Revisit last-iterate convergence of mSGD under milder requirement on step size

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding convergence of SGD-based optimization algorithms can help deal with enormous machine learning problems. To ensure last-iterate convergence of SGD and momentum-based SGD (mSGD), the existing studies usually constrain the step size ϵ_n to decay as $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$, which however is rather conservative and may lead to slow convergence in the early stage of the iteration. In this paper, we relax this requirement by studying an alternate step size for the mSGD. First, we relax the requirement of the decay on step size to $\sum_{n=1}^{+\infty} \epsilon_n^{2+\eta_0} < +\infty$ $(0 \leq \eta_0 < 1/2)$. This implies that a larger step size, such as $\epsilon_n = \frac{1}{\sqrt{n}}$ can be utilized for accelerating the mSGD in the early stage. Under this new step size and some common conditions, we prove that the gradient norm of mSGD for non-convex loss functions asymptotically decays to zero. In addition, we show that this step size can indeed help make the convergence into a neighborhood of the stationary points quicker in the early stage. In addition, we establish the convergence of mSGD under a constant step size $\epsilon_n \equiv \epsilon > 0$ by removing the common requirement in the literature on the strong convexity of the loss function. Some experiments are given to illustrate the developed results.

1 Introduction

The booming development of machine learning over the past decade relies on the employment of effective optimization algorithms. A large number of such optimization algorithms are based on gradient descent (GD). The optimization problem in machine learning can be cast as minimizing a loss function $g(\theta) \in \mathbb{R}$ over the choice of an N-dimensional real-valued parameter θ, i.e., by solving the problem

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^N} g(\theta). \tag{1}$$

This problem can be solved with a typical GD algorithm through an iteration of the form

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla_{\theta_n} g(\theta_n), \tag{2}$$

where θ_n is the estimate of θ^* at step n, ϵ_n is a positive step size (learning rate) to be designed, and $\nabla_{\theta_n} g(\theta_n)$ stands for the gradient of $g(\theta_n)$ at step n. Under certain technical conditions, θ_n in (2) can asymptotically (as $n \to \infty$) converge to the optimal solution θ^*. However, (2) is not efficient in machine leaning applications especially with enormous data. To accelerate (2), one attempt is stochastic gradient descent (SGD) originating from [1]. Instead of calculating $\nabla_{\theta_n} g(\theta_n)$ over all the data, SGD computes gradient estimate $\nabla_{\theta_n} g(\theta_n, \xi_n)$ by sampling a subset of data (using the distribution of the sampling noise $\{\xi_n\}$ and then updates the iterate as follows

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla_{\theta_n} g(\theta_n, \xi_n). \tag{3}$$

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In addition to the sampling noise, ξ_n can be used to model any external noises arising from the gradient computation.

However the price to pay for this data efficiency is that the employment of stochastic gradients (3) displays a relatively slow convergence rate during the learning process. In order to improve the convergence rate of SGD, momentum-based stochastic gradient descend (mSGD), which reduces the update variance by averaging the past gradients ([2]), has been proposed. An iteration in mSGD is as follows ([3, 4]):

$$v_n = \alpha v_{n-1} + \epsilon_n \nabla_{\theta_n} g(\theta_n, \xi_n), \quad \theta_{n+1} = \theta_n - v_n, \tag{4}$$

where $\alpha \in [0, 1)$ and $\epsilon_n > 0$ are momentum coefficient and step size (learning rate), respectively. An alternative formulation of mSGD, named stochastic heavy ball (SHB) [5, 6], has also been proposed as follows

$$v_n = \beta_n v_{n-1} + (1 - \beta_n) \nabla_{\theta_n} g(\theta_n, \xi_n), \quad \theta_{n+1} = \theta_n - \gamma_n v_n, \tag{5}$$

where $\beta_n \in (0, 1)$ and γ_n are momentum coefficient and step size. It has been shown that mSGD and SHB are essentially equivalent [7]. In recent years, mSGD has been widely employed in many applications of deep learning such as image classification [8], fault diagnosis [9], statistical image reconstruction [10], among others. Moreover, a number of variants on momentum have been emerging, see, e.g., synthesized Nesterov variants [11], robust momentum [12], and PID-control based methods [13]. The importance of momentum in deep learning has been illustrated in [4] through experiments.

There has been a long line of literature analyzing the convergence of SGD and mSGD algorithms. Regarding SGD, authors in [14] studied the last-iterate convergence when the step size is chosen to decay as $1/n$. Authors in [15, 16] established the last-iterate convergence of SGD for strongly convex loss functions without the bounded gradients assumption. For the normalized mSGD (SHB), Polyak [17, 2] and Kaniovski [18] established the convergence (subsequence convergence and convergence of time averages) properties for convex loss functions. Igor Gitman [5] provided convergence results of mSGD (SHB) for non-convex loss functions under the somewhat restrictive requirement for uniform boundedness of a noise term $\mathbb{E}(\|\nabla_{\theta_n} g(\theta_n, \xi_n) - \nabla_{\theta_n} g(\theta_n\|) \leq \delta$. It has been pointed out [19, 4] that the designs of momentum coefficients in [17, 2, 18, 5] may not be consistent with the requirements of some practical applications. The last-iterate convergence of mSGD for non-convex loss functions has recently been established in [20, 7] with the step size condition $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$, which has been traditionally utilized in stochastic optimization [1, 21, 16, 15]. Under the strong convexity condition, since this condition requires the step size to decay relatively fast, it may lead to slow convergence especially in the early stage of the iteration. With the requirement that the loss function is strongly convex, [22] showed that the last-iterate of mSGD can converge to a neighbor of the stationary point when the step size is a constant.

2 Problem of interest and contributions

In this paper, we consider the problem of last-iterate convergence of mSGD under relaxed requirements on step size. Specifically, we seek to relax the condition $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$ [1, 21, 16, 15, 7] that has been required in the literature for proving last-iterate convergence of mSGD. We consider two possible step sizes: step size that decay as $\epsilon_n = \frac{1}{\sqrt{n}}$ and constant step size $\epsilon_n \equiv \epsilon > 0$. Using such larger step size is expected to lead to the last iterate of mSGD (i.e., θ_n) converging faster. We note that since the SGD (3) is a special case of mSGD (4), the developed results in this paper for mSGD also work for SGD.

The contributions of this paper are summarized as follows:

1) Under the setting of step size decaying as $\epsilon_n = \frac{1}{\sqrt{n}}$, we prove that the gradient norm of mSGD for non-convex loss functions asymptotically decays to zero. This result is more general than mean-square gradient convergence established in [20]. In fact, we generalize the result for any step sizes of the form $\sum_{n=1}^{+\infty} \epsilon_n^{2+\eta_0} < +\infty$ $(0 \leq \eta_0 < 1/2)$, which is a superset of the widely required condition $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$ [1, 21, 16, 15, 7]. Given this condition, one can run mSGD by employing a larger step size like $\epsilon_n = \frac{1}{\sqrt{n}}$ for last-iterate convergence.

2) Under some mild conditions, we provide an estimate of the convergence rate of mSGD. Furthermore, we reveal that the new step size can indeed help to improve the convergence speed in the early stage of the iteration.

3) Under the setting of constant step size, first we prove that given any small neighborhood of stationary points, it is feasible to design a constant step size, such that there is a convergent iterate subsequence of mSGD staying within the neighborhood almost surely. In addition, we prove that the mean-square gradient can be arbitrary small by tunning the step size. Comparing with [22], we remove the requirement of strong convexity on loss function.

Regarding the definitions on sequence convergence, the following ones from literature are typical. For a stochastic variables sequence $\{\xi_n\} \in \mathbb{R}^N$ with 0 as the only limit point, the sequence is said to satisfy last-iterate convergence if $\lim_{n\to+\infty} \|\xi_n\| = 0$; and time-average (mean-square convergence) if $\lim_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(\|\xi_k\|^2) = 0$. It can be proved that last-iterate convergence implies time-average convergence under the general setting of the considered problem.

Paper outline. The rest of the paper is organized as follows. In Section 3, we provide the main results of the paper on the convergence of mSGD under decaying step size and constant step size respectively. In Section 4, two experiments are given. Conclusion is made in Section 5. The main proofs are given in Appendix.

3 Main results

In this section, we provide the main results of this paper on the last-iterate convergence for mSGD in (4) under two possible step sizes: step sizes that decay as $\epsilon_n = \frac{1}{\sqrt{n}}$ and constant step size $\epsilon_n \equiv \epsilon > 0$.

3.1 Convergence of mSGD with decaying step size

The following assumptions are needed in this paper.

Assumption 3.1. *Loss function $g(\theta)$ in (1) satisfies the following conditions:*

1. *Noise sequence $\{\xi_n\}$ are mutually independent and independent of θ_1 and v_0, such that $g(\theta) = \mathbb{E}_{\xi_n}\left(g(\theta, \xi_n)\right)$ for any $\theta \in \mathbb{R}^N$.*

2. *$g(\theta)$ is a non-negative and continuously differentiable function. The set of its stationary points $J = \{\theta | \|\nabla_\theta g(\theta)\| = 0\}$ is a bounded set which has only finite connected components $J_1, ..., J_n$. In addition, there is $\epsilon_1 > 0$, such that for any i and $0 < d(\theta, J_i) < \epsilon_1$, it holds that $\left|g(\theta) - g_i\right| \neq 0$, where $g_i = \{g(\theta) | \theta \in J_i\}$ is a constant.*

3. *There are two constants $M' > 0$ and $a' > 0$ such that for any $\theta \in \mathbb{R}^N$ and $n \in \mathbb{N}_+$,*
$$E_{\xi_n}\left(\left\|\nabla_\theta g(\theta, \xi_n)\right\|^2\right) \leq M'\left\|\nabla_\theta g(\theta)\right\|^2 + a'. \tag{6}$$

4. *The gradient $\nabla_\theta g(\theta)$ satisfies the Lipschitz condition, i.e., there is a constant $c > 0$, such that for any $x, y \in \mathbb{R}^N$,*
$$\left\|\nabla_x g(x) - \nabla_y g(y)\right\| \leq c\|x - y\|.$$

Conditions 1 and 4 in Assumption 3.1 are common in the literature [1, 21, 16, 15]. [7] required a condition on strong growth, i.e., $\mathbb{E}_{\xi_n} \|\nabla_{\xi_n} g(\theta, \xi_n)\|^2 \leq M_0 \|\nabla_\theta g(\theta)\|^2$, which however makes it close to the deterministic case. In contrast, condition 3 allows more randomness of data sampling. Condition 2 is a mild condition for the reasons as follows. In some works, the non-negative condition may be replaced by a lower bound condition $g(\theta) > \hat{l}_0 > -\infty$. These two conditions are essentially equivalent, since one can construct a new loss function $\overline{g} = g - \hat{l}_0$ under the lower bound condition, such that the new loss function is non-negative. The rest of Condition 2 is quite general, since it allows the loss function to have multiple stationary points and to be non-convex.

