Closing the Gap Between the Upper Bound and the Lower Bound of Adam's Iteration Complexity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, Arjevani et al. [1] establish a lower bound of iteration complexity for the first-order optimization under an L-smooth condition and a bounded noise variance assumption. However, a thorough review of existing literature on Adam's convergence reveals a noticeable gap: none of them meet the above lower bound. In this paper, we close the gap by deriving a new convergence guarantee of Adam, with only an L-smooth condition and a bounded noise variance assumption. Our results remain valid across a broad spectrum of hyperparameters. Especially with properly chosen hyperparameters, we derive an upper bound of iteration complexity of Adam and show that it meets the lower bound for first-order optimizers. To the best of our knowledge, this is the first to establish such a tight upper bound for Adam's convergence. Our proof utilizes novel techniques to handle the entanglement between momentum and adaptive learning rate and to convert the first-order term in the Descent Lemma to the gradient norm, which may be of independent interest.

1 Introduction

First-order optimizers, also known as gradient-based methods, make use of gradient (first-order derivative) information to find the minimum of a function. They have become a cornerstone of many machine learning algorithms due to the efficiency as only gradient informaiton is required, and the flexibility as gradients can be easily computed for any function represented as directed acyclic computational graph via auto-differentiation [2, 19].

Therefore, it is fundamental to theoretically uderstand the properties of these first-order methods. Recently, Arjevani et al. [1] establish a lower bound on the iteration complexity of stochastic first-order methods. Formally, for a well-studied setting where the objective is L-smooth and a stochastic oracle can query the gradient unbiasly with bounded variance (see Assumption 1 and 2), any stochastic first-order algorithm requires at least ε^{-4} queries (in the worst case) to find an ε-stationary point, i.e., a point with gradient norm at most ε. Arjevani et al. [1] further show the above lower bound is tight as it matches the existing upper bound of iteration complexity of SGD [1].

On the other hand, among first-order optimizers, Adam [16] becomes dominant in training state-of-the-art machine learning models [3, 15, 4, 11]. Compared to vanilla stochastic gradient descent (SGD), Adam consists of two more key components: (i) momentum to accumulate historical gradient information and (ii) adaptive learning rate to rectify coordinate-wise step sizes. The psedo-code of Adam is given as Algorithm 1. While the sophisticated design of Adam enables its empirical superiority, it brings great challenges for the theoretical analysis. After examining a series of theoretical works on the upper bound of iteration complexity of Adam [24, 9, 10, 27, 14, 21, 25], we find that none of them match the lower bound for first-order optimizers: they not only consume more queries than the lower bound to reach ϵ-stationary iterations but also requires additional assumptions.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

This theoretical mismatch becomes even more unnatural given the great empirical advantage of Adam over SGD, which incites us to think:

Is the gap between the upper and lower bounds for Adam a result of the inherent complexity induced by Adam's design, or could it be attributed to the proof techniques not being sharp enough?

This paper answers the above question, validating the latter hypothesis, by establishing a new upper bound on iteration complexity of Adam for a wide range of hyperparameters that cover typical choices. Specifically, our contribution can be summarized as follows:

- We examine existing works that analyze the iteration complexity of Adam, and find that none of them meets the lower bound of first-order optimization algorithms;
- We derive a new convergence guarantee of Adam with only assuming L-smooth condition and bounded variance assumption (Theorem 1), which holds for a wide range of hyperparameters covering typical choices;
- With chosen hyperparameters, we further tighten Theorem 1 and show that the upper bound on the iteration complexity of Adam meets the lower bound, closing the gap (Theorem 2). Our upper bound is tighter than existing results by a logarithmic factor, in spite of weaker assumption.

To the best of our knowledge, this work provide the first upper bound on the iteration complexity of Adam without additional assumptions other than L-smooth condition and bounded variance assumption. It is also the first upper bound matching the lower bound of first-order optimizers.

Organization of this paper. The rest of the paper is organized as follows: in Section 2, we first present the notations and settup of analysis in this paper ; in Section 3, we revisit the existing works on the iteration complexity of Adam; in Section 4, we present a convergence analysis of Adam with general hyperparameters (Theorem 1); in Section 5, we tighten Theorem 1 with a chosen hyperparameter, and derive an upper bound of Adam's iteration complexity which meets the lower bound; in Section 6, we discuss the limitation of our results; in Section 7, we discuss the related works.

2 Preliminary

The Adam algorithm is restated in Agorithm 1 for convenient reference. Note that compared to the orignal version of Adam in Kingma and Ba [16], the bias-correction terms are omitted to simplify the analysis, and our analysis can be immediately extended to the original version of Adam because the effect of bias-correction term decays exponentially. Also, in the original version of Adam, the adaptive learning rate is $\frac{\eta}{\sqrt{\boldsymbol{\nu}_t+\varepsilon\mathbb{1}_d}}$ instead of $\frac{\eta}{\sqrt{\boldsymbol{\nu}_t}}$. However, our setting is more challenging and our result can be easily extend to the original version of Adam, since the ε term makes the adaptive learning rate upper bounded and eases the analysis.

Algorithm 1 Adam

Input: Stochastic oracle $\boldsymbol{O}$, learning rate $\eta > 0$, initial point $\boldsymbol{w}_1 \in \mathbb{R}^d$, initial conditioner $\boldsymbol{\nu}_0 \in \mathbb{R}^+$, initial momentum $\boldsymbol{m}_0$, momentum parameter β_1, conditioner parameter β_2, number of epoch T

Sample $r \sim \text{Unif}\{1, \cdots, T\}$
For $t = 1 \to T$:
  Generate a random z_t, and query stochastic oracle $\boldsymbol{g}_t = \boldsymbol{O}_f(\boldsymbol{w}_t, z_t)$
  Calculate $\boldsymbol{\nu}_t = \beta_2 \boldsymbol{\nu}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^{\odot 2}$
  Calculate $\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$
  Update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{m}_t$
EndFor

Output: $\boldsymbol{w}_r$

Notations. For $a, b \in \mathbb{Z}^{\geq 0}$ and $a \leq b$, denote $[a, b] = \{a, a+1, \cdots, b-1, b\}$. For any two vectors $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$, denote $\boldsymbol{w} \odot \boldsymbol{v}$ as the Hadamard product (i.e., coordinate-wise multiplication) between $\boldsymbol{w}$ and $\boldsymbol{v}$. When analyzing Adam, we denote the true gradient at iteration t as $\boldsymbol{G}_t = \nabla f(\boldsymbol{w}_t)$, and the sigma algebra before iteration t as $\mathcal{F}_t = \sigma(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_{t-1})$. We denote conditional expectation as $\mathbb{E}^{|\mathcal{F}_t}[*] = \mathbb{E}[*|\mathcal{F}_t]$. We also use asymptotic notations $\boldsymbol{o}$, $\mathcal{O}$, Ω, and Θ, where $h_2(x) = \boldsymbol{o}_{x\to x_0}(h_1(x))$ means that $\lim_{x\to x_0} \frac{h_2(x)}{h_1(x)} = 0$ (when the context is clear, we abbreviate $x \to x_0$ and only use $\boldsymbol{o}(h_1(x))$); $h_2(x) = \mathcal{O}(h_1(x))$ means that there exists constant γ independent of x such that $h_2(x) \le \gamma g_1(x)$; $h_2(x) = \Omega(h_1(x))$ means that $h_1(x) = \mathcal{O}(h_2(x))$; and $h_2(x) = \Theta(h_1(x))$ means that $h_2(x) = \mathcal{O}(h_1(x))$ and $h_2(x) = \Omega(h_1(x))$.

Objective function. In this paper, we consider solving the following optimization problem: $\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$. We make the following assumption on the objective function f.

Assumption 1 (On objective function)**.** *We assume f is differentiable, and the gradient of f is L-Lipschitz.*

We denote the set of all objective functions satisfying Assumption 1 as $\mathcal{F}(L)$.

Stochastic oracle. As f is differentiable, we can utilize the gradient of f (i.e., ∇f) to solve the above optimization problem. However, the ∇f is usually expensive to compute. Instead, we query a stochastic estimation of ∇f through a stochastic oracle $\boldsymbol{O}$. Specifically, the stochastic oracle $\boldsymbol{O}$ consists of a distribution $\mathcal{P}$ over a measurable space $\mathcal{Z}$ and a mapping $\boldsymbol{O}_f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$. We make the following asssumption on $\boldsymbol{O}$.

Assumption 2 (On stochastic oracle)**.** *We assume that $\boldsymbol{O}$ is unbiased, i.e., $\forall \boldsymbol{w} \in \mathbb{R}^d$, $\mathbb{E}_{z\sim\mathcal{P}}\boldsymbol{O}_f(\boldsymbol{w}, z) = \nabla f(\boldsymbol{w})$. We further assume $\boldsymbol{O}$ has bounded variance, i.e., $\forall \boldsymbol{w} \in \mathbb{R}^d$, $\mathbb{E}_{z\sim\mathcal{P}}[\|\boldsymbol{O}_f(\boldsymbol{w}, z) - \nabla f(\boldsymbol{w})\|^2] \le \sigma^2$.*

We denote the set of all stochastic oracles satisfying Assumption 2 with variance bound σ^2 as $\mathfrak{O}(\sigma^2)$.

Adam belongs to first-order optimization algorithms, which is defined as follows:

Definition 1 (First-order optimization algorithm)**.** *An algorithm $\boldsymbol{A}$ is called a first-order optimization algorithm, if it takes an input $\boldsymbol{w}_1$ and hyperparameter θ, and produces a sequence of parameters as follows: first sample a random seed r from some distribution $\mathcal{P}_r$[1], set $\boldsymbol{w}_1^{\boldsymbol{A}(\theta)} = \boldsymbol{w}_1$ and then update the parameters as*

$$\boldsymbol{w}_{t+1}^{\boldsymbol{A}(\theta)} = \boldsymbol{A}_\theta^t(r, \boldsymbol{w}_1^{\boldsymbol{A}(\theta)}, \boldsymbol{O}_f(\boldsymbol{w}_1^{\boldsymbol{A}(\theta)}, z_1), \cdots, \boldsymbol{O}_f(\boldsymbol{w}_t^{\boldsymbol{A}(\theta)}, z_t)),$$

where $z_1, z_2, \cdots, z_t$ are sampled i.i.d. from $\mathcal{P}$.

Denote the set of all first-order optimization algorithms as $\mathcal{A}_{\text{first}}$. We next introduce *iteration complexity* to measure the convergence rate of optimization algorithms.

Definition 2 (Iteration complexity)**.** *The iteration complexity of first-order optimization algorithm $\boldsymbol{A}$ is defined as*

$$\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2) = \sup_{\boldsymbol{O}\in\mathfrak{O}(\sigma^2)} \sup_{f\in\mathcal{F}(L)} \sup_{\boldsymbol{w}_1 : f(\boldsymbol{w}_1)=\Delta} \inf_\theta \{T : \mathbb{E}\|\nabla f(\boldsymbol{w}_T^{\boldsymbol{A}(\theta)})\| \le \varepsilon\}.$$

Furthermore, the iteration complexity of the family of first-order optimization algorithms $\mathcal{A}_{\text{first}}$ is

$$\mathcal{C}_\varepsilon(\Delta, L, \sigma^2) = \sup_{\boldsymbol{O}\in\mathfrak{O}(\sigma^2)} \sup_{f\in\mathcal{F}(L)} \sup_{\boldsymbol{w}_1 : f(\boldsymbol{w}_1)=\Delta} \inf_{\boldsymbol{A}\in\mathcal{A}_{\text{first}}} \inf_\theta \{T : \mathbb{E}\|\nabla f(\boldsymbol{w}_T^{\boldsymbol{A}(\theta)})\| \le \varepsilon\}.$$

It should be noticed that the iteration complexity of the family of first-order optimization algorithms is a lower bound of the iteration complexity of a specific first-order optimization algorithm, i.e., $\forall \boldsymbol{A} \in \mathcal{A}_{\text{first}}, \mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2) \ge \mathcal{C}_\varepsilon(\Delta, L, \sigma^2)$.

