ON THE CONVERGENCE OF ADAGRAD-NORM FOR NON-CONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive optimizers have achieved significant success in deep learning by dynamically adjusting the learning rate based on iterative gradients. Compared to stochastic gradient descent (SGD), adaptive optimizers converge much faster in various deep-learning tasks. However, as a fundamental adaptive optimizer, the theoretical analysis of AdaGrad-Norm is inadequate, and there are many technical challenges regarding last-iterate convergence and average-iterate convergence rates for general non-convex loss functions. This paper aims to address these limitations and provides a comprehensive analysis of AdaGrad-Norm. We propose novel techniques that avoid the assumption of no saddle points and derive last-iterate convergence for both almost surely and mean-square senses. Furthermore, under milder conditions, we obtain the near-optimal and sub-optimal rates w.r.t averaged iterate in the expected sense and the almost surely sense, respectively. We relax one restrictive assumption of the uniformly bounded stochastic gradient used in existing high-probability convergence analysis. Moreover, the methodologies provided in this paper have the potential to contribute to further research on the convergence properties of other stochastic algorithms.

1 INTRODUCTION

Adaptive gradient methods (Duchi et al., 2011; Kingma & Ba, 2015) have achieved tremendous success in many fields of machine learning. It is observed that adaptive optimizers can achieve better performance than vanilla stochastic gradient descent (SGD) (Vaswani et al., 2017; Duchi et al., 2013; Lacroix et al., 2018; Dosovitskiy et al., 2021) in nonconvex optimization, thus become popular in deep learning. The intuitive explanation of its superiority compared to SGD is that the adaptive optimizers automatically adjust the learning rate based on past stochastic gradients.

AdaGrad (Duchi et al., 2011; McMahan & Streeter, 2010), as a fundamental adaptive learning rate algorithm, has attracted significant research attention in recent years. The norm version of AdaGrad (i.e., AdaGrad-Norm) as a single stepsize adaptation method is described as follows:

$$S_n = S_{n-1} + \left\|\nabla g(\theta_n, \xi_n)\right\|^2, \quad \theta_{n+1} = \theta_n - \frac{\alpha_0}{\sqrt{S_n}}\nabla g(\theta_n, \xi_n), \tag{1}$$

where $g(\theta)$ ($\theta \in \mathbb{R}^d$) is the loss function, $S_0 \geq 0$ is a pre-determined constant, α_0 is a positive constant, and $\nabla g(\theta, \xi_n)$ denotes an unbiased estimate of $\nabla g(\theta)$, i.e., $\mathbb{E}_{\xi_n}[\nabla g(\theta, \xi_n) \mid \mathcal{F}_n] = \nabla g(\theta)$ and the sequence $\{\xi_n\}$ is a sequence of independent random variables. We define a σ-filtration $\mathcal{F}_n := \sigma\{\theta_1, \xi_1, \xi_2, ..., \xi_{n-1}\}$. Despite its simple structure, the convergence results of AdaGrad-Norm on non-convex optimization are sparse and far from satisfactory, especially in the last-iterate sense and the average-iterate sense.

Jin et al. (2022) established almost surely convergence of AdaGrad-Norm in the sense of last-iterate, but heavily relied on the unrealistic assumption that the loss function does not have saddle points. This assumption does not hold in most deep learning applications, for example for neural networks with hidden layers. As a result, the analysis provided in Jin et al. (2022) is not applicable to general nonconvex loss functions with saddle points. It is crucial to explore alternative approaches for analyzing the convergence of AdaGrad-Norm in more general scenarios.

In terms of average-iterate convergence rate, most existing theoretical results for AdaGrad-Norm are based on strong assumptions (Ward et al., 2020; Défossez et al., 2020). For example, Ward et al. (2020); Défossez et al. (2020) assumed the uniform upper bound for all stochastic gradients. To the best of our knowledge, this assumption is often violated when stochastic gradients contain Gaussian random noise. Furthermore, even the conventional mini-batch stochastic gradient fails to satisfy this assumption when the loss function is quadratic (Wang et al., 2023). Recently, Faw et al. (2022); Wang et al. (2023) have removed the uniform boundedness assumption of stochastic gradients, however, they only achieve the convergence rates in the high-probability sense.

The goal of this paper is to address the limitations of existing results and provide a comprehensive analysis of the convergence properties of AdaGrad-Norm for general non-convex loss functions.

Technological Challenges. Despite the inherent simplicity structure of AdaGrad-Norm, investigating its convergence and convergence rate under general conditions poses a significant challenge. In this regard, we will highlight several major obstacles, of which only the first one has been effectively tackled in previous research.

(1) Learning rate $\alpha_0/\sqrt{S_n}$ and stochastic gradient $\nabla g(\theta_n, \xi_n)$ in AdaGrad-Norm are conditionally dependent on the σ-filtration $\mathcal{F}_n$. i.e., we cannot replace $\mathbb{E}\left(\frac{\nabla g(\theta_n)^\top \nabla g(\theta_n,\xi_n)}{\sqrt{S_n}}\middle|\mathcal{F}_n\right)$ with $\frac{\|\nabla g(\theta_n)\|^2}{\sqrt{S_n}}$. This challenge has been effectively resolved in (Jin et al., 2022; Faw et al., 2022; Wang et al., 2023). Faw et al. (2022) addressed this issue by scaling down $1/\sqrt{S_n}$ to $1/\sqrt{S_{n-1} + \|\nabla g(\theta_n)\|^2}$. In Jin et al. (2022); Wang et al. (2023), authors transformed $1/\sqrt{S_n}$ into $1/\sqrt{S_{n-1}} + 1/\sqrt{S_{n-1}} - 1/\sqrt{S_n}$ to obtain a new recurrence relation, where the conditional dependence issue no longer exists. The technique employed in Jin et al. (2022) to solve this issue is also utilized in the proof of this paper.

(2) The quadratic error term $\|\nabla g(\theta_n, \xi_n)\|^2/S_n$ generated by AdaGrad-Norm does not exhibit additivity, i.e., $\sum_{n=1}^{+\infty} \|\nabla g(\theta_n, \xi_n)\|^2/S_n = \Theta(\ln S_n) = +\infty$. The traditional proofs for the almost surely convergence at the last iterate, i.e., $\lim_{n\to+\infty} \|\nabla g(\theta_n)\| = 0 \ \ a.s.$, for SGD or SGD with momentum usually requires the summability of the quadratic error term. This is why the classical Robbins-Monro condition (Robbins & Siegmund, 1971; Jin et al., 2022; Lei et al., 2005; Li & Milzarek, 2022), i.e., $\sum_{n=1}^{+\infty} \epsilon_n = +\infty, \ \sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$, where $\{\epsilon_n\}_{n=1}^{+\infty}$ is the step size of SGD, arises. Under the Robbins-Monro condition and incorporating the boundedness of $\mathbb{E}(\|\nabla g(\theta_n, \xi_n)\|^2|\mathcal{F}_n)$, it is straightforward to establish the summability of this quadratic error term. However, this is not the same for AdaGrad-Norm. Jin et al. (2022) addressed this issue but relied on the assumption of the absence of saddle points in the loss function. Their approach cannot be applied to loss functions that do have saddle points. A detailed explanation of this issue is provided in the proof sketch of Theorem 3.1 in Section 4.

(3) Demonstrating the convergence in mean square of AdaGrad-Norm with respect to the last iterate, denoted as, $\lim_{n\to+\infty} \mathbb{E}\,\|\nabla g(\theta_n)\|^2 = 0$, encounters several challenges. Typically, for traditional SGD, it is easier to prove the last-iterate mean square convergence than the last-iterate almost surely convergence. By taking the mathematical expectation on both sides of the iteration equation associated with the loss function formed by SGD, we obtain

$$\mathbb{E}(g(\theta_{n+1})) \le \mathbb{E}(g(\theta_n)) - \epsilon_n\,\mathbb{E}\,\|\nabla g(\theta_n)\|^2 + \frac{\epsilon_n^2}{2}\,\mathbb{E}\,\|\nabla g(\theta_n, \xi_n)\|^2,$$

where ϵ_n is the learning rate of SGD. Regarding $\mathbb{E}\,\|\nabla g(\theta_n)\|^2$ as a unified quantity, we convert the original stochastic dynamical system into a deterministic dynamical system. Through further analysis, it is straightforward to derive the mean square convergence result in terms of the last iterate. However, this methodology is not applicable to AdaGrad-Norm. Since the learning rate of AdaGrad-Norm is a random variable, it is not allowed to move the step size $\alpha_0/\sqrt{S_n}$ outside the expectation when computing the mathematical expectation. While we may use operations such as *Hölder's Inequality*, it will introduce two inevitable issues. First, *employing Hölder's Inequality* introduces a change in order, which shifts our object from the second moment of the gradient norm to an alternative quantity. Second, after using *Hölder's Inequality*, our learning rate term incorporates the mathematical expectation, resulting in $\alpha_0/\,\mathbb{E}\sqrt{S_n}$, which makes the traditional SGD methods no longer inapplicable. Consequently, the only option is to first prove the last-iterate almost surely convergence and then establish the last-iterate mean-square convergence through *The Lebesgue's Dominated Convergence Theorem*. In other words, we need to prove that the expectation of the uniform upper bound of the gradient norm sequence, i.e., $\mathbb{E}(\sup_{n\ge1}\{\|\nabla g(\theta_n)\|^2\})$, is bounded, which is a challenging task. Due to this inherent difficulty, so far there has been no result regarding the mean square convergence of the last iterate of AdaGrad-Norm.

Contribution. In this paper, we overcome the aforementioned challenges, propose novel techniques, and derive convergence results for both last-iterate and average-iterate under mild conditions. Specifically, we make the following contributions:

(1) For the general nonconvex problems, we propose an innovative analytical perspective and demonstrate the almost surely convergence for AdaGrad-Norm at the last iterate. Since our approach does not examine the iterative characteristics of the dynamical system when the gradients are small, the summability of the squared learning rates is not necessary in our case. As a result, we can overcome the second challenge. Furthermore, our analysis does not depend on the no saddle point assumption required in (Jin et al., 2022), which is a significant improvement.

(2) We are the first to demonstrate the last-iterate mean-square convergence of AdaGrad-Norm under mild conditions. We propose a novel approach by splitting the dynamical system of AdaGrad-Norm into multiple sub-processes via many first entrance times. In this way, the expectation of the maximum value is proved to be finite and addresses the third challenge. Our analytical methodology has the potential applications to other algorithms.

(3) Based on the first two convergence results, we obtain a more exact estimate of S_n, i.e., $\mathbb{E}\, S_n = O(n)$. Utilizing this estimate and the martingale difference estimation lemma (Lemma A.4), we prove both the almost sure and expectation convergence rates for AdaGrad-Norm at the average iterate without assuming the uniform boundedness on stochastic gradients. Our results fill the gap in existing research (Ward et al., 2020; Faw et al., 2022; Wang et al., 2023) which only achieved a high probability convergence rate.

1.1 RELATED WORKS

Both Duchi et al. (2011) and McMahan & Streeter (2010) independently proposed AdaGrad for non-convex optimization. Since then, a series of studies have emerged, analyzing the convergence of AdaGrad-Norm on non-convex landscapes (Ward et al., 2020; Li & Orabona, 2019; Zou et al., 2018; Li & Orabona, 2020; Gadat & Gavra, 2020; Défossez et al., 2020; Kavis et al., 2022; Liu et al., 2022; Faw et al., 2022).

2 PROBLEM SETUP AND ASSUMPTIONS

Throughout the paper, we focus on the following nonconvex problem

$$\min_{\theta} g(\theta) \tag{2}$$

where $g : \mathbb{R}^d \to \mathbb{R}$ is a non-negative and continuously differentiable function and satisfies the following assumptions.

Assumption 2.1. *Loss function $g(\theta)$ satisfies the following conditions:*

(1) The loss function $g(\theta)$ is bounded for any θ that belongs to the gradient sub-level $J_\eta := \{\theta \mid \|\nabla g(\theta)\|^2 < \eta\}$ with some $\eta > 0$, i.e., $g(\theta) < +\infty,\ \forall\, \theta \in J_\eta$.

(2) The gradient $\nabla g(\theta)$ is $\mathcal{L}$-Lipschitz continuous, i.e., for any $x, y \in \mathbb{R}^d$,

$$\big\|\nabla g(x) - \nabla g(y)\big\| \le \mathcal{L}\|x - y\|.$$

(3) For two fixed constants $\sigma_0, \sigma_1 \ge 0$, the stochastic gradient $\nabla g(\theta, \xi_n)$ satisfies that

$$\mathbb{E}_{\xi_n}\Big(\big\|\nabla g(\theta, \xi_n)\big\|^2\Big) \le \sigma_0 \big\|\nabla g(\theta)\big\|^2 + \sigma_1 \tag{3}$$

for any $\theta \in \mathbb{R}^d$.

Assumption 2.1 is standard in the non-convex analysis and optimization, which can be found in many previous studies (Faw et al., 2022; Wang et al., 2023; Mertikopoulos et al., 2020). Regarding the first condition in Assumption 2.1 [1], we aim to exclude the existence of a stationary point at the point with infinite function value (i.e., $f(x) = \ln(x)$ $(x \to +\infty)$), which has been used in the literature (Mertikopoulos et al., 2020). Note that removing this assumption does not bring any substantial difficulties in the proofs of the theorems, but needs extra discussions on the stationary points at which the function value is infinite. However, the stationary points with the infinite function value are quite special and different from the infinitely distant stationary points with finite function values in logistic regression, i.e., $f(x) = e^{-x}$ $(x \to +\infty)$. Without this assumption, proving the statement would become extremely lengthy. Moreover, such functions have rarely appeared in machine learning. Therefore, we make the assumption to simplify the proof.

Jin et al. (2022) makes the assumption on the set of the stationary point to prove the last-iterate almost surely convergence of AdaGrad-Norm. However, such an assumption is not realistic. For loss functions with saddle points, unless the function is defined in one-dimensional space $\mathbb{R}$, we can always find an example that does not satisfy the assumption in Jin et al. (2022). In machine learning, except for problems such as linear regression or logistic regression, the existence of saddle points is very common in many applications. Our assumptions enable us to encompass almost any loss function in machine learning. Besides, Assumption 2.1(3) is commonly used in the analysis of SGD. Our Assumption 2.1(3) is much weaker than the uniformly bounded stochastic gradients (i.e., $\|\nabla g(\theta_n, \xi_n)\| < K < +\infty$ $a.s.$.) required in Ward et al. (2020).

3 Theoretical Results

In this section, we present the main convergence results of the AdaGrad-Norm algorithm for smooth nonconvex problems. The proof sketch of each result will be provided in Section 4.

The first result below demonstrates the almost surely convergence of AdaGrad-Norm at the last iterate, which is very challenging in the theoretical analysis of gradient-based methods.

Theorem 3.1. *Consider the AdaGrad-Norm algorithm defined in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we have*

$$\lim_{n\to\infty} \|\nabla g(\theta_n)\| = 0 \ \ a.s..$$

The description of the last iterate convergence provides a more accurate comprehension of the convergence properties of AdaGrad-Norm. This is because, in practice, we typically use the last iterate as the output, rather than the average iterate which is commonly studied in theoretical research. To the best of our knowledge, the literature on the convergence of AdaGrad-Norm in the almost sure sense is sparse. Moreover, we do not assume the absence of saddle points in the loss function, which makes a substantial improvement, compared to the results in Jin et al. (2022). The result of Theorem 3.1 is applicable to almost any loss function encountered in the machine learning regime.

The next theorem describes the convergence of the last iterate of the AdaGrad-Norm algorithm in the mean-square sense.

Theorem 3.2. *Consider the AdaGrad-Norm algorithm shown in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we have*

$$\lim_{n\to\infty} \mathbb{E}\,\|\nabla g(\theta_n)\|^2 = 0.$$

Theorem 3.2 provides the mean-square convergence of the last iterate of AdaGrad-Norm, which is a novel result, unveiling the uniform convergence of gradient norm convergence under the L_2 norm [2]. We are the first to use the approach by splitting the dynamical system of AdaGrad-Norm into multiple sub-processes through many first entrance times. The proof sketch of the method is provided in Section 4. This approach facilitates a deeper comprehension of the properties of AdaGrad-Norm and can also be applied to the study of other algorithms. We would like to clarify that the almost sure convergence does not imply mean-square convergence. To illustrate this, we

[1] Note that Assumption 2.1 (1) only concerns near-stationary points, not regions where $\|\nabla g(\theta)\|$ may be large. Meanwhile, compared to the assumption in Mertikopoulos et al. (2020), our assumption is weaker as we allow the existence of a stationary point at infinity with a finite loss function value.

[2] The L_2 norm of a random variable ζ is defined as $\sqrt{\mathbb{E}\,\|\zeta\|^2}$.

consider a sequence of random variables $\{\zeta_n\}_{n=1}^{+\infty}$, where $\mathbb{P}(\zeta_n = 0) = 1 - 1/n^2$ and $\mathbb{P}(\zeta_n = n^2) = 1/n^2$. According to *The B-C Lemma*, we can easily show that $\lim_{n\to+\infty} \zeta_n = 0$ almost surely. However, by calculating, we can see that $\mathbb{E}(\zeta_n) = 1$ for all $n > 0$.

As a direct byproduct of Theorem 3.2, we can obtain the following more accurate estimation for S_n:

Property 3.1. *Consider the AdaGrad-Norm algorithm in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we get that*

$$\mathbb{E}\, S_n = O(n),$$

where other hidden constant in $\mathcal{O}(n)$ is uniquely determined by α_0, c, $g(\theta_1)$, $\nabla g(\theta_1)$, and S_0.

Proof. Through Assumption 2.1 (3) and Theorem 3.2, we can clearly obtain the result. □

Faw et al. (2022); Wang et al. (2023) only obtained the estimation for S_n: $\mathbb{E}\sqrt{S_n} = O(\sqrt{n})$ to achieve the high probability result. However, this estimate is not enough to achieve convergence rates in the expectation sense, which is more difficult than in the high probability sense. In Property 3.1, we derive a more accurate estimation $\mathbb{E}\, S_n = O(n)$, rather than $\mathbb{E}\sqrt{S_n} = O(\sqrt{n})$ and achieve a result in the expectation sense in Theorem 3.4.

Furthermore, we present the almost surely convergence rates for the AdaGrad-Norm algorithm. It is worth noting that the convergence rates provided in this paper are based on the average-iterate sense, rather than the last-iterate convergence rates due to the milder conditions. To obtain convergence rates in the last-iterate sense, one usually needs more conditions to measure the relationship between loss function g and its gradient $\|\nabla g\|$, such as Polyak-Łojasiewicz (PL) condition or Kurdyka-Łojasiewicz (KL) condition. Since this paper focuses on the convergence properties of general non-convex functions, we do not provide the convergence rates of the last iterate here. The first convergence rate result is provided in the almost-surely sense.

Theorem 3.3. *Consider the AdaGrad-Norm algorithm in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we have*

$$\frac{1}{T}\sum_{k=2}^{T}\big\|\nabla g(\theta_k)\big\|^2 = \mathcal{O}\bigg(\frac{\ln^{\frac{3}{2}+\sigma} T}{\sqrt{T}}\bigg) \ \ (\forall\, \sigma > 0) \ \ a.s..$$

Theorem 3.3 presents the near-optimal convergence rate $\mathcal{O}\Big(\frac{\ln^{\frac{3}{2}+\sigma} T}{\sqrt{T}}\Big)$ in the almost-surely sense for AdaGrad-Norm. We are the first to demonstrate that AdaGrad-Norm converges in a near-optimal rate with probability one, while Faw et al. (2022); Wang et al. (2023) solely provide the high probability results. Moreover, unlike in Ward et al. (2020), we do not impose the restrictive requirement that $\|\nabla g(\theta_n, \xi_n)\|$ is uniformly bounded almost surely.

Theorem 3.4. *Consider the AdaGrad-Norm algorithm in Equation (1), if Assumption 2.1 holds, for any initial point $\theta_1 \in \mathbb{R}^d$, $S_0 \geq 0$, then we have*

$$\frac{1}{T}\sum_{n=1}^{\top}\mathbb{E}\,\|\nabla g(\theta_n)\|^2 = O\bigg(\frac{\ln^{\frac{2}{p}} T}{T^{\frac{1}{p}}}\bigg), \quad \forall\, p > 2.$$

Theorem 3.4 shows the convergence rate $O\left(\ln^{\frac{2}{p}} T/T^{\frac{1}{p}}\right)$ ($\forall p > 2$) for AdaGrad-Norm in the expectation sense. Note that the result of Theorem 3.4 in the expectation sense is different but not weaker than the almost surely result in Theorem 3.3 and high-probability results in Faw et al. (2022); Wang et al. (2023). The distinctions between Theorem 3.3 and 3.4 arise because the hidden constant in $O(\cdot)$ is regarded as a random variable. This hidden constant is almost surely bounded, but its distribution is unknown, so its expectation is not necessarily bounded especially when p approaches 2. On the other hand, in Faw et al. (2022); Wang et al. (2023), the authors may also obtain the expected result $\left(\mathbb{E}\sqrt{\sum_{n=1}^{\top}\|\nabla g(\theta_n)\|^2/T}\right)^2 = O(\ln T/\sqrt{T})$. However, this result does not induce the result w.r.t. $\sum_{n=1}^{\top}\mathbb{E}\,\|\nabla g(\theta_n)\|^2/T$ of Theorem 3.4. Ward et al. (2020) has achieved a near-optimal rate but is based on the uniformly bounded stochastic gradients assumption which is much stronger than ours, and this assumption will facilitate the proof. We will provide a detailed explanation of this situation in Appendix D.3.

4 PROOF SKETCH

In this section, we will describe the proof sketch of Theorems 3.1 and 3.2, summarize the limitations in previous approaches, and clarify the innovativeness of our methods.