Assumption 3.2. *In the mSGD (4), momentum coefficient $\alpha \in [0, 1)$ and the sequence of step size ϵ_n is positive and monotonically decreasing to zero, such that $\sum_{n=1}^{+\infty} \epsilon_n = +\infty$ and $\sum_{n=1}^{+\infty} \epsilon_n^{2+\eta_0} < +\infty$, where $0 \leq \eta_0 < 1/2$ is a constant.*

In Assumption 3.2, the momentum coefficient $\alpha \in [0, 1)$ is a constant. Compared with the setting in [6, 17, 18, 5], where $\alpha^{(n)}$ tends to 1 or 0, constant momentum coefficient is more general in practice [7, 19, 4]. Regarding the step size condition, it is more general than the one in many existing works on last-iterate convergence of SGD and mSGD [1, 21, 16, 15, 7, 13], i.e., $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$, which is obtained from Assumption 3.2 if $\eta_0 = 0$. Under the step size condition in Assumption 3.2, one can choose larger step size like $\epsilon_n = \frac{1}{\sqrt{n}}$, which however is not feasible in the commonly required condition $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$. Since the new step size can decay more slowly than the existing one, it provides more space for step size fine tinning, such that the algorithm can quickly converge to a neighborhood of the stationary point [23, 24, 22].

Then we attain the first main result in this paper on last-iterate convergence of mSGD with the new step size satisfying Assumption 3.2.

Theorem 3.1. *Consider the mSGD in* (4) *with any* $v_0 \in \mathbb{R}^N$ *and* $\theta_1 \in \mathbb{R}^N$. *Under Assumptions 3.1 and 3.2, the gradient norm tends to* 0 *almost surely, i.e.*

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

This theorem shows that the gradient norm of the last iteration θ_n is convergent to 0 almost surely.

Due to page constraint, the complete proof is given Appendix. Here, we provide a proof outline to present the main proof ideas.

Proof Sketch of Theorem 3.1: We aim to prove that $\nabla_{\theta_n} g(\theta_n) \to 0 \ a.s.$ via the following key steps.

Step 1: In this step we aim to prove $\mathbb{E}\left(\epsilon_n^{\eta_0} g(\theta_n)\right)$ is bounded. We present this result as Lemma B.8 in Appendix.

Step 2: We prove there exists a subsequence of $\nabla_{\theta_n} g(\theta_n)$ which is convergent to 0 with probability 1. It is attained by proving $\sum_{n=1}^{+\infty} \epsilon_n^{1+2\eta_0} \|\nabla_{\theta_n} g(\theta_n)\|^2 < +\infty \ a.s.$. We present this result as Lemma B.10 in Appendix.

Step 3: We aim to extend the subsequence convergence in Step 2 to the asymptotic convergence. The basic idea is to prove that the adjacent terms $g(\theta_n)$ and $g(\theta_{n+1})$ are "close" enough, such that

$$g(\theta_{n+1}) - g(\theta_n) \leq \hat{k} \epsilon_t^{1+2\eta_0} + Q_n,$$

where $\hat{k} > 0$ is a constant and $\{Q_n\}$ is a sequence, such that $\sum_{n=1}^{+\infty} Q_n$ convergence a.s.. This result is presented in Lemma B.11.

Step 4: Finally, we use some techniques to attain the result of asymptotic convergence $\nabla_{\theta_n} g(\theta_n) \to 0 \ a.s.$.

Technically, there are several big challenges when we try to extend the condition $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$ to $\sum_{n=1}^{+\infty} \epsilon_n^{2+\delta_0} < +\infty$ with $0 \leq \eta_0 < 1/2$. One of the challenges is: Usually when we try to prove $\sum_{n=1}^{+\infty} \zeta_n = \sum_{n=1}^{+\infty} \epsilon_n \nabla_{\theta_n} g(\theta_n)^T (\nabla_{\theta_n} g(\theta_n, \xi_n) - \nabla_{\theta_n} g(\theta_n))$ is convergent a.s., which is the vital step to get the last-iterate convergence, we use the $Martingale\ convergence\ theorem$, through which we just need to prove $\sum_{n=1}^{+\infty} \mathbb{E}(\zeta_n^2)$ is convergent. If the condition $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$ holds, we can use the following inequality to get $\sum_{n=1}^{+\infty} \mathbb{E}(\zeta_n^2) < \sum_{n=1}^{+\infty} \hat{K} \epsilon_n^2 < +\infty$ ($\hat{K}$ is a constant) directly. However, under the new condition $\sum_{n=1}^{+\infty} \epsilon_n^{2+\delta_0}$, we can not use this approach to prove it. Instead, we introduce a more technical approach to handle this problem.

Comparing with the corresponding result in [7] (Theorem 1 in [7]), we relax the condition on the decay of step size. With this new condition, one can utilize step size with slow decay in the iteration of mSGD. Let $\alpha = 0$ in mSGD, then we can get a similar result for SGD

Corollary 3.1. *Consider the SGD in* (3) *with any* $\theta_1 \in \mathbb{R}^N$. *Under Assumptions 3.1 and 3.2 with* $\alpha = 0$, *there is the gradient norm trend to* 0 *almost surely, i.e.*

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

In the following, we aim to show the last-iterate convergence rate under the new step size condition. Since $\|\nabla_{\theta_n} g(\theta_n)\|^2$ is a stochastic variable, it is not favorable to estimate the almost-sure convergence

rate of $\nabla_{\theta_n} g(\theta_n)$. Instead, the convergence rate of $\mathbb{E}\left(\|\nabla_\theta g(\theta_n)\|^2\right)$ is quite meaningful. In general, if we need to estimate the convergence rate of the last iterate, we usually need some extra assumptions to establish a quantitative relationship between g and ∇g. Existing works were established usually under the strong-convex assumption [25, 26, 13], local P-L condition [7] or convex loss function with some stability conditions, i.e., $\mathbb{E}\,\|\nabla_{\theta_n} g(\theta_n)\|^2 < \delta < +\infty$. These assumptions are all relatively strong to some extent, especially the stability condition which is very difficult to verify. In our paper, we need a milder assumption than the above conditions as follow.

Assumption 3.3. *Regarding the loss function $g(\theta)$ defined in (1), there exists $\delta_0 > 0$, $\hat{T} > 0$, such that for any $\theta \in \{\theta | \|\nabla_\theta g(\theta)\|^2/(g(\theta) - g^*) < \delta_0\}$, it holds that $g(\theta) < \hat{T}$, where $g^* = \inf_{\theta\in\mathbb{R}^N} g(\theta)$.*

This assumption requires the value of the loss function $g(\theta)$ to be bounded in the set $S = \{\theta | \|\nabla_\theta g(\theta)\|^2/(g(\theta) - g^*) < \delta_0\}$. The motivation of this assumption is: For a stochastic algorithm, there is always a probability at which the iterates are far away from the true value. Since second-derivative information will have a more positive effect on convergence (such as strongly convex loss functions), this condition restricts iterates which are far away due to extreme noise, so that it does not have a significant effect on global updates. This requirement is milder than the strong convexity condition, which is commonly used in existing works [25, 26, 13] for the analysis of the last-iterate convergence. In addition, this assumption allows the loss function to be non-convex, and does not need the local P-L condition used in [7] on the second derivative information near the stationary point. Note that Assumption 3.3 can be satisfied by many loss functions commonly encountered in machine learning. First of all, many loss functions usually used are bounded [27–30], and thus satisfy Assumption 3.3. For unbounded loss functions, there are a number of instances satisfying Assumption 3.3. For linear regression functions, since they are strongly convex, it is easy to verify the assumption. In the Appendix, we explain (not in a strict but an illustrating manner) this assumption can also be satisfied by a two layer neural network with Relu active function under a square loss function.

In the following theorem, we provide an estimate of the convergence rate for the mSGD in (4).

Theorem 3.2. *Consider the mSGD in (4) with any $v_0 \in \mathbb{R}^N$ under Assumptions 3.1–3.3. Then exists $a > 0$, for any $\|\nabla_{\theta_1} g(\theta_1)\|^2$, there is*

$$\mathbb{E}\left(\|\nabla_\theta g(\theta_n)\|^2\right) = O\Bigg(\left(e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{i=1}^n \epsilon_i}\right)\left(\sum_{i=1}^n \epsilon_i^2 e^{\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=1}^i \epsilon_k}\right)\Bigg), \tag{7}$$

where $\hat{c} = \min\{a/2T_0, 2\delta_0\}$.