3 None of existing upper bounds match the lower bound

In this section, we examine existing works that study the iteration complexity of Adam, and defer a discussion of other existing works to Appendix A. We find that none of them match the lower bound for first-order algorithms provided in [1] (restated as follows).

[1] Such a random seed allows sampling from all iterations to generate the final output of the optimization algorithm. As an example, Algorithm 1 set $\mathcal{P}_r$.

Proposition 1 (Theorem 3, [1]). *$\forall L, \Delta, \sigma^2 > 0$, we have $\mathcal{C}_\varepsilon(\Delta, L, \sigma^2) = \Omega(\frac{1}{\varepsilon^4})$.*

Note that in the above bound, we omit the dependence of the lower bound over Δ, L, and σ^2, which is a standard practice in existing works (see Cutkosky and Mehta [8], Xie et al. [23], Faw et al. [13] as examples) because the dependence over the accuracy ε can be used to derive how much additional iterations is required for a smaller target accuracy and is thus of more interest. In this paper, when we say "match the lower bound", we always mean that the upper bound has the same order of ε as the lower bound.

Generally speaking, existing works on the iteration complexity of Adam can be divided into two categories: they either (i) assume that gradient is universally bounded or (ii) make stronger assumptions on smoothness. Below we respectively explain how these two categories of works do not match the lower bound in [1].

The first line of works, including Zaheer et al. [24], De et al. [9], Défossez et al. [10], Zou et al. [27], Guo et al. [14], assume that the gradient norm of f is universally bounded, i.e., $\|\nabla f(\boldsymbol{w})\| \leq G$, $\forall \boldsymbol{w} \in \mathbb{R}^d$. In other words, what they consider is another iteration complexity defined as follows:

$$\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2, G) \triangleq \sup_{\boldsymbol{O}\in\mathfrak{O}(\sigma^2)} \sup_{f\in\mathcal{F}(L), \|\nabla f\|\leq G} \sup_{\boldsymbol{w}_1 : f(\boldsymbol{w}_1)=\Delta} \inf_\theta \{T : \mathbb{E}\|\nabla f(\boldsymbol{w}_T^{\boldsymbol{A}(\theta)})\| \leq \varepsilon\}.$$

This line of works do not match the lower bound due to the following two reasons: First of all, the upper bound they derive is $O(\frac{\log 1/\varepsilon}{\varepsilon^4})$, which has an additional $\log \varepsilon$ factor more than the lower bound; secondly, the bound they derive is for $\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2, G)$. Note that $\mathcal{F}(L) \cap \{f : \|\nabla f\| \leq G\}$ is a proper subset of $\mathcal{F}(L)$ for any G, where a simple example in $\mathcal{F}(L)$ but without bounded gradient is the quadratic function $f(x) = \|x\|^2$. Therefore, we have that

$$\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2) \geq \mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2, G), \quad \forall G \geq 0, \tag{1}$$

and thus the upper bound on $\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2, G)$ does not apply to $\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2)$. Moreover, their upper bound of $\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2, G)$ tends to ∞ as $G \to \infty$, which indicates that if following their analysis the upper bound of $\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2)$ would be infinity based on Eq. (1).

The second line of works includes Shi et al. [21], Zhang et al. [25], Wang et al. [22], which additionally assume a mean-squared smoothness property besides Assumption 1 and 2, i.e., $\mathbb{E}_{z\sim\mathcal{P}}\|\boldsymbol{O}_f(\boldsymbol{w}, z) - \boldsymbol{O}_f(\boldsymbol{v}, z)\|^2 \leq L\|\boldsymbol{w} - \boldsymbol{v}\|^2$. Denote $\tilde{\mathfrak{O}}(\sigma^2, L) \triangleq \{\boldsymbol{O} : \mathbb{E}_{z\sim\mathcal{P}}\|\boldsymbol{O}_f(\boldsymbol{w}, z) - \boldsymbol{O}_f(\boldsymbol{v}, z)\|^2 \leq L\|\boldsymbol{w} - \boldsymbol{v}\|^2, \forall \boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d\} \cap \mathfrak{O}(\sigma^2)$. The iteration complexity that they consider is defined as follows:

$$\tilde{\mathcal{C}}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2) = \sup_{\boldsymbol{O}\in\tilde{\mathfrak{O}}(\sigma^2, L)} \sup_{f\in\mathcal{F}(L)} \sup_{\boldsymbol{w}_1 : f(\boldsymbol{w}_1)=\Delta} \inf_\theta \{T : \mathbb{E}\|\nabla f(\boldsymbol{w}_T^{\boldsymbol{A}(\theta)})\| \leq \varepsilon\}.$$

The rate derived in [21, 25, 22] is $O(\frac{\log 1/\varepsilon}{\varepsilon^6})$, which is derived by minimizing the upper bounds in [21, 25, 22] with respect to the hyperparameter of adaptive learning rate β_2. According to [1], the lower bound of iteration complexity of $\tilde{\mathcal{C}}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2)$ is $\Omega(\frac{1}{\varepsilon^3})$ and smaller than the original lower bound $\Omega(\frac{1}{\varepsilon^4})$, resulting in an even larger gap between the upper bound and lower bound.

On the other hand, a concurrent work [17] which does not require bounded gradient assumption and mean-squared smoothness property but poses a stronger assumption on the stochastic oracle: the set of stochastic oracles they consider is $\tilde{\tilde{\mathfrak{O}}} = \{\boldsymbol{O} : \forall \boldsymbol{w} \in \mathbb{R}^d, \mathbb{E}_{z\sim\mathcal{P}}\boldsymbol{O}_f(\boldsymbol{w}, z) = \nabla f(\boldsymbol{w}), \mathbb{P}\left(\|\boldsymbol{O}_f(\boldsymbol{w}, z) - \nabla f(\boldsymbol{w})\|^2 \leq \sigma^2\right) = 1\}$. $\tilde{\tilde{\mathfrak{O}}}$ is a proper subset of $\mathfrak{O}$ because a simple example is that $\boldsymbol{O}_f(\boldsymbol{w}, z) = \nabla f(\boldsymbol{w}) + z$ where z is a standard gaussian variable. Therefore, their result does not provide a valid upper bound of $\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2)$.

4 Convergence analysis of Adam with only Assumptions 1 and 2

As discussed in Section 3, existing works on analyzing Adam require additional assumptions besides Assumption 1 and 2. In this section, we provide the first convergence analysis of Adam with only Assumption 1 and 2, which naturally gives an upper bound on the iteration complexity $\mathcal{C}_\varepsilon(\boldsymbol{A}, \Delta, L, \sigma^2)$. Specifically, we present the following theorem.

Theorem 1. *Let $\boldsymbol{A}$ be by Adam (Algorithm 1) and $\theta = (\eta, \beta_1, \beta_2)$ are the hyperparameters of $\boldsymbol{A}$. Let Assumption 1 and 2 hold. Then, if $0 \le \beta_1 < \beta_2 < 1$, we have*

$$
\begin{aligned}
\mathbb{E}\sum_{t=1}^{T} \|\nabla f(\boldsymbol{w}_t)\| \le & \sqrt{1-\beta_2}C_2 + \frac{2\sqrt{1-\beta_2}}{(1-\beta_1)\eta} C_1 d \ln\left(12C_2 + 2T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l} + (3-\beta_2)\sigma^2} + 4dC_1 \ln dC_1\right) \\
& + \sqrt{C_2 + \frac{2}{(1-\beta_1)\eta} C_1 d \ln\left(12C_2 + 2T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l} + (3-\beta_2)\sigma^2} + 4dC_1 \ln dC_1\right)} \\
& \times \sqrt{12C_2 + 2T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l} + (3-\beta_2)\sigma^2} + 4dC_1 \ln dC_1}. \qquad (2)
\end{aligned}
$$

where $\boldsymbol{\nu}_{0,l}$ is the l-th coordinate of $\boldsymbol{\nu}_0$,

$$
C_1 = \left(\frac{L}{2}\eta^2 + 2\frac{\sqrt{1-\beta_2}}{(1-\beta_1)^2}\eta\sigma + \frac{\eta^2\beta_1}{\sqrt{\beta_2}(1-\frac{\beta_1}{\sqrt{\beta_2}})} + L^2 \frac{\beta_1\eta^3(1-\beta_1)}{\beta_2(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})(1-\frac{\beta_1}{\beta_2})^2}\frac{d}{\sigma}\frac{(1-\beta_1)^2}{(1-\frac{\beta_1}{\sqrt{\beta_2}})^2}\right)\frac{1}{1-\beta_2}.
$$

and

$$
C_2 = \frac{2}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1) + \sum_{l=1}^{d} 2C_1\left(\mathbb{E}\ln\left(\frac{1}{\boldsymbol{\nu}_{0,l}}\right) - T\ln\beta_2\right)\right).
$$

A proof sketch is given in Section 4.2 and the full proof is deferred to Appendix.

The right-hand side in Eq. (2) looks messy at the first glance. We next explain Theorem 1 in detail and make the upper bound's dependence over hyperparameters crystally clear.

4.1 Discussion on Theorem 1

Required assumptions and conditions. As mentioned previously, Theorem 1 only requires Assumption 1 and 2, which aligns with the setting of the lower bound (Proposition 1). To our best knowledge, this is the first analysis of Adam without additional assumptions. Also, Theorem 1 holds for general choices of hyperparameters since the only condition posed on hyperparameters is $\beta_1 < \beta_2$. Such condition covers a wide range of hyperparameters, e.g., the default setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in PyTorch [19].

Dependence over β_2, η, and T. Here we consider the influence of β_2, η, and T while fixing β_1 constant (we will discuss the effect of β_1 in Section 6). With logarithmic factors ignored and coefficients hidden, C_1, C_2 and the right-hand-side of Eq. (2) can be rewritten with asymptotic notations as

$$
\begin{aligned}
C_1 &= \tilde{\mathcal{O}}\left(\frac{\eta}{\sqrt{1-\beta_2}} + \frac{\eta^3}{(1-\beta_2)^{\frac{3}{2}}}\right), \\
C_2 &= \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\beta_2}} + \frac{\eta^2}{(1-\beta_2)^{\frac{3}{2}}} + \frac{1}{\eta} + T\sqrt{1-\beta_2} + \frac{\eta^2 T}{(1-\beta_2)^{\frac{1}{2}}}\right), \\
\mathbb{E}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\| &= \tilde{\mathcal{O}}\left(\sqrt{1-\beta_2}C_2 + \frac{\sqrt{1-\beta_2}}{\eta}C_1 + \sqrt{C_2 + \frac{C_1}{\eta}}\sqrt{C_2 + T + C_1}\right),
\end{aligned}
$$

where $\tilde{\mathcal{O}}$ denotes $\mathcal{O}$ with logarithmic terms ignored. Consequently, the dependence of Eq. (2) over β_2, η and T becomes

$$
\begin{aligned}
\mathbb{E}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\| =& \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\beta_2}} + \frac{\eta^2}{(1-\beta_2)^{\frac{3}{2}}} + \frac{1}{\eta} + \frac{\eta^2 T}{(1-\beta_2)^{\frac{1}{2}}}\right) \\
&+ \tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{\sqrt[4]{1-\beta_2}} + \frac{\eta\sqrt{T}}{(1-\beta_2)^{\frac{3}{4}}} + \frac{\sqrt{T}}{\sqrt{\eta}} + T\sqrt[4]{1-\beta_2} + \frac{\eta T}{(1-\beta_2)^{\frac{1}{4}}}\right).
\end{aligned}
$$

Therefore, in order to ensure convergence, $\min_{t\in[T]}\mathbb{E}\|\boldsymbol{G}_t\|_1 \to 0$ as $T \to \infty$, a sufficient condition is that the right-hand-side of the above equation is $\boldsymbol{o}(T)$. Specifically, by choosing $\eta = \Theta(T^{-a})$ and $1-\beta_2 = \Theta(T^{-b})$, we obtain that

$$
\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\| = \tilde{\mathcal{O}}\left(T^{\frac{b}{2}-1} + T^{-2a+\frac{3}{2}b-1} + T^{a-1} + T^{-2a+\frac{1}{2}b} + T^{\frac{b}{4}-\frac{1}{2}} + T^{-a+\frac{3}{4}b-\frac{1}{2}} + T^{\frac{1}{2}a-\frac{1}{2}} + T^{-a+\frac{1}{4}b}\right).
$$

By simple calculation, we obtain that the right-hand side of the above inequality is $\boldsymbol{o}(1)$ as $T \to \infty$ if and only if $0 < \frac{b}{4} < a < 1$ and $3b - 4a < 2$. Moreover, the minimum of the right-hand side of the above inequality is $\tilde{\mathcal{O}}(\frac{1}{T^{\frac{1}{4}}})$, which is achieved at $a = \frac{1}{2}$ and $b = 1$. Such a minimum implies an upper bound of the iteration complexity which at most differs from the lower bound by logarithmic factors as solving $\tilde{\mathcal{O}}(\frac{1}{T^{\frac{1}{4}}}) = \varepsilon$ gives $T = \tilde{\mathcal{O}}(\frac{1}{\varepsilon^4})$. In Theorem 2, we will further remove the logarithmic factor by giving a refined proof when $a = \frac{1}{2}$ and $b = 1$ and close the gap between the upper and lower bounds.