4.1 PROOF SKETCH OF THEOREM 3.1

To demonstrate the almost surely convergence of AdaGrad-Norm (in Theorem 3.1), the main obstacle is to prove that the iterates sequence $\{\theta_n\}_{n=1}^{+\infty}$ will eventually fall within the vicinity of a connected component of the stationary point set J. We then proceed to narrow down the scope of this region to show that θ_n will ultimately converge to the stationary point set J almost surely.

Step 1: We establish a recursive inequality relationship of the loss functions g in adjacent iterative steps θ_i, θ_{i+1}, i.e.,

$$g(\theta_{i+1}) - g(\theta_i) \le -\frac{\alpha_0\|\nabla g(\theta_i)\|^2}{\sqrt{S_{i-1}}} + \alpha_0\frac{\|\nabla g(\theta_i)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_{i-1}} + \frac{c\alpha_0^2}{2}\frac{\mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_i} + P_i + Q_i + R_i, \tag{4}$$

where

$$P_i := \alpha_0\frac{\nabla g(\theta_i)^\top(\nabla g(\theta_i) - \nabla g(\theta_i, \xi_i))}{\sqrt{S_{i-1}}}, \quad Q_i := \alpha_0\frac{\|\nabla g(\theta_i)\| \cdot (\|\nabla g(\theta_i, \xi_i)\|^2) - \mathbb{E}(\|\nabla g(\theta_i, \xi_i)\|^2 | \mathcal{F}_i)}{S_{i-1}}$$

$$R_i := \frac{\mathcal{L}\alpha_0^2}{2}\frac{\|\nabla g(\theta_i, \xi_i)\|^2 - \mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_i}.$$

It is observed that when the gradient $\|\nabla g(\theta_n)\|$ is relatively large, i.e., $\forall\ u > 0$, $\|\nabla g(\theta_n)\|^2 > u$, the negative term $-\alpha_0\|\nabla g(\theta_i)\|^2/\sqrt{S_{i-1}}$ will dominate the right side of the inequality (4) as the iterations proceeds. The subsequent terms related to the square of the learning rate can be ignored. However, the martingale difference terms $P_i,\ Q_i,\ R_i$ may be positive and affect the negative term. Therefore, next step we aim to prove that the martingale difference term tends to zero.

Step 2: In order to prove the convergence of the martingale difference when $\|\nabla g(\theta_n)\|^2 > u$, i.e., $\sum_{i=1}^{+\infty} \mathbf{1}_{\|\nabla g(\theta_i)\|^2 > u}(P_i + Q_i + R_i)$ converges almost surely. We present the useful lemma below:

Lemma 4.1. *Suppose $\{\theta_n\}$ is a sequence generated by AdaGrad-norm, and Assumptions 2.1 holds. Then for given $S_0 \ge 0$ and for any $\forall n \in \mathbb{N}_+$, $\theta_1 \in \mathbb{R}^d$, and $\epsilon \in (0, \frac{1}{2})$, we have*

$$\sum_{k=3}^{n} \mathbb{E}\left(\frac{\left\|\nabla g(\theta_k)\right\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}\right) < +\infty.$$

This lemma was first proved in Jin et al. (2022) with the *no saddle point* condition. However, as we checked, this lemma does not really need this condition. For clarity, the complete proof is provided in Appendix B.1. Based on Lemma 4.1 and the convergence criterion for martingale difference (in Lemma A.2), we can conclude that $\sum_{i=1}^{+\infty} \mathbf{1}_{\|\nabla g(\theta_i)\|^2 > u}(P_i + Q_i + R_i)$ converges almost surely.

In Steps 1-2, when the gradient norm $\|\nabla g(\theta)\|^2$ is larger than any positive number u, the function value of g on each trajectory of AdaGrad-Norm eventually shows a decreasing trend. We expect that the decrease of function value will bring the iterate θ_n back to the region $\|\nabla g(\theta)\|^2 < u$. Then due to the arbitrariness of u, we can claim the convergence of $\|\nabla g\|$. However, in non-convex optimization, the main challenge is that the decrease of the loss function does not guarantee a corresponding decrease in its gradient. Our approach stands out from that of Jin et al. (2022) since this step. We will explain the inapplicability of Jin et al. (2022) to loss functions with saddle points.

Literature Review: based on loss function with no saddle points. The main result in Jin et al. (2022) is to prove that the gradient sequence $\{\nabla g(\theta_n)\}_{n=1}^{+\infty}$ crosses a given interval (e, o) in a finite number of times. To achieve this, the authors need to demonstrate the difference between $\|\nabla g(\theta_{n+1})\|^2$ and $\|\nabla g(\theta_n)\|^2$ becoming sufficiently small as the iterations progress. Jin et al. (2022) estimated $\|\nabla g(\theta_{n+1})\|^2 - \|\nabla g(\theta_n)\|^2$ through $g(\theta_{n+1}) - g(\theta_n)$ with an additional condition and then applied *Equation* (4). This condition supposes that when θ approaches J_i with sufficient proximity, the inequality $\|\nabla g(\theta)\|^2 \leq 2\mathcal{L}|g(\theta) - g_i|$ holds for a connected component J_i of a stationary point set J and $g_i := g(\theta)_{\theta \in J_i}$. However, this inequality does not hold near saddle points. For any neighborhood around a saddle point, we can always find a point with the same function value as the saddle point, resulting in the right-hand side of inequality being zero while the left-hand side is positive. Therefore, this method can not handle the presence of saddle points in the loss function. Next, we will introduce our method, which can resolve saddle points.

Step 3: The goal of this step is to show that the image set of the stationary points set $g(J) := \{g(\theta) \mid \theta \in J\}$ can be contained within at most a finite number of disjoint open intervals. Moreover, there exists a lower bound for the distance between any two open intervals, and the measure of each interval can be arbitrarily small. The main idea of the proof is as follows. First, we prove that the stationary points set J can only be divided into countably many connected components $\{J_i\}_{i=1}^{+\infty}$. Then, since each point on each connected component has the same value of the loss function, the set $g(J)$ has at most countably many elements. Next, we construct a sequence of disjoint open intervals $\mathcal{Y}_{x,\delta} := \bigcup_{n=1}^{+\infty} \big((x + (n-1)\delta, x + n\delta) \cup (x - (n-1)\delta, x - n\delta)\big)$, and show that there exists an x such that each open interval in $\mathcal{Y}_{x,\delta}$ does not intersect with set $g(J)$ (the proof is given in Appendix 3.1). By considering the continuity of ∇g and g, we conclude that the elements of set $g(J)$ are not dense in any open intervals. This implies that within each open interval in $\mathcal{Y}_{x,\delta}$, there exists at least one open interval $\mathcal{H}_{x,\delta,n}$ that does not contain any value from set $g(J)$. Furthermore, since $g(J)$ is a bounded set, the measure of all these open intervals $\mathcal{H}_{x,\delta,n}$ must have a minimum value δ_1. Because the value of δ can be arbitrarily small, we have achieved the goal of this step.

Next, we establish the result by demonstrating that $\{g(\theta_n)\}$ will eventually fall into one of the aforementioned open intervals.

Step 4: For any $\delta_0 > 0$. We first construct a subsequence $\{k_n\}_{n=1}^{+\infty}$ to record the boundary points between the sets $\|\nabla g(\theta_n)\|^2 \leq \delta_0/2$ and $\|\nabla g(\theta_n)\|^2 > \delta_0/2$ (the definition is provided in Appendix C). By the definition, the gradient of g at $\theta_{k_{2n-1}}$ and $\theta_{k_{2n}}$ must be greater than $\delta_0/2$. We then apply the inequality in *Equation* (4) to obtain

$$g(\theta_{k_{2n}}) \leq g(\theta_{k_{2n-1}+1}) - \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0 \|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} + \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0 \|\nabla g(\theta_j)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \big| \mathcal{F}_j\right)}{S_{j-1}}$$
$$+ \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{c\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \big| \mathcal{F}_i\right)}{S_j} + \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \big(P_j + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \geq \delta_0/2) \cdot (Q_j + R_j)\big).$$

In Step 2, the martingale difference sequences are proven to converge. According to *The Cauchy's Convergence Test*, they can be arbitrarily smaller than any given value. Since the distance between the two open intervals in Step 3 is at least δ, as the iteration progresses, $g(\theta_{k_{2n}})$ cannot be greater than $g(\theta_{k_{2n-1}}) + \delta$. We organize the open intervals in Step 3 in descending order based on the function values they encompass. As the iteration progresses, the index of the open interval where $g(\theta_{k_n})$ is located will definitely increase. According to the monotone convergence theorem, we can prove that $g(\theta_{k_n})$ will definitely fall within one of the open intervals. Then, using *Equation* (4) again, the gradient of points between $\theta_{k_{2n-1}}$ and $\theta_{k_{2n}}$ can be bounded by $O(\delta)$. Combining this result and the fact that the gradient of points between $\theta_{k_{2n-2}}$ and $\theta_{k_{2n-1}}$ are bounded by $\delta_0/2$, we can obtain $\limsup_{n\to+\infty} \|\nabla g(\theta_n)\|^2 < O(\delta_0, \delta)$. Based on the arbitrariness of δ_0 and δ, we prove the result.

4.2 Proof sketch of Theorem 3.2

Following the almost surely convergence of Theorem 3.1, we further can demonstrate the mean-square convergence of AdaGrad-Norm. According to *Lebesgue's Dominated Convergence Theorem*, to obtain mean-square convergence, we only need to find a h^* such that $\|\nabla g(\theta_n)\|^2 \leq h^*$ and $\mathbb{E}\left(|h^*|\right) < +\infty$. Since $\|\nabla g(\theta_n)\|^2 \leq \sup_{k\geq 1} \|\nabla g(\theta_k)\|^2$ always holds, our objective is to prove $\mathbb{E}\left(\sup_{k\geq 1} \|\nabla g(\theta_k)\|^2\right) < +\infty$. In this paper, we are the first to utilize the decomposition of the dynamic system formed by AdaGrad-Norm to prove $\mathbb{E}\left(\sup_{k\geq 1} \|\nabla g(\theta_k)\|^2\right) < +\infty$. In each decomposed sub-process, this dynamic system exhibits a form similar to that of the upper martingale.

Then, in each sub-process, we derive a local maximum by a derivation approach of *Doob's inequality*. Finally, we sum up all the local maxima of the sub-processes to obtain the global maximum.

Step 1: We construct a recursive expression with respect to $g^2(\theta_n)$, rather than $g(\theta)$ used in the proof of Theorem 3.1. This is creative and necessary, and further explanation is provided in **Step** 4. First, we obtain the following lemma for $g^2(\theta_n)$:

Lemma 4.2. *Suppose $\{\theta_n\}$ is a sequence generated by AdaGrad-Norm, and Assumption 2.1 holds. Then $\forall\, n \in \mathbb{N}_+, \forall\, \theta_1 \in \mathbb{R}^d, \forall\, u > 0$ as long as $\|\nabla g(\theta_n)\|^2 > u$, the following inequality holds*

$$\begin{aligned} g^2(\theta_{n+1}) - g^2(\theta_n) \le\ & \alpha_0(M+1)\left(\frac{g(\theta_{n-1})\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_n)\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\ & + \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_n} \\ & - \frac{\alpha_0 g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_n}} + \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\frac{g(\theta_n)\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}} \\ & + \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} + X_n, \end{aligned} \tag{5}$$

where $X_n := \frac{2\alpha_0 g(\theta_n)}{\sqrt{S_{n-1}}}\nabla g(\theta_n)^\top\big(\nabla g(\theta_n) - \nabla g(\theta_n,\xi_n)\big)$ and $M := 2\sigma_0 + 2(\sigma_1/u) - 1$.

Lemma 4.2 holds if $\|\nabla g(\theta_n)\|^2$ exceeds a given constant u. However, this condition may not be fulfilled in every iteration. Therefore, we need to consider the process in segments based on whether the gradient norm satisfies the condition $\|\nabla g(\theta)\|^2 > u$. Based on this, we construct a series of first entrance times in the next step.

Step 2: In this step, we create a sequence of stopping times. By Assumption 2.1(1) (let $u := \eta$): for any θ in $\{\theta \mid \|\nabla g(\theta)\|^2 < u\}$, we have $g(\theta)$ is bounded. This implies that there exists u_0 such that $\|\nabla g(\theta)\|^2 > u$ for all θ in $\{\theta \mid \hat{g}(\theta) > u_0\}$, where $\hat{g}_n := g^2(\theta_n) + \alpha_0(M+1)g(\theta_n)\|\nabla g(\theta_{n-1})\|^2/\sqrt{S_{n-1}}$. For any $\lambda > 0$, the aim of our proof is to calculate the probability $\mathbb{P}\left(\max_{1\le k\le n}\hat{g}(\theta_k) > \lambda\right)$ in order to obtain the probability $\mathbb{P}\left(\max_{1\le k\le n}\|\nabla g(\theta_k)\|^4 > \lambda\right)$. However, when the gradient norm is small, we do not have a recursive iteration formula similar to Lemma 4.2. However, the boundedness is automatically satisfied when the gradient norm is small. Thus, we need to decompose this process according to the following stopping time. We define events $\mathcal{C}_n := \{\|\nabla g(\theta_n)\|^2 > u\} \cap \{u_0 < \hat{g}(\theta_n) < \lambda\}$ and build a series of stopping times $\{\tau_i^{(\lambda)}\}_{i=1}^{+\infty}$ as follow:

$$\begin{aligned} &\tau_1^{(\lambda)} := \min\{k:\ k\ge 1,\ \mathcal{C}_k \text{ occurs}\},\ \ \tau_2^{(\lambda)} := \min\{k : k > \tau_1^{(\lambda)},\ \mathcal{C}_k \text{ does not occur}\}, ..., \\ &\tau_{2m-1}^{(\lambda)} := \min\{k : k > \tau_{2m-2}^{(\lambda)},\ \mathcal{C}_k \text{ occurs}\},\ \ \tau_{2m}^{(\lambda)} := \min\{k : k \ge \tau_{2m-1}^{(\lambda)},\ \mathcal{C}_k \text{ does not occur}\}. \end{aligned}$$

Then we can define another stopping times $\tau := \min\{k:\ \hat{g}(\theta_1) < \lambda, \hat{g}(\theta_2) < \lambda, ...,\ \hat{g}(\theta_k) < \lambda\}$. Between the stopping times $\tau_{2n-2}\wedge\tau$ and $\tau_{2n-1}\wedge\tau$, it is clear that $\|\nabla g(\theta_n)\|^2 < u$, and before time $\tau_{2n-1}\wedge\tau$, $\hat{g}(\theta_n)$ never exceeds u_0. Hence, we only need to calculate $\mathbb{P}(\max_{\tau_{2n-1}<k<\tau_{2n}}\hat{g}(\theta_k) > \lambda)$ and then sum up the above probabilities over n. In the next step, we will focus on calculating $\mathbb{P}(\max_{\tau_{2n-1}<k<\tau_{2n}}\hat{g}(\theta_k) > \lambda)$.

Step 3: To make it easier to understand, let's first ignore the left stopping time at the beginning of this step of the proof. We define events $\mathcal{B}_{i,k}$ and $\mathcal{B}'_{i,k}$ as follows

$$\mathcal{B}_{i,k} := \begin{cases} \{C_i \text{ does not occur}, \mathcal{C}_{i+1} \text{ occurs}, ...,\ \mathcal{C}_k \text{ occurs}\} & \text{for } k \ge i+1, \\ \{C_i \text{ does not occur}\} & \text{for } k \le i \end{cases} \quad \text{and} \quad \mathcal{B}'_{i,k} := \mathcal{B}_{i,k-1}/\mathcal{B}_{i,k}.$$

Then for any events $\mathcal{X} \in \mathcal{F}_i$, we get that:

$$\begin{aligned} &\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1}\right) \le -\,\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1} - \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_m\right) \\ &\quad + \mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}} + \frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right), \end{aligned}$$

where $e_0 \geq 2,\ e_1 \geq 2,\ r_0 > 0,\ r_1 > 0,\ q_0,\ q_1$, and β_0 are constants. The proof of the above inequality is quite complicated, and due to limited space, it cannot be fully explained here (the specific proof is given in Appendix D from (77) to (84)). We define $\tau^{(0)} := \min\{k : g(\theta_k) \geq \lambda\},\ \ \tau_m^{(0)} := \min\{k : g(\theta_k) \geq \lambda,\ \ k \geq \tau_{2m-1}^{(\lambda)} \wedge \tau\}$. Then, we let $\mathcal{X} = \{\tau \wedge \tau_{2m-1}^{(\lambda)} \wedge n = i\}$ and sum it up to obtain the result considering the left stopping time. We getting

$$\begin{aligned}\mathbb{E}(\hat{g}_{\tau_m^{(0)} \wedge n}) &< \sum_{i=\tau_{2n-2}^{(\lambda)}}^{n-1} \sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{n-1} \mathbb{E}\left(\mathbf{1}_{\{\tau\wedge\tau_{2m-1}^{(\lambda)}\wedge n=i\}\cap\mathcal{B}'_{i,m+1}} \hat{g}_{m+1}\right) \\ &\leq u_0\big(\mathbb{E}(\mathbf{1}_{\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}) - \mathbb{E}(\mathbf{1}_{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n})\big) \\ &+ \beta_0 \sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n} \mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}} + \frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right),\end{aligned} \tag{6}$$

where $e_0 \geq 2,\ e_1 \geq 2,\ r_0 > 0,\ r_1 > 0,\ q_0,\ q_1$ and $\delta_0 > 0$ are seven constants. Next, we will generalize the results from time $\tau_{2m-2} \wedge \tau \wedge n$ to time $\tau_{2m} \wedge \tau \wedge n$ to all times.

Step 4: Define $\overline{\|\nabla g(\theta_n)\|^2} := \sup_{1\leq k\leq n} \|\nabla g(\theta_n)\|^2$ and $\overline{g}_n := \sup_{1\leq k\leq n} g^2(\theta_k)$, for any $\lambda \geq u$, we have $\{\overline{\|\nabla g(\theta_n)\|^4} > \lambda\} \subset \{\overline{g}_n > 2c\lambda\} \cap \{\overline{\|\nabla g(\theta_n)\|^2} > u\} \subset \Big\{\sup_{1\leq k\leq n} g_k^2 > 2c\lambda\Big\}$. Using *The Markov's Inequality* and *Equation* (6) gives $\mathbb{P}\left(\overline{\|\nabla g(\theta_n)\|^4} > \lambda\right) \leq \frac{1}{2\mathcal{L}\lambda}\sum_{m=1}^{+\infty} \mathbb{E}(\hat{g}_{\tau_m^{(0)}\wedge n}) \leq \frac{K}{2c\lambda} \leq \frac{T}{\lambda}$, where $T > 0$ is a finite positive constant. The proof of this inequality is provided in Appendix D *Equation* 86. Then we estimate $\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right)$ and achieve that

$$\begin{aligned}&\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right) = \mathbb{E}\left(\sqrt{\overline{\|\nabla g(\theta_n)\|^4}}\right) = u + \int_{\lambda=u}^{+\infty} \lambda^{\frac{1}{2}-1}\,\mathbb{P}\left(\overline{\|\nabla g(\theta_m)\|^4} > \lambda\right) d\lambda \\ &\leq u + T\int_{\lambda=u}^{+\infty} \lambda^{-\frac{3}{2}} d\lambda = u + \frac{2T}{\sqrt{u}} < +\infty.\end{aligned}$$

Now we are able to address the question raised in **Step** 1 why we use $g^2(\theta)$ rather than $g(\theta)$. If $g(\theta)$ is studied in Step 1, we obtain $\mathbb{P}\left(\overline{\|\nabla g(\theta_n)\|^2} > \lambda\right) \leq O(1/\lambda)$ in Step 3, and further achieve that $\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right) < u + \int_{\lambda-u}^{+\infty} \lambda^{-1} d\lambda = +\infty$. Thus, this is not enough to guarantee bounded of $\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right)$. So far we have proven that $\mathbb{E}\left(\sup_{k\geq 1}\|\nabla g(\theta_k)\|^2\right) < +\infty$. By Theorem 3.1 and *The Lebesgue's Dominated Convergence Theorem*, we have proven $\mathbb{E}\left(\|\nabla g(\theta_k)\|^2\right) \to 0$.

4.3 PROOF SKETCH OF THEOREM 3.3 AND THEOREM 3.4

The proofs of Theorems 3.3 and 3.4 are relatively straightforward compared to those of Theorem 3.1 and Theorem 3.2. For brevity, we omit the proof sketch here, and the complete proofs can be found in Appendix D.1 and Appendix D.2.

5 CONCLUSION

In this paper, we effectively address several limitations of the theoretical analysis of AdaGrad-Norm. Specifically, we propose novel techniques that avoid the no saddle points assumption used in (Jin et al., 2022) and establish the last-iterate convergence in both almost surely and mean-square senses. Additionally, we demonstrate the near-optimal and sub-optimal convergence rates concerning the averaged iterate in the expectation sense and the almost surely sense, respectively. Moreover, we mitigate the uniform boundedness assumption on stochastic gradients, commonly used in existing high-probability convergence analysis. Furthermore, our approaches pave the way for exploring the convergence properties of other stochastic algorithms in future research.

REFERENCES

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, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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

John Duchi, Michael I Jordan, and Brendan McMahan. Estimation, optimization, and parallelism when data is sparse. *Advances in Neural Information Processing Systems*, 26:2832–2840, 2013.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. 2020.

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. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.

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

Ruinan Jin, Yu Xing, and Xingkang He. On the convergence of mSGD and AdaGrad for stochastic optimization. In *International Conference on Learning Representations*, 2022.

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.

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

Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *International Conference on Machine Learning*, pp. 2863–2872, 2018.

Guo Lei, Cheng Dai-Zhan, and Feng De-Xing. *Introduction to Control Theory: From Basic Concepts to Research Frontiers*. Beijing: Science Press, 2005.