Comparing with the result of the last-iterate convergence rate in [7], we remove the strong growth condition, i.e., $\mathbb{E}_{\xi_n}\|\nabla_\theta g(\theta, \xi_n)\|^2 \le M\|\nabla_\theta g(\theta)\|^2$, and relax the step size condition to be $\sum_{n=1}^{+\infty} \epsilon_n^{2+\delta_0}$ $(0 < \delta_0 \le 1/2)$. In the comparison with [20], we do not require the convexity of loss function or $\mathbb{E}\,\|\nabla_{\theta_n} g(\theta_n)\|^2 < G$. In addition, we relax the step size condition in [20]. There are quite a few results on the average iterates in the literature, such as [31, 32, 13]. In [31, 32, 13], the convergence rate $O(\frac{1}{\sqrt{n}})$ of mSGD is established with the step size $\frac{1}{\sqrt{n}}$. According to our result, let $\epsilon_n = \frac{1}{\sqrt{n}}$, if holds that

$$\mathbb{E}(\|\nabla_{\theta_n} g(\theta_n)\|^2) = O\Big(e^{-\sqrt{n}}\sum_{k=1}^n \frac{e^{\sqrt{k}}}{k}\Big) = O\Big(\frac{1}{\sqrt{n}}\Big).$$

It can be proved that the convergence rate of average iterates is also $O\Big(\frac{1}{\sqrt{n}}\Big)$.

From (7), we see that larger step size may not lead to larger convergence rate. This is reasonable, since the convergence rate reflects how quickly the iterate θ_n converges when $n \to +\infty$ and a larger step size usually makes the iterate converges to a neighborhood of stationary point quicker. We have the following theorem on this point.

Theorem 3.3. *Consider the mSGD in (4) with any $v_0 \in \mathbb{R}^N$. Under Assumptions 3.1–3.3, given any $a > 0$, if $\|\nabla_{\theta_1} g(\theta_1)\|^2 > a$, then it holds that*

$$P(\tau^{(a)} \ge n) = O\Big(e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{i=1}^n \epsilon_i}\Big),$$

where $\tau^{(a)} = \min_{n>0}\{\|\nabla_{\theta_n} g(\theta_n)\|^2 < a\}$ and $\hat{c} = \min\{a/2T_0, 2\delta_0\}$.

Proposition 3.1. *Consider the mSGD in* (4) *with any* $v_0 \in \mathbb{R}^N$*. Under Assumptions 3.1–3.3, given any* $a > 0$*, if* $\|\nabla_{\theta_1} g(\theta_1)\|^2 > a$*, then there exists* $\hat{k}_0 > 0$*, when* n *is sufficiently large, it holds that*

$$P(\tau^{(a)} \geq n) \geq \hat{k}_0 e^{-\frac{4c}{(1-\alpha)^2}\sum_{i=1}^{n}\epsilon_i},$$

where $\tau^{(a)} = \min_{n>0}\{\|\nabla_{\theta_n} g(\theta_n)\|^2 < a\}$.

Theorem 3.3 and Proposition 3.1 provides a probability description of how fast the iterates of mSGD converge into the preset neighborhood of stationary points. It can be illustrated as follows. It follows from Theorem 3.3 that

$$P(\tau^{(a)} < n) > 1 - \hat{k}_1 e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{i=1}^{n}\epsilon_i},$$

where $\hat{k}_1 > 0$ is a constant. From the inequality, as n increases, the event $\|\nabla_{\theta_n} g(\theta_n)\|^2 < a$ occurs with higher and higher probability tending to probability 1. The influence of momentum coefficient α and step size ϵ_n is also quantified. We note that with the measure in Theorem 3.3, the convergence to a neighbor of the stationary point is faster than under the traditional condition $\sum_{n=1}^{+\infty}\epsilon_n^2 < +\infty$ by setting the step size with a relatively slow decay speed as $\sum_{n=1}^{+\infty}\epsilon_n^{2+\delta_0}$ $(0 < \delta_0 \leq 1/2)$, such as $\epsilon_n = \frac{1}{\sqrt{n}}$ and $\eta_0 = 1/4$. Similar results (the large step size can make the iteration quickly converge to a neighbor of stationary point, but may converge slowly in the late stage of the algorithm) appear in [23, 24, 22] through experiments.

3.2 Convergence of mSGD with constant step size

In the previous subsection, we establish the last-iterate convergence of mSGD under a relaxed decay condition on step size. In this subsection, we study the convergence of mSGD under a constant step size $\epsilon_n \equiv \epsilon > 0$. We start with the following result on subsequence convergence.

Theorem 3.4. *Consider the mSGD in* (4) *with any* $v_0 \in \mathbb{R}^N$ *and* $\theta_1 \in \mathbb{R}^N$*. Under Assumption 3.1, for any* $\varphi > 0$*, there exists* $\mu_0^{(\varphi)} > 0$*, such that for any* $\epsilon_n \equiv \epsilon < \mu_0^{(\varphi)}$*, it holds that*

$$\left\|\nabla_{\theta_{k_n}^{(\varphi)}} g(\theta_{k_n}^{(\varphi)})\right\|^2 \leq \varphi, \text{ a.s.}$$

where $\{\theta_{k_n}^{(\varphi)}\}$ *is a subsequence* $\{\theta_n\}$.

Proof. Due to page constraint, the complete proof is given Appendix. Here, we provide a proof outline to present the main proof ideas.

To study the mSGD with constant step size, we notice the fact that the impact of the noise cannot be totally eliminated through the step size. But when $\|\nabla_{\theta_n} g(\theta_n)\|$ is relatively large (meaning the distance between θ_n and the stationary point θ^* is relatively large), it holds that $\mathbb{E}_{\xi_n}\left(\|\nabla_{\theta_n} g(\theta_n, \xi_n)\|^2\right) \leq M'\|\nabla_{\theta_n} g(\theta_n)\|^2 + a' \leq (M' + \hat{k})\|\nabla_{\theta_n} g(\theta_n)\|^2$. This implies that with a proper step size ϵ, the noise can be controlled. As a result, the algorithm can be stabilized around a stationary point. In the following, several key steps in the proof are provided.

Step 1: We define an event by $A_n^{(\varphi)} = \left\{\|\nabla_{\theta_1} g(\theta_1)\|^2 > \varphi,\ \|\nabla_{\theta_2} g(\theta_2)\|^2 > \varphi, \ldots,\ \|\nabla_{\theta_n} g(\theta_n)\|^2 > \varphi\right\}$ and a characteristic function of this event by $I_n^{(\varphi)}$.

Step 2: We prove that the probability of $A_n^{(\varphi)}$ is tending to 0 as $n \to \infty$. First, we attain a recursion formula:

$$\begin{aligned}
&\mathbb{E}\left(I_{t+1}^{(\varphi)} g(\theta_{t+1})\right) - \mathbb{E}\left(I_t^{(\varphi)} g(\theta_t)\right) \\
\leq &-\alpha^{t-1} I_1^{(\varphi)}\, \mathbb{E}\left(\nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1\right) - \sum_{i=2}^{t} \alpha^{t-i}(\epsilon - \hat{k}M'\epsilon^2)\, \mathbb{E}\left(I_{i-1}^{(\varphi)}\|\nabla_{\theta_i} g(\theta_i)\|^2\right).
\end{aligned}$$

By taking the sum of multiple inequalities as above, we have

$$\sum_{i=2}^{t} \mathbb{E}\left(I_{i-1}^{(\varphi)}\|\nabla_{\theta_i} g(\theta_i)\|^2\right) < +\infty.$$

This leads to

$$\mathbb{E}\left(I_{t-1}^{(\varphi)}\|\nabla_{\theta_t} g(\theta_t)\|^2\right) \to 0.$$

From the $Chebyshev\ inequality$, we attain $P(A_n^{(\varphi)}) \to 0$.

Step 3: Step 2 implies that with probability 1 that there exists at least one instant n at which $\|\nabla_{\theta_n} g(\theta_n)\|^2 \leq \varphi$. We note that for the time instant m at which $\|\nabla_{\theta_m} g(\theta_m)\|^2 > \varphi$, then we can use a new start point. By repeating the process of Step 2, we attain the conclusion of this theorem.

□

Theorem 3.4 shows that under the common conditions in the literature, the constant step size can guarantee the subsequence convergence of mSGD (4) in any small neighborhood of stationary point. In addition, we have a similar result on $\mathbb{E}\left(\|\nabla_{\theta_n} g(\theta_n)\|^2\right)$ as follows.

Proposition 3.2. *Suppose $\{\theta_n\}$ is a sequence generated by (4) with any $\theta_1 \in \mathbb{R}^N$ and $v_0 \in \mathbb{R}^N$. Under Assumptions 3.1 and 3.3, if $g(\theta)$ has only one optimal point $g^* = g(\theta^*)$ and $\inf_{d(\theta,\theta^*)>\delta_0} \|\nabla_\theta g(\theta)\|^2 > 0 \quad (\forall \delta > 0)$, then for any $\varphi_0 > 0$, there exist scalars $\mu_0^{(\varphi_0)} > 0$ and $n_0 > 0$, such that for any $\epsilon_n \equiv \epsilon < \mu_0^{(\varphi_0)}$ and $n > n_0$, it holds that*

$$\mathbb{E}\left(\|\nabla_{\theta_n} g(\theta_n)\|^2\right) \leq \varphi_0.$$

This result is a substantial extension of the existing results on SGD with constant step size to mSGD. Comparing with the results in [22], we remove the requirement of μ-strong convexity on loss function.

4 Experiments

Since this paper focuses on convergence analysis on the well-known mSGD, two experiments on regression and classification tasks are given respectively to show the effectiveness of the results.

4.1 Regression task

In this subsection, we study a house price prediction problem with neural networks trained by mSGD.

Network Architectures. We respectively employ a 3-layer and a 4-layer fully-connected neural network with ReLu active function and squared loss function.

Implementation. We implement two neural networks to train on this dataset using Keras. The first neural network consists of three fully connected layers with 13, 8, and 1 neurons. The second neural network consists of four fully connected layers with 13, 32, 16, 1 neurons respectively. We initialize the weights use glorot uniform algorithm. We use mSGD with a mini-batch size of 16, and use mean square error loss function to train the model. The momentum term coefficient is 0.9 and the models are trained for up to 1000 epochs, training takes about two hours at a time in 3080GPU. We do not use dropout.

