A Novel Convergence Analysis for Algorithms of the Adam Family

Zhishuai Guo[†] ZHISHUAI-GUO@UIOWA.EDU
Yi Xu[‡] YIXU@ALIBABA-INC.COM
Wotao Yin[‡] WOTAO.YIN@ALIBABA-INC.COM
Rong Jin[‡] JINRONG.JR@ALIBABA-INC.COM
Tianbao Yang[†] TIANBAO-YANG@UIOWA.EDU
[†]*Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA*
[‡]*Machine Intelligence Technology, Alibaba Group, Bellevue, WA 98004, USA*

Abstract

Since its invention in 2014, the Adam optimizer [12] has received tremendous attention. On one hand, it has been widely used in deep learning and many variants have been proposed, while on the other hand their theoretical convergence property remains to be a mystery. It is far from satisfactory in the sense that some studies require strong assumptions about the updates, which are not necessarily applicable in practice, while other studies still follow the original problematic convergence analysis of Adam, which was shown to be not sufficient to ensure convergence. Although rigorous convergence analysis exists for Adam, they impose specific requirements on the update of the adaptive step size, which are not generic enough to cover many other variants of Adam. To address theses issues, in this extended abstract, we present a simple and generic proof of convergence for a family of Adam-style methods (including Adam, AMSGrad, Adabound, etc.). Our analysis only requires an increasing or large "momentum" parameter for the first-order moment, which is indeed the case used in practice, and a boundness condition on the adaptive factor of the step size, which applies to all variants of Adam under mild conditions of stochastic gradients. We also establish a variance diminishing result for the used stochastic gradient estimators. Indeed, our analysis of Adam is so simple and generic that it can be leveraged to establish the convergence for solving a broader family of non-convex optimization problems, including min-max, compositional, and bilevel optimization problems. For the full (earlier) version of this extended abstract, please refer to [11].

1. Introduction

Stochastic adaptive methods originating from AdaGrad for convex minimization [7, 18] have attracted tremendous attention for stochastic non-convex optimization [2, 13, 17, 23, 28, 32, 34]. Adam [12] is an important variant of AdaGrad, which is widely used in practice for training deep neural networks, and many variants of Adam were proposed for improving its performance, e.g. [14, 17, 31]. Its analysis for non-convex optimization has also received a lot of attention [3]. For more generality, we consider a family of Adam-style algorithms. The update is given by

$$\text{Adam-style:} \quad \begin{cases} \mathbf{v}_{t+1} = (1-\beta_t)\mathbf{v}_t + \beta_t \mathcal{O}_F(\mathbf{x}_t), \\ \mathbf{u}_{t+1} = h_t(\mathcal{O}_F(\mathbf{x}_0), \ldots, \mathcal{O}_F(\mathbf{x}_t)), \\ \mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \dfrac{\mathbf{v}_{t+1}}{\sqrt{\mathbf{u}_{t+1}} + G_0}, t = 0, \ldots, T. \end{cases} \tag{1}$$

where h_t denotes an appropriate mapping function whose specific choices are given later.

One criticism of Adam is that it might not converge for some problems with inappropriate momentum parameters. In particular, the authors of AMSGrad [21] show that Adam with small momentum parameters can diverge for some problems. However, we notice that the failure of Adam shown in AMSGrad [21] and the practical success of Adam come from an inconsistent setting of the momentum parameter for the first-order moment. In practice, this momentum parameter (corresponding to $1 - \beta_t$ in (1)) is usually set to a large value (e.g., 0.9). However, in the failure case analysis of Adam [21] and many existing analysis of Adam and their variants [3, 14, 17, 22, 31], such momentum parameter is set as a small value or a decreasing sequence. We provide the first analysis of Adam and other variants with a more natural increasing or large momentum parameter for the first-order moment.

Several recent works have tried to prove the (non)-convergence of Adam. In particular, Zou et al. [35] establish some sufficient condition for ensuring Adam to converge. In particular, they choose to increase the momentum parameter for the second-order moment and establish a convergence rate in the order of $\log(T)/\sqrt{T}$, which was similarly established in [6] with some improvement on the constant factor. Zaheer et al. [31] show that Adam with a sufficiently large mini-batch size can converge to an accuracy level proportional to the inverse of the mini-batch size. Chen et al. [3] analyze the convergence properties for a family of Adam-style algorithms. However, their analysis requires a strong assumption of the updates to ensure the convergence, which does not necessarily hold as the authors give non-convergence examples. Different from these works, we give an alternative way to ensure Adam converge by using an increasing or large momentum parameter for the first-order moment without any restrictions on the momentum parameter for the second-order moment and without requiring a large mini-batch size. This seems more natural and consistent with the practice. Indeed, our analysis is applicable to a family of Adam-style algorithms, and is agnostic to the method for updating the normalization factor in the adaptive step size as long as it can be upper bounded. The large momentum parameter for the first-order moment is also the key part that differentiates our convergence analysis with existing non-convergence analysis of Adam [3, 12, 21], which require the momentum parameter for the first-order moment to be decreasing to zero or sufficiently small.

