t = \mathbb{E}[\langle \nabla f(\boldsymbol{x}_t), \Delta_t\rangle + L\|\Delta_t\|^2$. Then we have

$$\begin{aligned} f^* &\le \mathbb{E}[f x_{T+1})] \le f(x_1) + \sum_{t=1}^{T} M_t \\ &\le f(x_1) + \sum_{t=1}^{T} M_t \le f(x_1) + C_7\mathbb{E}\sum_{t=1}^{T}\left\|\frac{\sqrt{1-\theta_t}\bar{\boldsymbol{g}}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2 - \frac{1-\beta}{2}\sum_{t=1}^{T}\mathbb{E}\|\nabla f\left(\boldsymbol{x}_t\right)\|_{\hat{\eta}_t}^2. \end{aligned}$$

Using Lemma 38, 39 and 40 with rearranging the corresponding terms, we have

$$\begin{aligned}
&\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right]^2 \\
\leq &\left(\frac{\sqrt{T}\sqrt{2\sigma^2\theta+\epsilon s d}}{\sqrt{s}\alpha}+\sqrt{\frac{2\theta}{\alpha^2}}\mathbb{E}\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right)\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2_{\hat{\eta}_t}\right] \\
\leq &\left(\frac{\sqrt{T}\sqrt{2\sigma^2\theta+\epsilon s d}}{\sqrt{s}\alpha}+\sqrt{\frac{2\theta}{\alpha^2}}\mathbb{E}\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right)\frac{2}{1-\beta}\left(f(x_1)-f^*+C_7\mathbb{E}\sum_{t=1}^{T}\left\|\frac{\sqrt{1-\theta_t}\boldsymbol{g}_t}{\sqrt{\boldsymbol{v}_t}}\right\|^2\right) \\
\leq &\left(\frac{\sqrt{T}\sqrt{2\sigma^2\theta+\epsilon s d}}{\sqrt{s}\alpha}+\sqrt{\frac{2\theta}{\alpha^2}}\mathbb{E}\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right) \\
&\quad\frac{2}{1-\beta}\left(f(x_1)-f^*+C_7 d\theta+C_7 2d\sqrt[4]{1+\frac{2\sigma^2}{d\epsilon s}}+C_7\sqrt{\sqrt{\frac{2\theta}{d\epsilon T}}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right]}\right).
\end{aligned} \tag{85}$$

Before using Lemma 41, we list the order of 4 terms in Lemma 41 as follows:

$$\begin{aligned}
&(4BD)^2 = O\left(\sqrt{\frac{1}{d\epsilon T}}\right),\ 4BC = O(d+d(ds\epsilon)^{-1/4}), \\
&(4AD)^{2/3} = O(T^{1/6}(s^{-1/2}(\epsilon d)^{-1/4}+(\epsilon d)^{1/4})), \\
&\sqrt{4AC} = O(T^{1/4}d^{1/2}(s^{-1/4}+(\epsilon d)^{1/4})(1+(ds\epsilon)^{-1/8})).
\end{aligned}$$

Then, it holds that $\sqrt{T}\min_t\mathbb{E}\|\nabla f(\boldsymbol{x}_t)\| \leq \mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|^2}\right]$. By dividing $\sqrt{T}$ on both side, we can get the desired result. ■

References

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pages 404–413, 2018.

Anas Barakat and Pascal Bianchi. Convergence rates of a momentum algorithm with bounded adaptive step size for nonconvex optimization. In *Asian Conference on Machine Learning*, pages 225–240. PMLR, 2020.

Amitabh Basu, Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and ADAM in non-convex optimization and their comparison to nesterov acceleration on autoencoders. *arXiv preprint arXiv:1807.06766*, 2018.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569, 2018.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Guido Carnevale, Francesco Farina, Ivano Notarnicola, and Giuseppe Notarstefano. Distributed online optimization via gradient tracking with adaptive momentum. *arXiv preprint arXiv:2009.01745*, 2020.

Congliang Chen, Li Shen, Haozhi Huang, and Wei Liu. Quantized adam with error feedback. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–26, 2021.

Congliang Chen, Li Shen, Wei Liu, and Zhi-Quan Luo. Efficient-adam: Communication-efficient distributed adam with complexity analysis. *arXiv preprint arXiv:2205.14473*, 2022.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018a.

Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pages 119–128, 2020.

Zaiyi Chen, Tianbao Yang, Jinfeng Yi, Bowen Zhou, and Enhong Chen. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *arXiv preprint arXiv:1808.06296*, 2018b.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Timothy Dozat. Incorporating Nesterov momentum into Adam. *International Conference on Learning Representations Workshop*, 2016.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. page 14, 2012.

Haiwen Huang, Chang Wang, and Bin Dong. Nostalgic Adam: Weighing more of the past gradients when designing the adaptive learning rate. *arXiv preprint arXiv:1805.07557*, 2018.

Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Corinna Cortes, and Christopher JC Burges. Mnist handwritten digit database. 2010. *URL http://yann. lecun. com/exdb/mnist*, 2010.

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670, 2014.

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

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Mahesh Chandra Mukkamala and Matthias Hein. Variants of RMSProp and Adagrad with logarithmic regret bounds. In *International Conference on Machine Learning*, pages 2545–2553, 2017.

Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *arXiv preprint arXiv:1901.09109*, 2019.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Herbert Robbins and Sutton Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Yaowei Wang, Jinqiao Wang, and Ming Tang. Large batch optimization for object detection: Training coco in 12 minutes. In *European Conference on Computer Vision*, pages 481–496. Springer, 2020.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv preprint arXiv:1806.01811*, 2018.

Cong Xie, Oluwasanmi Koyejo, Indranil Gupta, and Haibin Lin. Local adaalter: Communication-efficient stochastic gradient descent with adaptive learning rates. *arXiv preprint arXiv:1911.09030*, 2019.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. *arXiv preprint arXiv:1808.10396*, 2018.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. *arXiv preprint arXiv:1905.03817*, 2019.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2018.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *arXiv preprint arXiv:1912.03194*, 2019.

Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018a.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. AdaShift: Decorrelation and convergence of adaptive learning rate methods. *arXiv preprint arXiv:1810.00143*, 2018b.

Fangyu Zou, Li Shen, Zequn Jie, Ju Sun, and Wei Liu. Weighted adagrad with unified momentum. *arXiv preprint arXiv:1808.03408*, 2018.

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*, pages 11127–11135, 2019.