4.2 Proof Sketch of Theorem 1

In this section, we demonstrate the proof idea of Theorem 1. Concretely, we sketch the proof by identifying two key challenges in the proof and provide our solutions respectively.

Challenge I: Disentangle the stochasticity in momentum and adaptive learning rate. According to the standard descent lemma, we have that

$$\begin{aligned}\mathbb{E}f(\boldsymbol{w}_{t+1}) &= f(\boldsymbol{w}_t) + \mathbb{E}\left[\langle \boldsymbol{G}_t, \boldsymbol{w}_{t+1} - \boldsymbol{w}_t\rangle + \frac{L}{2}\left\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\right\|^2\right] \\ &\le \mathbb{E}f(\boldsymbol{w}_t) + \underbrace{\mathbb{E}\left[\left\langle \boldsymbol{G}_t, -\eta\frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{m}_t\right\rangle\right]}_{\text{First Order}} + \underbrace{\frac{L}{2}\eta^2\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{m}_t\right\|^2}_{\text{Second Order}}\end{aligned} \tag{3}$$

The first challenge arises from bounding the "First Order" term above. To faciliate the understanding of the difficulty, we compare the "First Order" term of Adam to the corresponding "First Order" term of SGD, i.e., $-\eta\mathbb{E}\langle \boldsymbol{G}_t, \boldsymbol{g}_t\rangle$. By directly applying $\mathbb{E}^{|\mathcal{F}_t} g_t = \boldsymbol{G}_t$, we obtain that the "First-Order" term of SGD equals to $-\eta\mathbb{E}\|\boldsymbol{G}_t\|^2\rangle$. However, as for Adam, there are two folds of trouble: firstly, we do not know what $\mathbb{E}^{|\mathcal{F}_t}\frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{m}_t$ is, as the stochasticity in $\boldsymbol{m}_t$ and $\boldsymbol{\nu}_t$ entangles. Secondly, even without $\boldsymbol{\nu}_t$, it is unclear how $\mathbb{E}^{\mathcal{F}_t}\boldsymbol{m}_t$ aligns with $\boldsymbol{G}_t$ given the existence of $\boldsymbol{g}_{t-1}, \cdots, \boldsymbol{g}_1$ in $\boldsymbol{m}_t$.

Solution to Challenge I. For $i \in [1, t]$, we define a set of surrogate conditioner $\widetilde{\boldsymbol{\nu}}_t^i \triangleq \beta_2^i \boldsymbol{\nu}_{t-i} + \sum_{j=0}^{i-1}\beta_2^j(1-\beta_2)\boldsymbol{G}_{t-i+1}^{\odot 2} + (1-\beta_2)\sigma^2$, and $\widetilde{\boldsymbol{\nu}}_t^0 \triangleq \boldsymbol{\nu}_t$. Note that $\widetilde{\boldsymbol{\nu}}_t^i$ is measurable with respect to $\mathcal{F}_{t-i+1}$. The key idea of our solution is the following *peeling-off strategy*: starting from $\mathbb{E}[\langle \boldsymbol{G}_t, \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{m}_t\rangle]$, we replace $\boldsymbol{\nu}_t = \widetilde{\boldsymbol{\nu}}_t^0$ by $\widetilde{\boldsymbol{\nu}}_t^1$ (of course, such a replacement will bring a error term, which we temporily ignore and will consider it in the formal proof) and obtain $\mathbb{E}[\langle \boldsymbol{G}_t, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \boldsymbol{m}_t\rangle]$. As $\boldsymbol{m}_t = \beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$, we further have $\mathbb{E}[\langle \boldsymbol{G}_t, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \boldsymbol{m}_t\rangle] = \mathbb{E}[\langle \boldsymbol{G}_t, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot (1-\beta_1)\boldsymbol{g}_t\rangle] + \mathbb{E}[\langle \boldsymbol{G}_t - \boldsymbol{G}_{t-1}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \beta_1\boldsymbol{m}_{t-1}\rangle] + \mathbb{E}[\langle \boldsymbol{G}_{t-1}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \beta_1\boldsymbol{m}_{t-1}\rangle]$. As $\widetilde{\boldsymbol{\nu}}_t^1$ is measurable w.r.t. $\mathcal{F}_t$, we can then disentangle the stochasticity in g_t and $\boldsymbol{\nu}_t$, and the term $\mathbb{E}[\langle \boldsymbol{G}_t, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot (1-\beta_1)\boldsymbol{g}_t\rangle]$ equals to $\mathbb{E}[\langle \boldsymbol{G}_t, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot (1-\beta_1)\boldsymbol{G}_t\rangle]$, which is desired. The term $\mathbb{E}[\langle \boldsymbol{G}_t - \boldsymbol{G}_{t-2}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \beta_1\boldsymbol{m}_{t-1}\rangle]$ is small due to L-smooth condition. The term $\mathbb{E}[\langle \boldsymbol{G}_{t-1}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \beta_1\boldsymbol{m}_{t-1}\rangle]$ resembles $\mathbb{E}[\langle \boldsymbol{G}_t, \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{m}_t\rangle]$, and we can apply the methodology recursively to get $\mathbb{E}[\langle \boldsymbol{G}_{t-2}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^2}} \odot \beta_1^2\boldsymbol{m}_{t-2}\rangle]$, $\mathbb{E}[\langle \boldsymbol{G}_{t-3}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^3}} \odot \beta_1^3\boldsymbol{m}_{t-3}\rangle]$, and so on. All in all, the above methodology can be summarized as the following lemma.

Lemma 1. *Let all conditions in Theorem 1 hold. Denote $F_t^i \triangleq \mathbb{E}\langle \boldsymbol{G}_{t-i}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^i}} \odot \boldsymbol{m}_{t-i}\rangle$. Set $\boldsymbol{G}_0 \triangleq \boldsymbol{G}_1$ Then, $\forall t \ge 1$ and $i \in [0, t-1]$,*

$$\begin{aligned}F_t^i \ge& \beta_1 F_t^{i+1} + \frac{(1-\beta_1)}{2}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{G}_{t-i}\right\|^2\right] - \beta_1 L\mathbb{E}\left[\|\boldsymbol{w}_{t-i} - \boldsymbol{w}_{t-i-1}\|\left\|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i-1}\right\|\right] \\ &- \left(2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\sigma + L^2\frac{\eta^2(1-\beta_1)}{(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})\beta_2^i}\frac{i}{\sigma}d\right)\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i}}} \odot \boldsymbol{m}_{t-i}\right\|^2.\end{aligned}$$

The proof is deferred to Appendix C.1. We highlight here that despite the simple methodology above, the proof itself is highly non-trivial and technical. The core difficulty lies in handling the error introduced by approximating $\widetilde{\boldsymbol{\nu}}_t^i$ with $\widetilde{\boldsymbol{\nu}}_t^{i+1}$, where we need to bound the gap both between g_{t-i} and $\boldsymbol{G}_{t-i}$ and between $\boldsymbol{G}_{t-i}$ and $\boldsymbol{G}_{t-i+1}$.

Remark 1. *Our surrogate conditioners $\widetilde{\boldsymbol{\nu}}_t^i$ are novel. Previously, there are other surrogate conditioners in Défossez et al. [10], Zou et al. [27] which help to disentangle the stochasticity in $\boldsymbol{g}_t$ and $\boldsymbol{\nu}_t$. However, none of them can be applied in our setting because the bounded gradient assumption is required to use them, which is missed in our setting. Therefore, our surrogate conditioners may also shed light on the other analysis of Adam where no bounded gradient is assumed.*

Based on Lemma 1, we can estimate the "First-Order" term recursively. Combining the estimation of the "First-Order" term back to the descent lemma (Eq. (3)) and summing the descent lemma over t from 1 to T, we obtain

$$\sum_{t=1}^{T} \frac{(1-\beta_1)\eta}{2} \mathbb{E}\left[\left\| \frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \boldsymbol{G}_t \right\|^2\right] \leq f(\boldsymbol{w}_1) - \mathbb{E} f(\boldsymbol{w}_{T+1}) + \sum_{l=1}^{d} C_1 \left(\mathbb{E} \ln\left(\frac{\boldsymbol{\nu}_{T,l}}{\boldsymbol{\nu}_{0,l}}\right) - T \ln \beta_2 \right). \quad (4)$$

We then encounter the second challenge.

Challenge II: Convert Eq. (4) to a bound of gradient norm. Although we have bounded the sum of $\mathbb{E}[\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \boldsymbol{G}_t\|^2]$, we need to convert it into a bound of $\mathbb{E}[\|\boldsymbol{G}_t\|^2]$. In existing works [27, 10, 14] which assumes bounded gradient, such a conversion is straightforward because (their version of) $\widetilde{\boldsymbol{\nu}}_t^1$ is upper bounded. However, we do not assume bounded gradient and $\widetilde{\boldsymbol{\nu}}_t^1$ can be aribitrarily large, making $\mathbb{E}[\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \boldsymbol{G}_t\|^2]$ arbitrarily small than $\mathbb{E}[\|\boldsymbol{G}_t\|^2]$.