A key in the analysis is to carefully utilize the design of the stochastic estimator of the gradient. Traditional methods that simply use an unbiased gradient estimator of the objective function are not applicable to many problems and also suffer slow convergence due to large variance of the unbiased stochastic gradients. Recent studies in stochastic non-convex optimization have proposed better stochastic estimators of the gradient based on variance reduction technique (e.g., SPIDER, SARAH, STORM) [5, 9, 20, 26]. However, these estimators sacrifice generality as they require that the unbiased stochastic oracle is Lipschitz continuous with respect to the input, which prohibits many useful tricks in machine learning for improving **generalization** and **efficiency** (e.g., adding random noise to the stochastic gradient [19], gradient compression [1, 27, 33]). In addition, they also require computing stochastic gradients at two points per-iteration, making them further restrictive.

In Adam-style methods, the stochastic estimators are based on the moving average. In order to generate a sequence of iterates $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T\}$, we usually need to track another sequence of $\{g(\mathbf{x}_0), g(\mathbf{x}_1), \ldots, g(\mathbf{x}_T)\}$, where g is a Lipschitz continuous mapping that is useful for constructing the gradient of the objective function. However, $g(\mathbf{x}_t)$ can be only accessed through **an unbiased stochastic oracle** denoted by $\mathcal{O}_g$ such that for any input $\mathbf{x}$, it returns a random variable $\mathcal{O}_g(\mathbf{x})$ satisfying $\mathrm{E}[\mathcal{O}_g(\mathbf{x})] = g(\mathbf{x})$. For more generality, we do not assume that $\mathcal{O}_g$ is Lipschitz continuous with respect to the input even g is Lipschitz continuous. One example of such stochastic oracle is $O_g(\mathbf{x}) = g(\mathbf{x}; \zeta) + \xi$, where $\mathrm{E}_\zeta[g(\mathbf{x}; \zeta)] = g(\mathbf{x})$ and ξ is a zero-mean random noise (e.g., zero-mean

Table 1: Comparison with previous results. "mom. para." is short for momentum parameter. 1st and 2nd are short for first order and second order, respectively. "-" denotes no strict requirements and applicable to a range of updates. ↑ represents increasing as iterations and ↓ represents decreasing as iterations. ϵ denotes the target accuracy level for the objective gradient norm, i.e., $\mathrm{E}[\|\nabla F(\mathbf{x})\|] \leq \epsilon$.

Problem	Method	batch size	↑ or ↓ 1st mom. para.	↑ or ↓ 2nd mom. para.	Converge?
Non-convex (Adam-family)	This work	$O(1)$	↑	-	Yes
	[12]	$O(1)$	↓	constant	No
	[3]	$O(1)$	Non-↑	-	No
	[31]	$O(1/\epsilon^2)$	constant	↑	Yes
	[35]	$O(1)$	constant	↑	Yes
	[6]	$O(1)$	constant	↑	Yes

Gaussian noise). Therefore, the variance-reduced stochastic estimators based on SPIDER, SARAH or STORM are not applicable. Instead, we will consider another stochastic estimator based on moving average, i.e., we maintain and update a sequence of $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ by

$$\mathbf{z}_{t+1} = (1-\beta_t)\mathbf{z}_t + \beta_t \mathcal{O}_g(\mathbf{x}_t), \quad t = 0, \ldots, T. \tag{2}$$

We refer to this estimator sequence for tracking $\{g(\mathbf{x}_1), \ldots, g(\mathbf{x}_T)\}$ as stochastic moving average estimator (**SEMA**) in contrast to SPIDER/SARAH/STORM. It the literature, stochastic methods that employ the above estimator are usually referred to as momentum methods [3, 4, 16], with $1-\beta_t$ called the "momentum" parameter.

Besides in Adam-style methods, SEMA has been widely used in other stochastic non-convex optimization methods such as in stochastic compositional minimization [8, 10, 24, 25]. Although the SEMA has been widely used in practice, its power for solving a broad range of stochastic optimization problems has not been fully discovered. We present a simple and intuitive proof of convergence of a family of Adam-style algorithms with an increasing or large momentum parameter for the first-order moment, which include many variants such as Adam, AMSGrad, Adabound, AdaFom, etc. A surprising result is that Adam with an increasing or large momentum parameter for the first-order moment indeed converges at the same rate as SGD without any modifications on the update or restrictions on the "momentum" parameter for the second-order moment. In our analysis, β_t is decreasing in the same order of step size, which yields an increasing "momentum" parameter $1-\beta_t$. This increasing momentum parameter is more natural than the decreasing (or small) momentum parameter, which is indeed the reason that makes Adam diverge on some examples [21]. Our increasing/large "momentum" parameter $1-\beta_t$ is also consistent with that the large value close to 1 used in practice [12]. To the best of our knowledge, this is the first time that Adam was shown to converge for non-convex optimization with a more natural large "momentum" parameter for the first-order moment. We also prove that averaged variance of the stochastic estimator of the gradient decreases over time. A comparison of the results in this paper with existing results is summarized in Table 1. Moreover, our analysis can be extended to analyze the convergence of Adam-style algorithms for a broader family of non-convex optimization problems, including compositional optimization, min-max optimization and bilevel optimization problems [11].