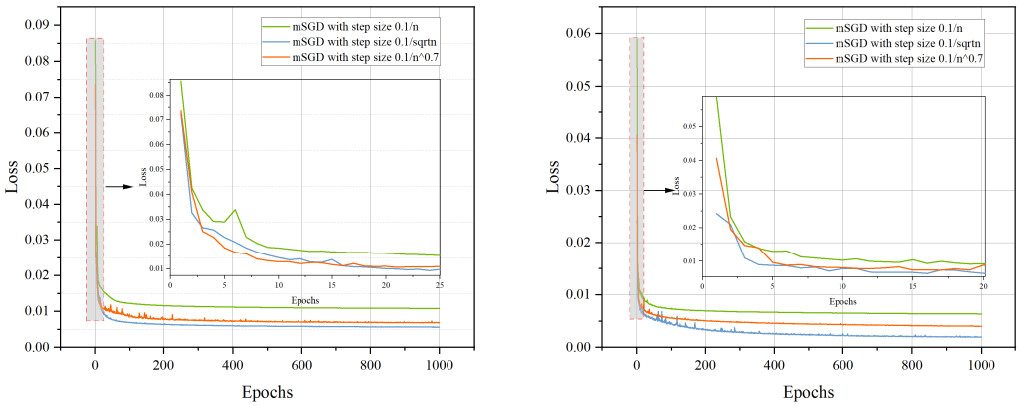
Dataset. The training dataset we use is from Boston House Price Dataset. This dataset belongs to the regression prediction dataset, which consists of 506 sets of data with dimension 13. During the training process, we randomly divide 30% as the test set and normalize the data before training.

Results. By using the mSGD in (4) with three different step sizes respectively, i.e., $\epsilon_n = \frac{1}{n}$, $\epsilon_n = \frac{1}{n^{0.7}}$, and $\epsilon_n = \frac{1}{n^{0.5}}$. The experiment results are given in Figure 1. The figure shows that 1) the loss decays to zero under the three step size setting; 2) $\epsilon_n = \frac{1}{n^{0.5}}$ can make the loss tend to a small neighborhood of zero fastest among the three step size settings. This conforms to the theoretical analysis in Theorems 3.1–3.3.

4.2 Classification task

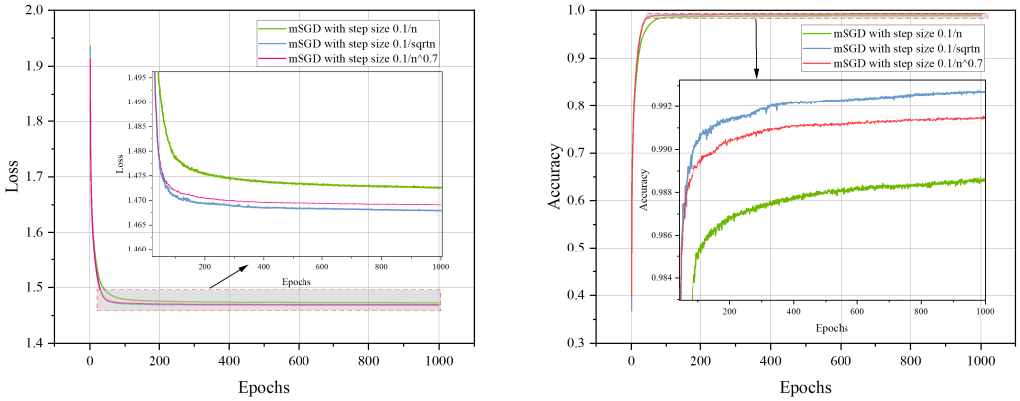
In this subsection, we study a image classification problem with neural networks trained by mSGD.

Network Architectures. We employ a 20-layer ResNet network. The convolutional layers mostly have 3×3 filters and follow two simple design rules: (i) for the same output feature map size, the layers have the same number of filters; (ii) if the feature map size is halved, the number of filters



(a) Three-layer neural network (b) Four-layer neural network

Figure 1: Training loss-iteration chart on the Boston house-price data.



(a) The training loss (b) Accuracy of ResNet20 on CIFAR-10

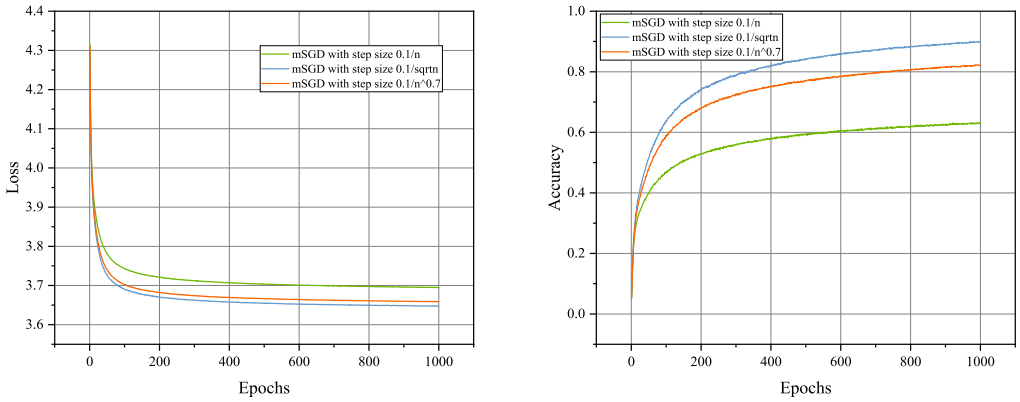
Figure 2: Training and prediction performance on CIFAR-10.

is doubled so as to preserve the time complexity per layer. We perform downsampling directly by convolutional layers that have a stride of 2 and padding with valid. The network ends with an average pooling layer and a 10-way fully-connected layer with softmax.

Implementation. We implement the ResNet20 network using Keras. We initialize the weights use glorot uniform algorithm. We use mSGD with a mini-batch size of 64, momentum coefficient 0.9 and use categorical crossentropy loss function to train the model. The models are trained for up to 1000 epochs, taking about two hours at a time in 3080GPU. We do not use dropout.

Dataset. We use two different datasets CIFAR-10 and CIFAR-100. CIFAR-10 consists of 50k training images and 10k testing images in 10 classes. CIFAR-10 is a color image dataset closer to ubiquitous objects, and the image size is 32x32x3. CIFAR-100 consists of 50k training images and 10k testing images in 100 classes. Each category contains 500 training images and 100 testing images.CIFAR-100 is a color image dataset closer to ubiquitous objects,and the image size is 32x32x3. Normalize the dataset between 0-1 before training.

Results. By using the mSGD in (4) with three different step sizes respectively, i.e., $\epsilon_n = \frac{1}{n}$, $\epsilon_n = \frac{1}{n^{0.7}}$, and $\epsilon_n = \frac{1}{n^{0.5}}$. The experiment results are given in Figure 2 and Figure 3. The figures show that 1) the loss decays to zero and the accuracy goes up to 1 under the three step size setting; 2) $\epsilon_n = \frac{1}{n^{0.5}}$ can make the loss tend to a small neighborhood of zero and make the accuracy tend to a small neighborhood of 1 fastest among the three step size settings, followed by the setting of $\epsilon_n = \frac{1}{n^{0.7}}$. These results conform to the theoretical analysis in Theorems 3.1–3.3.



(a) The training loss

(b) Accuracy of ResNet20 on CIFAR-100

Figure 3: Training and prediction performance on CIFAR-100.

5 Conclusion

In this paper, we studied the last-iterate convergence of momentum-based stochastic gradient descent (mSGD) under relaxed conditions on step sizes. First, under the relaxed condition $\sum_{n=1}^{+\infty} \epsilon_n^{2+\eta_0} < +\infty$ $(0 \leq \eta_0 < 1/2)$, we proved the last-iterate convergence of mSGD for non-convex loss functions. In addition, we showed that this step size can indeed help to improve the convergence speed in the early stage of the algorithm by quantifying the influence of the step size and momentum coefficient. This implies that a larger step size, such as $\epsilon_n = \frac{1}{\sqrt{n}}$ can be utilized for accelerating the mSGD. We also proved that the algorithm with a constant step size (i.e., $\epsilon_n \equiv \epsilon > 0$) can ensure the last-iterate convergence of mSGD without requiring the strong convexity assumption on loss function.

References

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[2] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics & Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[3] S. Zareba, A. Gonczarek, J. M. Tomczak, and J. Swiatek, "Accelerated learning for restricted Boltzmann machine with momentum term," in *Progress in Systems Engineering*, pp. 187–192, 2015.

[4] I. Sutskever, J. Martens, G. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, pp. 1139–1147, 2013.

[5] I. Gitman, H. Lang, P. Zhang, and L. Xiao, "Understanding the role of momentum in stochastic gradient methods," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] A. M. Gupal and L. G. Bazhenov, "Stochastic analog of the conjugate-gradient method," *Cybernetics*, vol. 8, no. 1, pp. 138–140, 1972.

[7] R. Jin, Y. Xing, and X. He, "On the convergence of mSGD and AdaGrad for stochastic optimization," in *International Conference on Learning Representations*, 2022.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[9] S. Tang, C. Shen, D. Wang, S. Li, W. Huang, and Z. Zhu, "Adaptive deep feature learning network with Nesterov momentum and its application to rotating machinery fault diagnosis," *Neurocomputing*, vol. 305, pp. 1–14, 2018.

[10] D. Kim, S. Ramani, and J. A. Fessler, "Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 167–178, 2014.

[11] M. O. Williams and T. E. Lovett, "Numerical diagnostics for systems of differential algebraic equations," *arXiv preprint arXiv:1602.07550*, 2016.

[12] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions," in *American Control Conference*, pp. 1376–1381, 2018.

[13] W. An, H. Wang, Q. Sun, J. Xu, Q. Dai, and L. Zhang, "A PID controller approach for stochastic optimization of deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8522–8531, 2018.

[14] O. Shamir, "Making gradient descent optimal for strongly convex stochastic optimization," *Omnipress*, 2011.

[15] L. M. Nguyen, P. H. Nguyen, P. Richtárik, K. Scheinberg, M. Taká, and M. V. Dijk, "New convergence aspects of stochastic gradient algorithms," 2018.