Solution to Challenge II. As this part involves coordinate-wise analysis, we define $\boldsymbol{g}_{t,l}$, $\boldsymbol{G}_{t,l}$, $\boldsymbol{\nu}_{t,l}$, and $\widetilde{\boldsymbol{\nu}}_{t,l}^1$ respectively as the l-th coordinate of $\boldsymbol{g}_t$, $\boldsymbol{G}_t$, $\boldsymbol{\nu}_t$, and $\widetilde{\boldsymbol{\nu}}_t^1$. To begin with, note that due to Cauchy's inequality and Hölder's inequality,

$$\left(\mathbb{E}\sum_{t=1}^{T} \|\boldsymbol{G}_t\|\right)^2 \leq \left(\sum_{t=1}^{T} \mathbb{E}\left[\left\| \frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}} \odot \boldsymbol{G}_t \right\|^2\right]\right) \left(\sum_{t=1}^{T} \mathbb{E}\left[\left\| \sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1} \right\|^2\right]\right). \quad (5)$$

Therefore, we only need to derive an upper bound of $\sum_{t=1}^{T} \mathbb{E}[\| \sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1} \|^2]$, which is achieved by the following divide-and-conque methodology. Firstly, when $|\boldsymbol{G}_{t,l}| \geq \sigma$, we can show $2\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2 \geq 2|\boldsymbol{G}_{t,l}|^2 \geq \mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2$. Then, by the concavity of $f(x) = \frac{x}{\sqrt{a+x}} (a > 0)$ and through a massive calculation, we obtain that

$$\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}} \mathbf{1}_{|\boldsymbol{G}_{t,l}| \geq \sigma}\right] \geq \frac{1}{3(1-\beta_2)} \mathbb{E}(\sqrt{\boldsymbol{\nu}_{t,l} + (1-\beta_2)\sigma^2} - \sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l} + (1-\beta_2)\sigma^2)}) \mathbf{1}_{|G_{t,l}| \geq \sigma},$$

and thus

$$\sum_{t=1}^{T} \mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right] \geq \sum_{t=1}^{T} \frac{1}{3(1-\beta_2)} \mathbb{E}(\sqrt{\boldsymbol{\nu}_{t,l} + (1-\beta_2)\sigma^2} - \sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l} + (1-\beta_2)\sigma^2)}) \mathbf{1}_{|G_{t,l}| \geq \sigma}.$$

Secondly, when $|\boldsymbol{G}_{t,l}| < \sigma$, define $\{\bar{\boldsymbol{\nu}}_{t,l}\}_{t=0}^{\infty}$ as $\bar{\boldsymbol{\nu}}_{0,l} = \boldsymbol{\nu}_{0,l}$, $\bar{\boldsymbol{\nu}}_{t,l} = \bar{\boldsymbol{\nu}}_{t-1,l} + |g_{t,l}|^2 \mathbf{1}_{|G_{t,l}| < \sigma}$. One can easily observe that $\bar{\boldsymbol{\nu}}_{t,l} \leq \boldsymbol{\nu}_{t,l}$, and thus

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left(\sqrt{\boldsymbol{\nu}_{t,l} + (1-\beta_2)\sigma^2} - \sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l} + (1-\beta_2)\sigma^2)}\right) \mathbf{1}_{|G_{t,l}| \geq \sigma} \\
\leq &\sum_{t=1}^{T} \mathbb{E}\left(\sqrt{\bar{\boldsymbol{\nu}}_{t,l} + (1-\beta_2)\sigma^2} - \sqrt{\beta_2(\bar{\boldsymbol{\nu}}_{t-1,l} + (1-\beta_2)\sigma^2)}\right) \\
= &\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{T,l} + (1-\beta_2)\sigma^2} + (1-\sqrt{\beta_2}) \sum_{t=1}^{T-1} \mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{t,l} + (1-\beta_2)\sigma^2} - \mathbb{E}\sqrt{\beta_2(\bar{\boldsymbol{\nu}}_{0,l} + (1-\beta_2)\sigma^2)}.
\end{aligned}$$

Putting the above two estimations together, we derive that

$$\sum_{t=1}^{T}\sum_{l=1}^{d}\mathbb{E}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}\le 3(1+\sqrt{\beta_2})\sum_{t=1}^{T}\sum_{l=1}^{d}\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right]+T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}.$$

The above methodology can be summarized as the following lemma.

Lemma 2. *Let all conditions in Theorem 1 hold. Then,*

$$\sum_{t=1}^{T}\sum_{l=1}^{d}\mathbb{E}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}\le 2T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}+4dC_1\ln dC_1+12C_2.$$

Based on Lemma 2, we can derive the estimation of $\sum_{t=1}^{T}\mathbb{E}[\|\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}\|^2]$ since $\widetilde{\boldsymbol{\nu}}_t^1$ is close to $\boldsymbol{\nu}_t$. The proof is then completed by combining the estimation of $\sum_{t=1}^{T}\mathbb{E}[\|\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}\|^2]$ and Eq. (5).

5 Gap-closing upper bound on the iteration complexity of Adam

In this section, based on a refined proof of Stage II of Theorem 1 (see Appendix C) under the specific case $\eta=\Theta(1/\sqrt{T})$ and $\beta_2=1-\Theta(1/T)$, we show that the logarithmic factor in Theorem 1 can be removed and the lower bound can be achieved. Specifically, we have the following theorem.

Theorem 2. *Let Assumption 1 and Assumption 2 hold. Then, select the hyperparameters of Adam as $\eta=\frac{a}{\sqrt{T}}$, $\beta_2=1-\frac{b}{T}$ and $\beta_1=c\beta_2$, where $a,b>0$ and $0\le c<1$ are independent of T. Then, let $\boldsymbol{w}_\tau$ be the output of Adam in Algorithm 1, and we have*

$$\mathbb{E}\|\nabla f(\boldsymbol{w}_\tau)\|\le\frac{1}{\sqrt[4]{T}}\sqrt{\frac{2}{\sqrt{b}}\left(D_1+2D_2\ln\left(\frac{2\sqrt{b}}{\sqrt{T}}D_1+\frac{4b}{T}D_2^2+\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+3b\sigma^2}\right)\right)}$$
$$\times\sqrt{\frac{2\sqrt{b}}{\sqrt{T}}D_1+\frac{4b}{T}D_2^2+\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+3b\sigma^2}}+\frac{1}{T}\left(D_1+2D_2\ln\left(\frac{2\sqrt{b}}{\sqrt{T}}D_1+\frac{4b}{T}D_2^2+\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+3b\sigma^2}\right)\right),$$

where

$$D_1\triangleq\frac{4\sqrt{b}}{a(1-c)}f(\boldsymbol{w}_1)+\sum_{l=1}^{d}\frac{2}{ab\sqrt{b}}\left(La^2+4\frac{a\sqrt{b}\sigma}{(1-c)^2}+2\frac{a^2c}{1-c}+2\frac{L^2ca^3d}{\sqrt{b}(1-c)^5\sigma}\right)\left(-\ln(\boldsymbol{\nu}_{0,l})+b\right),$$
$$D_2\triangleq d\frac{2}{ab\sqrt{b}}\left(La^2+4\frac{a\sqrt{b}\sigma}{(1-c)^2}+2\frac{a^2c}{1-c}+4\frac{L^2ca^3d}{\sqrt{b}(1-c)^5\sigma}\right).$$

As a result, let $\boldsymbol{A}$ be Adam in Algorithm 1, we have $\mathcal{C}_\varepsilon(\boldsymbol{A},\Delta,L,\sigma^2)=\mathcal{O}(\frac{1}{\varepsilon^4})$.

The proof of Theorem 2 is based on a refined solution of Challenge II in the proof of Theorem 1 under the specific hyperparameter settings, and we defer the concrete proof to Appendix D. Below we discuss on Theorem 2, comparing it with pratice, with Theorem 1 and existing convergence rate of Adam, and with the convergence rate of AdaGrad.

Alignment with the practical hyperparameter choice. The hyperparameter setting in Theorem 2 indicates that to achieve the lower bound of iteration complexity, we need to select small η and close-to-1 β_2, with less requirement over β_1. This agrees with the hyperparameter setting in deep learning libaries, for example, $\eta=10^{-3}$, $\beta_2=0.999$, and $\beta_1=0.9$ in PyTorch.

Comparison with Theorem 1 and existing works. To our best knowledge, Theorem 2 is the first to derive the iteration complexity $\mathcal{O}(\frac{1}{\varepsilon^4})$. Previously, the state-of-art iteration complexity is $\mathcal{O}(\frac{\log 1/\varepsilon}{\varepsilon^4})$ [10] where they additionally assume bounded gradient. Theorem 2 is also tight than Theorem 1 (while Theorem 1 holds for more general hyperparameter settings). As discussed in Section 4.1, if applying the hyperparameter setting in Theorem 2 (i.e., $\eta=\frac{a}{\sqrt{T}}$, $\beta_2=1-\frac{b}{T}$ and $\beta_1=c\beta_2$) to Theorem 1, we will obtain that $\mathbb{E}\|\nabla f(\boldsymbol{w}_\tau)\|\le\mathcal{O}(\text{poly}(\log T)/\sqrt[4]{T})$ and $\mathcal{C}_\varepsilon(\boldsymbol{A},\Delta,L,\sigma^2)=\mathcal{O}(\frac{\log 1/\varepsilon}{\varepsilon^4})$, which is worse than the upper bound in Theorem 2 and the lower bound in Proposition 1 by a logarithmic factor.

Comparison with AdaGrad. AdaGrad [12] is another popular adaptive optimizer. Under Assumptions 1 amd 2, the state-of-art iteration complexity of AdaGrad is $\mathcal{O}(\frac{\log 1/\varepsilon}{\varepsilon^4})$ [13], which is worse than Adam by a logarithmic factor. Here we show that such a gap may be not due to the limitation of analysis, and can be explained by analogizing AdaGrad to Adam without momentum as SGD with diminishing learning rate to SGD with constant learning rate. To start with, the update rule of AdaGrad is given as

$$\boldsymbol{\nu}_t = \boldsymbol{\nu}_{t-1} + \boldsymbol{g}_t^{\odot 2}, \boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{g}_t. \tag{6}$$

We first show that in Algorithm 1, if we allow the hyperparameters to be dynamical, i.e.,

$$\boldsymbol{\nu}_t = \beta_{2,t}\boldsymbol{\nu}_{t-1} + (1-\beta_{2,t})\boldsymbol{g}_t^{\odot 2}, \boldsymbol{m}_t = \beta_{1,t}\boldsymbol{m}_{t-1} + (1-\beta_{1,t})\boldsymbol{g}_t, \boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot \boldsymbol{m}_t, \tag{7}$$

then Adam is equivalent to AdaGrad by setting $\eta_t = \frac{\eta}{\sqrt{t}}$, $\beta_{1,t} = 0$, and $\beta_{2,t} = 1 - \frac{1}{t}$. Specifically, by setting $\boldsymbol{\mu}_t = t\boldsymbol{\nu}_t$ in Eq. (7), we have Eq. (7) is equivalent to with Eq. (6) (by replacing $\boldsymbol{\nu}_t$ by $\boldsymbol{\mu}_t$ in Eq. (6)). Comparing the above hyperparameter setting with that in Theorem 2, we see that the above hyperparameter setting can be obtained by changing T to t and setting $c = 0$ in Theorem 2. This is similar to the relationship between SGD with diminishing learning rate $\Theta(1/\sqrt{t})$ and SGD with diminishing learning rate $\Theta(1/\sqrt{T})$. Moreover, the iteration complexity of SGD with diminishing learning rate $\Theta(1/\sqrt{t})$ also has an additional logarithmic factor than SGD with constant learning rate, which may explain the gap between AdaGrad and Adam.

6 Limitations

Despite that our work provide the first result closing the upper bound and lower bound of the iteration complexity of Adam, there are several limitations listed as follows:

Dependence over the dimension d. The bounds in Theorem 1 and Theorem 2 is monotonously increasing with respect to d. This is undesired since the upper bound of iteration complexity of SGD is invariant with respect to d. Nevertheless, removing such an dependence over d is technically hard since we need to deal with every coordinate separately due to coodinate-wise learning rate, while the descent lemma does not hold for a single coordinate but combines all coordinates together. To our best knowledge, all existing works on the convergene of Adam also suffers from the same problem. We leave removing the dependence over d as an important future work.

No better result with momentum. It can be observed that in Theorem 1 and Theorem 2, the tightest bound is achieved when $\beta_1 = 0$ (i.e., no momentum is applied). This contradicts with the common wisdom that momentum helps to accelerate. Although the benefit of momentum is not very clear for simple optimizer SGD with momentum, we view this as a limitation of our work and defer proving the benefit of momentum in Adam as a future work.

7 Related works

Section 3 has provided a detailed discussion over existing convergence analysis of Adam. In this section, we briefly review other related works. Adam is proposed with a convergence analysis in online optimization [16]. The proof, however, is latter shown to be flawed in Reddi et al. [20] as it requires the adaptive learning rate of Adam to be non-increasing. This motivates a line of works modifying Adam to ensure convergence. The modifications include enforcing the adaptive learning rate to be non-increasing [20, 5], imposing upper bound and lower bound of the adaptive learning rate [18], and using different approach to estimate second-order momentum [26, 7]. Recently, Chen et al. [6] discover a new optimizer Lion through Symbolic Discovery, which uses sign operation to replace the adaptive learning rate in Adam, achieving comparable performance of Adam with less memory costs.

References

[1] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022.

[2] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

[6] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, Y. Liu, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023.

[7] M. Crawshaw, M. Liu, F. Orabona, W. Zhang, and Z. Zhuang. Robustness to unbounded smoothness of generalized signSGD. *arXiv preprint arXiv:2208.11195*, 2022.