Xiao Li and Andre Milzarek. A unified convergence theorem for stochastic optimization methods. *Advances in Neural Information Processing Systems*, 35:33107–33119, 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*, pp. 983–992. PMLR, 2019.

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

Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy Nguyen. On the convergence of adagrad (norm) on r^d: Beyond convexity, non-asymptotic rate and acceleration. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971.

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.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023.

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.

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

A USFUL LEMMAS

Lemma A.1. *(Lemma 10 in Jin et al. (2022)) Suppose that $f(x) \in C^1 \ (x \in \mathbb{R}^d)$ with $f(x) > -\infty$ and its gradient satisfies the following Lipschitz condition*

$$\|\nabla f(x) - \nabla f(y)\| \le \mathcal{L}\|x - y\|,$$

then $\forall\, x_0 \in \mathbb{R}^d$, there is

$$\|\nabla f(x_0)\|^2 \le 2\mathcal{L}\big(f(x_0) - f^*\big),$$

where $f^* = \inf_{x \in \mathbb{R}^d} f(x)$.

Lemma A.2. *(Theorem 4.2.1 in Lei et al. (2005)) Suppose that $\{X_n\} \in \mathbb{R}^d$ is an $\mathcal{L}_2$ martingale difference sequence, and $(X_n, \mathcal{F}_n)$ is an adaptive process. Then it holds that $\sum_{k=0}^{\infty} X_k < +\infty$ a.s., if there exists $p \in (0, 2)$, such that*

$$\sum_{n=1}^{\infty} \mathbb{E}(\|X_n\|^p) < +\infty, \quad \text{or} \quad \sum_{n=1}^{\infty} \mathbb{E}\left(\|X_n\|^p \big| \mathcal{F}_{n-1}\right) < +\infty. \quad a.s.$$

Lemma A.3. *(Lemma 6 in Jin et al. (2022)) Suppose that $\{X_n\} \in \mathbb{R}^d$ is a non-negative sequence of random variables, then it holds that $\sum_{n=0}^{\infty} X_n < +\infty$ a.s., if $\sum_{n=0}^{\infty} \mathbb{E}\left(X_n\right) < +\infty$.*

Lemma A.4. *(Lemma 4.2.13 in Lei et al. (2005)) Let $\{X_k, \mathcal{F}_k\}$ be a martingale difference sequence, where X_k can be a matrix. Let $(M_k, \mathcal{F}_k)$ be an adapted process, where M_k can be a matrix, and $\|M_k\| < +\infty$ almost surely for all k. If $\sup_n \mathbb{E}(\|X_{n+1}\| | \mathcal{F}_n) < +\infty$ a.s., then we have*

$$\sum_{k=0}^{n} M_k X_{k+1} = O\left(\left(\sum_{k=0}^{n} \|M_k\|\right) \ln^{1+\sigma}\left(\left(\sum_{k=0}^{n} \|M_k\|\right) + e\right)\right) \ (\forall\, \sigma > 0) \ a.s..$$

Lemma A.5. *Suppose $\{\theta_n\}$ is a sequence generated by AdaGrad-Norm in Equation (1), and Assumptions 2.1 holds. Then $\forall n \in \mathbb{N}_+$, $\forall \theta_1 \in \mathbb{R}^d$, $\forall \epsilon \in (0, \frac{1}{2})$, there exists $\zeta < +\infty$ such that*

$$\frac{g(\theta_{n+1}) - g^*}{S_{n+1}^{\epsilon}} \le \zeta < +\infty \ a.s.,$$

which $g^* = \inf_{\theta \in \mathbb{R}^d} g(\theta)$.

Lemma A.6. *Suppose that $\{\theta_n\}$ is a sequence generated by AdaGrad-Norm in Equation (1), and Assumption 2.1 holds. Then $\forall\ n \in \mathbb{N}_+$, $\forall\ \theta_1 \in \mathbb{R}^d$, as long as $\|\nabla g(\theta_n)\|^2 > u$, the following equation holds*

$$(M-1)\frac{\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}} + \frac{1}{M+1}\left(\frac{\mathbb{E}\left(\|\nabla g(\theta_n, \xi_n)\|^2 \big| \mathcal{F}_n\right)}{S_{n-1}}\right) - (M+1)\frac{\|\nabla g(\theta_{n-1})\|^2}{\sqrt{S_{n-1}}}$$
$$\le -\frac{\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}} + \left(2\Big(M + \frac{1}{2}\Big)^2 \alpha_0^2 \mathcal{L}^2 + \Big(M + \frac{1}{2}\Big)\mathcal{L}^2\alpha_0^2\right)\frac{\|\nabla g(\theta_{n-1}, \xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}},$$

where $M := 2\sigma_0 + 2(\sigma_1/u) - 1$.

B PROOFS OF LEMMAS IN SECTIONS 4 AND A

B.1 PROOF OF LEMMA 4.1

Proof. First of all, we consider the situation $\|\nabla g(\theta_n)\|^2 > \sigma_1$ (from here to *Equation* 22)

$$\begin{aligned} g(\theta_{n+1}) - g(\theta_n) &\le \nabla g(\theta_n)^\top(\theta_{n+1} - \theta_n) + \frac{\mathcal{L}\alpha_0^2}{2}\frac{\|\nabla g(\theta_n, \xi_n)\|^2}{S_n} \\ &= -\frac{\alpha_0 \nabla g(\theta_n)^\top \nabla g(\theta_n, \xi_n)}{\sqrt{S_n}} + \frac{\mathcal{L}\alpha_0^2}{2}\frac{\|\nabla g(\theta_n, \xi_n)\|^2}{S_n}. \end{aligned} \tag{7}$$

Note that

$$\begin{aligned}&\left\|\frac{1}{\sqrt{M_0+1}}\nabla g(\theta_n,\xi_n)-\sqrt{M_0+1}\nabla g(\theta_n)\right\|^2\\&=\frac{1}{M_0+1}\big\|\nabla g(\theta_n,\xi_n)\big\|^2+(M_0+1)\big\|\nabla g(\theta_n)\big\|^2-2\nabla g(\theta_n)^\top\nabla g(\theta_n,\xi_n),\end{aligned}\tag{8}$$

where $M_0=\sigma_0+2$ and σ_0 is defined in Assumption 2.1 (3). Substitute *Equation* 8 into *Equation* 7, then we get that

$$\begin{aligned}&g(\theta_{n+1})-g(\theta_n)\\&\leq-\frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{\sqrt{S_n}}+(M_0+1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right)\\&+\frac{\alpha_0}{2}\frac{1}{\sqrt{S_n}}\left\|\frac{1}{\sqrt{M_0+1}}\nabla g(\theta_n,\xi_n)-\sqrt{M_0+1}\nabla g(\theta_n)\right\|^2+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n}.\end{aligned}\tag{9}$$

Due to $S_n\geq S_{n-1}$, it follows that

$$\begin{aligned}&\frac{\alpha_0}{2}\frac{1}{\sqrt{S_n}}\left\|\frac{1}{\sqrt{M_0+1}}\nabla g(\theta_n,\xi_n)-\sqrt{M_0+1}\nabla g(\theta_n)\right\|^2\\&\leq\frac{\alpha_0}{2}\frac{1}{\sqrt{S_{n-1}}}\left\|\frac{1}{\sqrt{M_0+1}}\nabla g(\theta_n,\xi_n)-\sqrt{M_0+1}\nabla g(\theta_n)\right\|^2.\end{aligned}\tag{10}$$

Substitute *Equation* 10 into *Equation* 9, then we have

$$\begin{aligned}&g(\theta_{n+1})-g(\theta_n)\\&\leq-\frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{\sqrt{S_n}}+(M_0+1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right)\\&+\frac{\alpha_0}{2}\frac{1}{\sqrt{S_{n-1}}}\left\|\frac{1}{\sqrt{M_0+1}}\nabla g(\theta_n,\xi_n)-\sqrt{M_0+1}\nabla g(\theta_n)\right\|^2+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n}.\end{aligned}\tag{11}$$

Notice that

$$\begin{aligned}&\frac{\alpha_0}{2}\frac{1}{\sqrt{S_{n-1}}}\left\|\frac{1}{\sqrt{M_0+1}}\nabla g(\theta_n,\xi_n)-\sqrt{M_0+1}\nabla g(\theta_n)\right\|^2\\&=\frac{\alpha_0}{2}\frac{1}{\sqrt{S_{n-1}}}\left(\frac{1}{M_0+1}\big\|\nabla g(\theta_n,\xi_n)\big\|^2+(M_0+1)\big\|\nabla g(\theta_n)\big\|^2-2\nabla g(\theta_n,\xi_n)^\top\nabla g(\theta_n)\right)\\&=\frac{\alpha_0}{2}\frac{1}{\sqrt{S_{n-1}}}\left(\frac{1}{M_0+1}\big\|\nabla g(\theta_n,\xi_n)\big\|^2+(M_0+1)\big\|\nabla g(\theta_n)\big\|^2-2\big\|\nabla g(\theta_n)\big\|^2\right)\\&+\frac{\alpha_0}{\sqrt{S_{n-1}}}\nabla g(\theta_n)^\top\big(\nabla g(\theta_n)-\nabla g(\theta_n,\xi_n)\big).\end{aligned}\tag{12}$$

Substitute *Equation* 12 into *Equation* 11, and divide both sides of the inequality by S_n^{ϵ} ($\epsilon<\frac{1}{2}$), then we get

$$\begin{aligned}&\frac{g(\theta_{n+1})}{S_n^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\\&\leq-\frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}+(M_0+1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)\\&+\frac{\alpha_0}{2}\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}\left(\frac{1}{M_0+1}\big\|\nabla g(\theta_n,\xi_n)\big\|^2+(M_0+1)\big\|\nabla g(\theta_n)\big\|^2-2\big\|\nabla g(\theta_n)\big\|^2\right)\\&+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}}+\frac{\alpha_0}{S_{n-1}^{\frac{1}{2}+\epsilon}}\nabla g(\theta_n)^\top\big(\nabla g(\theta_n)-\nabla g(\theta_n,\xi_n)\big).\end{aligned}$$

Notice that $\frac{g(\theta_{n+1})}{S_n^{\epsilon}} > \frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}$, then we obtain

$$\begin{aligned}
&\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}} - \frac{g(\theta_n)}{S_n^{\epsilon}} \\
&\le -\frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}} + (M_0+1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right) \\
&+ \frac{\alpha_0}{2}\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}\left(\frac{1}{M_0+1}\big\|\nabla g(\theta_n,\xi_n)\big\|^2 + (M_0+1)\big\|\nabla g(\theta_n)\big\|^2 - 2\big\|\nabla g(\theta_n)\big\|^2\right) \\
&+ \frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}} + \frac{\alpha_0}{S_{n-1}^{\frac{1}{2}+\epsilon}}\nabla g(\theta_n)^\top\big(\nabla g(\theta_n) - \nabla g(\theta_n,\xi_n)\big).
\end{aligned} \tag{13}$$

Rearrange the above inequality, then it holds that

$$\begin{aligned}
&\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}} - \frac{g(\theta_n)}{S_n^{\epsilon}} \\
&\le -\frac{\alpha_0}{2}(M_0+1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}} \\
&+ \frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} + (M_0-1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right) + \frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}} + G_n^{(\epsilon)} \\
&= \frac{\alpha_0}{2}(M_0+1)\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right) \\
&+ \frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} + \frac{(M_0-1)\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} - \frac{(M_0+1)\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right) \\
&+ \frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}} + G_n^{(\epsilon)}
\end{aligned} \tag{14}$$

where $G_n^{(\epsilon)}$ is defined as follow

$$G_n^{(\epsilon)} = \frac{\alpha_0}{S_{n-1}^{\frac{1}{2}+\epsilon}}\nabla g(\theta_n)^\top\big(\nabla g(\theta_n) - \nabla g(\theta_n,\xi_n)\big).$$

Due to $\|\nabla g(\theta_n)\|^2 > \sigma_1$, we have

$$\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\Big|\mathcal{F}_{n-1}\right) \le \sigma_0\big\|\nabla g(\theta_n)\big\|^2 + \sigma_1 < (\sigma_0+1)\big\|\nabla g(\theta_n)\big\|^2. \tag{15}$$

Moreover, using the Taylor formula, we obtain

$$\begin{aligned}
\big\|\nabla g(\theta_n)\big\|^2 &= \big\|\nabla g(\theta_{n-1}) + \big(\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big)\big\|^2 \\
&= \big\|\nabla g(\theta_{n-1})\big\|^2 + 2\nabla g(\theta_{n-1})^\top\big(\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big) + \big\|\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big\|^2 \\
&\le \big\|\nabla g(\theta_{n-1})\big\|^2 + 2\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big\| + \big\|\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big\|^2.
\end{aligned}$$

Under Assumption 2.1 (3), we get that

$$\begin{aligned}
\big\|\nabla g(\theta_n)\big\|^2 &\le \big\|\nabla g(\theta_{n-1})\big\|^2 + 2\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big\| \\
&\quad + \big\|\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big\|^2 \\
&\le \big\|\nabla g(\theta_{n-1})\big\|^2 + \frac{2\alpha_0\mathcal{L}}{\sqrt{S_{n-1}}}\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\| \\
&\quad + \mathcal{L}^2\alpha_0^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}}.
\end{aligned} \tag{16}$$

From inequality $2a^\top b \leq \lambda\|a\|^2 + \frac{1}{\lambda}\|b\|^2 \ (\lambda > 0)$, it follows that

$$\begin{aligned}&(M_0-1)\big\|\nabla g(\theta_n)\big\|^2 + \big\|\nabla g(\theta_n)\big\|^2 \\ &\leq (M_0+1)\big\|\nabla g(\theta_{n-1})\big\|^2 - \frac{M_0-1}{4M_0-3}\big\|\nabla g(\theta_n)\big\|^2 + 4M_0^2\alpha_0^2\mathcal{L}^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}}.\end{aligned} \tag{17}$$

By substituting *Equation* 15 into *Equation* 17, we have

$$\begin{aligned}&(M_0-1)\big\|\nabla g(\theta_n)\big\|^2 + \frac{1}{M_0+1}\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\Big|\mathcal{F}_{n-1}\right) \\ &\leq (M_0+1)\big\|\nabla g(\theta_{n-1})\big\|^2 - \frac{M_0-1}{4M_0-3}\big\|\nabla g(\theta_n)\big\|^2 + 4M_0^2\alpha_0^2\mathcal{L}^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}}.\end{aligned} \tag{18}$$

Divide both sides of *Equation* 18 by $S_{n-1}^{\frac{1}{2}+\epsilon}$, and notice $\frac{M_0-1}{4M_0-3} > \frac{1}{5}$ since $M_0 > 2$, then it holds that

$$\begin{aligned}&\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} + \frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} \\ &\leq (M_0+1)\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{5S_{n-1}^{\frac{1}{2}+\epsilon}} + 4M_0^2\alpha_0^2\mathcal{L}^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}^{\frac{3}{2}+\epsilon}} + \frac{2}{\alpha_0}H_n^{(\epsilon)},\end{aligned} \tag{19}$$

where

$$H_n^{(\epsilon)} = \frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\Big|\mathcal{F}_{n-1}\right)}{S_{n-1}^{\frac{1}{2}+\epsilon}} - \frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right).$$

Making some simple transformations on *Equation* 19 leads to

$$\begin{aligned}&\frac{\alpha_0}{2}\left(\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} + \frac{(M_0-1)\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} - \frac{(M_0+1)\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right) \\ &\leq -\frac{\alpha_0}{10}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} + 2M_0^2\alpha_0^3\mathcal{L}^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}^{\frac{3}{2}+\epsilon}} + H_n^{(\epsilon)}.\end{aligned} \tag{20}$$

Substitute *Equation* 20 into *Equation* 14, then we get

$$\begin{aligned}\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}} - \frac{g(\theta_n)}{S_n^{\epsilon}} \leq{}& \frac{\alpha_0}{2}(M_0+1)\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right) - \frac{\alpha_0}{10}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} \\ &+ 2M_0^2\alpha_0^3\mathcal{L}^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}^{\frac{3}{2}+\epsilon}} \\ &+ \frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}} + G_n^{(\epsilon)} + H_n^{(\epsilon)}.\end{aligned} \tag{21}$$

It follows that

$$\begin{aligned}-\frac{\alpha_0}{10}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} &= -\frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} - \frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} \\ &= -\frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} + \frac{\alpha_0}{20}\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right) \\ &\quad - \frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}} \\ &\leq -\frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}} + \frac{\alpha_0}{20}\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right).\end{aligned} \tag{22}$$

Substituting *Equation* 22 into *Equation* 21 yields

$$\begin{aligned}
&\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\\
&\le\frac{\alpha_0}{2}(M_0+1)\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&-\frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0}{20}\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)\\
&+2M_0^2\alpha_0^3\mathcal{L}^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}^{\frac{3}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}}+G_n^{(\epsilon)}+H_n^{(\epsilon)}.
\end{aligned}\tag{23}$$

Then we calculate the inequality when $\|\nabla g(\theta_n)\|^2\le\sigma_1$, (from here to *Equation* 34) dividing both sides of *Equation* 7 by S_n^{ϵ} yields

$$\frac{g(\theta_{n+1})}{S_n^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\le-\frac{\alpha_0\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)}{S_n^{\frac{1}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}}.\tag{24}$$

Due to $S_{n+1}\ge S_n$, it holds that

$$\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\le-\frac{\alpha_0\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)}{S_n^{\frac{1}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}}.\tag{25}$$

Then we make some transformations to obtain that

$$\begin{aligned}
&-\frac{\alpha_0\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)}{S_n^{\frac{1}{2}+\epsilon}}\\
&=-\frac{\alpha_0\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)}{S_{n-1}^{\frac{1}{2}+\epsilon}}+\alpha_0\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&\le-\frac{\alpha_0\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)}{S_{n-1}^{\frac{1}{2}+\epsilon}}+\alpha_0\left(\frac{(M_0+1)\sigma_1}{2}+\frac{\big\|\nabla g(\theta_n)\big\|^2\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{2(M_0+1)\sigma_1}\right)\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&\le-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}+\frac{(M_0+1)\alpha_0\sigma_1}{2}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&+\left(\frac{\alpha_0}{2\sigma_1(M_0+1)}\big\|\nabla g(\theta_n)\big\|^2\,\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\Big|\mathcal{F}_{n-1}\right)\right)\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}+J_n^{(\epsilon)}+K_n^{(\epsilon)},
\end{aligned}\tag{26}$$

where

$$\begin{aligned}
J_n^{(\epsilon)}&=\frac{\alpha_0}{S_{n-1}^{\frac{1}{2}+\epsilon}}\left(\big\|\nabla g(\theta_n)\big\|^2-\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)\right)\\
K_n^{(\epsilon)}&=\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{2\sigma_1(M_0+1)S_{n-1}^{\frac{1}{2}+\epsilon}}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2-\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\Big|\mathcal{F}_{n-1}\right)\right).
\end{aligned}\tag{27}$$

Due to $\big\|\nabla g(\theta_n)\big\|^2\le\sigma_1$, we get

$$\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\Big|\mathcal{F}_{n-1}\right)\le\sigma_0\big\|\nabla g(\theta_n)\big\|^2+\sigma_1\le(M_0+1)\sigma_1.\tag{28}$$

Substitute it into *Equation* 26, then we get

$$\begin{aligned}\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\le&-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0\sigma_1(M_0+1)}{2}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\\&+\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{2S_{n-1}^{\frac{1}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}}+J_n^{(\epsilon)}+K_n^{(\epsilon)}\\&=-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{2S_{n-1}^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0\sigma_1(M_0+1)}{2}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\\&+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n^{1+\epsilon}}+J_n^{(\epsilon)}+K_n^{(\epsilon)}.\end{aligned}\tag{29}$$

We make some transformations on $-\frac{\alpha_0\|\nabla g(\theta_n)\|^2}{2S_{n-1}^{\frac{1}{2}+\epsilon}}$ to obtain that

$$\begin{aligned}&-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{2S_{n-1}^{\frac{1}{2}+\epsilon}}\le-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{20S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{20S_{n-1}^{\frac{1}{2}+\epsilon}}\\&=-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{20S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0}{20}\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right).\end{aligned}\tag{30}$$

Then we use inequality $2a^{\top}b\le\lambda\|a\|^2+\frac{1}{\lambda}\|b\|^2\ \ (\lambda>0)$ on *Equation* 16 to get

$$\begin{aligned}\big\|\nabla g(\theta_n)\big\|^2-\big\|\nabla g(\theta_{n-1})\big\|^2\le&\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{10(M_0+1)}+\frac{10\alpha_0^2\mathcal{L}^2(M_0+1)}{S_{n-1}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2\\&+\frac{\alpha_0^2\mathcal{L}^2}{S_{n-1}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2.\end{aligned}\tag{31}$$

Divide both sides of *Equation* 90 by $S_{n-1}^{\frac{1}{2}+\epsilon}$ and notice $S_{n-2}\le S_{n-1}\le S_n$, then we have

$$\begin{aligned}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\le&\frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{10S_{n-2}^{\frac{1}{2}+\epsilon}}+\frac{10\alpha_0^2\mathcal{L}^2(M_0+1)}{S_{n-1}^{1+\epsilon}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2\\&+\frac{\alpha_0^2\mathcal{L}^2}{S_{n-1}^{\frac{3}{2}+\epsilon}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2.\end{aligned}\tag{32}$$

Then Calculating $\frac{\alpha_0}{2}(M_0+1)$*Equation* 91 + *Equation* 30 gives

$$\begin{aligned}&-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{2S_{n-1}^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0(M_0+1)}{2}\left(\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)\\&\le-\frac{\alpha_0\big\|\nabla g(\theta_n)\big\|^2}{20S_{n-1}^{\frac{1}{2}+\epsilon}}+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}^{1+\epsilon}}+\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{n-1}^{\frac{3}{2}+\epsilon}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2\\&+\frac{\alpha_0}{20}\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right).\end{aligned}$$