2. Notations and Preliminaries

Notations and Definitions. Let $\|\cdot\|$ denote the Euclidean norm of a vector or the spectral norm of a matrix. Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix. A mapping h is L-Lipschitz continuous iff $\|h(\mathbf{x}) - h(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\|$ for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. A function F is called L-smooth if its gradient $\nabla F(\cdot)$ is L-Lipschitz continuous. A function g is λ-strongly convex iff $g(\mathbf{x}) \geq g(\mathbf{x}') + \nabla g(\mathbf{x}')^\top(\mathbf{x} - \mathbf{x}') + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}'\|^2$ for any $\mathbf{x}, \mathbf{x}'$. A function $g(\mathbf{y})$ is called λ-strongly concave if $-g(\mathbf{y})$ is λ-strongly convex. For a differentiable function $f(\mathbf{x}, \mathbf{y})$, we let $\nabla_x f(\mathbf{x}, \mathbf{y})$ and $\nabla_y f(\mathbf{x}, \mathbf{y})$ denote the partial gradients with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively. Denote by $\nabla f(\mathbf{x}, \mathbf{y}) = (\nabla_x f(\mathbf{x}, \mathbf{y})^\top, \nabla_y f(\mathbf{x}, \mathbf{y})^\top)^\top$. Let $\circ$ denote an element-wise product. We denote by $\mathbf{x}^2$, $\sqrt{\mathbf{x}}$ an element-wise square and element-wise square-root, respectively.

We will consider non-convex minimization (4), which has broad applications in machine learning. This paper focuses on theoretical analysis and our goal for these problems is to find an ϵ-stationary solution of the primal objective function $F(\mathbf{x})$ by using stochastic oracles.

Definition 1 *Consider a differentiable function $F(\mathbf{x})$, a randomized solution $\mathbf{x}$ is called an ϵ-stationary point if it satisfies $\|\nabla F(\mathbf{x})\| \leq \epsilon$.*

Before ending this section, we present the widely used stochastic momentum method for solving non-convex minimization problem $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$ through an unbiased stochastic oracle that returns a random variable $\mathcal{O}_F(\mathbf{x})$ for any $\mathbf{x}$ such that $\mathrm{E}[\mathcal{O}_F(\mathbf{x})] = \nabla F(\mathbf{x})$. For solving this problem, the stochastic momentum method (in particular stochastic heavy-ball (SHB) method) that employs the SEMA update is given by

$$\begin{cases} \mathbf{v}_{t+1} = (1-\beta)\mathbf{v}_t + \beta\mathcal{O}_F(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{v}_{t+1}, \quad t = 0, \ldots, T. \end{cases} \tag{3}$$

where $\mathbf{v}_0 = \mathcal{O}_F(\mathbf{x}_0)$. In the literature, $1-\beta$ is known as the momentum parameter and η is known as the step size or learning rate. It is notable that the stochastic momentum method can be also written as $\mathbf{z}_{t+1} = \beta\mathbf{z}_t - \eta\mathcal{O}_F(\mathbf{x}_t)$, and $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{z}_t$ [29], which is equivalent to the above update with some parameter change shown in Appendix A. The above method has been analyzed in various studies [10, 16, 29, 30]. Nevertheless, we will give a unified analysis for the Adam-family methods with a much more concise proof, which covers SHB as a special case. A core to the analysis is the use of a known variance recursion property of the SEMA estimator stated below.

Lemma 2 **(Variance Recursion of SEMA)***[Lemma 2, [24]] Consider a moving average sequence $\mathbf{z}_{t+1} = (1-\beta_t)\mathbf{z}_t + \beta_t\mathcal{O}_h(\mathbf{x}_t)$ for tracking $h(\mathbf{x}_t)$, where $\mathrm{E}_t[\mathcal{O}_h(\mathbf{x}_t)] = h(\mathbf{x}_t)$ and h is a L-Lipschitz continuous mapping. Then we have*

$$\mathrm{E}_t[\|\mathbf{z}_{t+1} - h(\mathbf{x}_t)\|^2] \leq (1-\beta_t)\|\mathbf{z}_t - h(\mathbf{x}_{t-1})\|^2 + 2\beta_t^2\mathrm{E}_t[\|\mathcal{O}_h(\mathbf{x}_t) - h(\mathbf{x}_t)\|^2] + \frac{L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2}{\beta_t}.$$

where E_t denotes the expectation conditioned on all randomness before $\mathcal{O}_h(\mathbf{x}_t)$.

We refer to the above property as variance recursion (VR) of the SEMA.

3. A Novel Analysis of Adam with a Large Momentum Parameter

In this section, we consider the standard stochastic non-convex minimization, i.e.,

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}), \tag{4}$$

Table 2: Different Adam-style methods and their satisfactions of Assumption 2

method	update for h_t	Additional assumption	c_l and c_u
SHB	$\mathbf{u}_{t+1} = 1, G = 0$	-	$c_l = 1, c_u = 1$
Adam	$\mathbf{u}_{t+1} = (1-\beta'_t)\mathbf{u}_t + \beta'_t \mathcal{O}_F^2(\mathbf{x}_t)$	$\|\mathcal{O}_F\|\infty \leq G$	$c_l \geq \frac{1}{G+G_0}, c_u \leq \frac{1}{G_0}$
AMSGrad	$\mathbf{u}'_{t+1} = (1-\beta'_t)\mathbf{u}'_t + \beta'_t \mathcal{O}_F^2(\mathbf{x}_t)$ $\mathbf{u}_{t+1} = \max(\mathbf{u}_t, \mathbf{u}'_{t+1})$	$\|\mathcal{O}_F\|\infty \leq G$	$c_l \geq \frac{1}{G+G_0}, c_u \leq \frac{1}{G_0}$
AdaFom (AdaGrad)	$\mathbf{u}_{t+1} = \frac{1}{t+1}\sum_{i=0}^t \mathcal{O}_F^2(\mathbf{x}_i)$	$\|\mathcal{O}_F\|\infty \leq G$	$c_l \geq \frac{1}{G+G_0}, c_u \leq \frac{1}{G_0}$
Adam$^+$	$\mathbf{u}_{t+1} = \|\mathbf{v}_{t+1}\|$	$\|\mathcal{O}_F\| \leq G$	$c_l \geq \frac{1}{\sqrt{G}+G_0}, c_u \leq \frac{1}{G_0}$
Adabound	$\mathbf{u}'_{t+1} = (1-\beta'_t)\mathbf{u}'_t + \beta'_t \mathcal{O}_F^2(\mathbf{x}_t)$ $\mathbf{u}_t = \Pi_{[1/c_u^2, 1/c_l^2]}[\mathbf{u}'_{t+1}], \quad G_0 = 0$	-	$c_l = c_l, c_u = c_u$

