Convergence of AdaGrad for Non-convex Objectives: Simple Proofs and Relaxed Assumptions

Bohan Wang BHWANGFY@GMAIL.COM
University of Science and Technology of China

Huishuai Zhang * HUZHANG@MICROSOFT.COM
Microsoft Research Asia

Zhi-Ming Ma MAZM@AMT.AC.CN
Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Wei Chen * CHENWEI2022@ICT.AC.CN
Institute of Computing Technology, Chinese Academy of Sciences

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We provide a simple convergence proof for AdaGrad optimizing non-convex objectives under only affine noise variance and bounded smoothness assumptions. The proof is essentially based on a novel auxiliary function ξ that helps eliminate the complexity of handling the correlation between the numerator and denominator of AdaGrad's update. Leveraging simple proofs, we are able to obtain tighter results than existing results (Faw et al., 2022) and extend the analysis to several new and important cases. Specifically, for the over-parameterized regime, we show that AdaGrad needs only $\mathcal{O}(\frac{1}{\varepsilon^2})$ iterations to ensure the gradient norm smaller than ε, which matches the rate of SGD and significantly tighter than existing rates $\mathcal{O}(\frac{1}{\varepsilon^4})$ for AdaGrad. We then discard the bounded smoothness assumption, and consider a more realistic assumption on smoothness called (L_0, L_1)-smooth condition, which allows local smoothness to grow with the gradient norm. Again based on the auxiliary function ξ, we prove that AdaGrad succeeds in converging under (L_0, L_1)-smooth condition as long as the learning rate is lower than a threshold. Interestingly, we further show that the requirement on learning rate under the (L_0, L_1)-smooth condition is necessary via proof by contradiction, in contrast with the case of uniform smoothness conditions where convergence is guaranteed regardless of learning rate choices. Together, our analyses broaden the understanding of AdaGrad, and demonstrate the power of the new auxiliary function in the investigations of AdaGrad.

Keywords: AdaGrad, Convergence Analysis

1. Introduction

Adaptive optimizers have been a great success in deep learning. Compared to stochastic gradient descent (SGD), adaptive optimizers use the gradient information of iterations to dynamically adjust the learning rate, which is observed to converge much faster than SGD in a wide range of deep learning tasks (Vaswani et al., 2017; Dosovitskiy et al., 2020; Yun et al., 2019). Such a superiority has attracted numerous researchers to analyze the behavior of adaptive optimizers.

AdaGrad (Duchi et al., 2011) is among the earliest adaptive optimizers and enjoys favorable convergence rate for online convex optimization. Specifically, the design of AdaGrad is quite simple: it tracks the gradient magnitudes of the past iterations and use its reciprocal to scale the current

* Corresponding Author

gradient. The pseudo-codes of the norm version of AdaGrad (i.e., AdaGrad-Norm) and AdaGrad are presented in Algorithm 1 and Algorithm 2, respectively.

Despite the popularity and the simplicity of AdaGrad, its theoretical analysis is not satisfactory when optimizing non-convex objectives. Specifically, until recently, Ward et al. (2020) analyze the convergence of AdaGrad-Norm and achieve $\mathcal{O}(\log T/\sqrt{T})$ rate. However, Their result is based on the assumption that the stochastic gradient g_t is uniformly bounded across the iterations, which does not hold even for quadratic functions, let alone deep neural networks. In comparison, the analysis of SGD does not require such an assumption.

Algorithm 1 AdaGrad-Norm

Input: Objective function $f(\boldsymbol{w})$, learning rate $\eta > 0$, initial point $\boldsymbol{w}_1 \in \mathbb{R}^d$, initial conditioner $\boldsymbol{\nu}_1 \in \mathbb{R}^+$

For $t = 1 \to \infty$:
- Generate stochastic gradient g_t
- Calculate $\boldsymbol{\nu}_t = \boldsymbol{\nu}_{t-1} + \|g_t\|^2$
- Update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{1}{\sqrt{\boldsymbol{\nu}_t}} g_t$

EndFor

Algorithm 2 AdaGrad

Input: Objective function $f(\boldsymbol{w})$, learning rate $\eta > 0$, initial point $\boldsymbol{w}_1 \in \mathbb{R}^d$, initial conditioner $\boldsymbol{\nu}_1 \in \mathbb{R}^{d,+}$

For $t = 1 \to \infty$:
- Generate stochastic gradient g_t
- Calculate $\boldsymbol{\nu}_t = \boldsymbol{\nu}_{t-1} + g_t^{\odot 2}$
- Update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot g_t$

EndFor

A very recent exception (Faw et al., 2022) relaxes the assumptions and proves that AdaGrad-Norm converges by only assuming uniformly bounded smoothness (c.f. our Assumption 1) and affine noise variance (c.f. our Assumption 2), which matches the conditions of SGD. However, the proof in (Faw et al., 2022) is rather complicated (around 30 pages), which is hard to understand the intuition behind and to extend to the analysis of other cases. Moreover, the convergence rate in (Faw et al., 2022) does not get better when strong growth condition holds (i.e., our Assumption 2 with $D_0 = 0$) while SGD does. We believe such a gap is vital as strong growth condition holds in over-parameterized models (Vaswani et al., 2019), which are widely adopted in deep learning.

We know that the convergence analysis of SGD under the same set of assumptions is quite simple. **What makes the analysis of AdaGrad so complicated?** We can understand the difficulty from the classical descent lemma

$$\mathbb{E}[f(\boldsymbol{w}_{t+1})|\mathcal{F}_t] \le f(\boldsymbol{w}_t) + \underbrace{\mathbb{E}\left[\langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle |\mathcal{F}_t\right]}_{\text{First Order}} + \underbrace{\frac{L}{2}\mathbb{E}\left[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 |\mathcal{F}_t\right]}_{\text{Second Order}}, \tag{1}$$

where $\mathcal{F}_t := \sigma(g_1, \cdots, g_{t-1})$ denotes the sigma field of the stochastic gradients up to $t-1$. Then

- for SGD, $\boldsymbol{w}_{t+1} - \boldsymbol{w}_t = -\eta g_t$ and hence the "First Order" term is $-\eta\|\nabla f(\boldsymbol{w}_t)\|^2$, which is negative and able to decrease the objective sufficiently,
- for AdaGrad(-Norm), $\boldsymbol{w}_{t+1} - \boldsymbol{w}_t = -\eta \frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}$. As $\boldsymbol{\nu}_t$ correlates with g_t, the "First Order" term does not admit a clear form.

To deal with the correlation in AdaGrad(-Norm), a common practice is to use a surrogate $\tilde{\boldsymbol{\nu}}_t$ of $\boldsymbol{\nu}_t$ (Ward et al., 2020; Défossez et al., 2020; Faw et al., 2022), which is measurable with respect to $\mathcal{F}_t$, to decompose the "First Order" term as follows,

$$\mathbb{E}\left[\langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t\rangle \,|\mathcal{F}_t\right] = \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_t), -\eta\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\rangle |\mathcal{F}_t\right] = \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_t), -\eta\frac{g_t}{\sqrt{\tilde{\boldsymbol{\nu}}_t}}\right\rangle |\mathcal{F}_t\right]$$
$$+ \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta g_t\left(\frac{1}{\sqrt{\tilde{\boldsymbol{\nu}}_t}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)\right\rangle |\mathcal{F}_t\right].$$

The first term equals $-\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\tilde{\boldsymbol{\nu}}_t}}$, which is negative and desired. However, the last term is an additional error term, which is very challenging to deal with. Existing results either assume bounded stochastic gradient to work around it (Ward et al., 2020), or resolve it through complicated analysis (Faw et al., 2022) (c.f. Section 3).

Contributions. In this paper, we propose a novel auxiliary function $\xi(t) = \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}$ for the convergence analysis of AdaGrad(-Norm), and show the error term can be bounded by $\mathbb{E}^{|\mathcal{F}_t}[\xi(t-1) - \xi(t)]$ (c.f. Lemma 4), which can be reduced by telescoping. As explained in Section 3, such a auxiliary function is rooted in the non-increasing nature of the adaptive learning rate $\frac{\eta}{\sqrt{\boldsymbol{\nu}_k}}$.

With the new and simplified proof, we are able to obtain stronger results for AdaGrad-Norm and extend the analysis to other important scenarios.

- Under strong growth condition (or the so-called over-parameterized regime), our convergence rate for AdaGrad-Norm is $\mathcal{O}(\frac{1}{T})$, which matches that of SGD and stronger than existing results (Faw et al., 2022). This demonstrates that AdaGrad-Norm converges faster in the over-parameterized regime than in the under-parameterized regime.

- We extend the analysis to AdaGrad by utilizing a coordinate version $\tilde{\xi}(t) = \sum_{l=1}^{d} \frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t,l}}}$ of $\xi(t)$ and obtain similar convergence. To the best of our knowledge, this is the first convergence result of AdaGrad without the requirement of bounded gradient norm. We also prove the convergence for randomly-reshuffled AdaGrad, which is the version of AdaGrad used in deep learning practice.

- We go beyond the uniform smoothness and consider a realistic non-uniformly smooth condition called (L_0, L_1)-smooth condition (Assumption 6). We prove that AdaGrad(-Norm) still converges under (L_0, L_1)-smooth condition, but requires the learning rate smaller than a threshold, whose necessity is conversely verified with a counterexample. Together, AdaGrad can converge under the non-uniform smoothness but may not be exactly tuning-free.

We have observed a concurrent work by (Faw et al., 2023), which also establishes the convergence of AdaGrad under (L_0, L_1)-smooth condition and affine noise variance assumption. Though results are similar and both appear in COLT 2023, there are some notable differences between their findings and ours. First, Faw et al. (2023) require either $D_1 < 2$ or an additional assumption on the objective function, whereas our result holds for all D_1 values without any additional assumption. Second, their result is based on a novel stopping time, while ours relies on a new auxiliary function. Lastly, Faw et al. (2023) establish a set of negative results for Clipped SGD and Sign SGD (with momentum) when analyzed under the (L_0, L_1)-smooth condition, highlighting the advantage of AdaGrad over these optimizers. All in all, both their work and ours complement each other, providing a more comprehensive understanding of AdaGrad.

Organization of this paper. The rest of this paper is organized as follows. In Section 2, we define notations and introduce assumptions; in Section 3, we describe the motivation to use the auxiliary function; in Section 4, we derive the convergence result of AdaGrad-Norm under L-smooth condition; in Section 5, we extend the result to AdaGrad; in Section 6, we analyze the convergence of AdaGrad(-Norm) under (L_0, L_1)-smooth condition; Section 7 presents the related works.

2. Preliminary

Notations. The following notations are used throughout this paper.

- (Vector operators) $\odot$ stands for the Hadamard product between vectors, and $g^{\odot 2} \triangleq g \odot g$. $\langle \boldsymbol{w}, \boldsymbol{v} \rangle$ stands for the L^2 inner product between $\boldsymbol{w}$ and $\boldsymbol{v}$, and $\|\boldsymbol{w}\| \triangleq \sqrt{\langle \boldsymbol{w}, \boldsymbol{w} \rangle}$.
- (Stochastic operators) $\mathcal{F}_t = \sigma(g_{t-1}, \cdots, g_1)$ stands for the sigma field of historical gradients up to time $t-1$ and thus $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$ is an adapted random process with respect to $\{\mathcal{F}_t\}_{t=1}^{\infty}$. For brevity, we abbreviate the expectation conditional on $\mathcal{F}_t$ as $\mathbb{E}^{|\mathcal{F}_t}[*] \triangleq \mathbb{E}[*|\mathcal{F}_t]$.

Assumptions. Throughout this paper, we assume that f is lower bounded. We also need the following assumptions:

Assumption 1 (L-smooth condition) *We assume that f is differentiable and its gradient satisfies that $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$, we have $\|\nabla f(\boldsymbol{w}_1) - \nabla f(\boldsymbol{w}_2)\| \le L\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$.*

Assumption 2 (Affine noise variance) *We assume that there exist positive constants D_0 and D_1 such that $\forall t \ge 1$, $\mathbb{E}^{|\mathcal{F}_t}[\|g_t\|^2] \le D_0 + D_1\|\nabla f(\boldsymbol{w}_t)\|^2$.*

To the best of our knowledge, the above two assumptions are the weakest requirements for the convergence of AdaGrad(-Norm) among the existing literature.

3. Motivation of the auxiliary function

As mentioned in Introduction, the main obstacle in the analysis of AdaGrad(-Norm) is to bound the error term $\mathbb{E}^{|\mathcal{F}_t}\langle \nabla f(\boldsymbol{w}_t), \eta g_t(\frac{1}{\sqrt{\tilde{\boldsymbol{\nu}}_t}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}})\rangle$. Most of the existing works assume that $\|g_t\|$ is uniformly bounded, and choose $\tilde{\boldsymbol{\nu}}_t = \boldsymbol{\nu}_{t-1}$. In this case, the error term can be shown to be as small as the "Second Order" term in Eq. (1) and can be further bounded. If the bounded gradient assumption is removed, Faw et al. (2022) shows that *most of the iterations are "good"*, in the sense that the error term is smaller than $\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\tilde{\boldsymbol{\nu}}_t}}$, which hence won't affect the negativity of the "First Order" term in Eq. (1) after decomposition. However, it is complicated to deal with the "bad" iterations, which occupies the main space of the proof in (Faw et al., 2022).