Move $\frac{\alpha_0(M_0+1)}{2}\left(\frac{\|\nabla g(\theta_n)\|^2}{S_n^{\frac{1}{2}+\epsilon}}-\frac{\|\nabla g(\theta_{n-1})\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)$ to the right-hand side of the above inequality, then we have

$$\begin{aligned}
-\frac{\alpha_0\|\nabla g(\theta_n)\|^2}{2S_{n-1}^{\frac{1}{2}+\epsilon}} \le& -\frac{\alpha_0\|\nabla g(\theta_n)\|^2}{20S_{n-1}^{\frac{1}{2}+\epsilon}}+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{1+\epsilon}}\\
&+\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{n-1}^{\frac{3}{2}+\epsilon}}\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2+\frac{\alpha_0}{20}\left(\frac{\|\nabla g(\theta_{n-1})\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\|\nabla g(\theta_n)\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)\\
&+\frac{\alpha_0(M_0+1)}{2}\left(\frac{\|\nabla g(\theta_{n-1})\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\|\nabla g(\theta_n)\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right). \qquad (33)
\end{aligned}$$

Substitute *Equation* 33 into *Equation* 29, then we have

$$\begin{aligned}
&\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\\
&\le -\frac{\alpha_0\|\nabla g(\theta_n)\|^2}{20S_{n-1}^{\frac{1}{2}+\epsilon}}+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{1+\epsilon}}\\
&+\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{n-1}^{\frac{3}{2}+\epsilon}}\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2+\frac{\alpha_0}{20}\left(\frac{\|\nabla g(\theta_{n-1})\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\|\nabla g(\theta_n)\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right) \qquad (34)\\
&+\frac{\alpha_0(M_0+1)}{2}\left(\frac{\|\nabla g(\theta_{n-1})\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\|\nabla g(\theta_n)\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)+\frac{\alpha_0\sigma_1(M_0+1)}{2}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&+\frac{\mathcal{L}\alpha_0^2}{2}\frac{\|\nabla g(\theta_n,\xi_n)\|^2}{S_n^{1+\epsilon}}+J_n^{(\epsilon)}+K_n^{(\epsilon)}.
\end{aligned}$$

Then we combine the events $\|\nabla g(\theta_n)\|^2\le\sigma_1$ and $\|\nabla g(\theta_n)\|^2>\sigma_1$,

$$\begin{aligned}
&\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\\
&=\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\right)\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\left(\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\right).
\end{aligned}$$

Through *Equation* 23, we get that

$$\begin{aligned}
&\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\right)\\
&\le -\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{\alpha_0\|\nabla g(\theta_n)\|^2}{20S_{n-1}^{\frac{1}{2}+\epsilon}}+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{1+\epsilon}}\\
&+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{n-1}^{\frac{3}{2}+\epsilon}}\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2\\
&+\frac{\alpha_0}{20}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{\|\nabla g(\theta_{n-1})\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\|\nabla g(\theta_n)\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right) \qquad (35)\\
&+\frac{\alpha_0(M_0+1)}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{\|\nabla g(\theta_{n-1})\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\|\nabla g(\theta_n)\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&+\frac{\alpha_0\sigma_1(M_0+1)}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&+\frac{\mathcal{L}\alpha_0^2}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{\|\nabla g(\theta_n,\xi_n)\|^2}{S_n^{1+\epsilon}}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}J_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}K_n^{(\epsilon)}.
\end{aligned}$$

Through *Equation* 34, we get

$$\begin{aligned}
&\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\left(\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\right)\\
&\le -\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\frac{\alpha_0}{20}\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0(M_0+1)}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\left(\frac{\left\|\nabla g(\theta_{n-1})\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&+\frac{\alpha_0}{20}\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\left(\frac{\left\|\nabla g(\theta_{n-1})\right\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)\\
&+2M_0^2\alpha_0^3\mathcal{L}^2\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_{n-1}),\xi_{n-1}\right\|^2}{S_{n-1}^{\frac{3}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_n,\xi_n)\right\|^2}{S_n^{1+\epsilon}}\\
&+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}G_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}H_n^{(\epsilon)}.
\end{aligned}\tag{36}$$

Calculating *Equation* $35+$ *Equation* 36, then we have

$$\begin{aligned}
&\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\right)+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\left(\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}\right)\\
&\le\frac{\alpha_0}{20}(\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1})\left(\frac{\left\|\nabla g(\theta_{n-1})\right\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)\\
&+\frac{\alpha_0}{2}(M_0+1)(\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1})\left(\frac{\left\|\nabla g(\theta_{n-1})\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&-\frac{\alpha_0}{20}\left(\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\right)\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}+2M_0^2\alpha_0^3\mathcal{L}^2\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_{n-1}),\xi_{n-1}\right\|^2}{S_{n-1}^{\frac{3}{2}+\epsilon}}\\
&+\frac{\mathcal{L}\alpha_0^2}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_n,\xi_n)\right\|^2}{S_n^{1+\epsilon}}+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{\left\|\nabla g(\theta_{n-1},\xi_{n-1})\right\|^2}{S_{n-1}^{1+\epsilon}}\\
&+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{n-1}^{\frac{3}{2}+\epsilon}}\left\|\nabla g(\theta_{n-1},\xi_{n-1})\right\|^2\\
&+\frac{\alpha_0\sigma_1(M_0+1)}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)+\frac{\mathcal{L}\alpha_0^2}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{\left\|\nabla g(\theta_n,\xi_n)\right\|^2}{S_n^{1+\epsilon}}\\
&+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}J_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}K_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}G_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}H_n^{(\epsilon)}.
\end{aligned}\tag{37}$$

Notice $\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\le 1$, then we get

$$\frac{\alpha_0\sigma_1(M_0+1)}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)\le\frac{\alpha_0\sigma_1(M_0+1)}{2}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right).\tag{38}$$

Substitute *Equation* 38 into *Equation* 37, then we get

$$\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_n)}{S_n^{\epsilon}}$$
$$\le\frac{\alpha_0}{20}\left(\frac{\left\|\nabla g(\theta_{n-1})\right\|^2}{S_{n-2}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)$$
$$+\frac{\alpha_0}{2}(M_0+1)\left(\frac{\left\|\nabla g(\theta_{n-1})\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)$$
$$-\frac{\alpha_0}{20}\frac{\left\|\nabla g(\theta_n)\right\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}+2M_0^2\alpha_0^3\mathcal{L}^2\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_{n-1}),\xi_{n-1}\right\|^2}{S_{n-1}^{\frac{3}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_n,\xi_n)\right\|^2}{S_n^{1+\epsilon}}$$
$$+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{\left\|\nabla g(\theta_{n-1},\xi_{n-1})\right\|^2}{S_{n-1}^{1+\epsilon}}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{n-1}^{\frac{3}{2}+\epsilon}}\left\|\nabla g(\theta_{n-1},\xi_{n-1})\right\|^2$$
$$+\frac{\alpha_0\sigma_1(M_0+1)}{2}\left(\frac{1}{S_{n-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)+\frac{\mathcal{L}\alpha_0^2}{2}\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}\frac{\left\|\nabla g(\theta_n,\xi_n)\right\|^2}{S_n^{1+\epsilon}}$$
$$+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}J_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2\le\sigma_1}K_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}G_n^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_n)\|^2>\sigma_1}H_n^{(\epsilon)}. \tag{39}$$

We make a summation of *Equation* 39 to get

$$\sum_{k=3}^{n}\left(\frac{g(\theta_{k+1})}{S_{k+1}^{\epsilon}}-\frac{g(\theta_k)}{S_k^{\epsilon}}\right)$$
$$\le\frac{\alpha_0}{20}\sum_{k=3}^{n}\left(\frac{\left\|\nabla g(\theta_{k-1})\right\|^2}{S_{k-2}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_k)\right\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}\right)$$
$$+\frac{\alpha_0}{2}(M_0+1)\sum_{k=3}^{n}\left(\frac{\left\|\nabla g(\theta_{k-1})\right\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}-\frac{\left\|\nabla g(\theta_k)\right\|^2}{S_k^{\frac{1}{2}+\epsilon}}\right)$$
$$-\frac{\alpha_0}{20}\sum_{k=3}^{n}\frac{\left\|\nabla g(\theta_k)\right\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}+2M_0^2\alpha_0^3\mathcal{L}^2\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_{k-1}),\xi_{k-1}\right\|^2}{S_{k-1}^{\frac{3}{2}+\epsilon}}$$
$$+\frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\left\|\nabla g(\theta_k,\xi_k)\right\|^2}{S_k^{1+\epsilon}}+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\sum_{k=2}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\left\|\nabla g(\theta_{k-1},\xi_{k-1})\right\|^2}{S_{k-1}^{1+\epsilon}}$$
$$+\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{k-1}^{\frac{3}{2}+\epsilon}}\left\|\nabla g(\theta_{k-1},\xi_{k-1})\right\|^2$$
$$+\frac{\alpha_0\sigma_1(M_0+1)}{2}\sum_{k=3}^{n}\left(\frac{1}{S_{k-1}^{\frac{1}{2}+\epsilon}}-\frac{1}{S_k^{\frac{1}{2}+\epsilon}}\right)+\frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\left\|\nabla g(\theta_k,\xi_k)\right\|^2}{S_k^{1+\epsilon}}$$
$$+\sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}\Big). \tag{40}$$

Then we get

$$\sum_{k=3}^{n}\frac{\left\|\nabla g(\theta_k,\xi_k)\right\|^2}{S_k^{1+\epsilon}}\le\int_{S_3}^{+\infty}\frac{1}{x^{1+\epsilon}}dx=\frac{1}{\epsilon S_3^{\epsilon}}, \tag{41}$$

and

$$\sum_{k=3}^{n}\frac{\left\|\nabla g(\theta_{k-1},\xi_{k-1})\right\|^2}{S_{k-1}^{1+\epsilon}}\le\int_{S_2}^{+\infty}\frac{1}{x^{1+\epsilon}}dx=\frac{1}{\epsilon S_2^{\epsilon}}, \tag{42}$$

and

$$\sum_{k=3}^{n}\frac{\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2}{S_{k-1}^{\frac{3}{2}+\epsilon}}\le\int_{S_2}^{+\infty}\frac{1}{x^{\frac{3}{2}+\epsilon}}dx=\frac{2}{(1+2\epsilon)S_2^{\frac{1}{2}+\epsilon}}. \tag{43}$$

Due to $\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\le 1$ and $\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\le 1$, we get that

$$\begin{aligned}
&2M_0^2\alpha_0^3\mathcal{L}^2\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_{k-1}),\xi_{k-1}\big\|^2}{S_{k-1}^{\frac{3}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k^{1+\epsilon}}\\
&+5\alpha_0^3\mathcal{L}^2(M_0+1)^2\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2}{S_{k-1}^{1+\epsilon}}\\
&+\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{k-1}^{\frac{3}{2}+\epsilon}}\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2+\frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k^{1+\epsilon}}\\
&\le\frac{4M_0^2\alpha_0^3\mathcal{L}^2}{(1+2\epsilon)S_2^{\frac{1}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2\epsilon S_3^{\epsilon}}+\frac{5\alpha_0^3\mathcal{L}^2(M_0+1)^2}{\epsilon S_2^{\epsilon}}+\frac{\alpha_0^3\mathcal{L}^2(M_0+1)}{(1+2\epsilon)S_2^{\frac{1}{2}+\epsilon}}+\frac{\mathcal{L}\alpha_0^2}{2\epsilon S_3^{\epsilon}}:=K.
\end{aligned} \tag{44}$$

Substituting *Equation* 44 into *Equation* 40 leads to

$$\begin{aligned}
&\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_3)}{S_3^{\epsilon}}\\
&\le\frac{\alpha_0}{20}\left(\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_1^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)+\frac{\alpha_0}{2}(M_0+1)\left(\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_2^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&-\frac{\alpha_0}{20}\sum_{k=3}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0\sigma_1(M_0+1)}{2}\left(\frac{1}{S_2^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)+K\\
&+\sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}\Big).
\end{aligned}$$

It is obvious that

$$\begin{aligned}
&\frac{\alpha_0}{20}\left(\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_1^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)+\frac{\alpha_0}{2}(M_0+1)\left(\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_2^{\frac{1}{2}+\epsilon}}-\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_n^{\frac{1}{2}+\epsilon}}\right)\\
&+\frac{\alpha_0\sigma_1(M_0+1)}{2}\left(\frac{1}{S_2^{\frac{1}{2}+\epsilon}}-\frac{1}{S_n^{\frac{1}{2}+\epsilon}}\right)+K\le\frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_1^{\frac{1}{2}+\epsilon}}+\frac{\alpha_0(M_0+1)}{2}\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_2^{\frac{1}{2}+\epsilon}}\\
&+\frac{\alpha_0\sigma_1(M_0+1)}{2S_2^{\frac{1}{2}+\epsilon}}+K:=L.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_3)}{S_3^{\epsilon}}\\
&\le-\frac{\alpha_0}{20}\sum_{k=3}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}+L+\sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}\\
&+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}\Big).
\end{aligned} \tag{45}$$

Note that $\{\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}\}$, $\{\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}\}$, $\{\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}\}$ and $\{\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}\}$ are all martingale difference sequences, thus it follows that $\mathbb{E}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}\Big)=0.$

Then we calculate mathematical expectation on *Equation* 45

$$\mathbb{E}\left(\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}-\frac{g(\theta_3)}{S_3^{\epsilon}}\right)\le -\frac{\alpha_0}{20}\,\mathbb{E}\left(\sum_{k=3}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}\right)+L.$$

That is

$$\mathbb{E}\left(\sum_{k=3}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}\right)<\frac{20}{\alpha_0}\left(\frac{g(\theta_3)}{S_3^{\epsilon}}+L\right)<+\infty.$$

With this, we have completed the proof. □

B.2 Proof of Lemma A.5

Proof. It follows from *Equation* 45 that

$$\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}\le\frac{g(\theta_3)}{S_3^{\epsilon}}+L+\sum_{k=3}^{n}\Bigg(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}$$
$$+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}\Bigg). \tag{46}$$

From *Equation* 27, we obtain

$$\sum_{k=3}^{n}\mathbb{E}\left(\big\|\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}\big\|^2\right)\le\sum_{k=3}^{n}\mathbb{E}\left(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\alpha_0^2}{S_{n-1}^{1+2\epsilon}}\Big(\big\|\nabla g(\theta_n)\big\|^2-\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)\Big)^2\right).$$

With inequality $(a+b)^2\le 2(a^2+b^2),\ 2a^{\top}b\le a^2+b^2\ \ (a,\ b>0)$, we get

$$\mathbb{E}\left(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\alpha_0^2}{S_{n-1}^{1+2\epsilon}}\Big(\big\|\nabla g(\theta_n)\big\|^2-\nabla g(\theta_n)^{\top}\nabla g(\theta_n,\xi_n)\Big)^2\right)$$
$$\le 2\,\mathbb{E}\left(\frac{\alpha_0^2\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{1+2\epsilon}}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\big\|\nabla g(\theta_n)\big\|^2\Big)\right)+2\,\mathbb{E}\left(\frac{\alpha_0^2\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}}{S_{n-1}^{1+2\epsilon}}\big\|\nabla g(\theta_n)\big\|^2\big\|\nabla g(\theta_n,\xi_n)\big\|^2\right)$$
$$\le 2(M+2)\,\mathbb{E}\left(\frac{\alpha_0^2\sigma_1\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{1+\epsilon}}\right)$$
$$\le 2(M+2)\alpha_0^2\sigma_1\,\mathbb{E}\left(\frac{\alpha_0^2\sigma_1\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}}\right)$$
$$< 40(M+2)\alpha_0\sigma_1\left(\frac{g(\theta_1)}{S_1^{\epsilon}}+L\right). \tag{47}$$

It follows from Lemma A.2 that $\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}$ is convergent a.s. Similarly, $\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}$, $\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}$ and $\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}$ are both convergent a.s. It follows that

$$\sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}G_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}H_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}J_k^{(\epsilon)}+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}K_k^{(\epsilon)}\Big)$$
$$<\xi^{'}<+\infty\ \ a.s..$$

Then we get

$$\frac{g(\theta_{n+1})}{S_{n+1}^{\epsilon}}\le\frac{g(\theta_3)}{S_3^{\epsilon}}+L+\xi^{'}<+\infty\ \ a.s.$$

For convenience, let $\xi=\frac{g(\theta_3)}{S_3^{\epsilon}}+L+\xi^{'}$. Thus, it holds that

$$\frac{g(\theta_{n+1})-g^{*}}{S_{n+1}^{\epsilon}}<\xi<+\infty\ \ a.s.. \tag{48}$$

□

B.3 Proof of Lemma A.6

Proof. We can find

$$\begin{aligned}\big\|\nabla g(\theta_n)\big\|^2 &= \big\|\nabla g(\theta_{n-1}) + \big(\nabla g(\theta_n) - \nabla g(\theta_{n-1})\big)\big\|^2 \\ &\leq \big\|\nabla g(\theta_{n-1})\big\|^2 + \frac{2\alpha_0\mathcal{L}}{\sqrt{S_{n-1}}}\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\| + \mathcal{L}^2\alpha_0^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}}.\end{aligned}$$

Then we multiple $M+\frac{1}{2}$ on the both side of above inequality, acquiring

$$\begin{aligned}\Big(M+\frac{1}{2}\Big)\big\|\nabla g(\theta_n)\big\|^2 \leq \Big(M+\frac{1}{2}\Big)\big\|\nabla g(\theta_{n-1})\big\|^2 + \Big(M+\frac{1}{2}\Big)\frac{2\alpha_0\mathcal{L}}{\sqrt{S_{n-1}}}\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\| \\ + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^2\mathcal{L}\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}}.\end{aligned}$$

Noting

$$\Big(M+\frac{1}{2}\Big)\frac{2\alpha_0\mathcal{L}}{\sqrt{S_{n-1}}}\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\| \leq \frac{1}{2}\big\|\nabla g(\theta_{n-1})\big\|^2 + 2\Big(M+\frac{1}{2}\Big)^2\alpha_0^2\mathcal{L}^2\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}}.$$

We get that

$$\Big(M+\frac{1}{2}\Big)\big\|\nabla g(\theta_n)\big\|^2 \leq (M+1)\big\|\nabla g(\theta_{n-1})\big\|^2 + \bigg(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^2\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^2\bigg)\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}},$$

that is

$$\begin{aligned}&(M-1)\big\|\nabla g(\theta_n)\big\|^2 + \frac{\sigma_0+\frac{\sigma_1}{u}}{M+1}\big\|\nabla g(\theta_n)\big\|^2 \\ &\leq -\big\|\nabla g(\theta_n)\big\|^2 + (M+1)\big\|\nabla g(\theta_{n-1})\big\|^2 + \bigg(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^2\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^2\bigg)\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}}.\end{aligned}$$

Then we multiple $1/\sqrt{S_{n-1}}$ on both side of above inequality, and noting when $\|\nabla g(\theta_n)\|^2 > u$, there is

$$\frac{\sigma_0+\frac{\sigma_1}{u}}{M+1}\big\|\nabla g(\theta_n)\big\|^2 \geq \frac{1}{M+1}E(\|\nabla g(\theta_n,\xi_n)\|^2|\mathcal{F}_n),$$

getting

$$\begin{aligned}&(M-1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_{n-1}}} + \frac{1}{M+1}\bigg(\frac{\mathbb{E}\left(\|\nabla g(\theta_n,\xi_n)\|^2|\mathcal{F}_n\right)}{S_{n-1}}\bigg) - (M+1)\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} \\ &\leq -\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_{n-1}}} + \bigg(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^2\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^2\bigg)\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}}.\end{aligned}$$

Thus, we have completed the proof. □

B.4 Proof of Lemma 4.2

Proof. we calculate $g^2(\theta_{n+1}) - g^2(\theta_n)$ as follow:

$$
\begin{aligned}
& g^2(\theta_{n+1}) - g^2(\theta_n) \\
& \leq -\frac{2\alpha_0 g(\theta_n)\nabla g(\theta_n)^\top \nabla g(\theta_n,\xi_n)}{\sqrt{S_n}} + \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} \\
& \leq -\alpha_0 g(\theta_n)\left(\frac{1}{M+1}\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{\sqrt{S_n}} + (M+1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\
& + \alpha_0 g(\theta_n)\frac{1}{\sqrt{S_{n-1}}}\Big\|\frac{1}{\sqrt{M+1}}\nabla g(\theta_n,\xi_n) - \sqrt{M+1}\nabla g(\theta_n)\Big\|^2 \\
& + \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} \\
& \leq \alpha_0 g(\theta_n)(M+1)\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\
& + \alpha_0 g(\theta_n)\left(\frac{1}{M+1}\frac{\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\big|\mathcal{F}_n\right)}{\sqrt{S_{n-1}}} + \frac{(M-1)\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_{n-1}}} - \frac{(M+1)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}}\right) \\
& + \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} + X_n,
\end{aligned}
\tag{49}
$$

where X_n is defined as follow

$$
X_n := \frac{2\alpha_0 g(\theta_n)}{\sqrt{S_{n-1}}}\nabla g(\theta_n)^\top\left(\nabla g(\theta_n) - \nabla g(\theta_n,\xi_n)\right),
$$

and $M := 2\sigma_0 + 2(\sigma_1/u) - 1$. For the term

$$
\alpha_0 g(\theta_n)(M+1)\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right),
$$

we have

$$
\begin{aligned}
& \alpha_0 g(\theta_n)(M+1)\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\
& = \alpha_0(M+1)\left(\frac{g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_{n+1})\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\
& + \alpha_0(M+1)\big|g(\theta_{n+1}) - g(\theta_n)\big|\frac{\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}} \\
& \leq \alpha_0(M+1)\left(\frac{g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_{n+1})\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\
& + \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_n} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_n}.
\end{aligned}
\tag{50}
$$