[16] L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takác, "SGD and Hogwild! convergence without the bounded gradients assumption," in *International Conference on Machine Learning*, pp. 3750–3758, 2018.

[17] B. T. Polyak, "Comparison of convergence rate of one-step and multistep optimization algorithms in the presence of noise," *Izv. Akad. Nauk SSSR, Tekh. Kibern.*, vol. 1, pp. 9–12, 1977.

[18] Y. M. Kaniovskii, "Behaviour in the limit of iterations of the stochastic two-step method," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 23, no. 1, pp. 13–20, 1983.

[19] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," 2018.

[20] X. Li, M. Liu, and F. Orabona, "On the last iterate convergence of momentum methods," 2021.

[21] H.-F. Chen, *Stochastic Approximation and Its Applications*. Springer Science & Business Media, 2006.

[22] Y. Liu, Y. Gao, and W. Yin, "An improved analysis of stochastic gradient descent with momentum," *arXiv preprint arXiv:2007.07989*, 2020.

[23] N. Iyer, V. Thejas, N. Kwatra, R. Ramjee, and M. Sivathanu, "Lrtuner: A learning rate tuner for deep neural networks," 2021.

[24] K. N. S. S. Sastry, "Step size matters in deep learning," 2018.

[25] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on Machine learning*, p. 116, 2004.

[26] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," *JMLR.org*, 2012.

[27] L. Yang, Z. Hui, and S. Yuan, "Bayes estimation of parameter of exponential distribution under a bounded loss function," *Research Journal of Mathematics Statistics*, vol. 5, no. 4, pp. 28–31, 2013.

[28] R. Leon and C. Wu, "A theory of performance measures in parameter design," *Statistica Sinica*, pp. 335–358, 1992.

[29] Berger and O. James, "Statistical decision theory and bayesian analysis," , vol. 10.1007/978-1-4757-4286-2, 2004.

[30] M. Towhidi and J. Behboodian, "Minimax estimation of a bounded parameter under some bounded loss functions," *Far East Journal of Theoretical Statistics*, vol. 6, no. 1, pp. 39–48, 2002.

[31] V. Mai and M. Johansson, "Convergence of a stochastic gradient method with momentum for nonsmooth nonconvex optimization," in *International Conference on Machine Learning*, pp. 6630–6639, 2020.

[32] T. Yang, Q. Lin, and Z. Li, "Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization," 2016.

[33] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Introductory Lectures on Convex Optimization: A Basic Course, 2004.

[34] Z.-z. Wang, Y. Dong, and F. Ding, "On almost sure convergence for sums of stochastic sequence," *Communications in Statistics-Theory and Methods*, vol. 48, no. 14, pp. 3609–3621, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A An illustration example for Assumption 3.3

In the following, we show that Assumption 3.3 can be met in simple neural networks by using a two layer neural network with $Relu$ active function under a square loss function. First we have that

$$g(\theta, W) = \frac{1}{N}\sum_{i=1}^{N}\big|y_i - \theta^T\sigma(Wx_i)\big|^2,$$

we can calculate its partial derivatives with respect to θ and W

$$\frac{\partial g(\theta, W)}{\partial \theta} = \frac{2}{N}\sum_{i=1}^{N}\big(y_i - \theta^T\sigma(Wx_i)\big)\sigma(Wx_i),$$

$$\frac{\partial g(\theta, W)}{\partial W} = \frac{2}{N}\sum_{i=1}^{N}\big(y_i - \theta^T\sigma(Wx_i)\big)(\theta \odot \mathbf{1}_{Wx_i\geq 0})x_i^T.$$

For illustration, we just consider the situation which only has one date, i.e., $N = 1$. We derive that

$$\left\|\frac{\partial g(\theta, W)}{\partial(\theta, W)}\right\|^2 = \left\|\frac{\partial g(\theta, W)}{\partial \theta}\right\|^2 + \left\|\frac{\partial g(\theta, W)}{\partial W}\right\|^2 = 4\big|y - \theta^T\sigma(Wx)\big|^2\|\sigma(Wx)\|^2$$
$$+ 4\big|y - \theta^T\sigma(Wx)\big|^2\|\theta \odot \mathbf{1}_{Wx_i\geq 0}\|^2\|x\|^2 = 4\big(\|\sigma(Wx)\|^2 + \|\theta \odot \mathbf{1}_{Wx_i\geq 0}\|^2\|x\|^2\big)\big|y - \theta^T\sigma(Wx)\big|^2.$$

As a result, $\|\nabla_\theta g(\theta)\|^2/g(\theta) = 4\big(\|\sigma(Wx)\|^2 + \|\theta \odot \mathbf{1}_{Wx_i\geq 0}\|^2\|x\|^2\big)$. When $4\big(\|\sigma(Wx)\|^2 + \|\theta \odot \mathbf{1}_{Wx_i\geq 0}\|^2\|x\|^2\big) < \delta_0$, we obtain $\|\theta^T\sigma(Wx)\|^2 \leq 1/2\big(\|\sigma(Wx)\|^2 + \|\theta \odot \mathbf{1}_{Wx_i\geq 0}\|^2\|x\|^2\big) \leq 2\delta_0$. This ensures that $g(\theta, W)$ bounded.

B Useful lemmas

To prove the main results, we need the following lemmas, where Lemmas B.1–B.5 are from the literature and Lemmas B.6–B.11 are first introduced in this paper with proofs given in the next section.

Lemma B.1. *(Lemma 1.2.3 in [33]) Suppose $f(\theta) \in C^1$ $(x \in \mathbb{R}^N)$ with gradient satisfying the following Lipschitz condition $\|\nabla f(x) - \nabla f(y)\| \leq c\|x - y\|$, where $c > 0$, then for any $x, y \in \mathbb{R}^N$, it holds that*

$$f(y) + \nabla f(y)^T(x - y) - \frac{c}{2}\|x - y\|^2 \leq f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{c}{2}\|x - y\|^2.$$

Lemma B.2. *(Lemma 10, [7]) Suppose $f(x) \in C^1$ $(x \in \mathbb{R}^N)$ with $f(x) > -\infty$ and its gradient satisfying the following Lipschitz condition $\|\nabla f(x) - \nabla f(y)\| \leq c\|x - y\|$, where $c > 0$, then $\forall\, x_0 \in \mathbb{R}^N$, there is*

$$\big\|\nabla f(x_0)\big\|^2 \leq 2c\big(f(x_0) - f^*\big),$$

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

Lemma B.3. *(Lemma 4, [7]) Suppose $f(x) \in C^1$ $(x \in \mathbb{R}^N)$ with gradient satisfying the following Lipschitz condition $\|\nabla f(x) - \nabla f(y)\| \leq c\|x - y\|$, where $c > 0$, and the set $S = \{x|\nabla f(x) = 0\}$ is bounded and only has finite connected components $\{S_1, S_2, \ldots, S_m\}$. Furthermore, assume there exists $\epsilon_1' > 0$, such that for any $i = 1, 2, ..., m$ and $x \in \{x|0 < d(x, S_i) < \epsilon_1'\}$, it holds that $\big|f(x) - f_i\big| \neq 0$, where $f_i = f(x)$ for $x \in S_i$. Then for any $i = 1, 2, ..., m$, if there is $\epsilon_0' > 0$ satisfying $d(x, S_i) < \epsilon_0'$, it follows that*

$$\big\|\nabla f(x)\big\|^2 \leq 2c\big|f(x) - f_i\big|.$$

Lemma B.4. *[34] Suppose that $\{x_n\} \in \mathbb{R}^N$ 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*

$$\sum_{n=1}^{\infty}\mathbb{E}(\|x_n\|^2) < +\infty, \quad \text{or} \quad \sum_{n=1}^{\infty}\mathbb{E}\left(\|x_n\|^2\big|\mathcal{F}_{n-1}\right) < +\infty. \quad a.s.$$

Lemma B.5. *(Lemma 6, [7]) Suppose that $\{x_n\} \in \mathbb{R}^N$ 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 B.6. *If $0 < \mu < 1$ and $0 < \sigma < 1$ $(\sigma \neq \mu)$ are two constants, then for any positive sequence $\{\psi_n\}$, there is*

$$\sum_{k=1}^{n}\mu^{n-k}\sum_{i=1}^{k}\sigma^{k-i}\psi_i = \frac{1 - \kappa^{(\omega_0-1)n}}{1 - \kappa^{(\omega_0-1)}}\sum_{i=1}^{n}\kappa^{n-i}\psi_i,$$

where $\kappa = \max\{\mu, \sigma\}$ and $\omega_0 = \log_\kappa \min\{\mu, \sigma\}$.

Lemma B.7. *Let $0 < \tau < 1$ be a constant, then for any sequence $\{\epsilon_n\}$ satisfying $\epsilon_n \to 0$, it holds that $\sum_{n=1}^{+\infty} c_n^{(\epsilon)} = 0$, where $c_n^{(\epsilon)} = \sum_{i=1}^{n} \tau^{n-i}\epsilon_i - (1/(1-\tau))\epsilon_n$.*

Lemma B.8. *Consider the mSGD in (4). If Assumptions 1–3 hold, then for any $\theta_1 \in \mathbb{R}^N$, $v_0 \in \mathbb{R}^N$ and $\eta_0 \geq \eta_0$, there is a scalar $T(\theta_1, v_0)$, such that $\mathbb{E}\left(\epsilon_n^{\eta_0} g(\theta_n)\right) < T(\theta_1, v_0)$ for any $n \geq 1$.*

Lemma B.9. *Suppose $\{v_n\}$ is a sequence generated by mSGD in (4). Under Assumptions 3.1–3.2, it holds that $\sum_{n=0}^{+\infty} \mathbb{E}\left(\epsilon_n^{2\eta_0}\|v_n\|^2\right) < c(v_0, \theta_1)$, and $\sum_{n=0}^{+\infty} \epsilon_n^{2\eta_0}\|v_n\|^2 < +\infty$ a.s., where $c(v_0, \theta_1)$ is a constant only related to v_0 and θ_1.*

Lemma B.10. *Suppose $\{\theta_n\}$ is a sequence generated by mSGD in (4). Under Assumptions 3.1–3.2, it holds that: for $n \geq 1$,*

$$\sum_{t=1}^{n} \epsilon_t^{1+2\eta_0}\, \mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t)\|^2\right) < B(v_0, \theta_1) < +\infty, \qquad \sum_{t=1}^{n} \epsilon_t^{1+2\eta_0} \|\nabla_{\theta_t} g(\theta_t)\|^2 < +\infty,$$

where $B(v_0, \theta_1) > 0$ is a constant only related to v_0 and θ_1.