Instead, to deal with the error term, we propose a simple auxiliary function $\xi(t)$ that can be canceled out during telescoping. The choice of $\xi(t)$ is motivated as follows. By choosing $\tilde{\boldsymbol{\nu}}_t = \boldsymbol{\nu}_{t-1}$, we find that the error term can be rewritten as

$$\begin{aligned}\mathbb{E}^{|\mathcal{F}_t}\left\langle \nabla f(\boldsymbol{w}_t), \eta g_t\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)\right\rangle \le& \eta\mathbb{E}^{|\mathcal{F}_t}\left[\|\nabla f(\boldsymbol{w}_t)\|\|g_t\|\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)\right] \\ =& \eta\mathbb{E}^{|\mathcal{F}_t}\left[\left(\frac{\|\nabla f(\boldsymbol{w}_t)\|\|g_t\|}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{\|\nabla f(\boldsymbol{w}_t)\|\|g_t\|}{\sqrt{\boldsymbol{\nu}_t}}\right)\right], \quad (2)\end{aligned}$$

where the inequality is due to the Cauchy-Schwarz inequality and $\boldsymbol{\nu}_t$ is non-decreasing. Note that if we have both $\|\nabla f(\boldsymbol{w}_t)\| \approx \|\nabla f(\boldsymbol{w}_{t-1})\|$ and $\|g_t\| \approx \|g_{t-1}\|$, the term (2) approximately equals to $\eta \mathbb{E}^{|\mathcal{F}_t} \left[\left(\frac{\|\nabla f(\boldsymbol{w}_{t-1})\| \|g_{t-1}\|}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{\|\nabla f(\boldsymbol{w}_t)\| \|g_t\|}{\sqrt{\boldsymbol{\nu}_t}} \right) \right]$. In this case, we can use $\hat{\xi}(t) = \frac{\|\nabla f(\boldsymbol{w}_t)\| \|g_t\|}{\sqrt{\boldsymbol{\nu}_t}}$ as an auxiliary function, and the sum of the expected error term satisfies

$$\sum_{t=1}^{T} \mathbb{E} \left\langle \nabla f(\boldsymbol{w}_t), \eta g_t \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \right) \right\rangle \lesssim \sum_{t=1}^{T} \mathbb{E} \left[\hat{\xi}(t-1) - \hat{\xi}(t) \right] = \hat{\xi}(0) - \mathbb{E}[\hat{\xi}(T)].$$

The RHS of the above inequality is bounded regardless of T. This is the motivation to use the auxiliary function. However, we do not have $\|g_t\| \approx \|g_{t-1}\|$ but only have $\|\nabla f(\boldsymbol{w}_t)\| \approx \|\nabla f(\boldsymbol{w}_{t-1})\|$ (due to bounded smoothness, i.e., Assumption 1). To resolve this challenge, we convert $\|g_t\|$ to $\|\nabla f(\boldsymbol{w}_t)\|$ by Assumption 2 in the above inequality, and use $\xi(t) \triangleq \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}$ instead of $\frac{\|\nabla f(\boldsymbol{w}_t)\| \|g_t\|}{\sqrt{\boldsymbol{\nu}_t}}$ as the auxiliary function. A formal statement of the above methodology can be seen in Lemma 4.

Remark 1 *Note that the above methodology is mainly based on that the adaptive learning rate is non-increasing. Therefore, we believe that similar approach can be applied to the analysis of other adaptive optimizers with non-increasing adaptive learning rates, such as AMSGrad.*

4. A refined convergence analysis of AdaGrad-Norm

In this section, we present our refined analysis of AdaGrad-Norm based on the auxiliary function $\xi(t)$. The refined convergence rate is given by the following theorem.

Theorem 2 *Let Assumptions 1 and 2 hold. Then, for AdaGrad-Norm with any learning rate $\eta > 0$, we have that with probability at least $1 - \delta$,*

$$\min_{t \in [T]} \|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{2\sqrt{2D_0}(2C_2 \ln(2\sqrt{2D_0 T} + C_3) + C_1)}{\sqrt{T}\delta^2} + \frac{C_3(C_1 + 2C_2 \ln(2\sqrt{2D_0 T} + C_3))}{T\delta^2},$$

where C_1, C_2, and C_3 are constants defined as $C_1 := 4(f(\boldsymbol{w}_1) - f^* + \frac{\eta D_1}{2} \frac{\|\nabla f(\boldsymbol{w}_0)\|^2}{\sqrt{\boldsymbol{\nu}_0}} + (2\eta (L\eta D_1)^2 + \eta D_1 (L\eta)^2 + \frac{\eta}{2} D_0) \frac{1}{\sqrt{\boldsymbol{\nu}_0}} - \frac{L}{2} \eta^2 \ln \boldsymbol{\nu}_0)/\eta$, $C_2 := 2L\eta$, and $C_3 := 4D_1 C_1 + 48 C_2 D_1 \ln(4C_2 D_1 + e) + 2\sqrt{\boldsymbol{\nu}_0}$.

Remark 3 *Faw et al. (2022) also prove that AdaGrad-Norm converges under Assumptions 1 and 2. Their rate is $\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(\boldsymbol{w}_t)\|^2 = \mathcal{O}(\frac{\log^{\frac{9}{4}} T}{\sqrt{T}})$. Compared to their result, our result has a tighter dependence over T. Moreover, when restricted to the strong growth condition, i.e., $D_0 = 0$, our result gives a rate $\mathcal{O}(\frac{1}{T})$, much faster than that in (Faw et al., 2022) and matching that of SGD. Such an improvement counts as strong growth condition characterizes the landscapes of over-parameterized models (Vaswani et al., 2019). Theorem 2 also shows that AdaGrad-Norm enjoys the tuning-free ability under L-smooth condition, i.e., it converges without tuning the learning rate.*

Proof of Theorem 2 The proof starts with the so-called expected descent lemma:

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

As discussed in Section 1, the "First Order" term does not have a simple form due to the correlation between g_t and $\boldsymbol{\nu}_t$. We follow the standard approach in existing literature to approximate $\boldsymbol{\nu}_t$ with the surrogate $\boldsymbol{\nu}_{t-1}$, which is measurable with respect to $\mathcal{F}_t$. The first-order term can then be decomposed into

$$\begin{aligned}\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), -\eta\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\rangle\right] = & \mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), -\eta\frac{g_t}{\sqrt{\boldsymbol{\nu}_{t-1}}}\right\rangle\right] + \mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}})g_t\right\rangle\right] \\ = & -\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)g_t\right\rangle\right]. \end{aligned} \tag{4}$$

The last term is an error term, coming from the gap between $\boldsymbol{\nu}_{t-1}$ and $\boldsymbol{\nu}_t$. Plugging Eq. (4) back to Eq. (3) and we obtain

$$\mathbb{E}^{|\mathcal{F}_t} f(\boldsymbol{w}_{t+1}) \leq f(\boldsymbol{w}_t) + \underbrace{-\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}}_{\text{First Order Main}} + \underbrace{\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)g_t\right\rangle\right]}_{\text{Error}} + \underbrace{\frac{L}{2}\eta^2\mathbb{E}^{|\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2}_{\text{Second Order}}.$$

The rest of the proof can be divided into two stages: in Stage I, We proceed by bounding the "Error" term through the auxiliary function $\xi(t) \triangleq \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}$, and bound $\sum_{t=1}^T \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}$. In Stage II, we convert the bound of $\sum_{t=1}^T \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}$ into the bound of $\sum_{t=1}^T \|\nabla f(\boldsymbol{w}_t)\|^2$.

Stage I: Bounding the "Error" term. The following lemma summarizes the intuition in Section 3.

Lemma 4 *Define an auxiliary function* $\xi(t) \triangleq \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}$, $t \geq 1$. *Then, the "Error" term can be bounded as*

$$\begin{aligned}&\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)g_t\right\rangle\right] \leq \frac{3}{4}\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}}D_0\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})^2}\right] \\ &+ \frac{\eta}{2}D_1\mathbb{E}^{|\mathcal{F}_t}\left[\xi(t-1) - \xi(t)\right] + \left(\eta(L\eta D_1)^2 + \frac{\eta}{2}D_1(L\eta)^2\right)\frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3}.\end{aligned}$$

Proof By a simple calculation, we have

$$\begin{aligned}&\left|\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)g_t\right\rangle\right]\right| = \left|\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_{t-1}})(\sqrt{\boldsymbol{\nu}_t})(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})}g_t\right\rangle\right]\right| \\ \leq & \eta\mathbb{E}^{|\mathcal{F}_t}\left[\|\nabla f(\boldsymbol{w}_t)\|\frac{\|g_t\|^3}{(\sqrt{\boldsymbol{\nu}_{t-1}})(\sqrt{\boldsymbol{\nu}_t})(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})}\right] \leq \eta\frac{\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}}}\right],\end{aligned}$$

where the first inequality is due to Cauchy-Schwarz inequality, and the second inequality is because $\boldsymbol{\nu}_t \geq \|g_t\|^2$. By the mean-value inequality ($2ab \leq a^2 + b^2$),

$$\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}}}\right] \leq \frac{1}{2}\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}}\left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}}}\right]\right)^2. \tag{5}$$

We focus on the last quantity $(\mathbb{E}^{|\mathcal{F}_t}[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}}}])^2$. By further applying Hölder's inequality,

$$\begin{aligned}\left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}}}\right]\right)^2 &\le \mathbb{E}^{|\mathcal{F}_t}\|g_t\|^2\cdot\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right]\\ &\le (D_0+D_1\|\nabla f(\boldsymbol{w}_t)\|^2)\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right],\end{aligned}$$

where in the last inequality we use Assumption 1. Plugging the above inequality back to Eq. (5), the "Error" term can be bounded as

$$\begin{aligned}&\left|\mathbb{E}^{|\mathcal{F}_t}\left[\langle\nabla f(\boldsymbol{w}_t),\eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}})g_t\rangle\right]\right| \le \frac{1}{2}\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}+\frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}}D_0\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right]\\ &+\frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}}D_1\|\nabla f(\boldsymbol{w}_t)\|^2\,\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right]. \qquad (6)\end{aligned}$$

In the RHS of the above equality, the first term is a negative half of the "First Order Main" term, and the second term is $\frac{1}{\eta L\sqrt{\boldsymbol{\nu}_{t-1}}}$ times of the "Second order" term, and thus is at the same order of the "Second order" term due to $\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}$ is upper bounded. We focus on the last term, and utilize the observation that

$$\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\le\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}\sqrt{\boldsymbol{\nu}_t}(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})}=\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}}.$$

Thus, the last term can be bounded as

$$\frac{1}{2}\frac{\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right]\le\frac{1}{2}\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2\mathbb{E}^{|\mathcal{F}_t}\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right),$$

which can be further decomposed into

$$\begin{aligned}&\frac{1}{2}\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2\mathbb{E}^{|\mathcal{F}_t}\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)\\ =&\frac{\eta}{2}D_1\mathbb{E}^{|\mathcal{F}_t}\left(\frac{\|\nabla f(\boldsymbol{w}_{t-1})\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}\right)+\frac{\eta}{2}D_1\frac{\|\nabla f(\boldsymbol{w}_t)\|^2-\|\nabla f(\boldsymbol{w}_{t-1})\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}.\end{aligned}$$

By Assumption 1, $\|\nabla f(\boldsymbol{w}_t)\|-\|\nabla f(\boldsymbol{w}_{t-1})\|\le\|\nabla f(\boldsymbol{w}_t)-\nabla f(\boldsymbol{w}_{t-1})\|\le L\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|$. Therefore,

$$\begin{aligned}&\frac{1}{2}\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2\mathbb{E}^{|\mathcal{F}_t}\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)\\ =&\frac{\eta}{2}D_1\mathbb{E}^{|\mathcal{F}_t}\left(\frac{\|\nabla f(\boldsymbol{w}_{t-1})\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}\right)+\frac{\eta}{2}D_1\frac{2L\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|+L^2\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\\ =&\frac{\eta}{2}D_1\mathbb{E}^{|\mathcal{F}_t}(\xi(t-1)-\xi(t))+\frac{\eta}{2}D_1\frac{2L\eta\frac{\|g_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}}+(L\eta\frac{\|g_{t-1}\|}{\sqrt{\boldsymbol{\nu}_{t-1}}})^2}{\sqrt{\boldsymbol{\nu}_{t-1}}},\end{aligned}$$

where the last inequality we use $\xi(t) \triangleq \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}$ and $\boldsymbol{w}_t - \boldsymbol{w}_{t-1} = \eta \frac{g_{t-1}}{\sqrt{\boldsymbol{\nu}_{t-1}}}$. Applying again the Cauchy-Schwarz inequality, we obtain

$$
\begin{aligned}
&\frac{\eta}{2} D_1 \mathbb{E}^{|\mathcal{F}_t} \|\nabla f(\boldsymbol{w}_t)\|^2 \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \right) \\
\le & \frac{\eta}{2} D_1 \mathbb{E}^{|\mathcal{F}_t} \left(\xi(t-1) - \xi(t)\right) + \frac{1}{4}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \eta (L\eta D_1)^2 \frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3} + \frac{\eta}{2} D_1 (L\eta)^2 \frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3}.
\end{aligned} \tag{7}
$$