Substitute *Equation* 50 into *Equation* 49, we acquiring

$$\begin{aligned}
&g^2(\theta_{n+1}) - g^2(\theta_n) \\
&\le \alpha_0(M+1)\left(\frac{g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_{n+1})\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\
&+ \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_n} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_n} \\
&+ \alpha_0 g(\theta_n)\left(\frac{1}{M+1}\frac{\mathbb{E}\left(\big\|\nabla g(\theta_n,\xi_n)\big\|^2\big|\mathcal{F}_n\right)}{\sqrt{S_{n-1}}} + \frac{(M-1)\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_{n-1}}} - \frac{(M+1)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}}\right) \\
&+ \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} + X_n,
\end{aligned} \tag{51}$$

Due to Lemma A.6, we have

$$\begin{aligned}
&(M-1)\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_{n-1}}} + \frac{1}{M+1}\left(\frac{\mathbb{E}\left(\|\nabla g(\theta_n,\xi_n)\|^2\big|\mathcal{F}_n\right)}{S_{n-1}}\right) - (M+1)\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} \\
&\le -\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_{n-1}}} + \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^2\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^2\right)\frac{\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}}.
\end{aligned} \tag{52}$$

Substitute *Equation* 52 into *Equation* 51, getting

$$\begin{aligned}
&g^2(\theta_{n+1}) - g^2(\theta_n) \le \alpha_0(M+1)\left(\frac{g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_{n+1})\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\
&+ \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_n} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_n} \\
&- \frac{\alpha_0 g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}} + \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\frac{g(\theta_n)\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}} \\
&+ \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} + X_n.
\end{aligned}$$

Thus, we have completed the proof. □

C Proof of Theorem 3.1

Proof. Through Lemma 4.1 and Lemma A.3, we know that

$$\sum_{k=3}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}} < +\infty \ \ a.s.\,. \tag{53}$$

Below, we divide the set of all tracks into the following two sets:

$$\mathcal{S}_1 := \big\{\lim_{n\to+\infty} S_n < +\infty\big\},\ \ \mathcal{S}_2 := \big\{\lim_{n\to+\infty} S_n = +\infty\big\}.$$

In $\mathcal{S}_1$, we have

$$\limsup_{n\to+\infty}\|\nabla g(\theta_n)\|^2 < \overline{M}\limsup_{n\to+\infty}\frac{\big\|\nabla g(\theta_n)\big\|^2}{S_{n-1}^{\frac{1}{2}+\epsilon}},$$

where $\overline{M} := 2\lim_{n\to+\infty} 1/S_{n-1}^{\frac{1}{2}+\epsilon}$. In $\mathcal{S}_1$, it is evident that we have

$$\lim_{n\to+\infty} 1/S_{n-1}^{\frac{1}{2}+\epsilon} = 0 \ \ a.s.,$$

so we have in $\mathcal{S}_1$, there is $\limsup_{n\to+\infty}\|\nabla g(\theta_n)\|^2=0\ \ a.s.$, that is $\limsup_{n\to+\infty}\|\nabla g(\theta_n)\|^2=0\ \ a.s.$. Next, let us consider the case in the set $\mathcal{S}_1$. In this set, we once again divide the tracks into two sets based on the convergence or divergence of the series $\sum_{n=1}^{+\infty}1/\sqrt{S_n}$, i.e.,

$$\mathcal{S}_{2,1}:=\bigg\{\lim_{n\to+\infty}S_n=+\infty\text{ and }\sum_{n=1}^{+\infty}\frac{1}{\sqrt{S_n}}<+\infty\bigg\},$$

$$\mathcal{S}_{2,2}:=\bigg\{\lim_{n\to+\infty}S_n=+\infty\text{ and }\sum_{n=1}^{+\infty}\frac{1}{\sqrt{S_n}}=+\infty\bigg\}.$$

Then we will demonstrate that $\mathcal{S}_{2,1}=\emptyset\ \ a.s.$. We know that for any $\epsilon\in(0,1/2)$, we can find a constant $\epsilon'>0$ such that $(\epsilon+\epsilon')\in(0,1/2)$ holds. Then we calculate the following series

$$\sum_{k=3}^{n}\frac{\|\nabla g(\theta_k,\xi_k)\|^2}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}=\sum_{k=3}^{n}\frac{\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right)}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}+\sum_{k=3}^{n}\frac{\|\nabla g(\theta_k,\xi_k)\|^2-\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right)}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}$$
$$\le\sigma_0\cdot\sum_{k=3}^{n}\frac{\mathbb{E}\|\nabla g(\theta_k)\|^2}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}+\sigma_1\cdot\sum_{k=3}^{n}\frac{1}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}+\sum_{k=3}^{n}\frac{\|\nabla g(\theta_k,\xi_k)\|^2-\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right)}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}.\tag{54}$$

Based on Lemma 53 and the properties of set $\mathcal{S}_{2,1}$, we can deduce that

$$\sigma_0\cdot\sum_{k=3}^{n}\frac{\mathbb{E}\|\nabla g(\theta_k)\|^2}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}+\sigma_1\cdot\sum_{k=3}^{n}\frac{1}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}<+\infty\ a.s..\tag{55}$$

Then, according to Lemma A.1 and A.5, we can verify the following inequality:

$$\sup_{n>0}\mathbb{E}\left(\left|\frac{\|\nabla g(\theta_n,\xi_n)\|^2-\mathbb{E}\left(\|\nabla g(\theta_n,\xi_n)\|^2\big|\mathcal{F}_n\right)}{S_{n-1}^{\epsilon'}}\right|\Bigg|\mathcal{F}_n\right)\le2\sup_{n>0}\frac{\sigma_0\cdot\|\nabla g(\theta_n)\|^2+\sigma_1}{S_{n-1}^{\epsilon}}$$
$$\le2\sup_{n>0}\frac{\sigma_0\cdot g(\theta_n)+\sigma_1}{S_{n-1}^{\epsilon}}<+\infty\ \ a.s..$$

Then, based on Lemma A.4 and the properties of set $\mathcal{S}_{2,1}$, we obtain

$$\sum_{k=3}^{n}\frac{\|\nabla g(\theta_k,\xi_k)\|^2-\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right)}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}<+\infty\ \ a.s..\tag{56}$$

Substitute *Equation* 55 and *Equation* 56 into *Equation* 54, we getting

$$\sum_{k=3}^{n}\frac{\|\nabla g(\theta_k,\xi_k)\|^2}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}<+\infty\ a.s..\tag{57}$$

Then we get

$$+\infty>\sum_{k=3}^{+\infty}\frac{\|\nabla g(\theta_k,\xi_k)\|^2}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}>\int_{S_2^{\frac12+\epsilon+\epsilon'}}^{+\infty}\frac{1}{x^{\frac12+\epsilon+\epsilon'}}=+\infty\ \ a.s.\ .$$

That means $\mathcal{S}_{2,1}=\emptyset\ \ a.s.$. Furthermore, we can get in $\mathcal{S}_2$, having $\sum_{n=1}^{+\infty}1/\sqrt{S_n}=+\infty\ \ a.s.$. Next, we aim to prove $\liminf_{n\to+\infty}\|\nabla g(\theta_n)\|^2=0\ \ a.s.$ by contradiction . For a given trajectory, we assume that $\liminf_{n\to+\infty}\|\nabla g(\theta_n)\|^2>l>0$. Then we get exists $N>0$, for any $n>N$, such that $\|\nabla g(\theta_n)\|^2>l_0/2$. Then we get

$$+\infty=\sum_{n=N+1}^{+\infty}\frac{1}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}<\frac{l_0}{2}\cdot\sum_{n=N+1}^{+\infty}\frac{\|\nabla g(\theta_k)\|^2}{S_{k-1}^{\frac12+\epsilon+\epsilon'}}<+\infty\ \ a.s.,$$

which is creates a contradiction. That means we have $\liminf_{n\to+\infty}\|\nabla g(\theta_n)\|^2=0\ \ a.s.$. Since we have already proven $\lim_{n\to+\infty}\|\nabla g(\theta_n)\|^2=0\ \ a.s.$ for any trajectory in set $\mathcal{S}_1$, we can conclude that for any trajectory, there is $\liminf_{n\to+\infty}\|\nabla g(\theta_n)\|^2=0\ \ a.s.$. Below, we focus on proving $\limsup_{n\to+\infty}\|\nabla g(\theta_n)\|^2=0\ \ a.s.$.

Let $J := \{\theta | \nabla g(\theta) = 0\}$. For any bounded closed set $\mathcal{T}$, we know that $\mathcal{T} \cap J$ is a bounded closed set. Next, for any positive number $\delta > 0$, we construct an open cover $\bigcup_{\theta \in \mathcal{T} \cap \Omega} U(\theta, \delta)$ for the set $\mathcal{T} \cap \Omega$. According to *The Heine-Borel Theorem*, we know that this open cover must have a finite subcover $\bigcup_{i=1}^{T_\delta} U(\theta_i, \delta) \supset \mathcal{T} \cap \Omega$. Then we construct a sequence of open sets $\{\mathcal{O}_i\}_{i=1}^{+\infty}$ as follows:

$$\mathcal{O}_1 := \bigcup_{i=1}^{T_1} U(\theta_{1,i}, 1),\ \mathcal{O}_2 := \mathcal{O}_1 \cap \bigcup_{i=1}^{T_{1/2}} U(\theta_{2,i}, 1/2), ...,\ \mathcal{O}_n := \mathcal{O}_{n-1} \cap \bigcup_{i=1}^{T_{1/n}} U(\theta_{n,i}, 1/n), ...$$

Obviously, the sequence of sets satisfies $\mathcal{O}_i \supset \mathcal{O}_{i+1}$ $(\forall\ i)$ and it is evident that $J = \lim_{n\to+\infty} \mathcal{O}_n$. For any given $\mathcal{O}_k$, we can express it as $\mathcal{O}_k := \bigcup_{t_1,t_2,...,t_k} \bigcap_{i=1}^{k} U(\theta_{i,t_i}, 1/i)$. It is easy to see that each $U(\theta_{i,t_i}, 1/i)$ is a convex set, so the intersection of them can generate at most a finite number of connected components. This proves that for $\mathcal{O}_k$, there are at most a finite number of connected components. This implies that $J \cap \mathcal{T}$ has at most countably many connected components. By taking $\mathcal{T} = U(0, N)$ and letting positive integer N tend to infinity, we can conclude that J has at most countably many connected components.

We assign these connect components as $\{J_i\}_{i=1}^{+\infty}$. It is easy to see that points located on the same connected component J_i have the same value of the loss function $g(\theta),\ \ \forall\ \theta \in J_i$. We denote these value as $\{g_i\}_{i=1}^{+\infty}$. Let us denote the set of all distinct values in the sequence $\{g_i\}_{i=1}^{+\infty}$ as $\mathcal{G}$.

Next, we define a family of open intervals $\mathcal{Y}_{x,\delta} := \bigcup_{n=1}^{+\infty} \big((x + (n-1)\delta, x + n\delta) \cup (x - (n-1)\delta, x - n\delta)\big)$. We then use a proof by contradiction to show that for any $\delta > 0$, there exists at least one $x \in \mathbb{R}$ such that $\mathcal{G} \subset \mathcal{Y}_{x,\delta}$. We assume that such an x does not exist. Then, we can deduce that for any $x \in (0, \delta/2)$, we can always find a $g_x \in \mathcal{G}$ such that $g_x \notin \mathcal{Y}_{x,\delta}$. Based on the properties of the set $\mathcal{Y}_{x,\delta}$, we can conclude that for different $x_0 \neq x_1$, we have $g_{x_0} \neq g_{x_1}$. This means that for every x in $(0, \delta/2)$, we can find an element in $\mathcal{G}$ that corresponds to x, and for different x, these corresponding g_x are distinct. This implies that the cardinality of set $\mathcal{G}$ is equal to the cardinality of set $(0, 1/2)$, which is $\aleph$. However, this contradicts the fact that $\mathcal{G}$ can have at most countably many elements. Therefore, we can conclude that for any $\delta > 0$ there exists at least one $x \in \mathbb{R}$ such that $\mathcal{G} \subset \mathcal{Y}_{x,\delta}$.

Next, we prove that the set of values corresponding to $\mathcal{G}$ cannot be dense in any open interval in $\mathbb{R}$. We use a proof by contradiction. Suppose there exists an open interval (a, b) such that $\mathcal{G}$ is dense in (a, b). Then, by the continuity of ∇g and g, we can conclude that $(a, b) \subset \mathcal{G}$. This implies that the cardinality of the $\mathcal{G}$ is $\aleph$, which contradicts countability. Therefore, we can conclude that $\mathcal{G}$ is not dense in any open interval in $\mathbb{R}$. This means that for every interval $(x + (n-1)\delta, x + n\delta)$, we can find an open interval $\mathcal{H}_{x,\delta,n} \subset (x + (n-1)\delta, x + n\delta)$, such that $\mathcal{H}_{x,\delta,n} \cap \mathcal{G} = \emptyset$. Without loss of generality, we can always find an interval with the maximum length among all such intervals in one $(x + (n-1)\delta, x + n\delta)$. Let the measure of this interval be denoted as $|\mathcal{H}_{x,\delta,n}|$. Due to $g(J) < +\infty$ (shown in Assumption 2.1), we can always find a δ_1 such that for all $\mathcal{H}_{x,\delta,n}$, we have $|\mathcal{H}_{x,\delta,n}| > \delta_1$.

For any δ, we construct the set $\mathcal{S} := \{\|\nabla g(\theta)\|^2 < \delta_0/2\}$. We define $g \circ \mathcal{S} := \{g(\theta) | \theta \in \mathcal{S}\}$. It is easy to see that for sufficiently small δ_0, we can always achieve $(g \circ \mathcal{S}) \cap \big(\bigcup_{n=0}^{+\infty} \mathcal{H}_{x,\delta,n}\big) \subset \bigcup_{n=0}^{+\infty} \mathcal{H}'_{x,\delta,n}$, where $\mathcal{H}'_{x,\delta,n}$ can at most be the left half of the interval $\mathcal{H}_{x,\delta,n}$, that is, there exists a $\hat{\delta}_0 > 0$ such that for any $\delta_0 < \hat{\delta}_0$, we have $(g \circ \mathcal{S}) \cap \big(\bigcup_{n=0}^{+\infty} \mathcal{H}_{x,\delta,n}\big) \subset \bigcup_{n=0}^{+\infty} \mathcal{H}'_{x,\delta,n}$, where $\mathcal{H}'_{x,\delta,n}$ can at most be the left half of the interval $\mathcal{H}_{x,\delta,n}$. Then we take $\delta_0 < \min\{\hat{\delta}_0, \delta/2\}$. Then due to $\liminf_{n\to+\infty} \|\nabla g(\theta_n)\|^2 = 0\ \ a.s.$, there exist infinitely many values $\|\nabla g(\theta_k)\|^2$ in the sequence $\{\|\nabla g(\theta_n)\|^2\}_{n=1}^{+\infty}$ such that $\|\nabla g(\theta_k)\|^2 < \delta_0/2$. This implies that there are infinitely many values of $\{\|\nabla g(\theta_n)\|^2\}_{n=1}^{+\infty}$ satisfying $\|\nabla g(\theta_k)\|^2 < \delta_0/2,\ \|\nabla g(\theta_{k+1})\|^2 \geq \delta_0/2$ and $\|\nabla g(\theta_{k'})\|^2 < \delta_0/2,\ \|\nabla g(\theta_{k'-1})\|^2 \geq \delta_0/2$. We can arrange these values in increasing order of their indices to form a subsequence $\{\|\nabla g(\theta_n)\|^2\}_{n=1}^{+\infty}$ of the sequence $\{\|\nabla g(\theta_{k_n})\|^2\}_{n=1}^{+\infty}$, which holds $\|\nabla g(\theta_{k_{2n-1}})\|^2 < \delta_0/2,\ \|\nabla g(\theta_{k_{2n-1}+1})\|^2 \geq \delta_0/2,\ \|\nabla g(\theta_{k_{2n}-1})\|^2 \geq \delta_0/2,\ \|\nabla g(\theta_{k_{2n}})\|^2 <$

$\delta_0/2$. Then we calculate $g(\theta_{i+1}) - g(\theta_i)$ as follow:

$$\begin{aligned}
&g(\theta_{i+1}) - g(\theta_i) = -\frac{\alpha_0 \nabla g(\theta_i)^\top \nabla g(\theta_i, \xi_i)}{\sqrt{S_i}} + \frac{\mathcal{L}\alpha_0^2}{2} \frac{\left\|\nabla g(\theta_i, \xi_i)\right\|^2}{S_i} \\
&= -\frac{\alpha_0 \nabla g(\theta_i)^\top \nabla g(\theta_i, \xi_i)}{\sqrt{S_{i-1}}} + \alpha_0 \nabla g(\theta_i)^\top \nabla g(\theta_i, \xi_i) \left(\frac{1}{\sqrt{S_{i-1}}} - \frac{1}{\sqrt{S_i}}\right) + \frac{\mathcal{L}\alpha_0^2}{2} \frac{\left\|\nabla g(\theta_i, \xi_i)\right\|^2}{S_i} \\
&\le -\frac{\alpha_0 \nabla g(\theta_i)^\top \nabla g(\theta_i, \xi_i)}{\sqrt{S_{i-1}}} + \alpha_0 \|\nabla g(\theta_i)\| \cdot \|\nabla g(\theta_i, \xi_i)\| \cdot \frac{\|\nabla g(\theta_i, \xi_i)\|^2}{\sqrt{S_{i-1}}\sqrt{S_i}(\sqrt{S_i} + \sqrt{S_{i-1}})} \\
&+ \frac{\mathcal{L}\alpha_0^2}{2} \frac{\left\|\nabla g(\theta_i, \xi_i)\right\|^2}{S_i} \le -\frac{\alpha_0 \nabla g(\theta_i)^\top \nabla g(\theta_i, \xi_i)}{\sqrt{S_{i-1}}} + \alpha_0 \frac{\|\nabla g(\theta_i)\| \cdot \|\nabla g(\theta_i, \xi_i)\|^2}{S_{i-1}} \\
&= -\frac{\alpha_0 \|\nabla g(\theta_i)\|^2}{\sqrt{S_{i-1}}} + \alpha_0 \frac{\|\nabla g(\theta_i)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\|^2 \middle| \mathcal{F}_i\right)}{S_{i-1}} + \frac{\mathcal{L}\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\|^2 \middle| \mathcal{F}_i\right)}{S_i} \\
&+ P_i + Q_i + R_i,
\end{aligned} \tag{58}$$

where

$$P_i := \alpha_0 \frac{\nabla g(\theta_i)^\top (\nabla g(\theta_i) - \nabla g(\theta_i, \xi_i))}{\sqrt{S_{i-1}}}, \ \ Q_i := \alpha_0 \frac{\|\nabla g(\theta_i)\| \cdot (\|\nabla g(\theta_i, \xi_i)\|^2) - \mathbb{E}(\|\nabla g(\theta_i, \xi_i)\|^2 | \mathcal{F}_i)}{S_{i-1}}$$

$$R_i := \frac{\mathcal{L}\alpha_0^2}{2} \frac{\|\nabla g(\theta_i, \xi_i)\|^2 - \mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\|^2 \middle| \mathcal{F}_i\right)}{S_i}.$$

Then for any $i \in (k_{2n-1}, k_{2n})$, we calculate $g(\theta_i) - g(\theta_{k_{2n-1}+1})$ as follow:

$$g(\theta_i) - g(\theta_{k_{2n-1}+1}) = \sum_{j=k_{2n-1}+1}^{i-1} \big(g(j+1) - g(j)\big).$$

By substituting *Equation* 58 into the above equation, we can obtain the following expression:

$$\begin{aligned}
&g(\theta_i) - g(\theta_{k_{2n-1}+1}) \\
&\le -\sum_{j=k_{2n-1}+1}^{i-1} \frac{\alpha_0 \|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} + \sum_{j=k_{2n-1}+1}^{i-1} \frac{\alpha_0 \|\nabla g(\theta_j)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_j\right)}{S_{j-1}} \\
&+ \sum_{j=k_{2n-1}+1}^{i-1} \frac{\mathcal{L}\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_i\right)}{S_j} + \sum_{j=k_{2n-1}+1}^{i-1} \big(P_j + Q_j + R_j\big).
\end{aligned}$$

Due to $\|\nabla g(\theta_i)\|^2 \ge \delta_0/2 \ \forall\, i \in (k_{2r_n-1}, k_{2r_n})$ we know $Q_i + R_i = \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot (Q_i + R_j)\ \ a.s.$, which conclude

$$\begin{aligned}
&g(\theta_i) - g(\theta_{k_{2n-1}+1}) \\
&\le -\sum_{j=k_{2n-1}+1}^{i-1} \frac{\alpha_0 \|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} + \sum_{j=k_{2n-1}+1}^{i-1} \frac{\alpha_0 \|\nabla g(\theta_j)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_j\right)}{S_{j-1}} \\
&+ \sum_{j=k_{2n-1}+1}^{i-1} \frac{\mathcal{L}\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_i\right)}{S_j} + \sum_{j=k_{2n-1}+1}^{i-1} \big(P_j + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot (Q_j + R_j)\big).
\end{aligned} \tag{59}$$