Lemma B.11. *Suppose $\{\theta_n\}$ is a sequence generated by mSGD in (4). Under Assumptions 3.1–3.2, it holds that: for $n \geq 1$,*

$$g(\theta_{n+1}) - g(\theta_n) \leq \hat{k}\epsilon_t^{1+2\eta_0} + Q_n, \tag{8}$$

where $\hat{k} > 0$ is a constant and $\{Q_n\}$ is a sequence, such that $\sum_{n=1}^{+\infty} Q_n$ exists and is finite a.s..

B.1 Proof of Lemma B.6

The proof of this lemma needs some identical transformations. We assume $\mu > \sigma$ (the case $\mu < \sigma$ is the same), and let $\omega_0 = \log_\mu \sigma > 1$. Then we derive

$$\begin{aligned}
&\sum_{k=1}^{n} \mu^{n-k} \sum_{i=1}^{k} \sigma^{k-i}\psi_i = \sum_{k=1}^{n}\sum_{i=1}^{k} \mu^{n-k}\sigma^{k-i}\psi_i = \sum_{i=1}^{n}\sum_{k=i}^{n} \mu^{n-k}\sigma^{k-i}\psi_i = \sum_{i=1}^{n}\sum_{k=i}^{n} \mu^{n-k+\omega_0(k-i)}\psi_i \\
&= \sum_{i=1}^{n} \mu^{n-\omega_0 i} \sum_{k=i}^{n} \mu^{(\omega_0-1)k}\psi_i = \sum_{i=1}^{n} \mu^{n-\omega_0 i} \frac{1-\mu^{(\omega_0-1)n}}{1-\mu^{(\omega_0-1)}} \mu^{(\omega_0-1)i}\psi_i = \frac{1-\mu^{(\omega_0-1)n}}{1-\mu^{(\omega_0-1)}} \sum_{i=1}^{n} \mu^{n-i}\psi_i.
\end{aligned}$$

B.2 Proof of Lemma B.7

We calculate $\sum_{t=1}^{n} c_t^{(\epsilon)}$ as follows

$$\begin{aligned}
\sum_{t=1}^{n} c_t^{(\epsilon)} &= \sum_{t=1}^{n}\left(\sum_{i=1}^{t} \tau^{t-i}\epsilon_i\right) - \sum_{t=1}^{n} \frac{1}{1-\tau}\epsilon_t = \sum_{i=1}^{n}\left(\sum_{t=i}^{n} \tau^{t-i}\right)\epsilon_i - \sum_{t=1}^{n} \frac{1}{1-\tau}\epsilon_t \\
&= \sum_{t=1}^{n}\left(\sum_{i=t}^{n} \tau^{i-t}\right)\epsilon_t - \sum_{t=1}^{n} \frac{1}{1-\tau}\epsilon_t = \sum_{t=1}^{n}\left(\sum_{i=0}^{n-t} \tau^{i}\right)\epsilon_t - \sum_{t=1}^{n} \frac{1}{1-\tau}\epsilon_t \\
&= \sum_{t=1}^{n}\left(\frac{1-\tau^{n-t+1}}{1-\tau}\right)\epsilon_t - \sum_{t=1}^{n} \frac{1}{1-\tau}\epsilon_t = -\frac{\tau}{1-\tau}\sum_{t=1}^{n} \tau^{n-t}\epsilon_t.
\end{aligned}$$

So we just need to prove $\lim_{n\to+\infty} \left|\sum_{t=1}^{n} \tau^{n-t}\epsilon_t\right| = 0$. It holds that

$$\left|\sum_{t=1}^{n} \tau^{n-t}\epsilon_t\right| = \left|\left(\sum_{t=1}^{n-[\sqrt{n}]} + \sum_{t=[n-\sqrt{n}]+1}^{n}\right)\tau^{n-t}\epsilon_t\right| \leq \left|\sum_{t=1}^{n-[\sqrt{n}]} \tau^{n-t}\epsilon_t\right| + \left|\sum_{t=[n-\sqrt{n}]+1}^{n} \tau^{n-t}\epsilon_t\right|.$$

Due to $n - \sqrt{n} \to +\infty$, $\forall\, \sigma_0 > 0$, $\exists\, N_0 > 0$, when $t > N_0 - [\sqrt{N_0}]$, it holds that $|\epsilon_t| < \sigma_0$. So when $n > N_0$, it holds that

$$\left|\sum_{t=n-[\sqrt{n}]+1}^{n} \tau^{n-t}\epsilon_t\right| < \sum_{t=n-[\sqrt{n}]+1}^{n} \tau^{n-t}\sigma_0 < \sigma_0 \sum_{t=0}^{+\infty} \tau^t = \frac{\sigma_0}{1-\tau}.$$

On the other hand

$$\left|\sum_{t=1}^{n-[\sqrt{n}]} \tau^{n-t}\epsilon_t\right| < |\epsilon_1| \sum_{t=1}^{n-[\sqrt{n}]} \tau^{n-t} = |\epsilon_1| \sum_{t=[\sqrt{n}]}^{n-1} \tau^t \to 0,$$

so $\forall\, \sigma_0 > 0$, $\exists\, N_1 > 1$, when $n > N_1$, it holds that $\left|\sum_{t=1}^{n-[\sqrt{n}]} \tau^{n-t}\epsilon_t\right| < \sigma_0$. Then when $n > \max\{N_0, N_1\}$, it holds that

$$\left|\sum_{t=1}^{n} \tau^{n-t}\epsilon_t\right| \leq \left|\sum_{t=1}^{n-[\sqrt{n}]} \tau^{n-t}\epsilon_t\right| + \left|\sum_{t=[n-\sqrt{n}]+1}^{n} \tau^{n-t}\epsilon_t\right| < \frac{2-\tau}{1-\tau}\sigma_0.$$

Due to the arbitrary of σ_0, we know

$$\sum_{t=1}^{n} \tau^{n-t} \epsilon_t \to 0.$$

Thus, it holds that $\sum_{t=1}^{n} c_t^{(\epsilon)} = 0$.

B.3 Proof of Lemma B.8

Recall the mSGD algorithm in (4)

$$v_n = \alpha v_{n-1} + \epsilon_n \nabla_{\theta_n} g(\theta_n, \xi_n) \tag{9}$$

$$\theta_{n+1} = \theta_n - v_n. \tag{10}$$

Equation (9) is equivalent to

$$v_n = \alpha v_{n-1} + \epsilon_n \nabla_{\theta_n} g(\theta_n) + \epsilon_n \big(\nabla_{\theta_n} g(\theta_n, \xi_n) - \nabla_{\theta_n} g(\theta_n)\big).$$

Under Assumption 3.1 4), it follows from Lemma B.1 that

$$-\nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} v_t - \frac{c}{2}\|v_t\|^2 \le g(\theta_{t+1}) - g(\theta_t) \le -\nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} v_t + \frac{c}{2}\|v_t\|^2. \tag{11}$$

In this subsection, we just use the right side of (11). The left side will be used in the next subsection. Consider $\nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} v_t$ in the following

$$\begin{aligned}
\nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} v_t &= (\nabla_{\theta_t} g(\theta_t))^{\mathsf{T}} (\alpha v_{t-1} + \epsilon_t \nabla_{\theta_n} g(\theta_t, \xi_t)) \\
&= \alpha(\nabla_{\theta_{t-1}} g(\theta_{t-1}) + \nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_{t-1}} g(\theta_{t-1}))^{\mathsf{T}} v_{t-1} + \epsilon_t \nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} \nabla_{\theta_t} g(\theta_t, \xi_t) \\
&= \alpha \nabla_{\theta_{t-1}} g(\theta_{t-1})^{\mathsf{T}} v_{t-1} + \alpha(\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_{t-1}} g(\theta_{t-1}))^{T} v_{t-1} + \epsilon_t \nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} \nabla_{\theta_t} g(\theta_t, \xi_t).
\end{aligned}$$

Recursively applying the above equation yields

$$\begin{aligned}
\nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} v_t &= \alpha^{t-1} \nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 + \sum_{i=1}^{t-1} \alpha^{t-i} (\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_{i-1}} g(\theta_{i-1}))^{T} v_{i-1} \\
&+ \sum_{i=2}^{t} \alpha^{t-i} \epsilon_i \nabla_{\theta_i} g(\theta_i)^{\mathsf{T}} \nabla_{\theta_i} g(\theta_i, \xi_i).
\end{aligned} \tag{12}$$

Substitute (12) into (11), then we have

$$\begin{aligned}
&\mathbb{E}\left(g(\theta_{t+1}) - g(\theta_t)\right) \\
&\le \alpha^{t-1} \mathbb{E}\left(\nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1\right) + c \sum_{i=1}^{t} \alpha^{t-i} \mathbb{E}\left(\|v_i\|^2\right) - \sum_{i=2}^{t} \alpha^{t-i} \epsilon_i \mathbb{E}\left\|\nabla_{\theta_i} g(\theta_i)\right\|^2 \\
&< -\alpha^{t-1} \mathbb{E}\left(\nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1\right) + c \sum_{i=1}^{t} \alpha^{t-i} \mathbb{E}\left(\|v_i\|^2\right).
\end{aligned} \tag{13}$$