[8] A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

[9] S. De, A. Mukherjee, and E. Ullah. Convergence guarantees for RMSProp and ADAM in non-convex optimization and an empirical comparison to Nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

[10] A. Défossez, L. Bottou, F. Bach, and N. Usunier. A simple convergence proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[12] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[13] M. Faw, I. Tziotis, C. Caramanis, A. Mokhtari, S. Shakkottai, and R. Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pages 313–355. PMLR, 2022.

[14] Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. A novel convergence analysis for algorithms of the Adam family. *arXiv preprint arXiv:2112.03459*, 2021.

[15] J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] H. Li, A. Jadbabaie, and A. Rakhlin. Convergence of Adam under relaxed assumptions. *arXiv preprint arXiv:2304.13972*, 2023.

[18] L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. volume 32, 2019.

[20] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[21] N. Shi, D. Li, M. Hong, and R. Sun. RMSprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021.

[22] B. Wang, Y. Zhang, H. Zhang, Q. Meng, Z.-M. Ma, T.-Y. Liu, and W. Chen. Provable adaptivity in Adam. *arXiv preprint arXiv:2208.09900*, 2022.

[23] X. Xie, P. Zhou, H. Li, Z. Lin, and S. Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022.

[24] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

[25] Y. Zhang, C. Chen, N. Shi, R. Sun, and Z.-Q. Luo. Adam can converge without any modification on update rules. *arXiv preprint arXiv:2208.09632*, 2022.

[26] Z. Zhou, Q. Zhang, G. Lu, H. Wang, W. Zhang, and Y. Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. *arXiv preprint arXiv:1810.00143*, 2018.

[27] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11127–11135, 2019.

A Related Works

B Auxilliary Lemmas

The following two lemmas are useful when bounding the second-order term.

Lemma 3. *Assume we have $0 < \beta_2 < 1$ and a sequence of real numbers $(a_n)_{n=1}^{\infty}$. Let $b_0 > 0$ and $b_n = \beta_2 b_{n-1} + (1-\beta_2)a_n^2$. Then, we have*

$$\sum_{n=1}^{T} \frac{a_n^2}{b_n} \le \frac{1}{1-\beta_2}\left(\ln\left(\frac{b_T}{b_0}\right) - T\ln\beta_2\right).$$

Lemma 4. *Assume we have $0 < \beta_1^2 < \beta_2 < 1$ and a sequence of real numbers $(a_n)_{n=1}^{\infty}$. Let $b_0 > 0$, $b_n = \beta_2 b_{n-1} + (1-\beta_2)a_n^2$, $c_0 = 0$, and $c_n = \beta_1 c_{n-1} + (1-\beta_1)a_n$. Then, we have*

$$\sum_{n=1}^{T} \frac{|c_n|^2}{b_n} \le \frac{(1-\beta_1)^2}{(1-\frac{\beta_1}{\sqrt{\beta_2}})^2(1-\beta_2)}\left(\ln\left(\frac{b_T}{b_0}\right) - T\ln\beta_2\right).$$

Proof. To begin with,

$$\frac{|c_n|}{\sqrt{b_n}} \le (1-\beta_1)\sum_{i=1}^{n} \frac{\beta_1^{n-i}|a_i|}{\sqrt{b_n}} \le (1-\beta_1)\sum_{i=1}^{n} \frac{\beta_1^{n-i}|a_i|}{\sqrt{b_n}} \le (1-\beta_1)\sum_{i=1}^{n}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{n-i}\frac{|a_i|}{\sqrt{b_i}}.$$

Applying Cauchy's inequality, we obtain

$$\begin{aligned}
&\frac{|c_n|^2}{b_n} \le (1-\beta_1)^2\left(\sum_{i=1}^{n}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{n-i}\frac{|a_i|}{\sqrt{b_i}}\right)^2 \\
\le&(1-\beta_1)^2\left(\sum_{i=1}^{n}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{n-i}\right)\left(\sum_{i=1}^{n}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{n-i}\frac{|a_i|^2}{b_i}\right) \le \frac{(1-\beta_1)^2}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\left(\sum_{i=1}^{n}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{n-i}\frac{|a_i|^2}{b_i}\right).
\end{aligned}$$

Summing the above inequality over n from 1 to T then leads to

$$\begin{aligned}
\sum_{n=1}^{T}\frac{|c_n|^2}{b_n} \le& \frac{(1-\beta_1)^2}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\sum_{n=1}^{T}\left(\sum_{i=1}^{n}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{n-i}\frac{|a_i|^2}{b_i}\right) = \frac{(1-\beta_1)^2}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\sum_{n=1}^{T}\frac{|a_n|^2}{b_n}\left(\sum_{i=0}^{T-n}\left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{i}\right) \\
\le& \frac{(1-\beta_1)^2}{(1-\frac{\beta_1}{\sqrt{\beta_2}})^2}\sum_{n=1}^{T}\frac{|a_n|^2}{b_n} \le \frac{(1-\beta_1)^2}{(1-\frac{\beta_1}{\sqrt{\beta_2}})^2(1-\beta_2)}\left(\ln\left(\frac{b_T}{b_0}\right) - T\ln\beta_2\right).
\end{aligned}$$

The proof is completed. □

The following lemma bound the update norm of Adam.

Lemma 5. *We have $\forall t \ge 1$, $|\boldsymbol{w}_{t+1,l} - \boldsymbol{w}_{t,l}| \le \eta\frac{1-\beta_1}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}$.*

Proof. We have that

$$\begin{aligned}
|\boldsymbol{w}_{t+1,l} - \boldsymbol{w}_{t,l}| = \eta\left|\frac{\boldsymbol{m}_{t,l}}{\sqrt{\boldsymbol{\nu}_{t,l}}}\right| \le& \eta\frac{\sum_{i=0}^{t-1}(1-\beta_1)\beta_1^i|\boldsymbol{g}_{t-i,l}|}{\sqrt{\sum_{i=0}^{t-1}(1-\beta_2)\beta_2^i|\boldsymbol{g}_{t-i,l}|^2 + \beta_2^t\boldsymbol{\nu}_{0,l}}} \\
\le& \eta\frac{1-\beta_1}{\sqrt{1-\beta_2}}\frac{\sqrt{\sum_{i=0}^{t-1}\beta_2^i|\boldsymbol{g}_{t-i,l}|^2}\sqrt{\sum_{i=0}^{t-1}\frac{\beta_1^{2i}}{\beta_2^i}}}{\sqrt{\sum_{i=0}^{t-1}\beta_2^i|\boldsymbol{g}_{t-i,l}|^2}} \le \eta\frac{1-\beta_1}{\sqrt{1-\beta_2}\sqrt{1-\frac{\beta_1^2}{\beta_2}}}.
\end{aligned}$$

Here the second inequality is due to Cauchy's inequality. The proof is completed. □

C Proof of Theorem 1

C.1 Proof of Lemma 1 and Lemma 2

Proof of Lemma 1. $\forall i \in [0, t-1]$, we have the following decomposition:

$$F_t^i = \underbrace{\mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i} \right\rangle\right]}_{(i)_t^i} + \underbrace{\mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i}, \left(\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i}}} - \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}}\right) \odot \boldsymbol{m}_{t-i} \right\rangle\right]}_{(ii)_t^i}.$$

As for $(i)_t^i$, according to the definition of $\boldsymbol{m}_{t-i}$, it can be lower bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i}, \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i} \right\rangle\right] = \mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i}, (1-\beta_1)\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{g}_{t-i} \right\rangle\right] + \mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i}, \beta_1\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i-1} \right\rangle\right] \\
=&\mathbb{E}\left[(1-\beta_1)\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{G}_{t-i}\right\|^2\right] + \mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i-1}, \beta_1\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i-1} \right\rangle\right] + \mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i} - \boldsymbol{G}_{t-i-1}, \beta_1\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i-1} \right\rangle\right] \\
\geq&\mathbb{E}\left[(1-\beta_1)\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{G}_{t-i}\right\|^2\right] + \mathbb{E}\left[\left\langle \boldsymbol{G}_{t-i-1}, \beta_1\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i-1} \right\rangle\right] - \beta_1 L\mathbb{E}\left[\|\boldsymbol{w}_{t-i} - \boldsymbol{w}_{t-i-1}\|\left\|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}} \odot \boldsymbol{m}_{t-i-1}\right\|\right],
\end{aligned}$$

where the last inequality is due to Assumption 1. As for $(ii)_t^i$, if $i = 0$, we have

$$\begin{aligned}
&\left|\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \boldsymbol{G}_t, \left(\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^0}} - \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}}\right) \odot \boldsymbol{m}_t \right\rangle\right]\right| \leq \sum_{l=1}^d |\boldsymbol{G}_{t,l}| \mathbb{E}^{|\mathcal{F}_t}\left[|\boldsymbol{m}_{t,l}|\left|\frac{1}{\sqrt{\boldsymbol{\nu}_{t,l}}} - \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right|\right] \\
=&\sum_{l=1}^d |\boldsymbol{G}_{t,l}| |\mathbb{E}^{|\mathcal{F}_t}\left[|\boldsymbol{m}_{t,l}|\frac{(1-\beta_2)\left||\boldsymbol{G}_{t,l}|^2 - |\boldsymbol{g}_{t,l}|^2\right| + (1-\beta_2)\sigma^2}{\sqrt{\boldsymbol{\nu}_{t,l}\widetilde{\boldsymbol{\nu}}_{t,l}^1}(\sqrt{\boldsymbol{\nu}_{t,l}} + \sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1})}\right] \\
\overset{(*)}{\leq}&\sum_{l=1}^d |\boldsymbol{G}_{t,l}| \mathbb{E}^{|\mathcal{F}_t}\left[||\boldsymbol{m}_{t,l}|\frac{(1-\beta_2)|\boldsymbol{G}_{t,l} - \boldsymbol{g}_{t,l}|(|\boldsymbol{G}_{t,l}| + |\boldsymbol{g}_{t,l}|) + (1-\beta_2)\sigma^2}{\sqrt{\boldsymbol{\nu}_{t,l}\widetilde{\boldsymbol{\nu}}_{t,l}^i}(\sqrt{\boldsymbol{\nu}_{t,l}} + \sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1})}\right] \\
\leq&\sum_{l=1}^d |\boldsymbol{G}_{t,l}| \mathbb{E}^{|\mathcal{F}_t}\left[|\boldsymbol{m}_{t,l}|\frac{\sqrt{1-\beta_2}|\boldsymbol{G}_{t,l} - \boldsymbol{g}_{t,l}| + \sqrt{1-\beta_2}\sigma}{\sqrt{\boldsymbol{\nu}_{t,l}\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right] \\
\overset{(\star)}{\leq}&\sum_{l=1}^d \frac{\sqrt{1-\beta_2}(1-\beta_1)|\boldsymbol{G}_{t,l}|^2}{4\sigma\widetilde{\boldsymbol{\nu}}_{t,l}^1}\left(\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{G}_{t,l} - \boldsymbol{g}_{t,l}|^2 + \sigma^2\right) + \sum_{l=1}^d 2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\mathbb{E}^{|\mathcal{F}_t}\sigma\frac{|\boldsymbol{m}_{t,l}|^2}{\boldsymbol{\nu}_{t,l}} \\
\leq&\sum_{l=1}^d \frac{(1-\beta_1)|\boldsymbol{G}_{t,l}|^2}{2\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}} + \sum_{l=1}^d 2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\mathbb{E}^{|\mathcal{F}_t}\sigma\frac{|\boldsymbol{m}_{t,l}|^2}{\boldsymbol{\nu}_{t,l}},
\end{aligned}$$

where inequality $(*)$ is due to the triangle inequality, and inequality $(\star)$ is due to the mean-value inequality, and the last inequality is due to Assumption 2.