Applying the above inequality back into Eq. (6), the "Error" term can be bounded as

$$
\begin{aligned}
&\mathbb{E}^{|\mathcal{F}_t}[\langle \nabla f(\boldsymbol{w}_t), \eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}) g_t \rangle] \le \frac{1}{2}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}} D_0 \mathbb{E}^{|\mathcal{F}_t} \left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})^2} \right] \\
&+ \frac{\eta}{2} D_1 \mathbb{E}^{|\mathcal{F}_t} \left(\xi(t-1) - \xi(t)\right) + \frac{1}{4}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \eta (L\eta D_1)^2 \frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3} + \frac{\eta}{2} D_1 (L\eta)^2 \frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3}.
\end{aligned}
$$

Rearranging the RHS of the above inequality leads to the claim. ■

Applying Lemma 4 back to the descent lemma, we then have

$$
\begin{aligned}
&\mathbb{E}^{|\mathcal{F}_t}[f(\boldsymbol{w}_{t+1})] \le f(\boldsymbol{w}_t) - \eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \frac{3}{4}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}} D_0 \mathbb{E}^{|\mathcal{F}_t} \left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})^2} \right] \\
&+ \frac{\eta}{2} D_1 \mathbb{E}^{|\mathcal{F}_t} \left(\xi(t-1) - \xi(t)\right) + \left(\eta (L\eta D_1)^2 + \frac{\eta}{2} D_1 (L\eta)^2 \right) \frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3} + \frac{L}{2}\eta^2 \mathbb{E}^{|\mathcal{F}_t} \left\| \frac{g_t}{\sqrt{\boldsymbol{\nu}_t}} \right\|^2 \\
=& f(\boldsymbol{w}_t) - \frac{1}{4}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}} D_0 \mathbb{E}^{|\mathcal{F}_t} \left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})^2} \right] + \frac{\eta}{2} D_1 \mathbb{E}^{|\mathcal{F}_t} \left(\xi(t-1) - \xi(t)\right) \\
&+ \left(\eta (L\eta D_1)^2 + \frac{\eta}{2} D_1 (L\eta)^2 \right) \frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3} + \frac{L}{2}\eta^2 \mathbb{E}^{|\mathcal{F}_t} \left\| \frac{g_t}{\sqrt{\boldsymbol{\nu}_t}} \right\|^2.
\end{aligned} \tag{8}
$$

Taking expectation with respect to $\mathcal{F}_t$ to the above inequality then leads to

$$
\begin{aligned}
\mathbb{E}[f(\boldsymbol{w}_{t+1})] \le & \mathbb{E}[f(\boldsymbol{w}_t)] - \frac{1}{4}\eta \mathbb{E}\left[\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \right] + \frac{\eta}{2} D_1 \mathbb{E}\left(\xi(t-1) - \xi(t)\right) \\
&+ \frac{\eta D_0}{2} \mathbb{E}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})^2} \right] + \frac{L}{2}\eta^2 \mathbb{E}\left\| \frac{g_t}{\sqrt{\boldsymbol{\nu}_t}} \right\|^2 + (\eta (L\eta D_1)^2 + \frac{\eta}{2} D_1 (L\eta)^2) \mathbb{E}\frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3}.
\end{aligned}
$$

The sum over t from 1 to T of the last three terms above can be bounded by

$$
\begin{aligned}
&\frac{\eta D_0}{2} \sum_{t=1}^{T} \mathbb{E}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}(\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}})^2} \right] + \sum_{t=1}^{T} \frac{L}{2}\eta^2 \mathbb{E}\left\| \frac{g_t}{\sqrt{\boldsymbol{\nu}_t}} \right\|^2 + (\eta (L\eta D_1)^2 + \sum_{t=1}^{T} \frac{\eta}{2} D_1 (L\eta)^2) \mathbb{E}\frac{\|g_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}^3} \\
\le & \left(2\eta (L\eta D_1)^2 + \eta D_1 (L\eta)^2) + \frac{\eta}{2} D_0 \right) \frac{1}{\sqrt{\boldsymbol{\nu}_0}} + \frac{L}{2}\eta^2 (\mathbb{E} \ln \boldsymbol{\nu}_T - \ln \boldsymbol{\nu}_0),
\end{aligned}
$$

where the inequality is due to that if $\{a_i\}_{i=0}^{\infty}$ is a series of non-negative real numbers with $a_0 > 0$, then $\sum_{t=1}^{T} \frac{a_t}{\sqrt{(\sum_{s=0}^{t} a_s)^3}} \le 2\frac{1}{\sqrt{a_0}}$, $\sum_{t=1}^{T} \frac{a_t}{\sum_{s=0}^{t} a_s} \le \ln \sum_{t=0}^{T} a_t - \ln a_0$, and $\sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=0}^{t} a_s}(\sqrt{\sum_{s=0}^{t-1} a_s} + \sqrt{\sum_{s=0}^{t} a_s})^2}$

$\leq \frac{1}{\sqrt{a}_0}$. Therefore, summing Eq. (8) over t from 1 to T leads to

$$\frac{1}{4}\eta\sum_{t=1}^{T}\mathbb{E}\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \leq f(\boldsymbol{w}_1) - \mathbb{E}[f(\boldsymbol{w}_T)] + \frac{\eta D_1}{2}\mathbb{E}[\xi(0) - \xi(t)]$$
$$+ \left(2\eta(L\eta D_1)^2 + \eta D_1(L\eta)^2) + \frac{\eta}{2}D_0\right)\frac{1}{\sqrt{\boldsymbol{\nu}_0}} + \frac{L}{2}\eta^2(\mathbb{E}\ln\boldsymbol{\nu}_T - \ln\boldsymbol{\nu}_0). \tag{9}$$

Applying the definition of C_1 and C_2, we have $\sum_{t=1}^{T}\mathbb{E}\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \leq C_1 + C_2\mathbb{E}\ln\boldsymbol{\nu}_T$. In Stage II, we translate such an inequality to the bound of $\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\|^2$.

Stage II: Bound $\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\|^2$. We bound $\mathbb{E}[\sqrt{\boldsymbol{\nu}_T}]$ by divide-and-conquer. We first consider the iterations satisfying $\nabla f(\boldsymbol{w}_t) > \frac{D_0}{D_1}$:

$$C_1 + C_2\mathbb{E}\ln\boldsymbol{\nu}_T \geq \sum_{t=1}^{T}\mathbb{E}\left[\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\right] \geq \sum_{t=1}^{T}\mathbb{E}\left[\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 > \frac{D_0}{D_1}}\right]$$
$$\geq \frac{1}{2D_1}\sum_{t=1}^{T}\mathbb{E}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 > \frac{D_0}{D_1}}\right] \geq \frac{1}{2D_1}\mathbb{E}\left[\frac{\sum_{t=1}^{T}\|g_t\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 > \frac{D_0}{D_1}}}{\sqrt{\boldsymbol{\nu}_T}}\right], \tag{10}$$

where in the third inequality, we use the following fact,

$$2D_1\|\nabla f(\boldsymbol{w}_t)\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 > \frac{D_0}{D_1}} \geq (D_0 + D_1\|\nabla f(\boldsymbol{w}_t)\|^2)\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 > \frac{D_0}{D_1}} \geq \mathbb{E}^{|\mathcal{F}_t}\|g_t\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 > \frac{D_0}{D_1}}.$$

We then consider the iterations satisfying $\nabla f(\boldsymbol{w}_t) \leq \frac{D_0}{D_1}$,

$$\frac{1}{2D_1}\sum_{t=1}^{T}\mathbb{E}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_T}}\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{D_0}{D_1}}\right] + \frac{1}{2D_1}\mathbb{E}\frac{\boldsymbol{\nu}_0}{\sqrt{\boldsymbol{\nu}_T}} \leq \frac{1}{2D_1}\mathbb{E}\left[\frac{\sum_{t=1}^{T}\|g_t\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{D_0}{D_1}} + \boldsymbol{\nu}_0}{\sqrt{\sum_{t=1}^{T}\|g_t\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{D_0}{D_1}} + \boldsymbol{\nu}_0}}\right]$$
$$= \frac{1}{2D_1}\mathbb{E}\sqrt{\sum_{t=1}^{T}\|g_t\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{D_0}{D_1}} + \boldsymbol{\nu}_0} \leq \frac{1}{2D_1}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\|g_t\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{D_0}{D_1}}\right] + \boldsymbol{\nu}_0}$$
$$\leq \frac{1}{2D_1}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}(D_1\|\nabla f(\boldsymbol{w}_t))\|^2 + D_0)\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{D_0}{D_1}} + \boldsymbol{\nu}_0\right]} \leq \frac{1}{2D_1}\sqrt{2D_0T + \boldsymbol{\nu}_0}. \tag{11}$$

Here in the second inequality we use Jensen's inequality, and in the third we use Assumption 2. Putting Eq. (10) and Eq. (11) together, we then have

$$\frac{1}{2D_1}\mathbb{E}[\sqrt{\boldsymbol{\nu}_T}] = \frac{1}{2D_1}\mathbb{E}\left[\frac{\sum_{t=1}^{T}\|g_t\|^2\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 > \frac{D_0}{D_1}}}{\sqrt{\boldsymbol{\nu}_T}}\right] + \frac{1}{2D_1}\sum_{t=1}^{T}\mathbb{E}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_T}}\mathbb{1}_{\|\nabla f(\boldsymbol{w}_t)\|^2 \leq \frac{D_0}{D_1}}\right] + \frac{1}{2D_1}\mathbb{E}\frac{\boldsymbol{\nu}_0}{\sqrt{\boldsymbol{\nu}_T}}$$
$$\leq \frac{1}{2D_1}\sqrt{2D_0T + \boldsymbol{\nu}_0} + C_1 + C_2\mathbb{E}\ln\boldsymbol{\nu}_T \leq \frac{1}{2D_1}\sqrt{2D_0T + \boldsymbol{\nu}_0} + C_1 + 2C_2\ln\mathbb{E}\sqrt{\boldsymbol{\nu}_T}.$$

Here in the last inequality we use Jensen's inequality. Solving the above inequality with respect to $\mathbb{E}[\sqrt{\boldsymbol{\nu}_T}]$, we have $\mathbb{E}[\sqrt{\boldsymbol{\nu}_T}] \leq 2\sqrt{2D_0T + \boldsymbol{\nu}_0} + 4D_1C_1 + 48C_2D_1\ln(4C_2D_1 + e)$. As

$$C_1 + 2C_2\ln\mathbb{E}\sqrt{\boldsymbol{\nu}_T} \geq \sum_{t=1}^{T}\mathbb{E}\left[\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\right] \geq \mathbb{E}\left[\frac{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_T}}\right] \geq \frac{\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\|^2}\right]^2}{\mathbb{E}[\sqrt{\boldsymbol{\nu}_T}]},$$

where the last inequality is due to Hölder's inequality. Applying the estimation of $\mathbb{E}\sqrt{\boldsymbol{\nu}_T}$, we obtain that $\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla f(\boldsymbol{w}_t)\|^2}\right]^2 \le (2\sqrt{2D_0T+\boldsymbol{\nu}_0}+C_3)(C_1+2C_2\ln(2\sqrt{2D_0T+\boldsymbol{\nu}_0}+C_3))$.

By further applying Markov's inequality, we conclude the proof. ■

Remark 5 *The proof in (Faw et al., 2022) can also be divided into two stages with similar goals. Our proof is simpler in both stages, and we discuss the reason here. As pointed out in Section 3, our proof in Stage I is simpler is due to the novel auxiliary function* ξ*. Moreover, our conclusion in Stage I is also stronger, which laid a better foundation for Stage II: Faw et al. (2022) can only derive the bound of* $\mathbb{E}\sum_{t\in\tilde{S}}\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\tilde{\boldsymbol{\nu}}_t}}$*, where* $\tilde{S}$ *is a subset of* $[T]$*. This raises additional challenges for Stage II in (Faw et al., 2022), as our divide-and-conquer technique can no longer be applied. Faw et al. (2022) resolve this through a recursively-improving technique, which not only require a complicated proof, but also entangles* D_0 *and* D_1 *and leads to a sub-optimal rate under strong growth condition.*

5. Extending the analysis to AdaGrad

In this section, we extend the convergence analysis of AdaGrad-Norm to AdaGrad. Such a result is attractive since AdaGrad is more commonly used in practice than AdaGrad-Norm. A natural hope is to prove the convergence of AdaGrad under the same set of assumptions in Theorem 2. However, challenge arises when we try to derive an AdaGrad version of Lemma 4. Concretely, the "First-Order Main" term becomes $\eta\sum_{l=1}^{d}\frac{\partial_i f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}}$ (we use $\boldsymbol{\nu}_{t,l}$ as the l-th coordinate of $\boldsymbol{\nu}_t$, and similar does $g_{t,l}$), while the bound of the "Error" term includes a term $\frac{\eta}{4}\sum_{l=1}^{d}\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}}$ and thus can not be controlled. Such a mismatch is due to that $\mathbb{E}[g_{t,i}^2]$ can only be bounded by the full gradient $\|\nabla f(\boldsymbol{w}_t)\|$ instead of the partial derivative of the corresponding coordinate $|\partial_i f(\boldsymbol{w}_t)|$ (see Appendix C.2 for details). Therefore, to derive the convergence of AdaGrad, we strengthen Assumption 2 to let the affine noise variance hold coordinate-wisely.