where F is smooth and is accessible only through an unbiased stochastic oracle. These conditions are summarized below for our presentation.

Assumption 1 *Regarding problem (4), the following conditions hold:*

- *∇F is L_F Lipschitz continuous;*
- *F is accessible only through an unbiased stochastic oracle that returns a random variable $\mathcal{O}_F(\mathbf{x})$ for any $\mathbf{x}$ such that $\mathrm{E}[\mathcal{O}_F(\mathbf{x})] = \nabla F(\mathbf{x})$, and $\mathcal{O}_F$ has a variance bounded by $\mathrm{E}[\|\mathcal{O}_F(\mathbf{x}) - \nabla F(\mathbf{x})\|^2] \leq \sigma^2(1 + c\|\nabla F(\mathbf{x})\|^2)$ for some $c > 0$.*

Remark: Note that the variance bounded condition is slightly weaker than the standard condition $\mathrm{E}[\|\mathcal{O}_F(\mathbf{x}) - \nabla F(\mathbf{x})\|^2] \leq \sigma^2$. An example of a random oracle that satisfies our condition but not the standard condition is $\mathcal{O}_F(\mathbf{x}) = d \cdot \nabla F(\mathbf{x}) \circ \mathbf{e}_i$, where $i \in \{1, \ldots, d\}$ is randomly sampled and $\mathbf{e}_i$ denotes the i-th canonical vector with only i-th element equal to one and others zero. For this oracle, we can show that $\mathrm{E}[\mathcal{O}_F(\mathbf{x})] = \nabla F(\mathbf{x})$ and $\mathrm{E}[\|\mathcal{O}_F(\mathbf{x}) - \nabla F(\mathbf{x})\|^2] \leq (d-1)\|\nabla F(\mathbf{x})\|^2$.

In the following we present a novel analysis of Adam-style methods based on VR of SEMA. The update rule of Adam-style rules has been given in (1). A key to our convergence analysis of Adam-style algorithms is the boundness of the step size scaling factor $\mathbf{s}_t = 1/(\sqrt{\mathbf{u}_{t+1}} + G_0)$, which is presented as an assumption below for more generality. We denote by $\tilde{\eta}_t = \eta_t \mathbf{s}_t$.

Assumption 2 *For the Adam-style algorithms in (1), we assume that $\mathbf{s}_t = 1/(\sqrt{\mathbf{u}_{t+1}} + G_0)$ is upper bounded and lower bounded, i.e., there exists $0 < c_l < c_u$ such that $c_l \leq \|\mathbf{s}_t\|_\infty \leq c_u$.*

Remark: Under the standard assumption $\|\mathcal{O}_F(\mathbf{x})\|_\infty \leq G$ [12, 21], we can see many variants of Adam will satisfy the above condition. Examples include Adam [12], AMSGrad [21], AdaFom [3], Adam$^+$ [15], whose $\mathbf{u}_t$ shown in Table 2 all satisfy the above condition under the bounded stochastic oracle assumption. Even if the condition $\|\mathcal{O}_F(\mathbf{x})\|_\infty \leq G$ is not satisfied, we can also use the clipping idea to make $\mathbf{u}_t$ bounded. This is used in Adabound [17], whose $\mathbf{u}_t$ is given by

$$\text{Adabound: } \mathbf{u}'_{t+1} = (1-\beta'_t)\mathbf{u}'_t + \beta'_t \mathcal{O}_F^2(\mathbf{x}_t), \quad \mathbf{u}_t = \Pi_{[1/c_u^2, 1/c_l^2]}[\mathbf{u}_{t+1}], \quad G_0 = 0,$$

where $c_l \leq c_u$ and $\Pi_{[a,b]}$ is a projection operator that projects the input into the range $[a, b]$. We summarize these updates and their satisfactions of Assumption 2 in Table 2. Note that SHB also satisfies Assumption 2 automatically.

To prove the convergence of the update (1). We first present a key lemma.

Lemma 3 *For $\mathbf{x}_{t+1} = \mathbf{x}_t - \tilde{\eta}_t \circ \mathbf{v}_{t+1}$ with $\eta_t c_l \leq \tilde{\eta}_{t,i} \leq \eta_t c_u$ and $\eta_t L_F \leq c_l/(2c_u^2)$, we have*

$$F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \frac{\eta_t c_u}{2}\|\nabla F(\mathbf{x}_t) - \mathbf{v}_{t+1}\|^2 - \frac{\eta_t c_l}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t c_l}{4}\|\mathbf{v}_{t+1}\|^2.$$

With the above lemmas, we can establish the following convergence of Adam-style algorithms.