We assign $R_n := \max_{k_{2n-1}<i<k_{2n}} \|\nabla g(\theta_i)\|^2,\ \tilde{R}_n := \max_{k_{2n-1}<i<k_{2n}} g(\theta_i)$. We get

$$R_n \le \frac{\sqrt{\delta_0}}{\sqrt{2}} + \mathcal{L} \sum_{i=k_{2n-1}}^{k_{2n}-1} \|\theta_{i+1} - \theta_i\| \le \frac{\sqrt{\delta_0}}{\sqrt{2}} + \mathcal{L}\alpha_0 \sum_{i=k_{2n-1}}^{k_{2n}-1} \frac{\mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\| \middle| \mathcal{F}_i\right)}{\sqrt{S_{i-1}}} + \sum_{i=k_{2n-1}}^{k_{2n}-1} T_i, \tag{60}$$

and

$$\tilde{R}_n \leq g(\theta_{k_{2n-1}}) + R_n \sum_{i=k_{2n-1}}^{k_{2n}-1} \|\theta_{i+1} - \theta_i\| \leq g(\theta_{k_{2n-1}}) + \alpha_0 R_n \sum_{i=k_{2n-1}}^{k_{2n}-1} \frac{\mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\| \middle| \mathcal{F}_i\right)}{\sqrt{S_{i-1}}} + \frac{R_n}{\mathcal{L}} \sum_{i=k_{2n-1}}^{k_{2n}-1} T_i < \frac{\sqrt{\delta_0}}{2}, \tag{61}$$

where

$$T_i := \frac{\mathcal{L}\alpha_0 \|\nabla g(\theta_i, \xi_i)\| - \mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\| \middle| \mathcal{F}_i\right)}{\sqrt{S_{i-1}}}.$$

Using *The Jensen's Inequality*, we know

$$\begin{aligned}
&\mathcal{L}\alpha_0 \sum_{i=k_{2n-1}}^{k_{2n}-1} \frac{\mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\| \middle| \mathcal{F}_i\right)}{\sqrt{S_{i-1}}} \leq \mathcal{L}\alpha_0 \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\sqrt{\mathbb{E}\left(\|\nabla g(\theta_i, \xi_i)\|^2 \middle| \mathcal{F}_i\right)}}{\sqrt{S_{i-1}}} \\
&\leq \mathcal{L}\alpha_0 \left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}} + \sqrt{\sigma_1} \right) \cdot \frac{1}{\sqrt{S_{k_{2n}}}} + \mathcal{L}\alpha_0 \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\sqrt{\sigma_0}\|\nabla g(\theta_i)\|}{\sqrt{S_{i-1}}} + \mathcal{L}\alpha_0 \sum_{i=k_{2n+1}}^{k_{2n}-1} \frac{\sqrt{\sigma_1}}{\sqrt{S_{i-1}}} \\
&\leq \mathcal{L}\alpha_0 \left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}} + \sqrt{\sigma_1} \right) \cdot \frac{1}{\sqrt{S_{k_{2n}}}} + \left(\frac{\sqrt{2}\mathcal{L}\alpha_0\sqrt{\sigma_0}}{\sqrt{\delta_0}} + \frac{2\mathcal{L}\alpha_0\sqrt{\sigma_1}}{\delta_0} \right) \cdot \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\|\nabla g(\theta_i)\|^2}{\sqrt{S_{i-1}}}.
\end{aligned} \tag{62}$$

Substitute *Equation* 62 into *Equation* 60, getting

$$R_n \leq \frac{\sqrt{\delta_0}}{\sqrt{2}} + \mathcal{L}\alpha_0 \left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}} + \sqrt{\sigma_1} \right) \frac{1}{\sqrt{S_{k_{2n}}}} + u \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_i)\|^2}{\sqrt{S_{i-1}}} + \sum_{i=k_{2n-1}}^{k_{2n}-1} T_i \tag{63}$$

and

$$\tilde{R}_n \leq g(\theta_{k_{2n-1}}) + R_n\alpha_0 \left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}} + \sqrt{\sigma_1} \right) \cdot \frac{1}{\sqrt{S_{k_{2n}}}} + \frac{uR_n}{\mathcal{L}} \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_i)\|^2}{\sqrt{S_{i-1}}} + \frac{R_n}{\mathcal{L}} \sum_{i=k_{2n-1}}^{k_{2n}-1} T_i, \tag{64}$$

where

$$u := \frac{\sqrt{2}\mathcal{L}\sqrt{\sigma_0}}{\sqrt{\delta_0}} + \frac{2\mathcal{L}\sqrt{\sigma_1}}{\delta_0}.$$

For *Equation* 59, we take $i = k_{2n-1}$, acquiring

$$\begin{aligned}
&g(\theta_{k_{2n}}) - g(\theta_{k_{2n-1}+1}) \\
&\leq -\sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} + \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_j)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_j\right)}{S_{j-1}} \\
&+ \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\mathcal{L}\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_i\right)}{S_j} + \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \left(P_j + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \geq \delta_0/2) \cdot (Q_j + R_j)\right).
\end{aligned} \tag{65}$$

Next we calculate *Equation* 63+$u\cdot$*Equation* 65, getting

$$\begin{aligned}
&R_n + u\big(g(\theta_{k_{2n}}) - g(\theta_{k_{2n-1}+1})\big) \le u \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_i)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_{i-1}} \\
&+ \frac{\sqrt{\delta_0}}{\sqrt{2}} + u \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\mathcal{L}\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_i} + \mathcal{L}\alpha_0\left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}} + \sqrt{\sigma_1}\right)\frac{1}{\sqrt{S_{k_{2n}}}} \\
&+ \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \big(uP_i + T_i + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot u(Q_i + R_i)\big)
\end{aligned} \tag{66}$$

and *Equation* 64+$uR_n/\mathcal{L}\cdot$*Equation* 65, getting

$$\begin{aligned}
&\tilde{R}_n + \frac{uR_n}{\mathcal{L}}\big(g(\theta_{k_{2n}}) - g(\theta_{k_{2n-1}+1})\big) \le \frac{uR_n}{\mathcal{L}} \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_i)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_{i-1}} \\
&+ g(\theta_{k_{2n-1}}) + \frac{uR_n}{\mathcal{L}} \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\mathcal{L}\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_i} \\
&+ R_n\alpha_0\left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}} + \sqrt{\sigma_1}\right)\frac{1}{\sqrt{S_{k_{2n}}}} \\
&+ \frac{uR_n}{\mathcal{L}} \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \big(uP_i + T_i + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot (Q_i + R_i)\big).
\end{aligned} \tag{67}$$

Then for P_i, we test following series, acquiring

$$\begin{aligned}
&\sum_{i=2}^{+\infty} \mathbb{E}\left(\|P_i\|^2 \big| \mathcal{F}_i\right) \le \sum_{i=2}^{+\infty} \frac{\|\nabla g(\theta_i)\|^2 \cdot (\sigma_0\|\nabla g(\theta_i)\|^2 + \sigma_1)}{S_{i-1}} \\
&\le \sigma_0 \sum_{n=1}^{+\infty} \frac{\|\nabla g(\theta_n)\|^4}{S_{n-1}} + \sigma_1 \sum_{n=1}^{+\infty} \frac{\|\nabla g(\theta_n)\|^2}{S_{n-1}} < +\infty \ a.s..
\end{aligned} \tag{68}$$

Through Lemma A.2, we know $\sum_{i=1}^{+\infty} uP_i$ is convergence almost surely. Similarly, we know $\sum_{i=1}^{+\infty} T_i$ is convergence almost sirely. For $\mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot Q_i$, we test $\sum_{i=1}^{n} \mathbb{E}(\mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot \|Q_i\| | \mathcal{F}_i)$, getting

$$\begin{aligned}
&\sum_{i=2}^{+\infty} \mathbb{E}\left(\mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot \|Q_i\| \big| \mathcal{F}_i\right) \\
&\le 2\sum_{i=2}^{+\infty} \frac{\sigma_0 \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2)\|\nabla g(\theta_i)\|^3}{S_{i-1}} + 2\sum_{i=2}^{+\infty} \frac{\sigma_1 \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2)}{S_{i-1}} \\
&\le \frac{2}{\sqrt{r}} \sum_{i=2}^{+\infty} \frac{\sigma_0\|\nabla g(\theta_i)\|^4}{S_{i-1}} + \frac{2}{r}\sum_{i=2}^{+\infty} \frac{\|\nabla g(\theta_i)\|^2}{S_{i-1}} \\
&\le \frac{2}{\sqrt{r}} \sum_{i=2}^{+\infty} \frac{\sigma_0\|\nabla g(\theta_i)\|^2}{S_{i-1}^{\frac{1}{2}+\epsilon}} + \frac{2}{r}\sum_{i=2}^{+\infty} \frac{\|\nabla g(\theta_i)\|^2}{S_{i-1}} < +\infty \ \ a.s.,
\end{aligned} \tag{69}$$

where the last inequality is derived from *Equation* 53, and the second-to-last inequality is derived from Lemma A.5. Through Lemma A.2, we know $\sum_{i=1}^{+\infty} \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2)\cdot uQ_i$ is convergence almost surely. Similarly, we get $\sum_{i=1}^{+\infty} \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot uR_i$ is convergence almost surely. As a result, we have

$$\sum_{i=1}^{+\infty} \big(uP_i + T_i + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \ge \delta_0/2) \cdot u(Q_i + R_i)\big)$$

is convergence almost surely. Then we use *The Cauchy's Convergence Test*, acquiring

$$\lim_{n\to+\infty}\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\big(uP_i+T_i+\mathbf{1}(\|\nabla g(\theta_i)\|^2\geq\delta_0/2)\cdot u(Q_i+R_i)\big)=0\ \ a.s..$$

Furthermore, according to the Lemma A.1 and Lemma A.5, it is evident that we have

$$\lim_{n\to+\infty}\frac{1}{\sqrt{S_{n-1}}}=0\ \ a.s..$$

Similarly, according to *Equation* 55, we can also obtain

$$\lim_{n\to+\infty}\left(u\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\alpha_0\|\nabla g(\theta_i)\|\cdot\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2\big|\mathcal{F}_i\right)}{S_{i-1}}+u\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{c\alpha_0^2}{2}\frac{\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2\big|\mathcal{F}_i\right)}{S_i}\right)$$
$$=0\ \ a.s.,$$

and

$$\lim_{n\to+\infty}c\alpha_0\sum_{i=k_{2n}}^{k_{2n}-1}\frac{\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|\big|\mathcal{F}_i\right)}{\sqrt{S_{i-1}}}=0\ \ a.s..$$

This implies that there exists a positive integer n_0 such that for any $n>n_0$, we have

$$\begin{aligned}&u\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\alpha_0\|\nabla g(\theta_i)\|\cdot\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2\big|\mathcal{F}_i\right)}{S_{i-1}}+c\alpha_0\left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}}+\sqrt{\sigma_1}\right)\frac{1}{\sqrt{S_{k_{2n}}}}\\&+\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\big(uP_i+T_i+\mathbf{1}(\|\nabla g(\theta_i)\|^2\geq\delta_0/2)\cdot u(Q_i+R_i)\big)<\Big(1-\frac{1}{\sqrt{2}}\Big)\sqrt{\delta_0},\end{aligned}\tag{70}$$

and

$$\begin{aligned}&\frac{u}{\mathcal{L}}\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\alpha_0\|\nabla g(\theta_i)\|\cdot\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2\big|\mathcal{F}_i\right)}{S_{i-1}}+\alpha_0\left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}}+\sqrt{\sigma_1}\right)\frac{1}{\sqrt{S_{k_{2n}}}}\\&+\frac{u}{\mathcal{L}}\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\big(uP_i+T_i+\mathbf{1}(\|\nabla g(\theta_i)\|^2\geq\delta_0/2)\cdot(Q_i+R_i)\big)<\frac{\delta_1}{2\sqrt{\delta_0}}.\end{aligned}\tag{71}$$

Since $(g\circ\mathcal{S})\cap\left(\bigcup_{n=0}^{+\infty}\mathcal{H}_{x,\delta,n}\right)=\emptyset$, we know that $g(\theta_{k_{2n-1}+1})$ and $g(\theta_{k_{2n-1}})$ must lie between two intervals $\mathcal{H}'_{x,\delta,n_1-1}$ and $\mathcal{H}'_{x,\delta,n_1}$. Now, we construct a mapping $g^{(2n-1)}:=x+n_1\delta$. Then through *Equation* 70 and *Equation* 71, we have for any $n>n_0$, there is

$$\begin{aligned}R_n&\leq u\big(g(\theta_{k_{2n-1}+1})-g(\theta_{k_{2n}})\big)\\&+u\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\alpha_0\|\nabla g(\theta_i)\|\cdot\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2\big|\mathcal{F}_i\right)}{S_{i-1}}+\frac{\sqrt{\delta_0}}{\sqrt{2}}\\&+u\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\mathcal{L}\alpha_0^2}{2}\frac{\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2\big|\mathcal{F}_i\right)}{S_i}+\mathcal{L}\alpha_0\left(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}}+\sqrt{\sigma_1}\right)\frac{1}{\sqrt{S_{k_{2n}}}}\\&+\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\big(uP_i+T_i+\mathbf{1}(\|\nabla g(\theta_i)\|^2\geq\delta_0/2)\cdot u(Q_i+R_i)\big)<\sqrt{\delta_0}\end{aligned}\tag{72}$$

and

$$\tilde{R}_n \leq \frac{uR_n}{\mathcal{L}}\big(g(\theta_{k_{2n-1}+1}) - g(\theta_{k_{2n}})\big) + \frac{uR_n}{\mathcal{L}}\sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_i)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \middle| \mathcal{F}_i\right)}{S_{i-1}}$$
$$+ g(\theta_{k_{2n-1}}) + \frac{uR_n}{\mathcal{L}}\sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\mathcal{L}\alpha_0^2}{2}\frac{\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \middle| \mathcal{F}_i\right)}{S_i}$$
$$+ R_n\alpha_0\bigg(\frac{\sqrt{\sigma_0}\sqrt{\delta_0}}{\sqrt{2}} + \sqrt{\sigma_1}\bigg)\frac{1}{\sqrt{S_{k_{2n}}}}$$
$$+ \frac{uR_n}{\mathcal{L}}\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\big(uP_i + T_i + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \geq \delta_0/2)\cdot(Q_i + R_i)\big) < g(\theta_{k_{2n-1}}) + \frac{\delta_1}{2}.$$

According to the definitions of $g^{(2n)}$ and $g^{(2n-1)}$, we know that in order for $g^{(2n)} > g^{(2n-1)}$ to hold, we must have at least $g^{(2n)} - g^{(2n-1)} > \delta_1/2$. However, since $\tilde{R}_n \leq g(\theta_{k_{2n-1}}) + \delta_1/2$, we can deduce that there must be $g^{(2n)} \leq g^{(2n-1)}$ $(\forall\, n > n_0)$. We have actually proven that $\{g^{(n)}\}_{n=n_0+1}^{+\infty}$ is a monotonic sequence. According to *The Monotone Convergence Theorem*, we can prove the existence of a limit for $\{g^{(n)}\}_{n=1}^{+\infty}$. Furthermore, since the sequence $\{g^{(n)}\}_{n=n_0+1}^{+\infty}$ only has a finite number of distinct values, we can assert that there exists an n'' such that for $n > n''$, we have $g^{(n)} \equiv g_*$.

Next, we will prove that the sequence of loss functions $\{g(\theta_n)\}_{n=1}^{+\infty}$ converges almost surely. For any $i+1 \in (k_{2n-1}, k_{2n})$, we have

$$\begin{aligned}
&g(\theta_{i+1}) - g(\theta_i)\\
&= \nabla g(\theta_i')^\top(\theta_{i+1} - \theta_i) = \nabla g(\theta_i)^\top(\theta_{i+1} - \theta_i) + (\nabla g(\theta_i') - \nabla g(\theta_i))^\top(\theta_{i+1} - \theta_i)\\
&= -\frac{\alpha_0 \nabla g(\theta_I)^\top \nabla g(\theta_i,\xi_i)}{\sqrt{S_i}} + (\nabla g(\theta_i') - \nabla g(\theta_i))^\top(\theta_{i+1} - \theta_i)\\
&= -\frac{\alpha_0\|\nabla g(\theta_i)\|^2}{\sqrt{S_{i-1}}} + \alpha_0\nabla g(\theta_i)^\top\nabla g(\theta_i,\xi_i)\bigg(\frac{1}{\sqrt{S_{i-1}}} - \frac{1}{\sqrt{S_i}}\bigg)\\
&+ (\nabla g(\theta_i') - \nabla g(\theta_i))^\top(\theta_{i+1} - \theta_i) + P_i,
\end{aligned} \tag{73}$$

where we utilized *The Lagrange's Mean Value Theorem* and θ_i' lies between θ_i and θ_{i+1}. P_i is defined in *Equation* 58. Note that *Equation* 73 does not overlap with *Equation* 58. Since *Equation* 58 contains an inequality, it will have an impact on subsequent estimations. We back to *Equation* 65, we have

$$\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\alpha_0\|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} \leq g(\theta_{k_{2n-1}+1}) - g(\theta_{k_{2n}}) + \sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\alpha_0\|\nabla g(\theta_j)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_j,\xi_j)\|^2 \middle| \mathcal{F}_j\right)}{S_{j-1}}$$
$$+ \sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\mathcal{L}\alpha_0^2}{2}\frac{\mathbb{E}\left(\|\nabla g(\theta_j,\xi_j)\|^2 \middle| \mathcal{F}_i\right)}{S_j} + \sum_{i=k_{2n-1}+1}^{k_{2n}-1}\big(P_j + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \geq \delta_0/2)\cdot(Q_j + R_j)\big).$$

It is easy to find that for any $n > \max\{n_0, n''\}$, we have

$$\sum_{i=k_{2n-1}+1}^{k_{2n}-1}\frac{\alpha_0\|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} < \Big(1 - \frac{1}{\sqrt{2}}\Big)\sqrt{\delta_0} + 2\delta < \Big(\frac{1}{\sqrt{2}} - \frac{1}{2}\Big)\sqrt{\delta} + 2\delta. \tag{74}$$

Based on *Equation* 73, we know for any $i+1 \in (k_{2n-1}, k_{2n})$, there is

$$\begin{aligned}
&g(\theta_{i+1}) - g(\theta_{k_{2n-1}})\\
&= -\sum_{j=k_{2n-1}}^{i}\frac{\alpha_0\|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} + \alpha_0\sum_{j=k_{2n-1}}^{i}\nabla g(\theta_j)^\top\nabla g(\theta_j,\xi_j)\bigg(\frac{1}{\sqrt{S_{j-1}}} - \frac{1}{\sqrt{S_j}}\bigg)\\
&+ \sum_{j=k_{2n-1}}^{i}(\nabla g(\theta_j') - \nabla g(\theta_j))^\top(\theta_{j+1} - \theta_j) + \sum_{j=k_{2n-1}}^{i}P_j,
\end{aligned}$$

which means when $n > \max\{n_0, n''\}$, there is

$$\begin{aligned}&|g(\theta_{i+1}) - g(\theta_{k_{2n-1}})|\\&\leq \sum_{j=k_{2n-1}}^{k_{2n}} \frac{\alpha_0\|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} + \alpha_0 \sum_{j=k_{2n-1}}^{i} \left|\nabla g(\theta_j)^\top \nabla g(\theta_j, \xi_j)\Big(\frac{1}{\sqrt{S_{j-1}}} - \frac{1}{\sqrt{S_j}}\Big)\right|\\&+ \sum_{j=k_{2n-1}}^{i} \big|(\nabla g(\theta_j') - \nabla g(\theta_j))^\top(\theta_{j+1} - \theta_j)\big| + \left|\sum_{j=k_{2n-1}}^{i} P_j\right|\\&< 2\Big(\frac{1}{\sqrt{2}} - \frac{1}{2}\Big)\sqrt{\delta} + 2\delta.\end{aligned}$$

That means for any $n > k_{2\max\{n_0, n''\}-1}$, there is

$$|g(\theta_n) - g_*| < 2\Big(\frac{1}{\sqrt{2}} - \frac{1}{2}\Big)\sqrt{\delta} + 3\delta.$$

By the arbitrariness of δ_0, we can prove that $\{g(\theta_n)\}_{n=1}^{+\infty}$ is convergence almost surely. We back to *Equation* 72, we have for any $n > n_0$.

$$\begin{aligned}&R_n \leq u\big|g(\theta_{k_{2n-1}+1}) - g(\theta_{k_{2n}})\big| + \frac{\sqrt{\delta_0}}{\sqrt{2}} + \Big(1 - \frac{1}{\sqrt{2}}\Big)\sqrt{\delta_0} = u\big(g(\theta_{k_{2n-1}+1}) - g(\theta_{k_{2n}})\big)\\&+ \sqrt{\delta_0}.\end{aligned} \tag{75}$$

Since we have already proven that the sequence $\{g(\theta_n)\}_{n=1}^{+\infty}$ convergence almost surely, according to *The Cauchy's Convergence Test*, we obtain the existence of n''', such that for $n > \max\{n''', n_0\}$, we have $\big|g(\theta_{k_{2n-1}+1}) - g(\theta_{k_{2n}})\big| < \delta_0/u$. Then we get when $n > \max\{n''', n_0\}$, we have

$$\begin{aligned}&R_n \leq u\big|g(\theta_{k_{2n-1}+1}) - g(\theta_{k_{2n}})\big| + \frac{\sqrt{\delta_0}}{\sqrt{2}} + \Big(1 - \frac{1}{\sqrt{2}}\Big)\sqrt{\delta_0} = u\big(g(\theta_{k_{2n-1}+1}) - g(\theta_{k_{2n}})\big) + \sqrt{\delta_0}\\&< \delta_0 + \sqrt{\delta_0},\end{aligned}$$

which implies that for any $i \in (k_{2n-1}, k_{2n})$, we have $\|\nabla g(\theta_i)\|^2 \leq R_n^2 \leq (\delta_0 + \sqrt{\delta_0})^2$. Consequently, we can prove that for any $n > k_{2\max\{n''', n_0\}-1}$, there is $\|\nabla g(\theta_n)\|^2 < (\delta_0 + \sqrt{\delta_0})^2$. By the arbitrariness of δ_0, we can prove that

$$\lim_{n \to +\infty} \big\|\nabla g(\theta_n)\big\| = 0 \ \ a.s. \tag{76}$$

With this, we complete the proof.