If $i > 0$, then

$$
\begin{aligned}
&\left|\mathbb{E}^{|\mathcal{F}_{t-i}}\left[\left\langle \boldsymbol{G}_{t-i}, \left(\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^i}} - \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}}\right) \odot \boldsymbol{m}_{t-i}\right\rangle\right]\right| \\
\le& \sum_{l=1}^d |\boldsymbol{G}_{t-i,l}| \mathbb{E}^{|\mathcal{F}_{t-i}}\left[|\boldsymbol{m}_{t-i,l}|\left|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i}} - \frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}}\right|\right] \\
\le& \sum_{l=1}^d |\boldsymbol{G}_{t-i,l}| \mathbb{E}^{|\mathcal{F}_{t-i}}\left[|\boldsymbol{m}_{t-i,l}| \frac{(1-\beta_2)\beta_2^i \left||\boldsymbol{G}_{t-i,l}|^2 - |\boldsymbol{g}_{t-i,l}|^2\right| + \sum_{j=0}^{i-1}\beta_2^j(1-\beta_2)\left||\boldsymbol{G}_{t-i,l}|^2 - |\boldsymbol{G}_{t-i+1,l}|^2\right|}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i \widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}\left(\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i} + \sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}]\right)}\right] \\
\le& \sum_{l=1}^d |\boldsymbol{G}_{t-i,l}| \mathbb{E}^{|\mathcal{F}_{t-i}}\left[|\boldsymbol{m}_{t-i,l}| \frac{(1-\beta_2)\beta_2^i |\boldsymbol{G}_{t-i,l} - \boldsymbol{g}_{t-i,l}|(|\boldsymbol{G}_{t-i,l}| + |\boldsymbol{g}_{t-i,l}|)}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i \widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}\left(\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i} + \sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{1}}\right)}\right] \\
&+ \sum_{l=1}^d |\boldsymbol{G}_{t-i,l}| \mathbb{E}^{|\mathcal{F}_{t-i}}\left[|\boldsymbol{m}_{t-i,l}| \frac{\sum_{j=0}^{i-1}\beta_2^j(1-\beta_2)||\boldsymbol{G}_{t-i,l}| - |\boldsymbol{G}_{t-i+1,l}||(|\boldsymbol{G}_{t-i,l}| + |\boldsymbol{G}_{t-i+1,l}|)}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i \widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}\left(\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i} + \sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{1}}\right)}\right]
\end{aligned}
$$

Applying Cauchy's inequality, we obtain the RHS of the above inequality is smaller than

$$
\begin{aligned}
& \sum_{l=1}^d |\boldsymbol{G}_{t-i,l}| \mathbb{E}^{|\mathcal{F}_{t-i}}\left[|\boldsymbol{m}_{t-i,l}| \frac{\sqrt{1-\beta_2}\sqrt{\beta_2^i}|\boldsymbol{G}_{t-i,l} - \boldsymbol{g}_{t-i,l}|}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i \widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}}\right] \\
&+ \sum_{l=1}^d |\boldsymbol{G}_{t-i,l}| \mathbb{E}^{|\mathcal{F}_{t-i}}\left[|\boldsymbol{m}_{t-i,l}| \frac{\sqrt{\sum_{j=0}^{i-1}\beta_2^j(1-\beta_2)|\boldsymbol{G}_{t-i,l} - \boldsymbol{G}_{t-i+1,l}|^2}}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^i \widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}}\right] \\
\overset{(\star)}{\le}& \sum_{l=1}^d \frac{\sqrt{1-\beta_2}(1-\beta_1)|\boldsymbol{G}_{t-i,l}|^2}{4\sigma \widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}\left(\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{G}_{t,l} - \boldsymbol{g}_{t,l}|^2\right) + \sum_{l=1}^d \frac{\sqrt{1-\beta_2}}{(1-\beta_1)}\beta_2^i \mathbb{E}^{|\mathcal{F}_{t-i}}\sigma\frac{|\boldsymbol{m}_{t-i,l}|^2}{\widetilde{\boldsymbol{\nu}}_{t,l}^i} \\
&+ \sum_{l=1}^d \frac{\sqrt{1-\beta_2}\sigma(1-\beta_1)|\boldsymbol{G}_{t-i,l}|^2}{4\widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}} + \sum_{l=1}^d \frac{1}{(1-\beta_1)}\mathbb{E}^{|\mathcal{F}_{t-i}}\frac{1}{\sigma}\frac{|\boldsymbol{m}_{t-i,l}|^2}{\widetilde{\boldsymbol{\nu}}_{t,l}^i}\left(\sum_{j=0}^{i-1}\beta_2^j\sqrt{1-\beta_2}|\boldsymbol{G}_{t-i,l} - \boldsymbol{G}_{t-i+1,l}|^2\right) \\
\le& \sum_{l=1}^d \frac{(1-\beta_1)|\boldsymbol{G}_{t-i,l}|^2}{2\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}} + \sum_{l=1}^d \frac{\sqrt{1-\beta_2}}{1-\beta_1}\mathbb{E}^{|\mathcal{F}_{t-i}}\sigma\beta_2^i\frac{|\boldsymbol{m}_{t-i,l}|^2}{\boldsymbol{\nu}_{t,l}} + L^2\frac{\eta^2\sqrt{1-\beta_2}}{(1-\beta_1)\beta_2^i}\frac{i}{\sigma}\mathbb{E}^{|\mathcal{F}_{t-i}}\left(\sum_{l=1}^d\frac{|\boldsymbol{m}_{t-i,l}|^2}{\boldsymbol{\nu}_{t-i,l}}\right)^2 \\
\overset{(\circ)}{\le}& \sum_{l=1}^d \frac{(1-\beta_1)|\boldsymbol{G}_{t-i,l}|^2}{2\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{i+1}}} + \sum_{l=1}^d \frac{\sqrt{1-\beta_2}}{(1-\beta_1)}\mathbb{E}^{|\mathcal{F}_{t-i}}\sigma\frac{|\boldsymbol{m}_{t-i,l}|^2}{\boldsymbol{\nu}_{t-i,l}} + L^2\frac{\eta^2(1-\beta_1)}{(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})\beta_2^i}\frac{i}{\sigma}d\mathbb{E}^{|\mathcal{F}_{t-i}}\left(\sum_{l=1}^d\frac{|\boldsymbol{m}_{t-i,l}|^2}{\boldsymbol{\nu}_{t-i,l}}\right).
\end{aligned}
$$

Here inequality $(\star)$ is due to the mean-value inequality, and inequality $(\circ)$ is due to Lemma 5. Putting the estimation of $(i)_t^i$ and $(ii)_t^i$ together completes the proof. □

Proof of Lemma 2. To begin with, we have that

$$
\sum_{t=1}^T \mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|G_{t,l}|\ge\sigma}\right] \le \sum_{t=1}^T \mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right]. \tag{8}
$$

On the other hand, we have that

$$
\begin{aligned}
&\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\ge\sigma} \ge \frac{\frac{2}{3}|\boldsymbol{G}_{t,l}|^2 + \frac{1}{3}\sigma^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\ge\sigma} \ge \frac{\frac{1}{3}\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2 + \frac{1-\beta_2}{3}\sigma^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\ge\sigma} \\
\ge& \frac{\frac{1}{3}\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2 + \frac{1-\beta_2}{3}\sigma^2}{\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l} + (1-\beta_2)\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2 + (1-\beta_2)\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\ge\sigma} \\
\ge& \mathbb{E}^{|\mathcal{F}_t}\frac{\frac{1}{3}|\boldsymbol{g}_{t,l}|^2 + \frac{1-\beta_2}{3}\sigma^2}{\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l} + (1-\beta_2)|\boldsymbol{g}_{t,l}|^2 + (1-\beta_2)\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\ge\sigma}.
\end{aligned}
$$

Here the last inequality is due to the concavity of $\frac{x}{\sqrt{x+a}}$ with respect to x. As a conclusion,

$$\sum_{t=1}^{T}\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}\right]\geq\sum_{t=1}^{T}\mathbb{E}\left[\frac{\left(\frac{1}{3}|\boldsymbol{g}_{t,l}|^2+\frac{1-\beta_2}{3}\sigma^2\right)}{\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)|\boldsymbol{g}_{t,l}|^2+(1-\beta_2)\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}\right]$$
$$\geq\frac{1}{3(1-\beta_2)}\sum_{t=1}^{T}\mathbb{E}\left(\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2)}\right)\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}.$$

On the other hand, as stated in Section 4.2, we define $\{\bar{\boldsymbol{\nu}}_{t,l}\}_{t=0}^{\infty}$ as $\bar{\boldsymbol{\nu}}_{0,l}=\boldsymbol{\nu}_{0,l}$, $\bar{\boldsymbol{\nu}}_{t,l}=\bar{\boldsymbol{\nu}}_{t-1,l}+|g_{t,l}|^2\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}$. One can easily observe that $\bar{\boldsymbol{\nu}}_{t,l}\leq\boldsymbol{\nu}_{t,l}$, and thus

$$\sum_{t=1}^{T}\mathbb{E}\left(\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2)}\right)\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}$$
$$=\sum_{t=1}^{T}\mathbb{E}\left(\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)|g_{t,l}|^2+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2)}\right)\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}$$
$$\leq\sum_{t=1}^{T}\mathbb{E}\left(\sqrt{\beta_2\bar{\boldsymbol{\nu}}_{t-1,l}+(1-\beta_2)|g_{t,l}|^2+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\bar{\boldsymbol{\nu}}_{t-1,l}+(1-\beta_2)\sigma^2)}\right)\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}$$
$$\leq\sum_{t=1}^{T}\mathbb{E}(\sqrt{\beta_2\bar{\boldsymbol{\nu}}_{t-1,l}+(1-\beta_2)|g_{t,l}|^2\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\bar{\boldsymbol{\nu}}_{t-1,l}+(1-\beta_2)\sigma^2)})$$
$$=\sum_{t=1}^{T}\mathbb{E}(\sqrt{\bar{\boldsymbol{\nu}}_{t,l}+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\bar{\boldsymbol{\nu}}_{t-1,l}+(1-\beta_2)\sigma^2)})$$
$$=\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{T,l}+(1-\beta_2)\sigma^2}+(1-\sqrt{\beta_2})\sum_{t=1}^{T-1}\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{t,l}+(1-\beta_2)\sigma^2}-\mathbb{E}\sqrt{\beta_2(\bar{\boldsymbol{\nu}}_{0,l}+(1-\beta_2)\sigma^2)}.$$

All in all, summing the above two inequalities together, we obtain that

$$\mathbb{E}\sqrt{\boldsymbol{\nu}_{T,l}+(1-\beta_2)\sigma^2}+(1-\sqrt{\beta_2})\sum_{t=1}^{T-1}\mathbb{E}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}-\mathbb{E}\sqrt{\beta_2(\boldsymbol{\nu}_{0,l}+(1-\beta_2)\sigma^2)}$$
$$=\sum_{t=1}^{T}\mathbb{E}(\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2)})$$
$$\leq\sum_{t=1}^{T}\mathbb{E}(\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2)})\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}$$
$$+\sum_{t=1}^{T}\mathbb{E}(\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}-\sqrt{\beta_2(\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2)})\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}$$
$$=3(1-\beta_2)\sum_{t=1}^{T}\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right]+\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{T,l}+(1-\beta_2)\sigma^2}+(1-\sqrt{\beta_2})\sum_{t=1}^{T-1}\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{t,l}+(1-\beta_2)\sigma^2}-\mathbb{E}\sqrt{\beta_2(\bar{\boldsymbol{\nu}}_{0,l}+(1-\beta_2)\sigma^2)}.$$

As $\mathbb{E}\sqrt{\boldsymbol{\nu}_{T,l}+(1-\beta_2)\sigma^2}\geq\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{T,l}+(1-\beta_2)\sigma^2}$ and $\mathbb{E}\sqrt{\boldsymbol{\nu}_{0,l}+(1-\beta_2)\sigma^2}=\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{0,l}+(1-\beta_2)\sigma^2}$, we obtain that

$$\begin{aligned}(1-\sqrt{\beta_2})\sum_{t=1}^{T}\mathbb{E}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}\leq&3(1-\beta_2)\sum_{t=1}^{T}\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right]+(1-\sqrt{\beta_2})\sum_{t=1}^{T}\mathbb{E}\sqrt{\bar{\boldsymbol{\nu}}_{t,l}+(1-\beta_2)\sigma^2}\\ \leq&3(1-\beta_2)\sum_{t=1}^{T}\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right]+(1-\sqrt{\beta_2})\sum_{t=1}^{T}\sqrt{\mathbb{E}\bar{\boldsymbol{\nu}}_{t,l}+(1-\beta_2)\sigma^2}\\ \leq&3(1-\beta_2)\sum_{t=1}^{T}\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\right]+(1-\sqrt{\beta_2})\sum_{t=1}^{T}\sqrt{\bar{\boldsymbol{\nu}}_{0,l}+(3-\beta_2)\sigma^2}.\end{aligned}\tag{9}$$