Theorem 1 *Let $\Delta_t = \|\mathbf{v}_{t+1} - \nabla F(\mathbf{x}_t)\|^2$ and $F(\mathbf{x}_0) - F_* \leq \Delta_F$ where $F_* = \min_{\mathbf{x}} F(\mathbf{x})$. Suppose Assumptions 1 and 2 hold. With $\beta_t = \beta \leq \frac{\epsilon^2 c_l}{12\sigma^2 c_u}$, $\eta_t = \eta \leq \min\{\frac{\beta\sqrt{c_l}}{2L_F\sqrt{c_u^3}}, \frac{1}{\sqrt{2}L_F c_u}, \frac{c_l}{2c_u^2 L_F}\}$, $T \geq \max\{\frac{6\Delta_0 c_u}{\beta\epsilon^2 c_l}, \frac{12\Delta_F}{\eta\epsilon^2 c_l}\}$, we have*

$$\mathrm{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq \epsilon^2, \quad \mathrm{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t\right] \leq 2\epsilon^2.$$

Remark: We can see that the Adam-style algorithms enjoy an oracle complexity of $O(1/\epsilon^4)$ for finding an ϵ-stationary solution. To the best of our knowledge, this is the first time that the Adam with a large momentum parameter $1-\beta$ was proven to converge. One can also use a decreasing step size $\eta_t \propto 1/\sqrt{t}$ and decreasing $\beta_t = 1/\sqrt{t}$ (i.e, increasing momentum parameter) and establish a rate of $\widetilde{O}(1/\sqrt{T})$ as stated below.

Theorem 2 *Let $\Delta_t = \|\mathbf{v}_{t+1} - \nabla F(\mathbf{x}_t)\|^2$ and $F(\mathbf{x}_0) - F_* \leq \Delta_F$ where $F_* = \min_{\mathbf{x}} F(\mathbf{x})$. Suppose Assumption 1 holds. With $c_1 = \min(1, \frac{1}{4c\sigma^2})$, $\beta_t = \frac{c_l}{8\sigma^2 c c_u\sqrt{t+1}}$, $\eta_t = \min\{\frac{\beta_t\sqrt{c_l}}{2L_F\sqrt{c_u^3}}, \frac{1}{2L_F c_u}, \frac{c_l}{2c_u^2 L_F}\}$ and $T \geq \widetilde{O}(\frac{c_u^5 c^2 L_F^2\sigma^4\Delta_F^2/c_l^5 + \Delta_0^2 c^2 c_u^4\sigma^4/c_l^4 + 1/c}{\epsilon^4})$, we have*

$$\mathrm{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq O(\epsilon^2), \quad \mathrm{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t\right] \leq O(\epsilon^2).$$

4. Conclusion & Discussion

In this paper, we have provided a simple and generic convergence analysis for a family of Adam-style methods for solving non-convex minimization problems. We leveraged the variance recursion of the stochastic moving average estimator and established the convergence of practically used Adam and its variants. Our results bring some new insights to make the Adam method converge or convergence better.

Indeed, the Lemma 3 paves the way for the convergence analysis of many Adam-style algorithms for solving a broader family of problems, including non-convex strongly concave min-max optimization problems, non-convex stochastic compositional optimization problems, and non-convex bilevel optimization problems. It is also worth mentioning that we can also prove a faster convergence of the Adam-style algorithms under a strong Polyak-Łojasiewicz condition. We will explore this direction and examine how it will affect the convergence rate in the future work.

Finally, it is worth mentioning that the oracle complexities established in this paper are optimal under a general stochastic unbiased oracle model. In addition, one can also replace the variance recursion of the stochastic moving average estimator by that of other stochastic estimators (e.g. STORM) to prove an optimal convergence under Lipchitz continuous oracle model.

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 1709–1720, 2017.

[2] Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3267–3275, 2020.

[3] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of A class of adam-type algorithms for non-convex optimization. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[4] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 2260–2268, 2020.

[5] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 15236–15245, 2019.

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

[7] 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.

[8] Yuri M. Ermoliev. Methods of stochastic programming. *Monographs in Optimization and Operations Research*, 1976.

[9] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 689–699, 2018.

[10] Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30 (1):960–979, 2020.

[11] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. On stochastic moving-average estimators for non-convex optimization. *arXiv preprint arxiv:2104.14840*, 2021.

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

[13] 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 (AISTATS)*, pages 983–992, 2019.

[14] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations (ICLR)*, 2020.

[15] Mingrui Liu, Wei Zhang, Francesco Orabona, and Tianbao Yang. Adam$^+$: A stochastic method with adaptive variance reduction. *arXiv preprint arXiv:2011.11985*, 2020.

[16] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, pages 18261–18271, 2020.

[17] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[18] H Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 109–123, 2004.

[19] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

[20] Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

[21] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *6th International Conference on Learning Representations (ICLR)*, 2018.

[22] Naichen Shi, Dawei Li, Mingyi Hong, and Sun Ruoyu. RMSprop converges with proper hyper-parameter. In *9th International Conference on Learning Representations (ICLR)*, 2021.

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

[24] Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[25] Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *The Journal of Machine Learning Research*, 18(1):3721–3743, 2017.