From Lemma B.2, it follows that $\mathbb{E}_{\xi_n}\left(\|\nabla_{\theta_n} g(\theta_n, \xi_n)\|^2\right) \le M' \|\nabla_{\theta_n} g(\theta_n)\|^2 + a' \le 2cM'\big(g(\theta_n) - g^*\big) + a' \le M(g(\theta_n) + 1)$ $(M\ always\ exists)$. Given $\forall\, \eta_0 > 0$, we denote $Z(t) := \prod_{k=t}^{+\infty}(1 + M_0 \epsilon_k^{2+\eta_0})$, where

$$M_0 = \frac{cM}{\alpha^{1-\delta}(1-\alpha^{\delta})(1-\alpha)^2},$$

in which $\delta > 0$ is a constant and M is defined above. Here we define M_0, $Z(t)$ and δ to ease the proof. From Assumption 3.2, it holds that $\sum_{t=1}^{+\infty} \epsilon_t^{2+\eta_0} < +\infty$, which ensures $\sum_{t=1}^{+\infty} M_0 \epsilon_t^{2+\eta_0} < +\infty$. From a general inequality $\ln(1+\theta) \le \theta$ for $\theta > -1$, we get

$$Z(t) \le \prod_{k=1}^{+\infty}(1 + M_0 \epsilon_k^{2+\eta_0}) = \mathbb{E}\left\{\sum_{k=1}^{\infty} \ln(1 + M_0 \epsilon_k^{2+\eta_0})\right\} \le \mathbb{E}\left\{\sum_{k=1}^{\infty} M_0 \epsilon_k^{2+\eta_0}\right\} < +\infty,$$

which means that $Z(t)$ is uniformly upper bounded. Then multiplying $Z(t+1)$ on both sides of (69) and taking the mathematical expectation yield

$$\begin{aligned}
&Z(t+1)\, \mathbb{E}\left(\epsilon_{t+1}^{\eta_0} g(\theta_{t+1}) - \epsilon_t^{\eta_0} g(\theta_t)\right) \\
&\le -Z(t+1)\alpha^{t-1} \epsilon_t^{\eta_0} \mathbb{E}\left(\nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1\right) + c \sum_{i=1}^{t} \alpha^{t-i} Z(i+1)\, \mathbb{E}\left(\epsilon_i^{\eta_0} \|v_i\|^2\right) \\
&\quad - Z(t+1)\epsilon_t^{\eta_0} \sum_{i=2}^{t} \alpha^{t-i} \epsilon_i \mathbb{E}\left\|\nabla_{\theta_i} g(\theta_i)\right\|^2 \\
&< -Z(t+1)\alpha^{t-1} \epsilon_t^{\eta_0} \mathbb{E}\left(\nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1\right) + c \sum_{i=1}^{t} \alpha^{t-i} Z(i+1)\, \mathbb{E}\left(\epsilon_i^{\eta_0} \|v_i\|^2\right),
\end{aligned} \tag{14}$$

where the first and second inequalities are respectively due to $Z(i+1) < Z(i)$ and

$$Z(t+1)\epsilon_t^{\eta_0} \sum_{i=2}^{t} \alpha^{t-i} \epsilon_i \, \mathbb{E} \left\| \nabla_{\theta_i} g(\theta_i) \right\|^2 > 0.$$

Next, we aim to analyze $c \sum_{i=1}^{t} \alpha^{t-i} Z(i+1) \, \mathbb{E} \left(\epsilon_t^{\eta_0} \|v_i\|^2 \right)$ in (14). It is proved in Appendix B.4 that

$$\begin{aligned}
& Z(i+1) \, \mathbb{E} \left(\epsilon_t^{\eta_0} \|v_i\|^2 \right) \\
& \leq \alpha^{(1+\delta)i} Z(1) \epsilon_t^{\eta_0} \, \mathbb{E} \|v_0\|^2 + \frac{1}{\alpha^{1-\delta}} \sum_{k=1}^{i} \alpha^{(1+\delta)(i-k)} \epsilon_k^{2+\eta_0} Z(k+1) \, \mathbb{E} \left(\|\nabla_{\theta_k} g(\theta_k) - \nabla_{\theta_k} g(\theta_k, \xi_k)\|^2 \right) \\
& - \frac{2}{\alpha^{1-\delta}} \sum_{k=1}^{i} \alpha^{(1+\delta)(i-k)} Z(k+1) \Big(\mathbb{E} \left(\epsilon_{k+1}^{1+\eta_0} g(\theta_{k+1}) \right) - \mathbb{E} \left(\epsilon_k^{1+\eta_0} g(\theta_k) \right) \Big).
\end{aligned} \tag{15}$$

Taking a weighted sum of (15) yields

$$\begin{aligned}
& \sum_{i=1}^{t} \alpha^{t-i} Z(i+1) \, \mathbb{E} \left(\epsilon_i^{\eta_0} \|v_i\|^2 \right) \\
\leq & \sum_{i=1}^{t} \alpha^{t-i} \alpha^{(1+\delta)i} Z(1) \, \mathbb{E} \left(\epsilon_0^{\eta_0} \|v_0\|^2 \right) \\
& + \sum_{i=1}^{t} \alpha^{t-i} \left(\frac{1}{\alpha^{1-\delta}} \sum_{k=1}^{i} \alpha^{(1+\delta)(i-k)} \epsilon_k^{2+\eta_0} Z(k+1) \, \mathbb{E} \left(\|\nabla_{\theta_k} g(\theta_k) - \nabla_{\theta_k} g(\theta_k, \xi_k)\|^2 \right) \right) \\
& - \sum_{i=1}^{t} \alpha^{t-i} \left(\frac{2}{\alpha^{1-\delta}} \sum_{k=1}^{i} \alpha^{(1+\delta)(i-k)} Z(k+1) \Big(\mathbb{E} \left(\epsilon_{k+1}^{1+\eta_0} g(\theta_{k+1}) \right) - \mathbb{E} \left(\epsilon_k^{1+\eta_0} g(\theta_k) \right) \Big) \right) \\
& := A + B + C.
\end{aligned} \tag{16}$$

Derive that

$$A = \left(\sum_{i=1}^{t} \alpha^{\delta i} \right) \alpha^t Z(1) \, \mathbb{E} \left(\epsilon_0^{\eta_0} \|v_0\|^2 \right) \leq \frac{\alpha^{t-1} \alpha^{\delta}}{1-\alpha^{\delta}} Z(1) \, \mathbb{E} \left(\|v_0\|^2 \right), \tag{17}$$

$$\begin{aligned}
B &= \frac{1}{\alpha^{1-\delta}} \sum_{i=1}^{t} \sum_{k=1}^{i} \alpha^{t-k+\delta(i-k)} \left(\epsilon_k^{2+\eta_0} Z(k+1) \, \mathbb{E} \left(\|\nabla_{\theta_k} g(\theta_k) - \nabla_{\theta_k} g(\theta_k, \xi_k)\|^2 \right) \right) \\
&\leq \frac{1}{\alpha^{1-\delta}(1-\alpha^{\delta})} \sum_{k=1}^{t} \alpha^{t-k} Z(k+1) \epsilon_k^{2+\eta_0} \, \mathbb{E} \left(\|\nabla_{\theta_k} g(\theta_k) - \nabla_{\theta_k} g(\theta_k, \xi_k)\|^2 \right),
\end{aligned} \tag{18}$$

$$\begin{aligned}
C &= -\frac{2}{\alpha^{1-\delta}} \sum_{k=1}^{t} \sum_{i=k}^{t} \alpha^{t-k+\delta(i-k)} Z(k+1) \Big(\mathbb{E} \left(\epsilon_{k+1}^{1+\eta_0} g(\theta_{k+1}) \right) - \mathbb{E} \left(\epsilon_k^{1+\eta_0} g(\theta_k) \right) \Big) \\
&= -\frac{2}{\alpha^{1-\delta}} \sum_{k=1}^{t} \left(\sum_{i=0}^{t-k} \alpha^{\delta i} \right) \alpha^{t-k} Z(k+1) \Big(\mathbb{E} \left(\epsilon_{k+1}^{1+\eta_0} g(\theta_{k+1}) \right) - \mathbb{E} \left(\epsilon_k^{1+\eta_0} g(\theta_k) \right) \Big).
\end{aligned} \tag{19}$$

Substituting (17)–(19) into (14) yields

$$\begin{aligned}
& Z(t+1) \, \mathbb{E} \Big(\left(\epsilon_{t+1}^{1+\eta_0} g(\theta_{t+1}) - \epsilon_t^{1+\eta_0} g(\theta_t) \right) \Big) \\
\leq & - Z(t+1) \alpha^{t-1} \epsilon_t^{1+\eta_0} \, \mathbb{E} \left(\nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 \right) + \frac{\alpha^{t-1} c \alpha^{\delta}}{1-\alpha^{\delta}} Z(1) \, \mathbb{E} \left(\epsilon_0^{1+\eta_0} \|v_0\|^2 \right) \\
& + \frac{c}{\alpha^{1-\delta}(1-\alpha^{\delta})} \sum_{i=1}^{t} \alpha^{t-i} Z(i+1) \epsilon_i^{2+\eta_0} \, \mathbb{E} \left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2 \right) \\
& - \frac{2c}{\alpha^{1-\delta}} \sum_{i=1}^{t} \left(\sum_{k=0}^{t-i} \alpha^{\delta k} \right) \alpha^{t-i} Z(i+1) \Big(\mathbb{E} \left(\epsilon_{i+1}^{1+\eta_0} g(\theta_{i+1}) \right) - \mathbb{E} \left(\epsilon_i^{1+\eta_0} g(\theta_i) \right) \Big).
\end{aligned} \tag{20}$$

Construct a sequence $\{V_n\}$ as follows

$$V_n = \sum_{t=1}^{n} \left(\frac{1}{2-\alpha} \right)^{n-t} Z(t+1) \, \mathbb{E} \Big(\left(\epsilon_{t+1}^{1+\eta_0} g(\theta_{t+1}) - \epsilon_t^{1+\eta_0} g(\theta_t) \right) \Big). \tag{21}$$