Assumption 3 (Coordinate-wise affine noise variance assumption) *We assume that there exist positive constants* D_0 *and* D_1 *such that* $\forall t\ge 1$ *and* $\forall i\in[d]$, $\mathbb{E}[|g_{t,i}|^2|\mathcal{F}_t]\le D_0+D_1\partial_i f(\boldsymbol{w}_t)^2$.

Note that Assumption 3 is *still general than most of the assumptions in existing works*. As an example, the bounded noise variance assumption is its special case. Next, we obtain the convergence result for AdaGrad as follows.

Theorem 6 (AdaGrad) *Let Assumptions 1 and 3 hold. Then, for AdaGrad with any learning rate* $\eta>0$*, we have that with probability at least* $1-\delta$, $\min_{t\in[T]}\|\nabla f(\boldsymbol{w}_t)\|^2=\mathcal{O}(\frac{1+\ln(1+\sqrt{D_0T})}{T\delta^2})+\mathcal{O}(\frac{\sqrt{D_0}(1+\ln(1+\sqrt{D_0T}))}{\sqrt{T}\delta^2})$.

The proof is a coordinate-wise version of the proof for Theorem 2 with some modifications, where we leverage a coordinate-wise version of $\xi(t)$, i.e., $\tilde{\xi}(t)=\sum_{l=1}^{d}\frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t,l}}}$ (please refer to Appendix C.1 for details).

We still seek to relax Assumption 3 back to Assumption 2. This is because Assumption 3 may preclude some basic objectives. We demonstrate this idea through the following example.

Example 1 *Consider the following linear regression problem:* $f(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)^2 = \|\boldsymbol{w}\|^2$*, where* $\mathcal{D}$ *is a standard Gaussian distribution over* $\mathbb{R}^d$ *with the absolute value of each coordinate truncated by* 1*. At point* $\boldsymbol{w}$*, define the stochastic gradient* $g(\boldsymbol{w})$ *as* $2\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}$*, where* $\boldsymbol{x}$ *is sampled according to* $\mathcal{D}$*. One can easily verify that* $g(\boldsymbol{w})$ *is an unbiased estimation of* $\nabla f(\boldsymbol{w})$*. For this example,* $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\|g(\boldsymbol{w})\|^2 = \Theta(\|\boldsymbol{w}\|^2)$ *and* $\|\nabla f(\boldsymbol{w})\|^2 = \Theta(\|\boldsymbol{w}\|^2)$*. Therefore, Assumption 2 holds with* $L_0 = 0$*. However,* $\|\partial_1 f(\boldsymbol{w})\|^2 = 4(\boldsymbol{w})_1^2$ *and* $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}(g(\boldsymbol{w}))_1^2 = \Theta(\|\boldsymbol{w}\|^2)$*, we see that* $\lim_{|(\boldsymbol{w})_2|\to\infty} \frac{\|\partial_1 f(\boldsymbol{w})\|^2}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}(g(\boldsymbol{w}))_1^2} = \infty$*, which violates Assumption 3.*

Algorithm 3 Randomly-reshuffled AdaGrad

Input: Objective function $f(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{w})$, learning rate $\eta > 0$, $\boldsymbol{w}_{1,1} \in \mathbb{R}^d$, $\boldsymbol{\nu}_{1,0} \in \mathbb{R}^{d,+}$

```
1: For t = 1 → ∞:
2:     For i = 1 → n:
3:         Uniformly sample {τ_{t,1}, ..., τ_{t,n}} as a random permutation of [n]
4:         Calculate g_{t,i} = ∇f_{τ_{t,i}}(w_{τ_{t,i}})
5:         Update ν_{t,i} = ν_{t,i-1} + g_{t,i}^{⊙2}
6:         Update w_{t,i+1} = w_{t,i} - η (1/√ν_{t,i}) ⊙ g_{k,i}
7:     EndFor
8:     Update w_{t+1,1} = w_{t,n+1}, ν_{t+1,0} = ν_{t,n}
9: EndFor
```

On the other hand, note that the above example obeys a stronger assumption on the smoothness, i.e, for every fixed x, the stochastic gradient $g(\boldsymbol{w})$ is globally Lipschitz. It is natural to ask whether we can relax Assumption 3 by strengthening the assumption on the smoothness. In Section 3, we explain that we use $\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}$ instead of $\frac{\|g_t\|\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}}$ as the auxiliary function due to $g_t \not\approx g_{t-1}$. Therefore, tightening the assumption on the smoothness may help us to ensure $g_t \approx g_{t-1}$, and we no longer need to bound g_t using Assumption 2, and as a result we will not encounter the mismatch between $\frac{\partial_i f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t-1,i}}}$ and $\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1,i}}}$. Motivated by the above example, we make the assumption that we have the access to an stochastic oracle $g(\boldsymbol{w}, \zeta)$, and g_t is generated by $g_t = g(\boldsymbol{w}_t, \zeta_t)$ where ζ_t is sampled independently from a distribution $\mathcal{D}$. We further assume that g is L Lipschitz with respect to $\boldsymbol{w}$ for a fixed ζ. Such an assumption is common in stochastic optimization literature (Yun et al., 2021; Shi and Li, 2021). Unfortunately, such assumptions are still not adequate to ensure $g_t \approx g_{t-1}$ since g_t and g_{t-1} may use different noise ζ. A good news is that, for the without-replacement version of AdaGrad (also called randomly-reshuffled AdaGrad. Please refer to Algorithm 5), every ζ appears once within one epoch and the above methodology can be used. Note that randomly-reshuffled AdaGrad is the version of AdaGrad commonly adopted in deep learning. Thus although we slightly change the analyzed algorithm, the problem we consider is still of significance.

As mentioned above, we require the following assumptions for the convergence of randomly-reshuffled AdaGrad.

Assumption 4 (Assumption 2, reformulated) *Let* $\boldsymbol{w}_{k,i}$ *and* $g_{k,i}$ *be the ones in Algorithm 5. Then, there exist constants* D_0 *and* D_1*, such that,* $\forall k, i$*,* $\mathbb{E}_{j\sim \textbf{\textit{Uniform}}(n)}\|\nabla f_j(\boldsymbol{w}_{k,i})\|^2 \le D_0 + D_1\|\nabla f(\boldsymbol{w}_{k,i})\|^2$*.*

Assumption 5 (Stochastic L-smooth condition) *We assume that* $\forall i \in [n]$*,* f_i *is differentiable and its gradient satisfies* $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$*, we have* $\|\nabla f_i(\boldsymbol{w}_1) - \nabla f_i(\boldsymbol{w}_2)\| \le L\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$*.*

Theorem 7 (Randomly-reshuffled AdaGrad) *Let Assumptions 4 and 5 hold. Then, for randomly-reshuffled AdaGrad with any* $\eta > 0$*,* $\min_{t\in[T]} \|\nabla f(\boldsymbol{w}_{t,0})\|^2 = \mathcal{O}(\frac{1+\ln(1+\sqrt{D_0}T)}{T}) + \mathcal{O}(\frac{\sqrt{D_0}(1+\ln(1+\sqrt{D_0}T))}{\sqrt{T}})$*.*

The proof utilizes a randomly-reshuffling version of $\xi(t)$, i.e. $\bar{\xi}(t) = \sum_{i=1}^{n}\sum_{l=1}^{d}\frac{|\partial_l f(\boldsymbol{w}_{t,1})||\partial_l f_{\tau_{t,i}}(\boldsymbol{w}_{t,i})|}{\sqrt{\boldsymbol{\nu}_{t,i,l}}}$, and we defer the details to Appendix C.3. Theorem 7 shows that randomly-reshuffled AdaGrad does converge under affine noise variance assumption, and extends our analysis techniques to new setting of AdaGrad.

6. Convergence of AdaGrad over non-uniformly smooth landscapes

So far, the characterizations of AdaGrad(-Norm) has closely matched those of SGD over the uniformly smooth landscape. However, in practice, the objective function is usually non-uniformly smooth. Simple examples include polynomial functions with degree larger than 2, and deep neural networks. A natural question is that whether AdaGrad still works well over non-smooth landscapes. In this section, we analyze AdaGrad(-Norm) under the (L_0, L_1)-smooth condition Zhang et al. (2019), which is considered as a preciser characterization of the landscape of neural networks through exhaustive experiments.

Assumption 6 ((L_0, L_1)-smooth condition) *We assume that $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. We further assume that there exists positive constants L_0 and L_1, such that if $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ satisfies $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \le \frac{1}{L_0}$, then $\|\nabla f(\boldsymbol{w}_1) - \nabla f(\boldsymbol{w}_2)\| \le (L_1\|\nabla f(\boldsymbol{w}_1)\| + L_0)\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$.*

Assumption 6 degenerates to Assumption 1 with $L = L_0$ if $L_1 = 0$. Therefore, Assumption 6 is more general than Assumption 1. Moreover, Assumption 6 holds for polynomials of any degree and even exponential functions. Zhang et al. (2019) demonstrate through extensive experiments that over the tasks where adaptive optimizers outperforms SGD, Assumption 6 is obeyed.

Theorem 8 *Let Assumptions 2 and 6 hold. Then, for AdaGrad-Norm with $\forall \eta < \frac{1}{L_1}\min\{\frac{1}{64D_1}, \frac{1}{8\sqrt{D_1}}\}$, we have that with probability at least $1-\delta$,*

$$\min_{t\in[T]}\|\nabla f(\boldsymbol{w}_t)\|^2 = \mathcal{O}\left(\frac{1+\ln(1+\sqrt{D_0 T})}{T\delta^2}\right) + \mathcal{O}\left(\frac{\sqrt{D_0}(1+\ln(1+\sqrt{D_0 T}))}{\sqrt{T}\delta^2}\right).$$

The proof can be found in Appendix D.1, the key insight of which is that the additional error terms caused by (L_0, L_1)-smooth condition is at the same order of the "Error" term in the proof of Theorem 2. Theorem 8 shows that AdaGrad-Norm can provably overcome the non-uniform smoothness in the objective function and converges. Similar result can be derived for AdaGrad. Compared to Theorem 2, we additionally require the learning rate to be lower than a threshold. To see such a requirement is not artificial due to the proof, we provide the following theorem.

Theorem 9 *For every learning rate $\eta > \frac{9\sqrt{5}}{2L_1}$, there exist a lower-bounded objective function f obeying Assumption 6 and a corresponding initialization point $\boldsymbol{w}_0$, such that AdaGrad with learning rate η and initialized at $\boldsymbol{w}_0$ diverges over f.*

The proof can be found in Appendix D.2. Theorem 9 shows that the learning rate requirement in Theorem 8 is tight w.r.t. L_1 up to constants. The tuning-free ability of AdaGrad under L-smooth condition, i.e, AdaGrad converges under any learning rate, has been considered to be superiority of AdaGrad over SGD. However, combining Theorem 8 and Theorem 9, we demonstrate that such a property is lost when under a more realistic assumption on the smoothness. On the other hand, Zhang et al. (2019) show that SGD converges arbitrarily slowly under (L_0, L_1)-smooth condition. Together with Theorem 8, we find another superiority of AdaGrad: it can provably overcome non-uniform smoothness but SGD can not.

7. Related works

Convergence of AdaGrad over non-convex landscapes. Duchi et al. (2011) and McMahan and Streeter (2010) simultaneously propose AdaGrad. Since then, there is a line of works analyzing the convergence of AdaGrad over non-convex landscapes (Ward et al., 2020; Li and Orabona, 2019; Zou et al., 2018; Li and Orabona, 2020; Défossez et al., 2020; Gadat and Gavra, 2020; Kavis et al., 2022; Faw et al., 2022). We summarize their assumptions and conclusions in Table 1.

Non-uniform smoothness. The convergence analysis of optimizers under non-uniform smoothness is initialized by (Zhang et al., 2019), who propose the (L_0, L_1)-smooth condition and verify its validity in deep learning. They further prove that Clipped SGD converges under such a condition. Since then, their analysis has been extended to other clipped optimizers (Zhang et al., 2020; Yang et al., 2022; Crawshaw et al., 2022). However, there is no such a result for AdaGrad.

8. Conclusion

In this paper, we analyze AdaGrad over non-convex landscapes. Specifically, we propose a novel auxiliary function to bound the error term that is brought by the update of AdaGrad. Based on this auxiliary function, we are able to significantly simplify the proof of AdaGrad-Norm and establish a tighter convergence rate in the over-parameterized regime. We further extend the analysis to AdaGrad and non-uniformly smooth landscapes through different variants of the auxiliary function. One future direction is to explore and compare the convergences of AdaGrad and Adam under the (L_0, L_1)-smooth condition, given the fact that convergences of AdaGard and SGD are clearly separated under this condition.

Acknowledgments

This work is founded by CAS Project for Young Scientists in Basic Research under Grant No. YSBR-034, Innovation Project of ICT CAS under Grants No. E261090.

References

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signSGD. *arXiv preprint arXiv:2208.11195*, 2022.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. *arXiv preprint arXiv:2202.05791*, 2022.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. *arXiv preprint arXiv:2302.06570*, 2023.