[26] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 2406–2416, 2019.

[27] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1306–1316, 2018.

[28] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6677–6686, 2019.

[29] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

[30] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7184–7193, 2019.

[31] Manzil Zaheer, Sashank J. Reddi, Devendra Singh Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 9815–9825, 2018.

[32] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[33] Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 4035–4043, 2017.

[34] Fangyu Zou and Li Shen. On the convergence of adagrad with momentum for training deep neural networks. *arXiv preprint arXiv:1808.03408*, 2(3):5, 2018.

[35] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11127–11135, 2019.

Appendix

Appendix A. Stochastic Momentum Method

In the literature [29], the stochastic heavy-ball method is written as:

$$\text{SHB:} \quad \begin{cases} \mathbf{v}'_{t+1} = \beta' \mathbf{v}'_t - \eta' \mathcal{O}_F(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{v}_{t+1}, \quad t = 0, \ldots, T. \end{cases} \tag{5}$$

To show the resemblance between the above update and the one in (3), we can transform them into one sequence update:

$$\begin{aligned} &(3): \mathbf{x}_{t+1} = \mathbf{x}_t - \eta\beta \mathcal{O}_F(\mathbf{x}_t) + (1-\beta)(\mathbf{x}_t - \mathbf{x}_{t-1}) \\ &\text{SHB: } \mathbf{x}_{t+1} = \mathbf{x}_t - \eta' \mathcal{O}_F(\mathbf{x}_t) + \beta'(\mathbf{x}_t - \mathbf{x}_{t-1}). \end{aligned}$$

We can see that SHB is equivalent to (3) with $\eta' = \eta\beta$ and $\beta' = (1-\beta)$.

Appendix B. Proof of Lemma 3

Proof Due to the smoothness of F, we can prove that under $\eta_t L_F \leq c_l/(2c_u^2)$

$$\begin{aligned} F(\mathbf{x}_{t+1}) &\leq F(\mathbf{x}_t) + \nabla F(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L_F}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= F(\mathbf{x}_t) - \nabla F(\mathbf{x}_t)^\top (\tilde{\eta}_t \circ \mathbf{v}_{t+1}) + \frac{L_F}{2}\|\tilde{\eta}_t \circ \mathbf{v}_{t+1}\|^2 \\ &\leq F(\mathbf{x}_t) + \frac{1}{2}\|\sqrt{\tilde{\eta}_t} \circ (\nabla F(\mathbf{x}_t) - \mathbf{v}_{t+1})\|^2 - \frac{1}{2}\|\sqrt{\tilde{\eta}_t} \circ \nabla F(\mathbf{x}_t)\|^2 + (\frac{L_F}{2}\|\tilde{\eta}_t \circ \mathbf{v}_{t+1}\|^2 - \frac{1}{2}\|\sqrt{\tilde{\eta}_t} \circ \mathbf{v}_{t+1}\|^2) \\ &\leq F(\mathbf{x}_t) + \frac{\eta_t c_u}{2}\|\nabla F(\mathbf{x}_t) - \mathbf{v}_{t+1}\|^2 - \frac{\eta_t c_l}{2}\|\nabla F(\mathbf{x}_t)\|^2 + \frac{\eta_t^2 c_u^2 L_F - \eta_t c_l}{2}\|\mathbf{v}_{t+1}\|^2 \\ &\leq F(\mathbf{x}_t) + \frac{\eta_t c_u}{2}\|\nabla F(\mathbf{x}_t) - \mathbf{v}_{t+1}\|^2 - \frac{\eta_t c_l}{2}\|\nabla F(\mathbf{x}_t)\|^2 - \frac{\eta_t c_l}{4}\|\mathbf{v}_{t+1}\|^2. \end{aligned}$$

■

Appendix C. Proof of Theorem 1

Proof By applying Lemma 2 to $\mathbf{v}_{t+1}$, we have

$$\mathrm{E}_t[\Delta_{t+1}] \leq (1-\beta)\Delta_t + 2\beta^2\sigma^2(1 + c\|\nabla F(\mathbf{x}_{t+1})\|^2) + \frac{L_F^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2}{\beta}.$$

Hence we have,

$$\mathrm{E}\left[\sum_{t=0}^{T}\Delta_t\right] \leq \mathrm{E}\left[\sum_{t=0}^{T}\frac{\Delta_t - \Delta_{t+1}}{\beta} + 2\beta\sigma^2(T+1) + 2\beta\sigma^2 c\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_{t+1})\|^2 + \sum_{t=0}^{T}\frac{L_F^2\eta^2 c_u^2\|\mathbf{v}_{t+1}\|^2}{\beta^2}\right].$$