□

D Proof of Theorem 3.2

Proof. Without loss of generality, in this proof, we assume that $\inf_{\theta\in\mathbb{R}^d} g(\theta) = 1$ (If this condition is not satisfied, we can construct a new loss function $g_{\text{new}} = g - \inf_{\theta\in\mathbb{R}^d} g(\theta) + 1$). In the proof of Theorem 3.1, we have actually shown that the sequence $\{\|\nabla g(\theta_n)\|^2\}_{n=1}^{+\infty}$ converges to 0 almost surely. According to *The Lebesgue's Dominated Convergence Theorem*, in order to obtain mean-square convergence, we only need to find a h^* such that it satisfies $\|\nabla g(\theta_n)\|^2 \leq h^*$ and $\mathbb{E}\left(|h^*|\right) < +\infty$. However, we know that for any $\|\nabla g(\theta_n)\|^2$, we always have $\|\nabla g(\theta_n)\|^2 \leq \sup_{k\geq 1} \|\nabla g(\theta_k)\|^2$. Therefore, the objective is to prove $\mathbb{E}\left(\sup_{k\geq 1} \|\nabla g(\theta_k)\|^2\right) < +\infty$. In the proof of Theorem 3.1, since we are considering convergence in the trajectory sense (almost surely sense), we do not need to consider the randomness of certain subscripts. However, in the discussion of mean-square convergence, we must also take into account the randomness of these subscripts.

Now we begin to prove this theorem. We divide the discussion of this problem into two cases. The first case is when $g(\theta)$ $(\theta \in \mathbb{R}^d)$ is bounded. In this case, it is evident from the inequality in Lemma A.1 that we can obtain $\mathbb{E}\left(\sup_{k\geq 1} \|\nabla g(\theta_k)\|^2\right) < +\infty$ immediately.

Next, we focus on the case where g is unbounded. From Lemma 4.2, for any u, we know that when $\|\nabla g(\theta_n)\|^2 > u$, there is

$$
\begin{aligned}
g^2(\theta_{n+1}) - g^2(\theta_n) \le\ & \alpha_0(M+1)\left(\frac{g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_{n+1})\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right)\\
& + \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_n} + c\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_n}\\
& - \frac{\alpha_0 g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}} + \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\frac{g(\theta_n)\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}}\\
& + \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} + X_n,
\end{aligned}
$$

where

$$X_n := \frac{2\alpha_0 g(\theta_n)}{\sqrt{n-1}}\nabla g(\theta_n)^\top(\nabla g(\theta_n) - \nabla g(\theta_n,\xi_n)).$$

We define a new object

$$\hat{g}_n := g^2(\theta_n) + \alpha_0(M+1)\frac{g(\theta_n)\|\nabla g(\theta_{n-1})\|^2}{\sqrt{S_{n-1}}}.$$

Obviously, we know $\hat{g}_n \in \mathcal{F}_n$, and

$$
\begin{aligned}
\hat{g}_{n+1} - \hat{g}_n \le\ & \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_n} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_n}\\
& - \frac{\alpha_0 g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}} + \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\frac{g(\theta_n)\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}}\\
& + \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} + X_n.
\end{aligned}
$$

Then we take $u := \eta$, (η shown in Assumption 2.1). Through we know that $\forall\, \theta \in \{\|\nabla g(\theta)\|^2 < u\}$, there is

$$g(\theta) < \hat{u}_0 < +\infty,$$

which means $\exists\, u_0, \forall\, \theta \in \{\theta|\hat{g}(\theta) > u_0\}$, there is $\|\nabla g(\theta)\|^2 > u$. Then for any $\lambda > 0$, we construct events $\mathcal{C}_n := \{\|\nabla g(\theta_n)\|^2 > u\} \cap \{u_0 < \hat{g}(\theta_n) < \lambda\}$. Then we construct a series of stopping times $\{\tau_i^{(\lambda)}\}_{i=1}^{+\infty}$ as follow:

$$
\begin{aligned}
&\tau_1^{(\lambda)} := \min\{k:\ k\ge 1,\ \mathcal{C}_k \text{ occurs}\},\ \ \tau_2^{(\lambda)} := \min\{k: k > \tau_1^{(\lambda)},\ \mathcal{C}_k \text{ does not occur}\},...,\\
&\tau_{2m-1}^{(\lambda)} := \min\{k: k > \tau_{2m-2}^{(\lambda)},\ \mathcal{C}_k \text{ occurs}\},\ \ \tau_{2m}^{(\lambda)} := \min\{k: k \ge \tau_{2m-1}^{(\lambda)},\ \mathcal{C}_k \text{ does not occur}\}.
\end{aligned}
$$

Obviously, we get for any i, j, there is $\tau_i^{(\lambda)} < \tau_j^{(\lambda)}$ and $\mathcal{F}_{\tau_i^{(\lambda)}} \subset \mathcal{F}_{\tau_j^{(\lambda)}}$. Then we define another stopping times

$$\tau := \min\{k:\ \hat{g}(\theta_1) < \lambda, \hat{g}(\theta_2) < \lambda, ...,\ \hat{g}(\theta_k) < \lambda\}\ ,\ \ \tau' := \{k: \theta_k \in \mathcal{R}\}\ ,\ \ \tau'' := \{k: \theta_k \in \mathcal{K}\}\ .$$

Then we define events

$$\mathcal{B}_{i,k} := \{C_i \text{ does not occur}, \mathcal{C}_{i+1} \text{ occurs}, ...,\ \mathcal{C}_k \text{ occurs}\}\ (k \ge i+1),$$

$$\mathcal{B}_{i,k} := \{\mathcal{C}_i \text{ does not occur}\}\ (k \le i),\ \ \mathcal{B}'_{i,k} := \mathcal{B}_{i,k-1}/\mathcal{B}_{i,k}.$$

Then for any events $\mathcal{X} \in \mathcal{F}_i$, we have

$$\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_{m+1} - \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_m \leq \alpha_0(M+1)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^3 \cdot \|\nabla g(\theta_m, \xi_m)\|}{S_m}$$
$$+ \mathcal{L}\alpha_0^2(M+1)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^2 \cdot \|\nabla g(\theta_m, \xi_m)\|}{S_m}$$
$$- \nabla\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\alpha_0 g(\theta_m)\|\nabla g(\theta_m)\|^2}{\sqrt{S_{m-1}}}$$
$$+ \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{g(\theta_m)\|\nabla g(\theta_{m-1}, \xi_{m-1})\|^2}{S_{m-1}^{\frac{3}{2}}}$$
$$+ \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\big(4\|\nabla g(\theta_m)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_m)\big)\frac{\big\|\nabla g(\theta_m, \xi_m)\big\|^2}{S_m} + \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}X_m + \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}Y_m. \tag{77}$$

The reason the above inequality holds is due to the fact that where $\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}$, we have $\|\nabla g(\theta_m)\|^2 > u$. For the left side of the above inequality, we notice

$$\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_{m+1} = (\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}} - \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}})\hat{g}_{m+1} + \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1}$$
$$= \mathbf{1}_{\mathcal{X}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1} + \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1}.$$

Then we acquire

$$\begin{aligned}
&\mathbf{1}_{\mathcal{X}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1} + \big(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1} - \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_m\big)\\
&\leq \alpha_0(M+1)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^3 \cdot \|\nabla g(\theta_m, \xi_m)\|}{S_m}\\
&+ \mathcal{L}\alpha_0^2(M+1)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^2 \cdot \|\nabla g(\theta_m, \xi_m)\|}{S_m} - \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\alpha_0 g(\theta_m)\|\nabla g(\theta_m)\|^2}{\sqrt{S_{m-1}}}\\
&+ \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{g(\theta_m)\|\nabla g(\theta_{m-1}, \xi_{m-1})\|^2}{S_{m-1}^{\frac{3}{2}}}\\
&+ \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\big(4\|\nabla g(\theta_m)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_m)\big)\frac{\big\|\nabla g(\theta_m, \xi_m)\big\|^2}{S_m} + \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}X_m.
\end{aligned} \tag{78}$$

For convenient, we assign

$$\begin{aligned}
A_{i,m} &:= \alpha_0(M+1)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^3 \cdot \|\nabla g(\theta_m, \xi_m)\|}{S_m}\\
&+ \mathcal{L}\alpha_0^2(M+1)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^2 \cdot \|\nabla g(\theta_m, \xi_m)\|}{S_m}\\
&+ \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{g(\theta_m)\|\nabla g(\theta_m, \xi_m)\|^2}{S_{m-1}^{\frac{3}{2}}}\\
&+ \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\big(4\|\nabla g(\theta_m)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_m)\big)\frac{\big\|\nabla g(\theta_{m-1}, \xi_{m-1})\big\|^2}{S_{m-1}}.
\end{aligned} \tag{79}$$

Then take the mathematical expectation on *Equation* 78, noting $\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}} \in \mathcal{F}_m$, we getting

$$\begin{aligned}
&\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1}\right) \leq -\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1} - \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_m\right)\\
&- \mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\alpha_0 g(\theta_m)\|\nabla g(\theta_m)\|^2}{\sqrt{S_{m-1}}}\right) + \mathbb{E}(A_{i,m}) + 0 + 0.
\end{aligned} \tag{80}$$

For $\mathbb{E}(A_{i,m})$, we have

$$
\begin{aligned}
&\mathbb{E}(A_{i,m}) \leq \alpha_0(M+1)\,\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^3\cdot(\sqrt{\sigma_0}\|\nabla g(\theta_m)\|+\sqrt{\sigma_1})}{S_{m-1}}\right)\\
&+\mathcal{L}\alpha_0^2(M+1)\,\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{\|\nabla g(\theta_m)\|^2\cdot(\sqrt{\sigma_0}\|\nabla g(\theta_m)\|+\sqrt{\sigma_1})}{S_{m-1}}\right)\\
&+\left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2+\Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\frac{g(\theta_m)(\sigma_0\|\nabla g(\theta_m)\|^2+\sigma_1)}{S_{m-1}^{\frac{3}{2}}}\right)\\
&+\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\big(4\|\nabla g(\theta_m)\|^2+4\mathcal{L}\alpha_0+2\mathcal{L}\alpha_0^2g(\theta_m)\big)\frac{(\sigma_0\|\nabla g(\theta_m)\|^2+\sigma_1)}{S_{m-1}}\right).
\end{aligned}
$$

Now let us simplify the inequality above. We notice that where $\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}} = \mathbf{1}_{\mathbb{R}^d}$, we have $u < \|\nabla g(\theta_m)\|^2$, and combine Lemma A.1, that, at the beginning of our proof, we assumed $g(\theta) \geq 1$, we getting

$$
\begin{aligned}
&\sigma_0\|\nabla g(\theta_m)\|^2+\sigma_1 \leq \Big(\sigma_0+\frac{\sigma_1}{u}\Big)\|\nabla g(\theta_m)\|^2,\\
&\text{and}\\
&\|\nabla g(\theta_m)\|^2\big(\sigma_0\|\nabla g(\theta_m)\|^2+\sigma_1\big) \leq \frac{1}{2\mathcal{L}}\Big(\sigma_0+\frac{\sigma_1}{u}\Big)\|\nabla g(\theta_m)\|^2,\\
&\text{and}\\
&\|\nabla g(\theta_m)\|^3(\sqrt{\sigma_0}\|\nabla g(\theta_m)\|+\sqrt{\sigma_1}) \leq \frac{1}{2\mathcal{L}}\left(\sqrt{\sigma_0}+\sqrt{\frac{\sigma_1}{u}}\right)g(\theta_m)\|\nabla g(\theta_m)\|^2,\\
&\text{and}\\
&\|\nabla g(\theta_m)\|^2(\sqrt{\sigma_0}\|\nabla g(\theta_m)\|+\sqrt{\sigma_1}) \leq \frac{1}{\sqrt{2\mathcal{L}}}\left(\sqrt{\sigma_0}+\sqrt{\frac{\sigma_1}{u}}\right)g(\theta_m)\|\nabla g(\theta_m)\|^2\ .
\end{aligned}
\tag{81}
$$

Substitute *Equation* 81 into *Equation* 80, yielding

$$
\begin{aligned}
&\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1}\right) \leq -\,\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1}-\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_m\right)\\
&-\mathbb{E}\left(\mathbf{1}_{S_{m-1}\geq 4\beta_0^2/\alpha_0^2}\cdot\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}g(\theta_m)\|\nabla g(\theta_m)\|^2\left(\frac{\alpha_0}{2\sqrt{S_{m-1}}}-\frac{\beta_0}{S_{m-1}}\right)\right)\\
&+\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(A_{i,m}-\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}g(\theta_m)\|\nabla g(\theta_m)\|^2\frac{\alpha_0}{2\sqrt{S_{m-1}}}\right)\right),
\end{aligned}
\tag{82}
$$

where β_0 is defined as:

$$
\begin{aligned}
\beta_0 &:= \alpha_0(M+1)\left(1+\left(\frac{1}{2\mathcal{L}}+\frac{\mathcal{L}\alpha_0}{\sqrt{2\mathcal{L}}}\right)\left(\sqrt{\sigma_0}+\sqrt{\frac{\sigma_1}{u}}\right)\right)\\
&+\Big(\sigma_0+\frac{\sigma_1}{u}\Big)\frac{1}{\sqrt{S_0}}\left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2+\Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)+\frac{4}{2\mathcal{L}}\Big(\sigma_0+\frac{\sigma_1}{u}\Big)\\
&+(4\mathcal{L}\alpha_0+2\mathcal{L}\alpha_0^2)\Big(\sigma_0+\frac{\sigma_1}{u}\Big).
\end{aligned}
\tag{83}
$$

Then we have

$$
-\,\mathbb{E}\left(\mathbf{1}_{S_{m-1}\geq 4\beta_0^2/\alpha_0^2}\cdot\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}g(\theta_m)\|\nabla g(\theta_m)\|^2\left(\frac{\alpha_0}{2\sqrt{S_{m-1}}}-\frac{\beta_0}{S_{m-1}}\right)\right) \leq 0.
$$

Substitute above inequality into *Equation* 82, we concluding

$$
\begin{aligned}
&\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1}\right) \leq -\,\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1}-\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_m\right)\\
&+\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(A_{i,m}-\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}g(\theta_m)\|\nabla g(\theta_m)\|^2\frac{\alpha_0}{2\sqrt{S_{m-1}}}\right)\right).
\end{aligned}
$$

Combine *Equation* 79, we acquire

$$\begin{aligned}&\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(A_{i,m}-\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}g(\theta_m)\|\nabla g(\theta_m)\|^2\frac{\alpha_0}{2\sqrt{S_{m-1}}}\right)\right)\\&\le\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right),\end{aligned}\tag{84}$$

where $e_0\ge 2,\ e_1\ge 2,\ r_0>0,\ r_1>0$ and $q_0,\ q_1$ are six constants. We know that the necessary condition for the above inequality to hold is $m\ge i+1$. In order to facilitate the subsequent stacking, we also need an inequality for the case when $m=i$. Upon observation, we find that $\mathbf{1}_{\mathcal{B}_{i,i+1}}\le\mathbf{1}_{\mathcal{B}_{i,i}\cap\{\|\nabla g(\theta_i)\|>u-\mathcal{L}\alpha_0\}}$, which means

$$\mathbf{1}_{\mathcal{B}'_{i,i+1}}\hat{g}_{m+1}+\mathbf{1}_{\mathcal{B}_{i,i+1}}\hat{g}_{m+1}-\mathbf{1}_{\mathcal{B}_{i,i}}\hat{g}_m\le\mathbf{1}_{\mathcal{B}_{i,i}\cap\{\|\nabla g(\theta_i)\|>\sqrt{\eta}-\mathcal{L}\alpha_0\}}\big(\hat{g}_{m+1}-\hat{g}_m\big).$$

Then, next, we can use the inequality *Equation* 5 when $u:=(\sqrt{\eta}-\mathcal{L}\alpha_0)^2$ to complete the inequality. We define

$$\tau^{(0)}:=\min\{k:g(\theta_k)\ge\lambda\},\ \ \tau_m^{(0)}:=\min\{k:g(\theta_k)\ge\lambda,\ \ k\ge\tau_{2m-1}^{(\lambda)}\wedge\tau\}.$$

We take $\mathcal{X}=\{\tau\wedge\tau_{2m-1}^{(\lambda)}\wedge n=i\}$, and make a sum, getting

$$\begin{aligned}&\mathbb{E}(\hat{g}_{\tau_m^{(0)}\wedge n})<\sum_{i=\tau_{2n-2}^{(\lambda)}}^{n-1}\ \sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{n-1}\mathbb{E}\left(\mathbf{1}_{\{\tau\wedge\tau_{2m-1}^{(\lambda)}\wedge n=i\}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1}\right)\\&\le u_0\big(\mathbb{E}(\mathbf{1}_{\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n})-\mathbb{E}(\mathbf{1}_{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n})\big)\\&+\beta_0\sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n}\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right),\end{aligned}\tag{85}$$

where $\sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n}\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right)$ defined as

$$\begin{aligned}&\sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n}\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right.\\&=\mathbb{E}\left(\sum_{m=1}^{\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right)\\&-\mathbb{E}\left(\sum_{m=1}^{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n}\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right).\end{aligned}$$

Next we define $\overline{\|\nabla g(\theta_n)\|^2}:=\sup_{1\le k\le n}\|\nabla g(\theta_n)\|^2$ and $\overline{g}_n:=\sup_{1\le k\le n}g^2(\theta_k)$. We can get for any $\lambda\ge u$, there is

$$\{\overline{\|\nabla g(\theta_n)\|^4}>\lambda\}\subset\{\overline{g}_n>2c\lambda\}\cap\{\overline{\|\nabla g(\theta_n)\|^2}>u\}\subset\Big\{\sup_{1\le k\le n}g_k^2>2c\lambda\Big\}.$$

Using *The Markov's Inequality* and *Equation* 85, we get

$$\begin{aligned}
&\mathbb{P}\left(\overline{\|\nabla g(\theta_n)\|^4} > \lambda\right) \leq \frac{1}{2\mathcal{L}\lambda}\,\mathbb{E}(g_{\tau^{(0)}\wedge n}) \leq \frac{1}{2\mathcal{L}\lambda}\,\mathbb{E}(\hat{g}_{\tau^{(0)}\wedge n}) = \frac{1}{2\mathcal{L}\lambda}\sum_{m=1}^{+\infty}\mathbb{E}(\hat{g}_{\tau_m^{(0)}\wedge n}) \\
&\leq \frac{1}{2\mathcal{L}\lambda}\Bigg(u_0\,\mathbb{E}(\mathbf{1}_{\tau_1^{(\lambda)}}) \\
&+\sum_{m=1}^{+\infty}\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right)\Bigg) \\
&= \frac{1}{2\mathcal{L}\lambda}\left(u_0+\sum_{m=1}^{+\infty}\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right)\right).
\end{aligned} \tag{86}$$

Noting when $S_{m-1} < 4\beta_0^2/\alpha_0^2$, we have

$$\sum_{m=1}^{+\infty}\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}}+\frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right) < K < +\infty.$$

Substitute the bound of $g(\theta_k)$ into *Equation* 86, and use the result presented in Lemma 4.1, we get

$$\mathbb{P}\left(\overline{\|\nabla g(\theta_n)\|^4} > \lambda\right) \leq \frac{K}{2\mathcal{L}\lambda} \leq \frac{T}{\lambda},$$

where $T > 0$ is a finite positive constant. Then we calculate $\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right)$, acquiring

$$\begin{aligned}
&\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right) = \mathbb{E}\left(\sqrt{\overline{\|\nabla g(\theta_n)\|^4}}\right) = u + \int_{\lambda=u}^{+\infty}\lambda^{\frac{1}{2}-1}\,\mathbb{P}\left(\overline{\|\nabla g(\theta_m)\|^4} > \lambda\right)d\lambda \\
&\leq u + T\int_{\lambda=u}^{+\infty}\lambda^{-\frac{3}{2}}d\lambda = u + \frac{2T}{\sqrt{u}} < +\infty.
\end{aligned}$$

Then, according to *The Lebesgue's Monotone Convergence Theorem*, we know that $\lim_{n\to+\infty}\overline{\|\nabla g(\theta_n)\|^2} = \sup_{n\geq 1}\|\nabla g(\theta_n)\|^2\ a.s.$, and

$$\lim_{n\to+\infty}\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right) = \mathbb{E}\left(\sup_{n\geq 1}\|\nabla g(\theta_n)\|^2\right),$$

which implies

$$\mathbb{E}\left(\sup_{n\geq 1}\|\nabla g(\theta_n)\|^2\right) \leq u + \frac{2T}{\sqrt{u}} < +\infty.$$