Sébastien Gadat and Ioana Gavra. Asymptotic study of stochastic adaptive algorithm in non-convex landscape. 2020.

Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. *arXiv preprint arXiv:2204.02833*, 2022.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pages 983–992. PMLR, 2019.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

Naichen Shi and Dawei Li. Rmsprop converges with proper hyperparameter. In *International conference on learning representation*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

Xiaodong Yang, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Normalized/clipped SGD with perturbation for differentially private non-convex optimization. *arXiv preprint arXiv:2206.13033*, 2022.

Chulhee Yun, Shashank Rajput, and Suvrit Sra. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. *arXiv preprint arXiv:2110.10342*, 2021.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019.

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

Appendix A. Table listing existing works on convergence of AdaGrad over non-convex landscapes

	Allow unbounded gradient	Not delayed$^{(a)}$	Tuning-free learning rate	Additional comments
Ward et al. (2020)	×	✓	✓	AdaGrad-Norm
Li and Orabona (2019)	✓	×$^{(b)}$	×	AdaGrad-Norm, sub-gaussian noise
Zou et al. (2018)	×	✓	✓	AdaGrad
Li and Orabona (2019)	✓	×	×	AdaGrad, sub-gaussian noise
Défossez et al. (2020)	×	✓	✓	AdaGrad
Gadat and Gavra (2020)	×	✓	×	AdaGrad, asymptotic
Kavis et al. (2022)	×	✓	✓	AdaGrad-Norm$^{(c)}$
Faw et al. (2022)	✓	✓	✓	AdaGrad-Norm

Table 1: Summary of existing convergence result of AdaGrad. **None of these works show that AdaGrad converges at rate $\frac{1}{T}\|\nabla f(\boldsymbol{w}_t)\|^2 = \mathcal{O}(1/T)$ when over-parameterized.** We provide some explanation on the upper footmarks: (a): "delayed" refers to Delayed AdaGrad, where $\boldsymbol{\nu}_t$ does not contain the information of g_t; (b): Li and Orabona (2019) also change the degree of $\boldsymbol{\nu}_t$ in the adaptive learning rate from $-\frac{1}{2}$ to $-(\frac{1}{2}+\varepsilon)$; (c): the bound in (Kavis et al., 2022) has logarithmic dependence on the probability margin.

Appendix B. Preparations

This section collects technical lemmas that will be used latter.

Lemma 10 *Let $\{a_i\}_{i=0}^{\infty}$ be a series of non-negative real numbers with $a_0 > 0$. Then, the following inequalities hold:*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{(\sum_{s=0}^{t} a_s)^3}} \le 2\frac{1}{\sqrt{a_0}}, \quad \sum_{t=1}^{T} \frac{a_t}{\sum_{s=0}^{t} a_s} \le \ln \sum_{t=0}^{T} a_t - \ln a_0,$$

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=0}^{t} a_s}(\sqrt{\sum_{s=0}^{t-1} a_s} + \sqrt{\sum_{s=0}^{t} a_s})^2} \le \frac{1}{\sqrt{a_0}}.$$

Proof [Proof of Lemma 10] The first inequality is because $\sum_{t=1}^{T} \frac{a_t}{\sqrt{(\sum_{s=0}^{t} a_s)^3}} \le \int_{x=a_0}^{\sum_{t=0}^{T} a_t} \frac{1}{\sqrt{x}^3} \mathrm{d}x$. The second inequality is because $\sum_{t=1}^{T} \frac{a_t}{\sum_{s=0}^{t} a_s} \le \int_{x=a_0}^{\sum_{t=0}^{T} a_t} \frac{1}{x} \mathrm{d}x$. The third inequality is due to

$$\frac{a_t}{\sqrt{\sum_{s=0}^{t} a_s}(\sqrt{\sum_{s=0}^{t-1} a_s} + \sqrt{\sum_{s=0}^{t} a_s})^2} \le \frac{a_t}{\sqrt{\sum_{s=0}^{t} a_s}\sqrt{\sum_{s=0}^{t-1} a_s}(\sqrt{\sum_{s=0}^{t-1} a_s} + \sqrt{\sum_{s=0}^{t} a_s})}$$
$$= \frac{1}{\sqrt{\sum_{s=0}^{t-1} a_s}} - \frac{1}{\sqrt{\sum_{s=0}^{t} a_s}}.$$

The proof is completed. ■

Appendix C. Analysis of AdaGrad

C.1. Proof of Theorem 6

Before the proof, we define $g_{k,l}$ as the l-th component of g_k and $\boldsymbol{\nu}_{k,l}$ as the l-th component of $\boldsymbol{\nu}_k$.

Proof [Proof of Theorem 6] As the proof highly resembles that of Theorem 6, we only highlight several key steps with the rest of the details omitted.

By the descent lemma, we have that

$$\mathbb{E}^{|\mathcal{F}_t} f(\boldsymbol{w}_{t+1}) \leq f(\boldsymbol{w}_t) + \underbrace{-\eta \left\langle \frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} \odot \nabla f(\boldsymbol{w}_t), \nabla f(\boldsymbol{w}_t) \right\rangle}_{\text{First Order Main}} + \underbrace{\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right) \odot g_t \right\rangle\right]}_{\text{Error}}$$
$$+ \underbrace{\frac{L}{2}\eta^2 \mathbb{E}^{|\mathcal{F}_t} \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_t}} \odot g_t \right\|^2}_{\text{Second Order}}. \tag{12}$$

The "Error" term can be expanded as

$$\mathbb{E}^{|\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right) \odot g_t \right\rangle\right] = \sum_{l=1}^{d} \mathbb{E}^{|\mathcal{F}_t}\left[\eta \partial_l f(\boldsymbol{w}_t) \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{t,l}}}\right) g_{t,l}\right].$$

For each l, the RHS of the above inequality can then be bounded as

$$\mathbb{E}^{|\mathcal{F}_t}\left[\eta \partial_l f(\boldsymbol{w}_t) \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{t,l}}}\right) g_{t,l}\right] \leq \frac{1}{2}\eta \frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} + \frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} \left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{\sqrt{\boldsymbol{\nu}_{t,l}} + \sqrt{\boldsymbol{\nu}_{t-1,l}}}\right]\right)^2,$$

where by Assumption 3, $\left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{\sqrt{\boldsymbol{\nu}_{t,l}}+\sqrt{\boldsymbol{\nu}_{t-1,l}}}\right]\right)^2$ can be bounded by

$$\left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{\sqrt{\boldsymbol{\nu}_{t,l}} + \sqrt{\boldsymbol{\nu}_{t-1,l}}}\right]\right)^2 \leq (D_0 + D_1 \partial_l f(\boldsymbol{w}_t)^2) \mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{(\sqrt{\boldsymbol{\nu}_{t,l}} + \sqrt{\boldsymbol{\nu}_{t-1,l}})^2}\right].$$

Then, following the similar routine as the proof of Theorem 2 and leveraging the potential function $\tilde{\xi}_l(t) \triangleq \frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t,l}}}$, we have

$$\mathbb{E}^{|\mathcal{F}_t}\left[\eta \partial_l f(\boldsymbol{w}_t) \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{t,l}}}\right) g_{t,l}\right] \leq \frac{3}{4}\eta \frac{|\partial_l f(\boldsymbol{w}_t)|^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} + \frac{1}{2}\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} D_0 \mathbb{E}^{|\mathcal{F}_t}\left[\frac{\|g_{t,l}\|^2}{(\sqrt{\boldsymbol{\nu}_{t,l}} + \sqrt{\boldsymbol{\nu}_{t-1,l}})^2}\right]$$
$$+ \frac{\eta}{2} D_1 \mathbb{E}^{|\mathcal{F}_t}\left(\tilde{\xi}_l(t-1) - \tilde{\xi}_l(t)\right) + \left(\eta (L\eta D_1)^2 + \frac{\eta}{2} D_1 (L\eta)^2\right) \frac{1}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} \odot g_{t-1} \right\|^2$$

Applying the above inequality back to Eq. (12) and summing Eq. (12) over t from 1 to T, we obtain that

$$\frac{1}{4}\eta \sum_{t=1}^{T} \sum_{l=1}^{d} \mathbb{E} \frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}} \leq f(\boldsymbol{w}_1) - \mathbb{E}[f(\boldsymbol{w}_T)] + \sum_{l=1}^{d} \frac{\eta D_1}{2} \mathbb{E}[\tilde{\xi}_l(0) - \tilde{\xi}_l(t)] + \frac{\eta}{2} D_0 \frac{1}{\sqrt{\boldsymbol{\nu}_0}}$$
$$+ \left(\frac{1}{\min_l \sqrt{\boldsymbol{\nu}_{0,l}}} \left(\eta (L\eta D_1)^2 + \frac{\eta}{2} D_1 (L\eta)^2\right) + \frac{L}{2}\eta^2\right) (\mathbb{E} \ln \boldsymbol{\nu}_T - \ln \boldsymbol{\nu}_0).$$

Define

$$\tilde{C}_2 \triangleq 4\left(\frac{1}{\min_l \sqrt{\boldsymbol{\nu}_{0,l}}}\left(\eta(L\eta D_1)^2+\frac{\eta}{2}D_1(L\eta)^2\right)+\frac{L}{2}\eta\right),$$

$$\tilde{C}_1 \triangleq \frac{4}{\eta}f(\boldsymbol{w}_1)-\frac{4}{\eta}\mathbb{E}[f(\boldsymbol{w}_T)]+\sum_{l=1}^{d}2D_1\mathbb{E}[\tilde{\xi}_l(0)-\tilde{\xi}_l(T)]+2D_0\frac{1}{\sqrt{\boldsymbol{\nu}_0}}-\tilde{C}_2\ln\boldsymbol{\nu}_0.$$

Then, following the same routine as Stage II in the proof of Theorem 2, we have

$$\begin{aligned}\frac{1}{2D_1}\mathbb{E}\left[\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{T,l}}\right] \le& \frac{1}{2D_1}\sum_{l=1}^{d}\sqrt{2D_0T+\boldsymbol{\nu}_{0,l}}+\tilde{C}_1+\tilde{C}_2\mathbb{E}\sum_{l=1}^{d}\ln\boldsymbol{\nu}_{T,l}\\ \le& \frac{1}{2D_1}\sum_{l=1}^{d}\sqrt{2D_0T+\boldsymbol{\nu}_{0,l}}+\tilde{C}_1+2d\tilde{C}_2\ln\mathbb{E}\sum_{l=1}^{d}\sqrt{\boldsymbol{\nu}_{T,l}}.\end{aligned}$$

The rest of the proof flows just as Theorem 6. The proof is completed. ■

C.2. Explanation of why Theorem 6 uses Assumption 3

Here we provide a more detailed explanation of why Theorem 6 can not be established using Assumption 3.

In the proof of the previous section, we see that by Assumption 3,

$$\left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{\sqrt{\boldsymbol{\nu}_{t,l}}+\sqrt{\boldsymbol{\nu}_{t-1,l}}}\right]\right)^2 \le (D_0+D_1\partial_l f(\boldsymbol{w}_t)^2)\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{(\sqrt{\boldsymbol{\nu}_{t,l}}+\sqrt{\boldsymbol{\nu}_{t-1,l}})^2}\right].$$

The term $\partial_l f(\boldsymbol{w}_t)^2$ latter results in a $\frac{1}{4}\eta\frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}}$ term in the bound of the "Error" term. However, if we replace Assumption 3 by Assumption 2, the bound of $\left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{\sqrt{\boldsymbol{\nu}_{t,l}}+\sqrt{\boldsymbol{\nu}_{t-1,l}}}\right]\right)^2$ change to

$$\left(\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{\sqrt{\boldsymbol{\nu}_{t,l}}+\sqrt{\boldsymbol{\nu}_{t-1,l}}}\right]\right)^2 \le (D_0+D_1\|\nabla f(\boldsymbol{w}_t)\|^2)\mathbb{E}^{|\mathcal{F}_t}\left[\frac{g_{t,l}^2}{(\sqrt{\boldsymbol{\nu}_{t,l}}+\sqrt{\boldsymbol{\nu}_{t-1,l}})^2}\right],$$

and the term $\frac{1}{4}\eta\frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}}$ changes to $\frac{1}{4}\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}}$ correspondingly. Such a term is no longer smaller than the absolute value of the "First Order Main" term $-\eta\sum_{l=1}^{d}\frac{\partial_l f(\boldsymbol{w}_t)^2}{\sqrt{\boldsymbol{\nu}_{t-1,l}}}$, and the proof can not go on.

C.3. Proof of Theorem 7

Before the proof, we define several notations. We define $\boldsymbol{\nu}_{k,i,l}$ as the l-th component of $\boldsymbol{\nu}_{k,i}$, $\forall k\ge 1, i\in[n], l\in[d]$. Similarly, we define $g_{k,i,l}$ as the l-th component of $g_{k,i}$. We define $\tilde{\tau}_{k,i}$ as the inverse index of $\tau_{k,i}$, in the sense that $\tau_{k,\tilde{\tau}_{k,i}}=i$, $\forall k\ge 1, i\in[n]$. One can easily check that $\{\tilde{\tau}_{k,i}\}_{i=1}^{n}=[n]$, $\forall k\ge 1$.