Leveraging Eq. (4), we then obtain that

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{l=1}^{d}\mathbb{E}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}\\
\le& 3(1+\sqrt{\beta_2})\sum_{t=1}^{T}\mathbb{E}\left[\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{1}}}\right]+\sum_{t=1}^{T}\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}\\
\le& \frac{12}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1)+2\sum_{l=1}^{d}C_1\left(\mathbb{E}\ln\left(\frac{\sqrt{\boldsymbol{\nu}_{T,l}+(1-\beta_2)\sigma^2}}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)\right)+T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}\\
\le& \frac{12}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1)+2\sum_{l=1}^{d}C_1\left(\mathbb{E}\ln\left(\frac{\sum_{t=1}^{T}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)\right)+T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}\\
\le& \frac{12}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1)+2\sum_{l=1}^{d}C_1\left(\ln\left(\frac{\mathbb{E}\sum_{t=1}^{T}\sum_{m=1}^{d}\sqrt{\boldsymbol{\nu}_{t,m}+(1-\beta_2)\sigma^2}}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)\right)+T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2},
\end{aligned}
$$

where in the last inequality we use the concavity of $h(x)=\ln x$. Solving the above inequality with respect to $\sum_{t=1}^{T}\sum_{l=1}^{d}\mathbb{E}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}$ then gives

$$
\begin{aligned}
\sum_{t=1}^{T}\sum_{l=1}^{d}\mathbb{E}\sqrt{\boldsymbol{\nu}_{t,l}+(1-\beta_2)\sigma^2}\le& 2T\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}+4dC_1\ln dC_1\\
&+\frac{24}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1)+2\sum_{l=1}^{d}C_1\left(\ln\left(\frac{1}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)\right).
\end{aligned}
$$

The proof is then completed.

□

C.2 Proof of Theorem 1

Proof of Theorem 1. As stated in Section 4.2, the proof involves solving two key challenges. We respectively divide the proof into two stages according to the challenges.

Stage I. Based on Lemma 1, we can estimate $\mathbb{E}\langle\boldsymbol{G}_t,\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t}}\odot\boldsymbol{m}_t\rangle=F_t^0$ recursively. Specifically, we have

$$
\begin{aligned}
F_t^0\ge&\sum_{i=0}^{t-1}\beta_1^i\left(\frac{(1-\beta_1)}{2}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^{i+1}}}\odot\boldsymbol{G}_{t-i}\right\|^2\right]-\beta_1\mathbb{E}\left[\|\boldsymbol{w}_{t-i}-\boldsymbol{w}_{t-i-1}\|\left\|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}}\odot\boldsymbol{m}_{t-i-1}\right\|\right]\right.\\
&\left.-\left(2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\sigma+L^2\frac{\eta^2(1-\beta_1)}{(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})\beta_2^i}\frac{i}{\sigma}d\right)\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i}}}\odot\boldsymbol{m}_{t-i}\right\|^2\right)\\
\ge&\frac{(1-\beta_1)}{2}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right]-\sum_{i=0}^{t-1}\beta_1^i\left(\beta_1\mathbb{E}\left[\|\boldsymbol{w}_{t-i}-\boldsymbol{w}_{t-i-1}\|\left\|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}}\odot\boldsymbol{m}_{t-i-1}\right\|\right]\right.\\
&\left.+\left(2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\sigma+L^2\frac{\eta^2(1-\beta_1)}{(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})\beta_2^i}\frac{i}{\sigma}d\right)\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i}}}\odot\boldsymbol{m}_{t-i}\right\|^2\right)
\end{aligned}
$$

Applying the above inequality back to Eq. (3) then gives

$$
\begin{aligned}
&\mathbb{E}f(\boldsymbol{w}_{t+1})\\
\leq&\mathbb{E}f(\boldsymbol{w}_t)-\frac{(1-\beta_1)\eta}{2}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right]+\frac{L}{2}\eta^2\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\odot\boldsymbol{m}_t\right\|^2+\eta\sum_{i=0}^{t-1}\beta_1^i\left(\beta_1\mathbb{E}\left[\|\boldsymbol{w}_{t-i}-\boldsymbol{w}_{t-i-1}\|\right.\right.\\
&\times\left.\left\|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}}\odot\boldsymbol{m}_{t-i-1}\right\|\right]+\left(2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\sigma+L^2\frac{\eta^2(1-\beta_1)}{(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})\beta_2^i}\frac{i}{\sigma}d\right)\mathbb{E}\left.\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i}}}\odot\boldsymbol{m}_{t-i}\right\|^2\right).
\end{aligned}
$$

Summing the above inequality with respect to t then gives

$$
\begin{aligned}
&\mathbb{E}f(\boldsymbol{w}_{T+1})\\
\leq&f(\boldsymbol{w}_1)-\sum_{t=1}^{T}\frac{(1-\beta_1)\eta}{2}\mathbb{E}\left[\left\|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right]+\left(\frac{L}{2}\eta^2+2\frac{\sqrt{1-\beta_2}}{(1-\beta_1)^2}\eta\sigma+\frac{\eta^2\beta_1}{\sqrt{\beta_2}(1-\frac{\beta_1}{\sqrt{\beta_2}})}\right.\\
&+\left.L^2\frac{\beta_1\eta^3(1-\beta_1)}{\beta_2(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})(1-\frac{\beta_1}{\beta_2})^2}\frac{d}{\sigma}\right)\sum_{t=1}^{T}\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\odot\boldsymbol{m}_t\right\|^2.
\end{aligned}
$$

Here the inequality is due to

$$
\begin{aligned}
2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\eta\sum_{t=1}^{T}\sum_{i=0}^{t-1}\beta_1^i\mathbb{E}\sigma\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i}}}\odot\boldsymbol{m}_{t-i}\right\|^2=&2\frac{\sqrt{1-\beta_2}}{1-\beta_1}\eta\sigma\sum_{i=1}^{T}\sum_{t=i}^{T}\beta_1^{t-i}\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\odot\boldsymbol{m}_t\right\|^2\\
\leq&2\frac{\sqrt{1-\beta_2}}{(1-\beta_1)^2}\eta\sigma\sum_{i=1}^{T}\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_i}}\odot\boldsymbol{m}_i\right\|^2,
\end{aligned}
$$

$$
\begin{aligned}
&\eta\sum_{t=1}^{T}\sum_{i=0}^{t-1}\beta_1^{i+1}\mathbb{E}\left[\|\boldsymbol{w}_{t-i}-\boldsymbol{w}_{t-i-1}\|\left\|\frac{1}{\sqrt{\widetilde{\boldsymbol{\nu}}_t^{i+1}}}\odot\boldsymbol{m}_{t-i-1}\right\|\right]\\
\leq&\eta\sum_{t=1}^{T}\sum_{i=0}^{t-1}\frac{\beta_1^{i+1}}{\sqrt{\beta_2^{i+1}}}\mathbb{E}\left[\|\boldsymbol{w}_{t-i}-\boldsymbol{w}_{t-i-1}\|\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i-1}}}\odot\boldsymbol{m}_{t-i-1}\right\|\right]\\
=&\eta^2\sum_{t=1}^{T}\sum_{i=0}^{t-1}\frac{\beta_1^{i+1}}{\sqrt{\beta_2^{i+1}}}\mathbb{E}\left[\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i-1}}}\odot\boldsymbol{m}_{t-i-1}\right\|^2\right]=\eta^2\sum_{i=0}^{T-1}\sum_{t=i+1}^{T}\frac{\beta_1^{t-i}}{\sqrt{\beta_2^{t-i}}}\mathbb{E}\left[\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_i}}\odot\boldsymbol{m}_i\right\|^2\right]\\
\leq&\frac{\eta^2\beta_1}{\sqrt{\beta_2}(1-\frac{\beta_1}{\sqrt{\beta_2}})}\sum_{i=0}^{T-1}\mathbb{E}\left[\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_i}}\odot\boldsymbol{m}_i\right\|^2\right]=\frac{\eta^2\beta_1}{\sqrt{\beta_2}(1-\frac{\beta_1}{\sqrt{\beta_2}})}\sum_{i=1}^{T-1}\mathbb{E}\left[\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_i}}\odot\boldsymbol{m}_i\right\|^2\right],
\end{aligned}
$$

and

$$
\begin{aligned}
&L^2\frac{\eta^3(1-\beta_1)}{(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})}\frac{d}{\sigma}\sum_{t=1}^{T}\sum_{i=0}^{t-1}\frac{\beta_1^i}{\beta_2^i}i\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_{t-i}}}\odot\boldsymbol{m}_{t-i}\right\|^2\\
=&L^2\frac{\eta^3(1-\beta_1)}{(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})}\frac{d}{\sigma}\sum_{i=1}^{T}\sum_{t=i}^{T}\frac{\beta_1^{t-i}}{\beta_2^{t-i}}(t-i)\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_i}}\odot\boldsymbol{m}_i\right\|^2\leq L^2\frac{\beta_1\eta^3(1-\beta_1)}{\beta_2(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})(1-\frac{\beta_1}{\beta_2})^2}\frac{d}{\sigma}\sum_{i=1}^{T}\mathbb{E}\left\|\frac{1}{\sqrt{\boldsymbol{\nu}_i}}\odot\boldsymbol{m}_i\right\|^2.
\end{aligned}
$$

Applying Lemma 4, we obtain that

$$
\begin{aligned}
&\mathbb{E}f(\boldsymbol{w}_{T+1})\\
\leq&f(\boldsymbol{w}_1)+\sum_{l=1}^{d}\left(\frac{L}{2}\eta^2+2\frac{\sqrt{1-\beta_2}}{(1-\beta_1)^2}\eta\sigma+\frac{\eta^2\beta_1}{\sqrt{\beta_2}(1-\frac{\beta_1}{\sqrt{\beta_2}})}+L^2\frac{\beta_1\eta^3(1-\beta_1)}{\beta_2(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})(1-\frac{\beta_1}{\beta_2})^2}\frac{d}{\sigma}\frac{(1-\beta_1)^2}{(1-\frac{\beta_1}{\sqrt{\beta_2}})^2}\right)\frac{1}{1-\beta_2}\\
&\times\left(\mathbb{E}\ln\left(\frac{\boldsymbol{\nu}_{T,l}}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)-\sum_{t=1}^{T}\frac{(1-\beta_1)\eta}{2}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right].
\end{aligned}
$$

The proof of Stage I is completed.