Then, based on the almost surely convergence result in Theorem 3.1 and *The Lebesgue's Dominated Convergence Theorem*, we can obtain the mean-squre convergence, i.e.,

$$\lim_{n\to\infty}\mathbb{E}\,\|\nabla g(\theta_n)\|^2 = 0.$$

With this, we have completed the proof. □

D.1 PROOF OF THEOREM 3.3

Proof. Through *Equation* 40, we get

$$\frac{\alpha_0}{20}\sum_{k=2}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}} \le \frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_1^{\frac{1}{2}+\epsilon}} + \frac{\alpha_0}{2}(M_0+1)\frac{\big\|\nabla g(\theta_2)\big\|^2}{S_2^{\frac{1}{2}+\epsilon}} + \frac{g(\theta_2)}{S_2^{\epsilon}}$$
$$+ \frac{\alpha_0\sigma_1(M_0+1)}{2S_2^{\frac{1}{2}+\epsilon}} + 4M_0^2\alpha_0^3\mathcal{L}^2\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_{k-1}),\xi_{k-1}\big\|^2}{S_{k-1}^{\frac{3}{2}+\epsilon}}$$
$$+ \frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k^{1+\epsilon}} + 5\alpha_0^3\mathcal{L}^2(M_0+1)^2\sum_{k=2}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2}{S_{k-1}^{1+\epsilon}}$$
$$+ \sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{k-1}^{\frac{3}{2}+\epsilon}}\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2$$
$$+ \frac{\alpha_0\sigma_1(M_0+1)}{2}\sum_{k=3}^{n}\left(\frac{1}{S_{k-1}^{\frac{1}{2}+\epsilon}} - \frac{1}{S_k^{\frac{1}{2}+\epsilon}}\right) + \frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k^{1+\epsilon}}$$
$$+ \sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}G_k^{(\epsilon)} + \mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}H_k^{(\epsilon)} + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}J_k^{(\epsilon)} + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}K_k^{(\epsilon)}\Big),$$

where $M_0 := \sigma_0 + 2$. For each fixed n, taking the limit as $\epsilon \to 0$ on both sides of the above inequality, we obtain:

$$\frac{\alpha_0}{20}\sum_{k=2}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{\sqrt{S_{k-1}}} \le \frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_2)\big\|^2}{\sqrt{S_1}} + \frac{\alpha_0}{2}(M_0+1)\frac{\big\|\nabla g(\theta_2)\big\|^2}{\sqrt{S_2}} + \frac{g(\theta_2)}{S_2^{\epsilon}}$$
$$+ \frac{\alpha_0\sigma_1(M_0+1)}{2S_2^{\frac{1}{2}+\epsilon}} + 4M_0^2\alpha_0^3\mathcal{L}^2\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_{k-1}),\xi_{k-1}\big\|^2}{S_{k-1}^{\frac{3}{2}+\epsilon}}$$
$$+ \frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k} + 5\alpha_0^3\mathcal{L}^2(M_0+1)^2\sum_{k=2}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2}{S_{k-1}}$$
$$+ \sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{k-1}^{\frac{3}{2}}}\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2$$
$$+ \frac{\alpha_0\sigma_1(M_0+1)}{2}\sum_{k=3}^{n}\left(\frac{1}{\sqrt{S_{k-1}}} - \frac{1}{\sqrt{S_k}}\right) + \frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k}$$
$$+ \sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}G_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}H_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}J_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}K_k\Big), \tag{87}$$

where

$$\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}G_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}H_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}J_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}K_k$$
$$:= \frac{\alpha_0}{\sqrt{S_{k-1}}}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\nabla g(\theta_k)^\top\big(\nabla g(\theta_k) - \nabla g(\theta_k,\xi_k)\big)$$
$$+ \frac{\alpha_0}{2}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\left(\frac{1}{M_0+1}\frac{\mathbb{E}\left(\big\|\nabla g(\theta_k,\xi_k)\big\|^2\Big|\mathcal{F}_{k-1}\right)}{\sqrt{S_{k-1}}} - \frac{1}{M_0+1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{\sqrt{S_{k-1}}}\right)$$
$$+ \frac{\alpha_0}{\sqrt{S_{k-1}}}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\left(\big\|\nabla g(\theta_k)\big\|^2 - \nabla g(\theta_k)^\top\nabla g(\theta_k,\xi_k)\right)$$
$$+ \frac{\alpha_0\big\|\nabla g(\theta_k)\big\|^2}{2\sigma_1(M_0+1)\sqrt{S_{k-1}}}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}\left(\big\|\nabla g(\theta_k,\xi_k)\big\|^2 - \mathbb{E}\left(\big\|\nabla g(\theta_k,\xi_k)\big\|^2\Big|\mathcal{F}_{k-1}\right)\right).$$

Accroding Lemma 4.1 and A.3, we know

$$\sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}H_k+\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}G_k+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}J_k+\mathbf{1}_{\|\nabla g(\theta_k)\|^2\le\sigma_1}K_k\Big)<+\infty\ a.s.,$$

Substitute above result into *Equation* 87,acquiring

$$\sum_{n=2}^{T}\frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_{n-1}}}<U_1\ln(S_T)+U_2\ \ a.s.,$$

where U_1 is a constant and $U_2<+\infty\ a.s.$ is a random variable. Then we set $n:=T$, getting

$$\frac{1}{T}\sum_{n=2}^{T}\big\|\nabla g(\theta_n)\big\|^2<\frac{U_1\sqrt{S_T}\ln(S_T)+U_2\sqrt{S_T}}{T}\ \ a.s., \tag{88}$$

For S_T, we have

$$S_T=\sum_{k=1}^{T}\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right)+\sum_{k=1}^{\top}(\|\nabla g(\theta_k,\xi_k)\|^2-\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right))).$$

By the result of Theorem 3.1, we can verify that the martingale difference sequence $\{\|\nabla g(\theta_k,\xi_k)\|^2-\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right)\}_{k=1}^{+\infty}$ satisfies the conditions of Lemma A.4. Therefore, we have:

$$\sum_{k=1}^{T}(\|\nabla g(\theta_k,\xi_k)\|^2-\mathbb{E}\left(\|\nabla g(\theta_k,\xi_k)\|^2\big|\mathcal{F}_k\right)=O(T\ln^{1+\sigma}T)\ \ a.s..$$

Substitute it into *Equation* 88, we get

$$\frac{1}{T}\sum_{k=1}^{T}\big\|\nabla g(\theta_k)\big\|^2=O\bigg(\frac{\ln^{\frac{3}{2}+\sigma}T}{\sqrt{T}}\bigg)\ \ (\forall\ \sigma>0)\ \ a.s..$$

□

D.2 PROOF OF THEOREM 3.4

Proof. Without loss of generality, in this proof, we assume that $\inf_{\theta\in\mathbb{R}^d}g(\theta)=1$ (If this condition is not satisfied, we can construct a new loss function $g_{\text{new}}=g-\inf_{\theta\in\mathbb{R}^d}g(\theta)+1$).

The proof of the convergence rate in this theorem is much simpler compared to the previous two theorems. In this theorem, the key step is to provide an estimate for $g^2(\theta_{n+1})-g^2(\theta_n)$. Since we have already provided an estimate when $\|\nabla g(\theta_n)\|^2>u$ in Lemma 4.2, we only need to provide another estimate when $\|\nabla g(\theta_n)\|^2\le u$, We have

$$\begin{aligned}
&g^2(\theta_{n+1})-g^2(\theta_n)\\
&\le-\frac{2\alpha_0 g(\theta_n)\nabla g(\theta_n)^\top\nabla g(\theta_n,\xi_n)}{\sqrt{S_n}}+\left(4\|\nabla g(\theta_n)\|^2+4\mathcal{L}\alpha_0+2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n}\\
&\le-\frac{2\alpha_0 g(\theta_n)\nabla g(\theta_n)^\top\nabla g(\theta_n,\xi_n)}{\sqrt{S_{n-1}}}+2\alpha_0 g(\theta_n)\nabla g(\theta_n)^\top\nabla g(\theta_n,\xi_n)\bigg(\frac{1}{\sqrt{S_{n-1}}}-\frac{1}{\sqrt{S_n}}\bigg)\\
&+\left(4\|\nabla g(\theta_n)\|^2+4\mathcal{L}\alpha_0+2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n}.
\end{aligned} \tag{89}$$

Under Assumption 2.1, we get that

$$\begin{aligned}
&\big\|\nabla g(\theta_n)\big\|^2\le\big\|\nabla g(\theta_{n-1})\big\|^2\\
&+2\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_n)-\nabla g(\theta_{n-1})\big\|+\big\|\nabla g(\theta_n)-\nabla g(\theta_{n-1})\big\|^2\\
&\le\big\|\nabla g(\theta_{n-1})\big\|^2+\frac{2\alpha_0\mathcal{L}}{\sqrt{S_{n-1}}}\big\|\nabla g(\theta_{n-1})\big\|\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|\\
&+\mathcal{L}^2\alpha_0^2\frac{\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2}{S_{n-1}}.
\end{aligned}$$

Then we use inequality $2a^\top b \le \lambda\|a\|^2 + \frac{1}{\lambda}\|b\|^2 \ (\lambda > 0)$ to get

$$\begin{aligned}&\big\|\nabla g(\theta_n)\big\|^2 - \big\|\nabla g(\theta_{n-1})\big\|^2 \le \frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{10(M+1)} \\ &+ \frac{10\alpha_0^2\mathcal{L}^2(M+1)}{S_{n-1}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2 + \frac{\alpha_0^2\mathcal{L}^2}{S_{n-1}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2.\end{aligned} \tag{90}$$

Divide both sides of *Equation* 90 by $\sqrt{S_{n-1}}$ and notice $S_{n-2} \le S_{n-1} \le S_n$, then we have

$$\begin{aligned}&-\frac{1}{M+1}\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{10\sqrt{S_{n-2}}} - \frac{10\alpha_0^2\mathcal{L}^2(M+1)}{S_{n-1}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2 \\ &-\frac{\alpha_0^2\mathcal{L}^2}{S_{n-1}^{\frac{3}{2}}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2 \le \frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}.\end{aligned} \tag{91}$$

On the other hand, we have

$$\begin{aligned}&\alpha_0 g(\theta_n)(M+1)\left(\frac{\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\ &= \alpha_0(M+1)\left(\frac{g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_{n+1})\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\ &+ \alpha_0(M+1)\big|g(\theta_{n+1}) - g(\theta_n)\big|\frac{\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}} \\ &\le \alpha_0(M+1)\left(\frac{g(\theta_{n-1})\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_n)\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\ &+ \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}}.\end{aligned} \tag{92}$$

Combine *Equation* 92 and *Equation* 91, we getting

$$\begin{aligned}&-\frac{\alpha_0 g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{10\sqrt{S_{n-2}}} - \frac{10\alpha_0^3\mathcal{L}^2(M+1)^2}{S_{n-1}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2 \\ &-\frac{\alpha_0^3 c^2(M+1)}{S_{n-1}^{\frac{3}{2}}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2 \\ &\le \alpha_0(M+1)\left(\frac{g(\theta_{n-1})\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_n)\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\ &+ \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}}.\end{aligned} \tag{93}$$

Substitute *Equation* 93 into *Equation* 89, acquiring

$$\begin{aligned}&g^2(\theta_{n+1}) - g^2(\theta_n) \\ &\le -\frac{2\alpha_0 g(\theta_n)\nabla g(\theta_n)^\top\nabla g(\theta_n,\xi_n)}{\sqrt{S_{n-1}}} + 2\alpha_0 g(\theta_n)\nabla g(\theta_n)^\top\nabla g(\theta_n,\xi_n)\left(\frac{1}{\sqrt{S_{n-1}}} - \frac{1}{\sqrt{S_n}}\right) \\ &+ \big(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\big)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} \\ &+ \alpha_0(M+1)\left(\frac{g(\theta_{n-1})\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_n)\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\ &+ \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2\cdot\|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}} \\ &+ \frac{\alpha_0 g(\theta_n)\big\|\nabla g(\theta_{n-1})\big\|^2}{10\sqrt{S_{n-2}}} + \frac{10\alpha_0^3\mathcal{L}^2(M+1)^2}{S_{n-1}}\big\|\nabla g(\theta_{n-1},\xi_{n-1})\big\|^2.\end{aligned}$$

Now that we have obtained an estimate for $g^2(\theta_{n+1}) - g^2(\theta_n)$ under the condition $\|\nabla g(\theta_n)\|^2 \leq u$. Combining Lemma A.6, we can now derive the estimate for any arbitrary $\theta_n \in \mathbb{R}^d$ as follow:

$$\begin{aligned}
&\mathbb{E}\left(g^2(\theta_{n+1})\right) - \mathbb{E}\left(g^2(\theta_n)\right) \\
&= \mathbb{E}\left(\mathbf{1}_{\|\nabla g(\theta_n)\|^2 \leq u}(g^2(\theta_{n+1}) - g^2(\theta_n))\right) + \mathbb{E}\left(\mathbf{1}_{\|\nabla g(\theta_n)\|^2 > u}(g^2(\theta_{n+1}) - g^2(\theta_n))\right) \\
&\leq -\frac{\alpha_0}{2}\,\mathbb{E}\left(\frac{g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}}\right) + k_0\,\mathbb{E}\left(\frac{\|\nabla g(\theta_n,\xi_n)\|^2}{S_n}\right) \\
&\alpha_0(M+1)\,\mathbb{E}\left(\frac{g(\theta_n)\|\nabla g(\theta_{n-1})\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_{n+1})\|\nabla g(\theta_n)\|^2}{\sqrt{S_n}}\right) + k_1\,\mathbb{E}\left(\frac{1}{\sqrt{S_{n-1}}} - \frac{1}{\sqrt{S_n}}\right).
\end{aligned}$$

Due to the complexity of the specific expressions for k_0 and k_1, we will not provide a detailed description here. Both of these numbers are polynomial functions of S_0, $\mathcal{L}$, u, and M. Here, we explain why we assume $g \geq 1$. In the derivation process, we encounter the term $-g(\theta_n)\|\nabla g(\theta_n)\|^2 + \|\nabla g(\theta_n)\|^2$. By assuming $g \geq 1$, we ensure that this term can be less than 0. Then, we sum up the above inequality and use Property 3.1 to obtain:

$$\begin{aligned}
&\frac{\alpha_0}{2}\sum_{n=1}^{T}\mathbb{E}\left(\frac{g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}}\right) \leq A_0 + k_0\,\mathbb{E}(\ln S_T) \leq A_0 + k_0 \ln\left(\mathbb{E}(S_T)\right) \\
&= O(\ln T),
\end{aligned}$$

where A_0 is a constant determined by $g(\theta_1)$ and $\|\nabla g(\theta_1)\|$. Noting Lemma A.1, we have

$$\frac{\alpha_0}{4\mathcal{L}}\sum_{n=1}^{T}\mathbb{E}\left(\frac{\|\nabla g(\theta_n)\|^4}{\sqrt{S_{n-1}}}\right) \leq \frac{\alpha_0}{2}\sum_{n=1}^{T}\mathbb{E}\left(\frac{g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}}\right) = O(\ln T).$$

According to *Equation* (87), we know that there is

$$\sum_{n=1}^{T}\mathbb{E}\left(\frac{\|\nabla g(\theta_n)\|^2}{\sqrt{S_{n-1}}}\right) \leq A_1 + k_1 \ln(\mathbb{E}(S_T)) = O(\ln T),$$

where A_1, K_1 is two constants. Then for any $2 \leq p \leq 4$, we have

$$\sum_{n=1}^{T}\mathbb{E}\left(\frac{\|\nabla g(\theta_n)\|^p}{\sqrt{S_{n-1}}}\right) \leq Q_1 + Q_2 \ln(\mathbb{E}(S_T)),$$

where $Q_1 > 0$ and $Q_2 > 0$ are two constants that are uniformly bounded for all $p \in (2, 4]$. Through *The Hölder's Inequality*, and using Property 3.1, we have

$$\begin{aligned}
&\sum_{n=1}^{T}\mathbb{E}\left(\frac{\|\nabla g(\theta_n)\|^p}{\sqrt{S_{n-1}}}\right) = \sum_{n=1}^{T}\frac{1}{\mathbb{E}^{\frac{p}{2}(1-\frac{2}{p-2})}(S_{n-1}^{\frac{1}{p}})}\left(\left(\mathbb{E}\left(\left(\frac{\|\nabla g(\theta_n)\|^2}{S_{n-1}^{\frac{1}{p}}}\right)^{\frac{p}{2}}\right)\right)^{\frac{2}{p}} \cdot (\mathbb{E}(S_{n-1}^{\frac{1}{p-2}}))^{1-\frac{2}{p}}\right)^{\frac{p}{2}} \\
&\geq \sum_{n=1}^{T}\frac{1}{\mathbb{E}^{\frac{p}{2}(1-\frac{2}{p})}(S_{n-1}^{\frac{1}{p-2}})}\left(\mathbb{E}\,\|\nabla g(\theta_n)\|^2\right)^{\frac{p}{2}} \geq p_0\sum_{n=1}^{T}\frac{1}{\sqrt{n}}\left(\mathbb{E}\,\|\nabla g(\theta_n)\|^2\right)^{\frac{p}{2}} \\
&= p_0\frac{1}{\left(\sum_{n=1}^{T} n^{\frac{1}{p-2}}\right)^{\frac{p}{2}(1-\frac{2}{p})}}\left(\left(\sum_{n=1}^{T} n^{\frac{1}{p-2}}\right)^{1-\frac{2}{p}} \cdot \left(\sum_{n=1}^{T}\left(\frac{1}{n^{\frac{1}{p}}}\,\mathbb{E}\,\|\nabla g(\theta_n)\|^2\right)^{\frac{p}{2}}\right)^{\frac{2}{p}}\right)^{\frac{p}{2}} \\
&\geq \frac{p_1}{n^{\frac{p-1}{2}}}\left(\sum_{n=1}^{T}\mathbb{E}\,\|\nabla g(\theta_n)\|^2\right)^{\frac{p}{2}},
\end{aligned} \tag{94}$$

where $p > 0,\ p_1 > 0$ are two constants. Then we acquire

$$\frac{1}{T}\sum_{n=1}^{T}\mathbb{E}\,\|\nabla g(\theta_n)\|^2 = O\left(\frac{\ln^{\frac{2}{p}} T}{T^{\frac{1}{p}}}\right).$$

Therefore, we complete the proof. □

D.3 DISCUSSION ON THE ASSUMPTION OF UNIFORMLY BOUNDED STOCHASTIC GRADIENT

In this appendix, we will demonstrate the statement in the introduction that the bounded stochastic gradient assumption, i.e., $\exists\ K > 0, \forall\ \ n \geq 1$, such that $\|\nabla g(\theta_n, \xi_n)\|^2 < K < +\infty$, greatly simplifies the proof and circumvents some of the key challenges. Back to equation 87, we have

$$\frac{\alpha_0}{20}\sum_{k=2}^{n}\frac{\big\|\nabla g(\theta_k)\big\|^2}{\sqrt{S_{k-1}}} \leq \frac{\alpha_0}{20}\frac{\big\|\nabla g(\theta_2)\big\|^2}{\sqrt{S_1}} + \frac{\alpha_0}{2}(M_0+1)\frac{\big\|\nabla g(\theta_2)\big\|^2}{\sqrt{S_2}} + \frac{g(\theta_2)}{S_2^{\epsilon}}$$
$$+ \frac{\alpha_0\sigma_1(M_0+1)}{2S_2^{\frac{1}{2}+\epsilon}} + 4M_0^2\alpha_0^3\mathcal{L}^2\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_{k-1}),\xi_{k-1}\big\|^2}{S_{k-1}^{\frac{3}{2}+\epsilon}}$$
$$+ \frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k} + 5\alpha_0^3\mathcal{L}^2(M_0+1)^2\sum_{k=2}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\leq\sigma_1}\frac{\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2}{S_{k-1}}$$
$$+ \sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2\leq\sigma_1}\frac{(M_0+1)\alpha_0^3\mathcal{L}^2}{2S_{k-1}^{\frac{3}{2}}}\big\|\nabla g(\theta_{k-1},\xi_{k-1})\big\|^2$$
$$+ \frac{\alpha_0\sigma_1(M_0+1)}{2}\sum_{k=3}^{n}\left(\frac{1}{\sqrt{S_{k-1}}} - \frac{1}{\sqrt{S_k}}\right) + \frac{\mathcal{L}\alpha_0^2}{2}\sum_{k=3}^{n}\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}\frac{\big\|\nabla g(\theta_k,\xi_k)\big\|^2}{S_k}$$
$$+ \sum_{k=3}^{n}\Big(\mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}G_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2>\sigma_1}H_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\leq\sigma_1}J_k + \mathbf{1}_{\|\nabla g(\theta_k)\|^2\leq\sigma_1}K_k\Big).$$

We take the mathematical expectation, getting

$$\sum_{k=2}^{n}\mathbb{E}\left(\frac{\big\|\nabla g(\theta_k)\big\|^2}{\sqrt{S_{k-1}}}\right) \leq O\Bigg(\mathbb{E}\left(\sum_{k=1}^{n}\frac{\|\nabla g(\theta_k.\xi_k)\|^2}{S_k}\right)\Bigg) \leq O(\ln\mathbb{E}\,S_n).$$

Then if we have the condition $\|\nabla g(\theta_k,\xi_k)\|^2 < K$, we can immediately acquire

$$\sum_{k=2}^{n}\mathbb{E}\left(\frac{\big\|\nabla g(\theta_k)\big\|^2}{\sqrt{K}\cdot\sqrt{n}}\right) \leq \sum_{k=2}^{n}\mathbb{E}\left(\frac{\big\|\nabla g(\theta_k)\big\|^2}{\sqrt{S_{k-1}}}\right) \leq O\Bigg(\mathbb{E}\left(\sum_{k=1}^{n}\frac{\|\nabla g(\theta_k.\xi_k)\|^2}{S_k}\right)\Bigg)$$
$$\leq O(\ln\mathbb{E}\,S_n) < O(\ln n).$$

Then we can get the near-optimal rate

$$\frac{1}{T}\sum_{n=1}^{T}\mathbb{E}\,\|\nabla g(\theta_n)\|^2 = O\left(\frac{\ln T}{\sqrt{T}}\right).$$

It can be observed that, without the bounded stochastic gradient assumption, even though we have proven $\|\nabla g(\theta_n)\| \to 0\ a.s.,\ \mathbb{E}\,\|\nabla g(\theta_n)\|^2 \to 0$ in Theorems 3.1 and Theorem 3.2, we still cannot guarantee that the following inequality holds:

$$\sum_{k=2}^{n}\mathbb{E}\left(\frac{\big\|\nabla g(\theta_k)\big\|^2}{\cdot\sqrt{n}}\right) \leq K'\sum_{k=2}^{n}\mathbb{E}\left(\frac{\big\|\nabla g(\theta_k)\big\|^2}{\sqrt{S_{k-1}}}\right),$$

where $K' < +\infty$ is a constant. Therefore, we can only obtain sub-optimal results.