Adding the above inequality with Lemma 3, we have

$$
\begin{aligned}
&\frac{\eta c_l}{2}\mathrm{E}\left[\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq F(\mathbf{x}_0) - F_* - \frac{\eta c_l}{4}\sum_{t=0}^{T}\|\mathbf{v}_{t+1}\|^2 \\
&+ \frac{\eta c_u}{2}\mathrm{E}\left[\sum_{t=0}^{T}\frac{\Delta_t - \Delta_{t+1}}{\beta} + 2\beta\sigma^2(T+1) + 2\beta\sigma^2 c\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_{t+1})\|^2 + \sum_{t=0}^{T}\frac{L_F^2\eta^2 c_u^2\|\mathbf{v}_{t+1}\|^2}{\beta^2}\right] \\
&\leq F(\mathbf{w}_0) - F_* - \frac{\eta c_l}{4}\sum_{t=0}^{T}\|\mathbf{v}_{t+1}\|^2 \\
&+ \frac{\eta c_u}{2}\mathrm{E}\left[\sum_{t=0}^{T}\frac{\Delta_t - \Delta_{t+1}}{\beta} + 2\beta\sigma^2(T+1) + 4\beta\sigma^2 c\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2 + 4\beta\sigma^2 c L_F^2\eta^2 c_u^2\|\mathbf{v}_{t+1}\|^2\right. \\
&\left.+ \sum_{t=0}^{T}\frac{L_F^2\eta^2 c_u^2\|\mathbf{v}_{t+1}\|^2}{\beta^2}\right]
\end{aligned}
$$

Let $L_F^2\eta^2c_u^3/(2\beta^2) \leq c_l/8$ (i.e., $\eta \leq \frac{\beta\sqrt{c_l}}{2L_F\sqrt{c_u^3}}$) and $2\beta\sigma^2 c \leq c_l/(4c_u)$, $2\beta\sigma^2 c L_F^2\eta^2 c_u^3 \leq c_l/8$ (i.e, $\eta L_F \leq \frac{1}{\sqrt{2}c_u}$), we have

$$
\frac{1}{T+1}\mathrm{E}\left[\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq \frac{\Delta_0 c_u}{\beta T c_l} + \frac{2(F(\mathbf{x}_0) - F_*)}{\eta c_l T} + 2\beta\sigma^2\frac{c_u}{c_l} + \frac{1}{2}\frac{1}{T+1}\mathrm{E}\Big[\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\Big].
$$

As a result,

$$
\frac{1}{T+1}\mathrm{E}\left[\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq \frac{2\Delta_0 c_u}{\beta T c_l} + \frac{4(F(\mathbf{x}_0) - F_*)}{\eta c_l T} + 4\beta\sigma^2\frac{c_u}{c_l}.
$$

With $\beta \leq \frac{\epsilon^2 c_l}{12\sigma^2 c_u}$, $T \geq \max\{\frac{6\Delta_0 c_u}{\beta\epsilon^2 c_l}, \frac{12\Delta_F}{\eta\epsilon^2 c_l}\}$, we conclude the proof for the first part. For the second part, we have

$$
\begin{aligned}
\mathrm{E}\left[\sum_{t=0}^{T}\Delta_t\right] &\leq \frac{\Delta_0}{\beta} + \beta\sigma^2(T+1) + \frac{c_l}{2c_u}\mathrm{E}\Big[\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\Big] + \mathrm{E}\Big[\sum_{t=0}^{T}\frac{c_l}{2c_u}\|\mathbf{v}_{t+1}\|^2\Big] \\
&\leq \frac{\Delta_0}{\beta} + 2\beta\sigma^2(T+1) + \frac{1}{2}\mathrm{E}\Big[\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\Big] + \mathrm{E}\Big[\sum_{t=0}^{T}\frac{1}{2}\Delta_t\Big]
\end{aligned}
$$

As a result,

$$
\mathrm{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t\right] \leq \frac{2\Delta_0}{\beta T} + 4\beta\sigma^2 + \mathrm{E}\Big[\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\Big] \leq 2\epsilon^2.
$$

■

Appendix D. Poof of Theorem 2

Proof By applying Lemma 2 to $\mathbf{v}_{t+1}$, we have

$$\mathrm{E}_t[\Delta_{t+1}] \le (1-\beta_t)\Delta_t + 2\beta_t^2\sigma^2(1+c\|\nabla F(\mathbf{x}_{t+1})\|^2) + \frac{L_F^2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2}{\beta_t}. \tag{6}$$

Hence we have

$$\mathrm{E}\left[\sum_{t=0}^{T}\beta_t\Delta_t\right] \le \mathrm{E}\left[\sum_{t=0}^{T}[\Delta_t-\Delta_{t+1}] + \sum_{t=0}^{T}2\beta_t^2\sigma^2(1+c\|\nabla F(\mathbf{x}_{t+1})\|^2) + \sum_{t=0}^{T}\frac{L_F^2\eta_t^2c_u^2\|\mathbf{v}_{t+1}\|^2}{\beta_t}\right]. \tag{7}$$