Proof By Assumption 5 and the triangle inequality, f satisfies L-smooth condition. By applying the descent lemma to the change of parameters within a whole epoch, we have

$$f(\boldsymbol{w}_{k+1,1}) \leq f(\boldsymbol{w}_{k,1}) + \langle \nabla f(\boldsymbol{w}_{k,1}), \boldsymbol{w}_{k+1,1} - \boldsymbol{w}_{k,1} \rangle + \frac{L}{2} \|\boldsymbol{w}_{k+1,1} - \boldsymbol{w}_{k,1}\|^2. \tag{13}$$

We temporarily focus on the first order term $\langle \nabla f(\boldsymbol{w}_{k,1}), \boldsymbol{w}_{k+1,1} - \boldsymbol{w}_{k,1} \rangle$ as the second order term is as simple to be bounded as those in the proofs of Theorem 2 and Theorem 6. The $\boldsymbol{w}_{k+1,1} - \boldsymbol{w}_{k,1}$ in the first order term can be rewritten as

$$\begin{aligned}
&\boldsymbol{w}_{k+1,1} - \boldsymbol{w}_{k,1} = -\eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}} \odot g_{k,i} \\
=& -\eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot g_{k,i} + \eta \sum_{i=1}^{n} \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}}\right) \odot g_{k,i} \\
=& -\eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) + \eta \sum_{i=1}^{n} \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}}\right) \odot g_{k,i} + \eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot (\nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) - g_{k,i}) \\
=& -\eta \frac{n}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) + \eta \sum_{i=1}^{n} \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}}\right) \odot g_{k,i} + \eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot (\nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) - g_{k,i})
\end{aligned}$$

Thus, the first order term can be rewritten as

$$\begin{aligned}
&\langle \nabla f(\boldsymbol{w}_{k,1}), \boldsymbol{w}_{k+1} - \boldsymbol{w}_{k,1} \rangle \\
=& \langle \nabla f(\boldsymbol{w}_{k,1}), -\eta \frac{n}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \rangle + \langle \nabla f(\boldsymbol{w}_{k,1}), \eta \sum_{i=1}^{n} \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}}\right) \odot g_{k,i} \rangle \\
&+ \langle \nabla f(\boldsymbol{w}_{k,1}), \eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot (\nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) - g_{k,i}) \rangle \\
=& -\eta n \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \left\langle \nabla f(\boldsymbol{w}_{k,1}), \eta \sum_{i=1}^{n} \left(\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}}\right) \odot g_{k,i} \right\rangle \\
&+ \langle \nabla f(\boldsymbol{w}_{k,1}), \eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot (\nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) - g_{k,i}) \rangle.
\end{aligned}$$

We then tackle the three terms respectively and denote $F_1 = -n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2$, $F_2 = \langle \nabla f(\boldsymbol{w}_{k,1}), \eta \sum_{i=1}^{n} (\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}}) \odot g_{k,i} \rangle$, and $F_3 = \langle \nabla f(\boldsymbol{w}_{k,1}), \eta \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot (\nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) -$

$g_{k,i})\rangle$. F_2 can be bounded as

$$
\begin{aligned}
&|F_2| \\
=&|\langle \nabla f(\boldsymbol{w}_{k,1}), \eta \sum_{i=1}^{n} (\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i}}}) \odot g_{k,i} \rangle| = |\eta \sum_{l=1}^{d} \partial_l f(\boldsymbol{w}_{k,1}) \sum_{i=1}^{n} (\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i,l}}}) g_{k,i,l}| \\
\le&\eta \sum_{l=1}^{d} \sum_{i=1}^{n} |\partial_l f(\boldsymbol{w}_{k,1}) (\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i,l}}}) g_{k,i,l}| \overset{(\star)}{=} \eta \sum_{l=1}^{d} \sum_{i=1}^{n} (\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i,l}}}) |\partial_l f(\boldsymbol{w}_{k,1})| |g_{k,i,l}| \\
=&\eta \sum_{l=1}^{d} \sum_{i=1}^{n} (\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,i,l}}}) |\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_{\tau_{k,i}}(\boldsymbol{w}_{k,i})| = \eta \sum_{l=1}^{d} \sum_{i=1}^{n} (\frac{1}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}}) |\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})| \\
\le&\eta \sum_{l=1}^{d} \sum_{i=1}^{n} (\frac{|\partial_l f(\boldsymbol{w}_{k-1,1})| |\partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}}) \\
&+ \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \frac{|\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) - \partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} \\
&+ \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \frac{|\partial_l f(\boldsymbol{w}_{k,1}) - \partial_l f(\boldsymbol{w}_{k-1,1})| |\partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k,1,l}}},
\end{aligned}
$$

where Eq. $(\star)$ is due to $\boldsymbol{\nu}_{k,i,l}$ is non-decreasing with respect to k and i. Since $\forall k \ge 2$,

$$
\|\nabla f(\boldsymbol{w}_{k,1}) - \nabla f(\boldsymbol{w}_{k-1,1})\|^2 \le L^2 \|\boldsymbol{w}_{k,1} - \boldsymbol{w}_{k-1,1}\|^2 \le nL^2 \sum_{j=1}^{n} \|\boldsymbol{w}_{k-1,j+1} - \boldsymbol{w}_{k-1,j}\|^2,
$$

and

$$
\begin{aligned}
\|\nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) - \nabla f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})\|^2 \le& L^2 \|\boldsymbol{w}_{k,\tilde{\tau}_{k,i}} - \boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}}\|^2 \\
\le& 2nL^2 \sum_{j=1}^{n} \|\boldsymbol{w}_{k-1,j+1} - \boldsymbol{w}_{k-1,j}\|^2 + 2nL^2 \sum_{j=1}^{n} \|\boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j}\|^2,
\end{aligned}
$$

then $\forall k \geq 2$ we can further bound $|F_2|$ as

$$
\begin{aligned}
&|F_2| \\
\leq & \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{k-1,1})||\partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})||\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}} \right) \\
& + \sum_{i=1}^{n} \frac{1}{2} (\eta L \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \frac{\eta}{L \min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \| \nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) - \nabla f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}}) \|^2) \\
& + \sum_{i=1}^{n} \frac{1}{2} (\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}}) \right\|^2 + \frac{1}{L} \| \nabla f(\boldsymbol{w}_{k,1}) - \nabla f(\boldsymbol{w}_{k-1,1}) \|^2) \\
\leq & \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{k-1,1})||\partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})||\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}} \right) \\
& + \sum_{i=1}^{n} \frac{1}{2} \left(\eta L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \frac{2\eta n L}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \sum_{j=1}^{n} \| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \|^2 \right) \\
& + \sum_{i=1}^{n} \frac{1}{2} \left(\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}}) \right\|^2 + \left(1 + \frac{2\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \right) nL \sum_{j=1}^{n} \| \boldsymbol{w}_{k-1,j+1} - \boldsymbol{w}_{k-1,j} \|^2 \right) \\
\overset{(\circ)}{\leq} & \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{k-1,1})||\partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k-1,\tilde{\tau}_{k-1,i},l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})||\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}} \right) \\
& + \sum_{i=1}^{n} \frac{1}{2} \left(\eta L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \frac{2\eta n L}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \sum_{j=1}^{n} \| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \|^2 \right) \\
& + \sum_{i=1}^{n} \frac{1}{2} \left(\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}}) \right\|^2 + \left(1 + \frac{2\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \right) nL \sum_{j=1}^{n} \| \boldsymbol{w}_{k-1,j+1} - \boldsymbol{w}_{k-1,j} \|^2 \right),
\end{aligned}
$$

where Inequality $(\circ)$ uses the non-increasing property of $\boldsymbol{\nu}_{k,i,l}$ with respect to i.

When $k = 1$, we directly bound $|F_2|$ by

$$
|F_2| \leq \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{k,1})||\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})||\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}} \right).
$$

Meanwhile, F_3 can be bounded by

$$\begin{aligned}
|F_3| \leq& \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} |\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) - g_{k,i,l}| \\
=& \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} |\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) - \partial_l f_{\tau_{k,i}}(\boldsymbol{w}_{k,i})| \\
\leq& \sum_{i=1}^{n} (\frac{1}{4}\eta L \| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \|^2 + \frac{\eta}{L \min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \| \nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,1}) - \nabla f_{\tau_{k,i}}(\boldsymbol{w}_{k,i}) \|^2) \\
\leq& \frac{n}{4} \eta L \| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \|^2 + \sum_{i=1}^{n} \frac{\eta n L}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \sum_{j=1}^{n} \| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \|^2 \\
=& \frac{n}{4} \eta L \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \frac{\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \sum_{j=1}^{n} n^2 L \| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \|^2.
\end{aligned}$$

As a conclusion, when $k \geq 2$, the first order term can be bounded as

$$\begin{aligned}
&\langle \nabla f(\boldsymbol{w}_{k,1}), \boldsymbol{w}_{k+1} - \boldsymbol{w}_{k,1} \rangle \leq F_1 + |F_2| + |F_3| \\
\leq& - n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \eta \sum_{l=1}^{d} \sum_{i=1}^{n} (\frac{|\partial_l f(\boldsymbol{w}_{k-1,1})| |\partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k-1,\tilde{\tau}_{k-1,i},l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}}) \\
&+ \frac{3}{4} n\eta L \| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \|^2 + \sum_{i=1}^{n} \frac{1}{2} \eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}}) \right\|^2 \\
&+ \left(\frac{1}{2} + \frac{\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \right) n^2 L \sum_{j=1}^{n} \| \boldsymbol{w}_{k-1,j+1} - \boldsymbol{w}_{k-1,j} \|^2 + \frac{2\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} n^2 L \sum_{j=1}^{n} \| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \|^2.
\end{aligned}$$

When $k = 1$, the first order term can be bounded as

$$\begin{aligned}
&\langle \nabla f(\boldsymbol{w}_{k,1}), \boldsymbol{w}_{k+1} - \boldsymbol{w}_{k,1} \rangle \leq F_1 + |F_2| + |F_3| \\
\leq& - n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \eta \sum_{l=1}^{d} \sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})| |\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}} \right) \\
&+ \frac{n}{2} \eta L \| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \|^2 + \frac{\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}} \sum_{j=1}^{n} n^2 L \| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \|^2.
\end{aligned}$$

Besides, the second order term can be bounded as

$$\frac{L}{2} \| \boldsymbol{w}_{k+1,1} - \boldsymbol{w}_{k,1} \|^2 \leq \frac{nL}{2} \sum_{j=1}^{n} \| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \|^2.$$

Combining the estimations of the first order term and the second order term together into the descent lemma and summing over k then leads to

$$
\begin{aligned}
& f(\boldsymbol{w}_{T+1,1}) \\
\leq & f(\boldsymbol{w}_{1,1}) - \frac{1}{4}\sum_{k=1}^{T} n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 \\
& + \eta \sum_{l=1}^{d}\sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{1,1})||\partial_l f_i(\boldsymbol{w}_{1,\tilde{\tau}_{1,i}})|}{\sqrt{\boldsymbol{\nu}_{1,1,l}}} - \frac{|\partial_l f(\boldsymbol{w}_{1,1})||\partial_l f_i(\boldsymbol{w}_{1,\tilde{\tau}_{1,i}})|}{\sqrt{\boldsymbol{\nu}_{1,\tilde{\tau}_{1,i},l}}} \right) \\
& + \eta \sum_{k=2}^{T}\sum_{l=1}^{d}\sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{k-1,1})||\partial_l f_i(\boldsymbol{w}_{k-1,\tilde{\tau}_{k-1,i}})|}{\sqrt{\boldsymbol{\nu}_{k-1,\tilde{\tau}_{k-1,i},l}}} - \frac{|\partial_l f(\boldsymbol{w}_{k,1})||\partial_l f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})|}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i},l}}} \right) \\
& + \sum_{k=1}^{T}\sum_{i=1}^{n} \frac{1}{2}\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,1}}} \odot \nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) \right\|^2 + \left(1 + \frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) n^2 L \sum_{k=1}^{T}\sum_{j=1}^{n} \|\boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j}\|^2 \\
= & f(\boldsymbol{w}_{1,1}) - \sum_{k=1}^{T} n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 \\
& + \eta \sum_{l=1}^{d}\sum_{i=1}^{n} \left(\frac{|\partial_l f(\boldsymbol{w}_{1,1})||\partial_l f_i(\boldsymbol{w}_{1,\tilde{\tau}_{1,i}})|}{\sqrt{\boldsymbol{\nu}_{1,1,l}}} - \frac{|\partial_l f(\boldsymbol{w}_{T,1})||\partial_l f_i(\boldsymbol{w}_{T,\tilde{\tau}_{T,i}})|}{\sqrt{\boldsymbol{\nu}_{T,\tilde{\tau}_{T,i},l}}} \right) \\
& + \sum_{k=1}^{T}\sum_{i=1}^{n} \frac{1}{2}\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k+1,1}}} \odot \nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) \right\|^2 + \left(1 + \frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) n^2 L \sum_{k=1}^{T}\sum_{j=1}^{n} \|\boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j}\|^2.
\end{aligned}
$$

We respectively bound the last two terms.