Stage II. According to Cauchy's inequality, we have

$$\left(\mathbb{E}\sum_{t=1}^{T}\|\boldsymbol{G}_t\|_1\right)^2 \leq \left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right]\right)\left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}\right\|^2\right]\right). \tag{10}$$

Meanwhile, by Lemma 2, we have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\left\|\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}\right\|^2\right] =&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l=1}^{d}\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)|\boldsymbol{G}_{t,l}|^2+(1-\beta_2)\sigma^2}\right]\\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l=1}^{d}\left(\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2}+\sqrt{1-\beta_2}|\boldsymbol{G}_{t,l}|\right)\right]\\
=&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{l=1}^{d}\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2}+\sum_{t=1}^{T}\sqrt{1-\beta_2}\|\boldsymbol{G}_t\|_1\right]\\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{1-\beta_2}\|\boldsymbol{G}_t\|_1\right]+2T\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}+4dC_1\ln dC_1\\
&+\frac{24}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1)+2\sum_{l=1}^{d}C_1\left(\ln\left(\frac{1}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)\right).
\end{aligned}$$

Combining the above inequality and Eq. (10) gives

$$\begin{aligned}
\left(\mathbb{E}\sum_{t=1}^{T}\|\boldsymbol{G}_t\|_1\right)^2 \leq& \frac{2}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1)+\sum_{l=1}^{d}C_1\left(\mathbb{E}\ln\left(\frac{\boldsymbol{\nu}_{T,l}}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)\right)\\
&\times\left(\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{1-\beta_2}\|\boldsymbol{G}_t\|_1\right]+2T\sqrt{\boldsymbol{\nu}_{0,l}+(3-\beta_2)\sigma^2}+4dC_1\ln dC_1\right.\\
&\left.+\frac{24}{(1-\beta_1)\eta}\left(f(\boldsymbol{w}_1)+2\sum_{l=1}^{d}C_1\left(\ln\left(\frac{1}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)\right)\right).
\end{aligned}$$

Solving the above quadratic inequality with respect to $\mathbb{E}\sum_{t=1}^{T}\|\boldsymbol{G}_t\|_1$ then completes the proof.

$\square$

D Proof of Theorem 2

Proof. According to Stage I in the proof of Theorem 1, we obtain

$$\begin{aligned}
&\mathbb{E}f(\boldsymbol{w}_{T+1})\\
\leq&f(\boldsymbol{w}_1)+\sum_{l=1}^{d}\left(\frac{L}{2}\eta^2+2\frac{\sqrt{1-\beta_2}}{(1-\beta_1)^2}\eta\sigma+\frac{\eta^2\beta_1}{\sqrt{\beta_2}(1-\frac{\beta_1}{\sqrt{\beta_2}})}+L^2\frac{\beta_1\eta^3(1-\beta_1)}{\beta_2(1-\beta_2)^{\frac{1}{2}}(1-\frac{\beta_1^2}{\beta_2})(1-\frac{\beta_1}{\beta_2})^2}\frac{d}{\sigma}\frac{(1-\beta_1)^2}{(1-\frac{\beta_1}{\sqrt{\beta_2}})^2}\right)\frac{1}{1-\beta_2}\\
&\times\mathbb{E}\left(\ln\left(\frac{\boldsymbol{\nu}_{T,l}}{\boldsymbol{\nu}_{0,l}}\right)-T\ln\beta_2\right)-\sum_{t=1}^{T}\frac{(1-\beta_1)\eta}{2}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right].
\end{aligned}$$

Applying the definition of η, β_1, and β_2, we obtain that

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right]\leq\frac{2\sqrt{T}}{\sqrt{b}}\left(D_1+\frac{D_2}{d}\sum_{l=1}^{d}\mathbb{E}\ln\boldsymbol{\nu}_{T,l}\right). \tag{11}$$

Meanshile, we have that

$$
\begin{aligned}
&\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma} \geq \frac{\frac{1}{2}\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}\\
=&\frac{\frac{1}{2}\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2}{\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\mathbb{E}^{|\mathcal{F}_t}|\boldsymbol{g}_{t,l}|^2+(1-\beta_2)\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}\\
\geq&\frac{1}{2}\mathbb{E}^{|\mathcal{F}_t}\frac{|\boldsymbol{g}_{t,l}|^2}{\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)|\boldsymbol{g}_{t,l}|^2+(1-\beta_2)\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}\\
\geq&\frac{1}{2\sqrt{1-\beta_2}}\mathbb{E}^{|\mathcal{F}_t}\frac{|\boldsymbol{g}_{t,l}|^2}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma},
\end{aligned}
$$

where the last inequality is due to that

$$
\begin{aligned}
&\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)|\boldsymbol{g}_{t,l}|^2=(1-\beta_2)\sum_{s=1}^t\beta_2^{t-s}|\boldsymbol{g}_{s,l}|^2+\beta_2^t\boldsymbol{\nu}_{0,l}\\
\leq&(1-\beta_2)\sum_{s=1}^T|\boldsymbol{g}_{s,l}|^2+\boldsymbol{\nu}_{0,l}. \qquad (12)
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
&\frac{\sigma^2+\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}}+\sum_{t=1}^T\mathbb{E}\frac{|\boldsymbol{g}_{t,l}|^2}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}\\
\leq&\frac{\sigma^2+\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}}+\sum_{t=1}^T\mathbb{E}\frac{|\boldsymbol{g}_{t,l}|^2}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2\mathbf{1}_{|\boldsymbol{G}_{s,l}|<\sigma}+\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}\\
=&\mathbb{E}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2\mathbf{1}_{|\boldsymbol{G}_{s,l}|<\sigma}+\sigma^2}\leq\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\mathbb{E}\sum_{s=1}^T|g_{s,l}|^2\mathbf{1}_{|\boldsymbol{G}_{s,l}|<\sigma}+\sigma^2}\\
\leq&\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+2\sigma^2T+\sigma^2}.
\end{aligned}
$$

Conclusively, we obtain

$$
\begin{aligned}
&\mathbb{E}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}\\
=&\frac{\sigma^2+\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}}+\sum_{t=1}^T\mathbb{E}\frac{|\boldsymbol{g}_{t,l}|^2}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|<\sigma}\\
&+\sum_{t=1}^T\mathbb{E}\frac{|\boldsymbol{g}_{t,l}|^2}{\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^T|g_{s,l}|^2+\sigma^2}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}\\
\leq&\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+2\sigma^2T+\sigma^2}+2\sqrt{1-\beta_2}\sum_{t=1}^T\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\geq\sigma}.
\end{aligned}
$$

Summing the above inequality with respect to l then gives

$$
\begin{aligned}
&\sum_{l=1}^{d}\mathbb{E}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^{T}|g_{s,l}|^2+\sigma^2}\\
\le&\sum_{l=1}^{d}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+2\sigma^2T+\sigma^2}+2\sqrt{1-\beta_2}\sum_{l=1}^{d}\sum_{t=1}^{T}\frac{|\boldsymbol{G}_{t,l}|^2}{\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}}\mathbf{1}_{|\boldsymbol{G}_{t,l}|\ge\sigma}\\
\le&\sum_{l=1}^{d}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+2\sigma^2T+\sigma^2}\\
&+\frac{4\sqrt{b}}{a(1-c)}f(\boldsymbol{w}_1)+\sum_{l=1}^{d}\frac{2}{ab\sqrt{b}}\left(La^2+4\frac{a\sqrt{b}\sigma}{(1-c)^2}+2\frac{a^2c}{1-c}+2\frac{L^2ca^3d}{\sqrt{b}(1-c)^5\sigma}\right)\left(\mathbb{E}\ln\left(\frac{\boldsymbol{\nu}_{T,l}}{\boldsymbol{\nu}_{0,l}}\right)+b\right)\\
=&\sum_{l=1}^{d}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+2\sigma^2T+\sigma^2}+\sum_{l=1}^{d}\frac{2}{ab\sqrt{b}}\left(La^2+4\frac{a\sqrt{b}\sigma}{(1-c)^2}+2\frac{a^2c}{1-c}+4\frac{L^2ca^3d}{\sqrt{b}(1-c)^5\sigma}\right)\mathbb{E}\ln\left(\sqrt{\boldsymbol{\nu}_{T,l}}\right)\\
&+\frac{4\sqrt{b}}{a(1-c)}f(\boldsymbol{w}_1)+\sum_{l=1}^{d}\frac{2}{ab\sqrt{b}}\left(La^2+4\frac{a\sqrt{b}\sigma}{(1-c)^2}+2\frac{a^2c}{1-c}+2\frac{L^2ca^3d}{\sqrt{b}(1-c)^5\sigma}\right)(-\ln(\boldsymbol{\nu}_{0,l})+b)\\
\le&\sum_{l=1}^{d}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+2\sigma^2T+\sigma^2}\\
&+d\frac{2}{ab\sqrt{b}}\left(La^2+4\frac{a\sqrt{b}\sigma}{(1-c)^2}+2\frac{a^2c}{1-c}+4\frac{L^2ca^3d}{\sqrt{b}(1-c)^5\sigma}\right)\mathbb{E}\ln\left(\sum_{l=1}^{d}\sqrt{1-\beta_2}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^{T}|g_{s,l}|^2+\sigma^2}\right)\\
&+\frac{4\sqrt{b}}{a(1-c)}f(\boldsymbol{w}_1)+\sum_{l=1}^{d}\frac{2}{ab\sqrt{b}}\left(La^2+4\frac{a\sqrt{b}\sigma}{(1-c)^2}+2\frac{a^2c}{1-c}+2\frac{L^2ca^3d}{\sqrt{b}(1-c)^5\sigma}\right)(-\ln(\boldsymbol{\nu}_{0,l})+b)\\
\le&\sum_{l=1}^{d}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+3\sigma^2T}+D_1+D_2\ln\left(\mathbb{E}\sum_{l=1}^{d}\sqrt{1-\beta_2}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^{T}|g_{s,l}|^2+\sigma^2}\right),
\end{aligned}
$$

where the second inequality is due to Eq. (11), the second-to-last inequality is due to Eq. (12), and the last inequality is due to Jensen's inequality. Solving the above ineqaulity with respect to $\sqrt{1-\beta_2}\sum_{l=1}^{d}\mathbb{E}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^{T}|g_{s,l}|^2+\sigma^2}$ then gives

$$
\begin{aligned}
\sqrt{1-\beta_2}\sum_{l=1}^{d}\mathbb{E}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^{T}|g_{s,l}|^2+\sigma^2}\le&2\sqrt{1-\beta_2}D_1+4\sqrt{1-\beta_2}D_2\ln(1+\sqrt{1-\beta_2}D_2)\\
&+\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+3b\sigma^2}.
\end{aligned}
$$

Therefore, by Cauchy's inequality, we have

$$
\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{G}_t\|_1\right]^2\le\left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{\sqrt[4]{\widetilde{\boldsymbol{\nu}}_t^1}}\odot\boldsymbol{G}_t\right\|^2\right]\right)\left(\sum_{t=1}^{T}\sum_{l=1}^{d}\mathbb{E}\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^1}\right).
$$

Since

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{l=1}^{d}\sqrt{\widetilde{\boldsymbol{\nu}}_{t,l}^{1}} \le \sum_{t=1}^{T}\sum_{l=1}^{d}\left(\sqrt{\beta_2\boldsymbol{\nu}_{t-1,l}+(1-\beta_2)\sigma^2}+\sqrt{(1-\beta_2)}|\boldsymbol{G}_{t,l}|\right)\\
\le& T\sum_{l=1}^{d}\sqrt{1-\beta_2}\sqrt{\frac{\boldsymbol{\nu}_{0,l}}{1-\beta_2}+\sum_{s=1}^{T}|g_{s,l}|^2+\sigma^2}+\sum_{t=1}^{T}\sum_{l=1}^{d}\sqrt{(1-\beta_2)}|\boldsymbol{G}_{t,l}|\\
\le& T\left(2\sqrt{1-\beta_2}D_1+4\sqrt{1-\beta_2}D_2\ln(1+\sqrt{1-\beta_2}D_2)+\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+3b\sigma^2}\right)+\sum_{t=1}^{T}\sqrt{(1-\beta_2)}\|\boldsymbol{G}_t\|_1,
\end{aligned}
$$

we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{G}_t\|_1\right]^2\\
\le&\left(T\left(2\sqrt{1-\beta_2}D_1+4\sqrt{1-\beta_2}D_2\ln(1+\sqrt{1-\beta_2}D_2)+\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{0,l}+3b\sigma^2}\right)+\sum_{t=1}^{T}\sqrt{(1-\beta_2)}\mathbb{E}\|\boldsymbol{G}_t\|_1\right)\\
&\times\frac{2\sqrt{T}}{\sqrt{b}}\left(D_1+\frac{D_2}{d}\sum_{l=1}^{d}\mathbb{E}\ln\boldsymbol{\nu}_{T,l}\right).
\end{aligned}
$$

Solving the above inequality with respect to $\sum_{t=1}^{T}\mathbb{E}\left\|\boldsymbol{G}_t\right\|_1$ completes the proof. $\square$