Combining this with Lemma 2,

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=0}^{T}\frac{\eta_t c_l}{2}\|\nabla F(\mathbf{x}_t)\|^2\right]\\
&\le \mathrm{E}\Bigg[\sum_{t=0}^{T}[F(\mathbf{x}_t)-F(\mathbf{x}_{t+1})] - \sum_{t=0}^{T}\frac{\eta_t c_l}{4}\|\mathbf{v}_{t+1}\|^2\\
&\quad + \frac{\eta_1 c_u}{2\beta_1}\left[\sum_{t=0}^{T}(\Delta_t-\Delta_{t+1}) + \sum_{t=0}^{T}2\beta_t^2\sigma^2(1+c\|\nabla F(\mathbf{x}_{t+1})\|^2) + \sum_{t=0}^{T}\frac{L_F^2\eta_t^2c_u^2\|\mathbf{v}_{t+1}\|^2}{\beta_t}\right]\Bigg]\\
&\le \mathrm{E}\Bigg[\sum_{t=0}^{T}[F(\mathbf{x}_t)-F(\mathbf{x}_{t+1})] - \sum_{t=0}^{T}\frac{\eta_t c_l}{4}\|\mathbf{v}_{t+1}\|^2\\
&\quad + \frac{\eta_1 c_u}{2\beta_1}\Bigg[\sum_{t=0}^{T}(\Delta_t-\Delta_{t+1}) + \sum_{t=0}^{T}2\beta_t^2\sigma^2 + \sum_{t=0}^{T}4\beta_t^2\sigma^2c\|\nabla F(\mathbf{x}_t)\|^2) + \sum_{t=0}^{T}4\beta_t^2\sigma^2cL_F^2\eta_t^2c_u^2\|\mathbf{v}_{t+1}\|^2\\
&\qquad + \sum_{t=0}^{T}\frac{L_F^2\eta_t^2c_u^2\|\mathbf{v}_{t+1}\|^2}{\beta_t}\Bigg]\Bigg]\\
&\le \mathrm{E}\left[F(\mathbf{x}_0)-F_* + \frac{\eta_1c_u\Delta_0}{2\beta_1} + \frac{2c_u\eta_1}{\beta_1}\sum_{t=0}^{T}\beta_t^2\sigma^2 + \sum_{t=0}^{T}\frac{\eta_tc_l}{4}\|\nabla F(\mathbf{x}_t)\|^2)\right],
\end{aligned} \tag{8}$$

where the last inequality holds because $\frac{2\eta_1c_u}{\beta_1}\beta_t^2\sigma^2c \le \frac{\eta_tc_l}{4}$, $\frac{2\eta_1}{\beta_1}\beta_t^2\sigma^2cL_F^2\eta_t^2c_u^3 \le \frac{\eta_tc_l}{8}$ and $\frac{\eta_1}{2\beta_1}\frac{L_F^2\eta_t^2c_u^3}{\beta_t} \le \frac{\eta_tc_l}{8}$. Hence,

$$\begin{aligned}
\mathrm{E}[\sum_{t=0}^{T}\eta_Tc_l\|\nabla F(\mathbf{x}_t)\|^2] &\le \mathrm{E}[\sum_{t=0}^{T}\eta_tc_l\|\nabla F(\mathbf{x}_t)\|^2]\\
&\le \mathrm{E}\left[4(F(\mathbf{x}_0)-F_*) + \frac{2\eta_1c_u\Delta_0}{\beta_1} + \sum_{t=0}^{T}\frac{8c_u\eta_1}{\beta_1}\beta_t^2\sigma^2\right]\\
&\le \mathrm{E}\left[4(F(\mathbf{x}_0)-F_*) + \frac{\sqrt{c_l}\Delta_0}{L_F\sqrt{c_u}} + \sum_{t=0}^{T}\frac{4\sqrt{c_l}}{L_F\sqrt{c_u}}\beta_t^2\sigma^2\right].
\end{aligned} \tag{9}$$

Thus,

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] &\leq \mathrm{E}\left[\frac{4(F(\mathbf{x}_0)-F(\mathbf{x}_*))}{\eta_T c_l(T+1)} + \frac{\Delta_0}{L_F\sqrt{c_l c_u}\eta_T(T+1)} + \sum_{t=0}^{T}\frac{4\beta_t^2}{\eta_T L_F\sqrt{c_l c_u}(T+1)}\sigma^2\right] \\
&\leq \frac{4\Delta_F}{\eta_T c_l(T+1)} + \frac{\Delta_0}{L_F\sqrt{c_l c_u}\eta_T(T+1)} + \frac{4\sigma^2}{\eta_T L_F\sqrt{c_l c_u}(T+1)}\sum_{t=0}^{T}\beta_t^2 \\
&\leq \frac{4\Delta_F}{\eta_T c_l(T+1)} + \frac{\Delta_0}{L_F\sqrt{c_l c_u}\eta_T(T+1)} + \frac{c_l^2}{16\eta_T\sigma^2 L_F c^2\sqrt{c_l c_u^5}(T+1)}\ln(T+2).
\end{aligned} \tag{10}$$

Setting $T \geq \widetilde{O}(\frac{c_u^5 c^2 L_F^2\sigma^4\Delta_F^2/c_l^5+\Delta_0^2 c^2 c_u^4\sigma^4/c_l^4+1/c}{\epsilon^4})$, we conclude the proof for the first part. For the second part, we have

$$\mathrm{E}\left[\sum_{t=0}^{T}\beta_t\Delta_t\right] \leq \Delta_0 + \frac{c_l^2}{16\sigma^4 c^2 c_u^2}\ln(T+2) + \frac{1}{2}\mathrm{E}\left[\sum_{t=0}^{T}\beta_t\|\nabla F(\mathbf{x}_t)\|^2\right] + \mathrm{E}\left[\sum_{t=0}^{T}\frac{1}{2}\beta_t\Delta_t\right]. \tag{11}$$

Then,

$$\mathrm{E}\left[\frac{1}{T+1}\sum_{t=0}^{T}\Delta_t\right] \leq \frac{2\Delta_0}{\beta_T T} + \frac{c_l^2}{8\sigma^4 c^2 c_u^2(T+1)}\ln(T+2) + \frac{1}{2}\mathrm{E}\left[\sum_{t=0}^{T}\beta_t\|\nabla F(\mathbf{x}_t)\|^2\right], \tag{12}$$

which concludes the proof of the second part. ■