The term $\sum_{k=1}^{T}\sum_{i=1}^{n}\frac{1}{2}\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k+1,1}}} \odot \nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) \right\|^2$ can be bounded as

$$
\begin{aligned}
& \sum_{k=1}^{T}\sum_{i=1}^{n} \frac{1}{2}\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k+1,1}}} \odot \nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) \right\|^2 \leq \sum_{k=1}^{T}\sum_{i=1}^{n} \frac{1}{2}\eta^2 L \left\| \frac{1}{\sqrt{\boldsymbol{\nu}_{k,\tilde{\tau}_{k,i}}}} \odot \nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) \right\|^2 \\
= & \sum_{k=1}^{T}\sum_{i=1}^{n} \frac{1}{2} L \left\| \boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j} \right\|^2. \qquad (14)
\end{aligned}
$$

Thus, we only need to bound the last term in order to bound the first two terms. The third term $\sum_{k=1}^{T}\left(1 + \frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) nL^2 \sum_{j=1}^{n} \|\boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j}\|^2$ can be bounded as

$$
\begin{aligned}
\sum_{k=1}^{T}\left(1 + \frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) nL^2 \sum_{j=1}^{n} \|\boldsymbol{w}_{k,j+1} - \boldsymbol{w}_{k,j}\|^2 = & \left(1 + \frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) nL^2 \sum_{k=1}^{T}\sum_{j=1}^{n}\sum_{l=1}^{d} \frac{g_{k,i,l}^2}{\boldsymbol{\nu}_{k,i,l}} \\
\leq & \left(1 + \frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) nL^2 \sum_{l=1}^{d} \ln \frac{\boldsymbol{\nu}_{T,n,l}}{\boldsymbol{\nu}_{1,0,l}}.
\end{aligned}
$$

Combining the above estimations together, we obtain

$$\begin{aligned} f(\boldsymbol{w}_{T+1,1}) \leq& f(\boldsymbol{w}_{1,1}) - \frac{1}{4}\sum_{k=1}^{T} n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 + \eta \sum_{l=1}^{d}\sum_{i=1}^{n} \frac{|\partial_l f(\boldsymbol{w}_{1,1})||\partial_l f_i(\boldsymbol{w}_{1,\tilde{\tau}_{1,i}})|}{\sqrt{\boldsymbol{\nu}_{1,1,l}}} \\ &+ \left(\left(1+\frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) nL^2 + \frac{1}{2}L\right)\sum_{l=1}^{d}(\ln \boldsymbol{\nu}_{T,n,l} - \ln \boldsymbol{\nu}_{1,0,l}). \end{aligned} \tag{15}$$

Define

$$C_0 \triangleq f(\boldsymbol{w}_{1,1}) - f(\boldsymbol{w}_{T+1,1}) + \eta \sum_{l=1}^{d}\sum_{i=1}^{n} \frac{|\partial_l f(\boldsymbol{w}_{1,1})||\partial_l f_i(\boldsymbol{w}_{1,\tilde{\tau}_{1,i}})|}{\sqrt{\boldsymbol{\nu}_{1,1,l}}} + \left(\left(1+\frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) nL^2 + \frac{1}{2}L\right)\sum_{l=1}^{d} \ln \boldsymbol{\nu}_{1,0,l},$$

$$C_1 \triangleq \left(1+\frac{3\eta}{\min_l \sqrt{\boldsymbol{\nu}_{1,0,l}}}\right) nL^2 + \frac{1}{2}L.$$

Based on these notations, Eq. (15) can be abbreviated as

$$\sum_{k=1}^{T} n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 \leq 4C_0 + 4C_1 \sum_{l=1}^{d} \ln \boldsymbol{\nu}_{T,n,l}. \tag{16}$$

Since

$$\left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 = \sum_{l=1}^{d} \frac{\partial_l f(\boldsymbol{w}_{k,1})^2}{\sqrt{\boldsymbol{\nu}_{k,1,l}}} \geq \sum_{l=1}^{d} \frac{\partial_l f(\boldsymbol{w}_{k,1})^2}{\sqrt{\sum_{l'=1}^{d} \boldsymbol{\nu}_{k,1,l'}}} = \frac{\|\nabla f(\boldsymbol{w}_{k,1})\|^2}{\sqrt{\sum_{l'=1}^{d} \boldsymbol{\nu}_{k,1,l'}}},$$

the LHS of Eq. (16) can be lower bounded as

$$\begin{aligned} &\sum_{k=1}^{T} n\eta \left\| \frac{1}{\sqrt[4]{\boldsymbol{\nu}_{k,1}}} \odot \nabla f(\boldsymbol{w}_{k,1}) \right\|^2 \geq \sum_{k=1}^{T} n\eta \frac{\|\nabla f(\boldsymbol{w}_{k,1})\|^2}{\sqrt{\sum_{l'=1}^{d} \boldsymbol{\nu}_{k,1,l'}}} \geq n\eta \frac{\sum_{k=1}^{T}\|\nabla f(\boldsymbol{w}_{k,1})\|^2}{\sqrt{\sum_{l'=1}^{d} \boldsymbol{\nu}_{T,1,l'}}} \\ &\overset{(\star)}{\geq} n\eta \frac{\frac{1}{nD_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{w}_{k,1})\|^2 - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d} \boldsymbol{\nu}_{T,1,l'}}} \\ &\geq n\eta \frac{\frac{1}{2nD_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})\|^2 - \frac{1}{nD_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}}) - \nabla f_i(\boldsymbol{w}_{k,1})\|^2 - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d} \boldsymbol{\nu}_{T,1,l'}}} \\ &\geq n\eta \frac{\frac{1}{2nD_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})\|^2 - \frac{L^2}{nD_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\|\boldsymbol{w}_{k,\tilde{\tau}_{k,i}} - \boldsymbol{w}_{k,1}\|^2 - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d} \boldsymbol{\nu}_{T,1,l'}}} \end{aligned}$$

$$\geq n\eta \frac{\frac{1}{2nD_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})\|^2 - \frac{L^2}{D_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{w}_{k,j+1}-\boldsymbol{w}_{k,j}\|^2 - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,1,l'}}}$$

$$= n\eta \frac{\frac{1}{2nD_1}\sum_{k=1}^{T}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{w}_{k,\tilde{\tau}_{k,i}})\|^2 - \frac{nL^2}{D_1}\sum_{k=1}^{T}\sum_{j=1}^{n}\|\boldsymbol{w}_{k,j+1}-\boldsymbol{w}_{k,j}\|^2 - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,1,l'}}}$$

$$\geq n\eta \frac{\frac{1}{2nD_1}(\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,1,l'} - \sum_{l'=1}^{d}\boldsymbol{\nu}_{1,0,l'}) - \frac{nL^2}{D_1}\sum_{k=1}^{T}\sum_{j=1}^{n}\|\boldsymbol{w}_{k,j+1}-\boldsymbol{w}_{k,j}\|^2 - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,1,l'}}}.$$

As a conclusion,

$$n\eta \frac{\frac{1}{2nD_1}(\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,n,l'} - \sum_{l'=1}^{d}\boldsymbol{\nu}_{1,0,l'}) - \frac{nL^2}{D_1}\sum_{k=1}^{T}\sum_{j=1}^{n}\|\boldsymbol{w}_{k,j+1}-\boldsymbol{w}_{k,j}\|^2 - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,1,l'}}}$$

$$\leq 4C_0 + 4C_1\sum_{l=1}^{d}\ln \boldsymbol{\nu}_{T,n,l}.$$

Echoing Eq. (14), we then obtain

$$n\eta \frac{\frac{1}{2nD_1}(\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,n,l'} - \sum_{l'=1}^{d}\boldsymbol{\nu}_{1,0,l'}) - \frac{nL^2}{D_1}\eta^2\sum_{l=1}^{d}\ln\frac{\boldsymbol{\nu}_{T,n,l}}{\boldsymbol{\nu}_{1,0,l}} - \frac{D_0}{D_1}T}{\sqrt{\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,1,l'}}}$$

$$\leq 4C_0 + 4C_1\sum_{l=1}^{d}\ln \boldsymbol{\nu}_{T,n,l}.$$

Solving the above inequality with respect to $\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,n,l'}$ leads to

$$\sum_{l'=1}^{d}\boldsymbol{\nu}_{T,n,l'} = \mathcal{O}(1) + \mathcal{O}(D_0 T).$$

Applying the above inequality back to Eq. (16) completes the proof. ■

Appendix D. Analysis under (L_0, L_1)-smooth condition

D.1. Proof of Theorem 8

Before the formal proof of Theorem 8, we first introduce the version of descent lemma under (L_0, L_1)-smooth condition.

Lemma 11 (Descent lemma under (L_0, L_1)-smooth condition) *Let Assumption 6 holds. Then, if $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ satisfies $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \leq \frac{1}{L_1}$, then*

$$f(\boldsymbol{w}_1) \leq f(\boldsymbol{w}_2) + \langle \nabla f(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle + \frac{(L_0 + L_1\|\nabla f(\boldsymbol{w}_2)\|)}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2.$$

The proof bears great similarity to that of the descent lemma under L-smooth condition and we omit it here. Interested readers can refer to (Zhang et al., 2020, Lemma A.3) for details.

Proof [Proof of Theorem 8] Since we set $\eta \le \frac{1}{L_1}$ and $\|\frac{\nabla f(\boldsymbol{w}_t)}{\sqrt{\boldsymbol{\nu}_t}}\| \le 1$, we have

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| = \eta \left\| \frac{\nabla f(\boldsymbol{w}_t)}{\sqrt{\boldsymbol{\nu}_t}} \right\| \le \frac{1}{L_1}.$$

Therefore, Lemma 11 can be applied to $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}$, taking expectation to which leads to

$$\mathbb{E}_{\mathcal{F}_t} f(\boldsymbol{w}_{t+1}) \le f(\boldsymbol{w}_t) + \underbrace{\mathbb{E}_{\mathcal{F}_t}\langle \nabla f(\boldsymbol{w}_t), -\eta \frac{\nabla f(\boldsymbol{w}_t)}{\sqrt{\boldsymbol{\nu}_t}}\rangle}_{\text{First Order}} + \underbrace{\frac{L_0 + L_1\|\nabla f(\boldsymbol{w}_t)\|}{2}\eta^2 \mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2}_{\text{Second Order}}.$$

We then decompose the "First Order" term into the "First Order Main" term and the "Error" term as follows.

$$\begin{aligned}
\mathbb{E}_{\mathcal{F}_t} f(\boldsymbol{w}_{t+1}) \le & f(\boldsymbol{w}_t) + \underbrace{-\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}}_{\text{First Order Main}} + \underbrace{\mathbb{E}_{\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}})g_t \right\rangle\right]}_{\text{Error}} \\
& + \underbrace{\frac{L_0 + L_1\|\nabla f(\boldsymbol{w}_t)\|}{2}\eta^2 \mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2}_{\text{Second Order}} \\
\le & f(\boldsymbol{w}_t) + \underbrace{-\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}}_{\text{First Order Main}} + \underbrace{\mathbb{E}_{\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}})g_t \right\rangle\right]}_{\text{Error}} \\
& + \underbrace{\frac{L_0}{2}\eta^2 \mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2}_{\text{Second Order I}} + \underbrace{\frac{L_1\|\nabla f(\boldsymbol{w}_t)\|}{2}\eta^2 \mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2}_{\text{Second Order II}}.
\end{aligned} \tag{17}$$

Here the second inequality is obtained by directly expanding the "Second Order" term. The "Second Order I" term takes the same form of the "Second Order" term in the proof of Theorem 2, and it can be handled in the same way as the proof of Theorem 2. So does the "First Order Main" term. As for the "Error" term, following the same routine as the proof of Lemma 4, we obtain that

$$\left|\mathbb{E}_{\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}})g_t \right\rangle\right]\right| \le \eta \frac{\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}} \mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}}}\right].$$

Meanwhile, since $\boldsymbol{\nu}_t$ is non-decreasing with respect to t, we have that

$$\frac{L_1\|\nabla f(\boldsymbol{w}_t)\|}{2}\eta^2 \mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2 \le L_1\eta^2 \frac{\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}} \mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t} + \sqrt{\boldsymbol{\nu}_{t-1}}}\right].$$

One can easily observe that the RHSs of the above two inequalities have the same form except the coefficients! Thus, the "Error" term plus the "Second Order II" can be bounded by

$$
\begin{aligned}
&\underbrace{\mathbb{E}_{\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}}) g_t\right\rangle\right]}_{\text{Error}}+\underbrace{\frac{L_1\|\nabla f(\boldsymbol{w}_t)\|}{2}\eta^2\mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2}_{\text{Second Order II}}\\
\le&(\eta+L_1\eta^2)\frac{\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}}}\right]\le 2\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}}}\right],
\end{aligned}
$$

where in the last inequality we use $\eta L_1 \le 1$. By the mean-value inequality ($2ab \le a^2+b^2$),

$$
\begin{aligned}
2\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}}}\right]\le&\frac{1}{2}\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}+2\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}}\left(\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}}}\right]\right)^2\\
\le&\frac{1}{2}\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}+2\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}_{\mathcal{F}_t}\|g_t\|^2\cdot\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right]\\
\le&\frac{1}{2}\eta\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}+2\frac{\eta}{\sqrt{\boldsymbol{\nu}_{t-1}}}(D_0+D_1\|\nabla f(\boldsymbol{w}_t)\|^2)\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right],
\end{aligned}
$$

where the second inequality is due to Hölder's inequality, and in the last inequality we use Assumption 1. Similar to the proof of Theorem 2, we focus on the term $2D_1\|\nabla f(\boldsymbol{w}_t)\|^2\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right]$, since the rest of the terms can be easily bounded. Such a term can be bounded as

$$
\begin{aligned}
&2\frac{\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\mathbb{E}_{\mathcal{F}_t}\left[\frac{\|g_t\|^2}{(\sqrt{\boldsymbol{\nu}_t}+\sqrt{\boldsymbol{\nu}_{t-1}})^2}\right]\\
\le&2\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2\mathbb{E}_{\mathcal{F}_t}\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)\\
=&2\eta D_1\mathbb{E}_{\mathcal{F}_t}\left(\frac{\|\nabla f(\boldsymbol{w}_{t-1})\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}\right)+2\eta D_1\frac{\|\nabla f(\boldsymbol{w}_t)\|^2-\|\nabla f(\boldsymbol{w}_{t-1})\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}.
\end{aligned}
$$

By Assumption 6, $\|\nabla f(\boldsymbol{w}_t)\|-\|\nabla f(\boldsymbol{w}_{t-1})\|\le\|\nabla f(\boldsymbol{w}_t)-\nabla f(\boldsymbol{w}_{t-1})\|\le(L_0+L_1\|\nabla f(\boldsymbol{w}_t)\|)\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|$. Therefore,

$$
\begin{aligned}
&2\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2\mathbb{E}_{\mathcal{F}_t}\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)\\
=&2\eta D_1\frac{2(L_0+L_1\|\nabla f(\boldsymbol{w}_t)\|)\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|+(L_0+L_1\|\nabla f(\boldsymbol{w}_t)\|)^2\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\\
&+2\eta D_1\mathbb{E}_{\mathcal{F}_t}\left(\frac{\|\nabla f(\boldsymbol{w}_{t-1})\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}-\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}\right)\\
\le&2\eta D_1\frac{2(L_0+L_1\|\nabla f(\boldsymbol{w}_t)\|)\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|+2(L_0^2+L_1^2\|\nabla f(\boldsymbol{w}_t)\|^2)\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}\\
&+2\eta D_1\mathbb{E}_{\mathcal{F}_t}\left(\xi(t-1)-\xi(t)\right).
\end{aligned}
$$

where the last inequality we use $\xi(t) \triangleq \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_t}}$ and the mean-value inequality $(L_0+L_1\|\nabla f(\boldsymbol{w}_t)\|)^2 \le 2(L_0^2 + L_1^2\|\nabla f(\boldsymbol{w}_t)\|^2)$. Reorganize the RHS of the above inequality then leads to

$$\begin{aligned}
&2\eta D_1 \frac{2(L_0 + L_1\|\nabla f(\boldsymbol{w}_t)\|)\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\| + 2(L_0^2 + L_1^2\|\nabla f(\boldsymbol{w}_t)\|^2)\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \\
&+ 2\eta D_1 \mathbb{E}_{\mathcal{F}_t}\left(\xi(t-1) - \xi(t)\right) \\
=&4\eta D_1 L_0 \frac{\|w_t - w_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 4\eta D_1 L_1 \frac{\|w_t - w_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 4L_0^2\eta D_1 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \\
&+ 4L_1^2\eta D_1 \frac{\|w_t - w_{t-1}\|^2\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 2\eta D_1 \mathbb{E}_{\mathcal{F}_t}\left(\xi(t-1) - \xi(t)\right).
\end{aligned}$$

In the RHS of the above inequality, the first term can be bounded by the mean-value inequality as

$$4\eta D_1 L_0 \frac{\|w_t - w_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|}{\sqrt{\boldsymbol{\nu}_{t-1}}} \le \frac{\eta}{4}\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 16 D_1^2 L_0^2 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}.$$

By $\eta \le \frac{1}{64 D_1 L_1}$, the second term can be bounded as

$$4\eta D_1 L_1 \frac{\|w_t - w_{t-1}\|\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \le \frac{1}{16}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}.$$

By $\eta \le \frac{1}{8\sqrt{D_1} L_1}$, the forth term can be bounded as

$$4\eta D_1 L_1^2 \frac{\|w_t - w_{t-1}\|^2\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \le \frac{1}{16}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}.$$

Therefore, $2\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2 \mathbb{E}_{\mathcal{F}_t}\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right)$ can be bounded as

$$\begin{aligned}
&2\eta D_1\|\nabla f(\boldsymbol{w}_t)\|^2 \mathbb{E}_{\mathcal{F}_t}\left(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}}\right) \\
\le&\frac{3}{8}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 4L_0^2\eta D_1 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 16 D_1^2 L_0^2 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \\
&+ 2\eta D_1 \mathbb{E}_{\mathcal{F}_t}\left(\xi(t-1) - \xi(t)\right)
\end{aligned}$$

All in all, we conclude that the "Error" term plus the "Second Order II" term can be bounded as

$$\begin{aligned}
&\underbrace{\mathbb{E}_{\mathcal{F}_t}\left[\left\langle \nabla f(\boldsymbol{w}_t), \eta(\frac{1}{\sqrt{\boldsymbol{\nu}_{t-1}}} - \frac{1}{\sqrt{\boldsymbol{\nu}_t}})g_t \right\rangle\right]}_{\text{Error}} + \underbrace{\frac{L_1\|\nabla f(\boldsymbol{w}_t)\|}{2}\eta^2 \mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2}_{\text{Second Order II}} \\
\le&\frac{7}{8}\eta \frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 4L_0^2\eta D_1 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 16 D_1^2 L_0^2 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} \\
&+ 2\eta D_1 \mathbb{E}_{\mathcal{F}_t}\left(\xi(t-1) - \xi(t)\right).
\end{aligned}$$

Applying the above inequality back to Eq. (17), we conclude that

$$\mathbb{E}_{\mathcal{F}_t} f(\boldsymbol{w}_{t+1}) \leq f(\boldsymbol{w}_t) - \frac{\eta}{8}\frac{\|\nabla f(\boldsymbol{w}_t)\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + \frac{L_0}{2}\eta^2 \mathbb{E}_{\mathcal{F}_t}\left\|\frac{g_t}{\sqrt{\boldsymbol{\nu}_t}}\right\|^2 + 4L_0^2\eta D_1 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}}$$
$$+ 16D_1^2 L_0^2 \frac{\|w_t - w_{t-1}\|^2}{\sqrt{\boldsymbol{\nu}_{t-1}}} + 2\eta D_1 \mathbb{E}_{\mathcal{F}_t}\left(\xi(t-1) - \xi(t)\right).$$

The above inequality takes the same form of Eq. (8) except the coefficients. The rest of the proof follow the same routine as the rest of the proof of Theorem 2.

The proof is completed. ■

D.2. Proof of Theorem 9

Before the formal proof, we first present some notations.
Constants. $\forall t \in \mathbb{Z}^+$, we define $a_t = \eta \frac{2^t}{1+\cdots+4^t} = \eta \frac{2^t}{\sqrt{\frac{4^{t+1}-1}{3}}}$. One can easily observe that a_t is increasing with respect to t and thus $a_t \geq a_1 = \eta\frac{2}{\sqrt{5}} > \frac{9}{L_1}$, where the last inequality we use $\eta > \frac{4\sqrt{5}}{L_1}$. We then define $S_{2k-1} \triangleq \sum_{t=1}^{k} a_{2t-1}$ and $S_{2k} \triangleq \sum_{t=1}^{k} a_{2t}$ with $S_0 = 0$ and $S_{-1} = 0$.
Line segments. We define the following line segments: $\mathcal{C}_{2k-1} = \{(x, S_{2k-2}) : x \in [S_{2k-3}, S_{2k-1})\}$, and $\mathcal{C}_{2k} = \{(S_{2k-1}, y) : y \in [S_{2k-2}, S_{2k})\}$. Define $\mathcal{C} = \cup_{k=1}^{\infty}\mathcal{C}_k$.
Proof [Proof of Theorem 9] We first define a function over $\mathcal{C}$ as follows. Define $f((0,0)) = \frac{2}{L_1}$, and

$$\nabla f|_{\mathcal{C}_{2k-1}}(x, S_{2k-2}) = \begin{cases} (2^{2k-1}, 0) + L_1(x - S_{2k-3})(-2^{2k-2}, 2^{2k-2}),\ x \in [S_{2k-3}, S_{2k-3} + \frac{4}{L_1}) \\ (-2^{2k-1}, 2^{2k}) + (x - S_{2k-3} - \frac{4}{L_1})(-2^{2k-2}, 0),\ x \in [S_{2k-3} + \frac{4}{L_1}, S_{2k-3} + \frac{8}{L_1}) \\ (0, 2^{2k}),\ x \in [S_{2k-3} + \frac{8}{L_1}, S_{2k-1}) \end{cases}.$$

$$\nabla f|_{\mathcal{C}_{2k}}(S_{2k}, y) = \begin{cases} (0, 2^{2k}) + L_1(y - S_{2k-2})(2^{2k-1}, -2^{2k-1}),\ y \in [S_{2k-2}, S_{2k-2} + \frac{4}{L_1}) \\ (2^{2k+1}, -2^{2k}) + (y - S_{2k-2} - \frac{4}{L_1})(0, -2^{2k-1}),\ y \in [S_{2k-2} + \frac{4}{L_1}, S_{2k-2} + \frac{8}{L_1}) \\ (2^{2k+1}, 0),\ y \in [S_{2k-2} + \frac{8}{L_1}, S_{2k}) \end{cases}.$$

To begin with, one can easily observe that $\nabla f((S_{2k-1}, S_{2k-2})) = (0, 2^{2k})$ and $\nabla f((S_{2k-1}, S_{2k})) = (2^{2k+1}, 0)$. We then prove that $f|_{\mathcal{C}}$ obeys $(0, L_1)$-smooth condition: as the length of $\mathcal{C}_k$ is a_k, which is longer than $\frac{8}{L_1}$, if $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ belong to $\mathcal{C}$ and satisfy $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \leq \frac{1}{L_1}$, it must be that either there exists a k, such that $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{C}_k$, or there exists a k such that $\boldsymbol{w}_1 \in \mathcal{C}_k$ and $\boldsymbol{w}_2 \in \mathcal{C}_{k+1}$ (or there exists a k such that $\boldsymbol{w}_1 \in \mathcal{C}_{k+1}$ and $\boldsymbol{w}_2 \in \mathcal{C}_k$, which can be tackled in the same way due to the symmetry between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$). For the former case, without the loss of generality we assume that k is odd and equals $2k' - 1$. We then observe that the gradient norm is no smaller than $2^{2k'-2}\sqrt{2}$ when $\boldsymbol{w} \in [S_{2k'-3}, S_{2k'-3} + \frac{4}{L_1}] \times \{S_{2k'-2}\}$, while the norm of directional Hessian is $L_1 2^{2k'-2}\sqrt{2}$. Similarly, we observe that the gradient norm is no smaller than $2^{2k'}$ when

$\boldsymbol{w} \in [S_{2k'-3} + \frac{4}{L_1}, S_{2k'-1}] \times \{S_{2k'-2}\}$, while the norm of directional Hessian is no larger than $L_1 2^{2k'-2}$. Therefore, we conclude that $\|\nabla f(\boldsymbol{w}_1) - \nabla f(\boldsymbol{w}_2)\| \le L_1 \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \|\nabla f(\boldsymbol{w}_1)\|$. As for the latter case, without the loss of generality we still assume $k = 2k' - 1$, and $\boldsymbol{w}_1 \in \mathcal{C}_{2k'-1}$ and $\boldsymbol{w}_2 \in \mathcal{C}_{2k'}$. Since $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \le \frac{1}{L_1}$, we must have $\boldsymbol{w}_1 \in [S_{2k'-3} + \frac{8}{L_1}, S_{2k'-1}] \times \{S_{2k'-2}\}$. Thus,

$$
\begin{aligned}
&\|\nabla f(\boldsymbol{w}_1) - \nabla f(\boldsymbol{w}_2)\| \\
=&\|\nabla f((S_{2k'-1}, S_{2k'})) - \nabla f(\boldsymbol{w}_2)\| \\
\le& L_1 \|\nabla f(\boldsymbol{w}_2)\| \|(S_{2k'-1}, S_{2k'}) - \boldsymbol{w}_2\| \le L_1 \min\{\|\nabla f(\boldsymbol{w}_2)\|, \|\nabla f(\boldsymbol{w}_1)\|\} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|.
\end{aligned}
$$

This proves that $f|_{\mathcal{C}}$ obeys $(0, L_1)$-smooth condition.

Also, by induction, one can show that $f|_{\mathcal{C}} \ge 0$. Together, we can then interpolate $f|_{\mathcal{C}}$ to get f, which satisfies Assumption 6 and $f \ge 0$. Furthermore, running AdaGrad-Norm on f with the initial point $\boldsymbol{w}_1 = (0, 0)$ leads to that $\boldsymbol{w}_{2k-1} = (S_{2k-3}, S_{2k-2})$ and $\boldsymbol{w}_{2k} = (S_{2k-1}, S_{2k-2})$, and $\|\nabla f(\boldsymbol{w}_k)\| = 2^k$.

The proof is completed. ■