By substituting (20) into (21) following the way of (17)–(19), we have

$$
\begin{aligned}
v_n \leq & \frac{\alpha^n(2-\alpha)}{1-\alpha} Z(1)\epsilon_0^{1+\eta_0} \left| \mathbb{E}\left(\nabla_{\theta_1} g(\theta_1)^\mathsf{T} v_1\right)\right| + \frac{c\alpha^n\alpha^\delta(2-\alpha)}{(1-\alpha)(1-\alpha^\delta)} Z(1)\,\mathbb{E}\left(\epsilon_0^{1+\eta_0}\|v_0\|^2\right) \\
& + \frac{c}{\alpha^{1-\delta}(1-\alpha^\delta)(1-\alpha)^2} \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-i} Z(t+1)\epsilon_t^{2+\eta_0}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t,\xi_t)\|^2\right) \\
& - \frac{2c}{\alpha^{1-\delta}(1-\alpha^\delta)} \sum_{t=1}^{n} f(n-t)\left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\Big(\mathbb{E}\left(\epsilon_{t+1}^{1+\eta_0} g(\theta_{t+1})\right) - \mathbb{E}\left(\epsilon_t^{1+\eta_0} g(\theta_t)\right)\Big),
\end{aligned}
$$

where $f(n-t)$ is defined as follows,

$$
f(n-t) = \sum_{k=1}^{n-t} \left(\alpha(2-\alpha)\right)^k - \alpha^\delta \sum_{k=1}^{n-t} \left(\alpha^{1+\delta}(2-\alpha)\right)^k.
$$

Move the last term to the left-hand side of the above inequality, then we have

$$
\begin{aligned}
& \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\,\mathbb{E}\left(e_{t+1}^{(n)} g(\theta_{t+1}) - e_t^{(n-1)} g(\theta_t)\right) \\
\leq & \frac{\alpha^n(2-\alpha)}{1-\alpha} Z(1)\left|\mathbb{E}\left(\epsilon_0^{\eta_0}\nabla_{\theta_1} g(\theta_1)^\mathsf{T} v_1\right)\right| + \frac{c\alpha^n\alpha^\delta(2-\alpha)}{(1-\alpha)(1-\alpha^\delta)} Z(1)\,\mathbb{E}\left(\epsilon_0^{\eta_0}\|v_0\|^2\right) \\
& + \frac{c}{\alpha^{1-\delta}(1-\alpha^\delta)(1-\alpha)^2} \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\epsilon_t^{2+\eta_0}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t,\xi_t)\|^2\right),
\end{aligned} \tag{22}
$$

where

$$
e_t^{(n-1)} = \left(\epsilon_t^{\eta_0} + \frac{2c\epsilon_t^{1+\eta_0} f(n-t)}{\alpha^{1-\delta}(1-\alpha^\delta)}\right). \tag{23}
$$

Because of $\alpha < 1$, it holds that $f(n-t) > 0$ and $e_t^{(n-1)} > 1$. It follows from Assumption 3.2 that

$$
\begin{aligned}
& \frac{c}{\alpha^{1-\delta}(1-\alpha^\delta)(1-\alpha)^2} \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\epsilon_t^{2+\eta_0}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t,\xi_t)\|^2\right) \\
\leq & \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1) M_0 \epsilon_t^{2+\eta_0} e_t^{(n-1)}\left(1 + \mathbb{E}(g(\theta_t))\right),
\end{aligned} \tag{24}
$$

where

$$
M_0 = \frac{cM}{\alpha^{1-\delta}(1-\alpha^\delta)(1-\alpha)^2}.
$$

Calculate $f(n-t)$, then we obtain

$$
f(n-t) = \sum_{k=1}^{n-t} \left(\alpha(2-\alpha)\right)^k - \alpha^\delta \sum_{k=1}^{n-t} \left(\alpha^{1+\delta}(2-\alpha)\right)^k > 0.
$$

It holds that

$$
\begin{aligned}
& \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\,\mathbb{E}\left(e_{t+1}^{(n)} g(\theta_{t+1}) - e_t^{(n-1)} g(\theta_t)\right) \\
& - \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1) M_0 \epsilon_t^{2+\eta_0} e_t^{(n-1)}\left(1 + \mathbb{E}(g(\theta_t))\right) \\
= & \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\,\mathbb{E}\left(e_{t+1}^{(n)}\left(1 + g(\theta_{t+1})\right)\right) \\
& - \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)(1 + M_0\epsilon_t^{2+\eta_0})\,\mathbb{E}\left(e_t^{(n-1)}\left(1 + g(\theta_t)\right)\right) \\
& - \frac{c}{\alpha^{1-\delta}(1-\alpha^\delta)} \sum_{t=1}^{n} f(n-t)\left(\frac{1}{2-\alpha}\right)^{n-t} \left(\epsilon_{t+1}^{1+\eta_0} - \epsilon_t^{1+\eta_0}\right) \\
> & \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\,\mathbb{E}\left(e_{t+1}^{(n)}\left(1 + g(\theta_{t+1})\right)\right) \\
& - \sum_{t=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)(1 + M_0\epsilon_t^{2+\eta_0})\,\mathbb{E}\left(e_t^{(n-1)}\left(1 + g(\theta_t)\right)\right).
\end{aligned} \tag{25}
$$

Substituting (24) and (25) into (22) yields

$$\begin{aligned}
&\sum_{t=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\,\mathbb{E}\left(e_{t+1}^{(n)}\big(1+g(\theta_{t+1})\big)\right)\\
&-\sum_{t=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)(1+M_0\epsilon_t^{2+\eta_0})\,\mathbb{E}\left(e_t^{(n-1)}\big(1+g(\theta_t)\big)\right)\\
\leq&\frac{\alpha^n(2-\alpha)}{1-\alpha}Z(1)\epsilon_0^{1+\eta_0}\left|\mathbb{E}\left(\nabla_{\theta_1}g(\theta_1)^{\mathsf{T}}v_1\right)\right| + \frac{c\alpha^n\alpha^\delta(2-\alpha)}{2(1-\alpha)(1-\alpha^\delta)}Z(1)\,\mathbb{E}\left(\epsilon_0^{1+\eta_0}\|v_0\|^2\right).
\end{aligned} \tag{26}$$

According to the definition of $Z(t)$, we have

$$Z(t)=\prod_{k=t}^{+\infty}(1+M_0\epsilon_k^2)=(1+M_0\epsilon_t^{2+\eta_0})\prod_{k=t+1}^{+\infty}(1+M_0\epsilon_k^{2+\eta_0})=(1+M_0\epsilon_t^{2+\eta_0})Z(t+1).$$

Let

$$F_n:=\sum_{t=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-t} Z(t+1)\,\mathbb{E}\left(e_{t+1}^{(n)}\big(1+g(\theta_{t+1})\big)\right),$$

and it follows from (26) that

$$F_n-F_{n-1}\leq\widetilde{k}_0(\theta_1,v_0)\left(\left(\frac{1}{2-\alpha}\right)^{n-1}+\alpha^n\right), \tag{27}$$

where $\widetilde{k}_0(\theta_1,v_0)$. Denote $p=\mathbb{E}\left\{\sum_{k=1}^{\infty}M_0\epsilon_k^{2+\eta_0}\right\}$. By taking the summation of (27), we obtain

$$F_n\leq F_1+\widetilde{k}_0\sum_{t=1}^{+\infty}\left(\left(\frac{1}{2-\alpha}\right)^{n-1}+\alpha^n\right).$$

Thus, we derive that

$$\begin{aligned}
\epsilon_{n+1}^{\eta_0}\,\mathbb{E}\left(g(\theta_{n+1})\right)<e_{n+1}^{n}\,\mathbb{E}\left(1+g(\theta_{n+1})\right)&<p\left(\frac{1}{2-\alpha}\right)^{n-n}Z(n+1)\,\mathbb{E}\left(e_{n+1}^{(n)}\big(1+g(\theta_{n+1})\big)\right)\\
&<p\sum_{t=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-t}Z(t+1)\,\mathbb{E}\left(e_{t+1}^{(n)}\big(1+g(\theta_{t+1})\big)\right)\\
&<pF_n<T(\theta_1,v_0),
\end{aligned}$$

where $T(\theta_1,v_0)$ is a constant determined by v_0 and θ_1.

B.4 Proof of (15)

We consider

$$\begin{aligned}
&\epsilon_{i+1}^{1+\eta_0}g(\theta_{i+1})-\epsilon_i^{1+\eta_0}g(\theta_i)\\
=&\,\epsilon_i^{1+\eta_0}\big(g(\theta_{i+1}-g(\theta_i)\big)+(\epsilon_{i+1}^{1+\eta_0}-\epsilon_i^{1+\eta_0})g(\theta_i)\\
\leq&\,\epsilon_i^{1+\eta_0}\big(g(\theta_{i+1})-g(\theta_i)\big)\\
\leq&-\epsilon_i^{1+\eta_0}\nabla_{\theta_i}g(\theta_i)^{\mathsf{T}}v_i+\frac{c}{2}\epsilon_i^{1+\eta_0}\|v_i\|^2\\
=&\,(-\epsilon_i^{1+\eta_0}\nabla_{\theta_i}g(\theta_i,\xi_i)^{\mathsf{T}}v_i-(\epsilon_i^{1+\eta_0}\nabla_{\theta_i}g(\theta_i)-\epsilon_i^{1+\eta_0}\nabla_{\theta_i}g(\theta_i,\xi_i))^{\mathsf{T}}v_i+\frac{c}{2}\epsilon_i^{1+\eta_0}\|v_i\|^2\\
=&\,\alpha\epsilon_i^{1+\eta_0}v_i^{\mathsf{T}}v_{i-1}-\epsilon_i^{1+\eta_0}\|v_i\|^2+\epsilon_i^{2+\eta_0}\left\|\nabla_{\theta_i}g(\theta_i)-\nabla_{\theta_i}g(\theta_i,\xi_i)\right\|^2\\
&-\left(\epsilon_i^{1+\eta_0}\alpha v_{i-1}+\epsilon_i^{2+\eta_0}\nabla_{\theta_i}g(\theta_i)\right)^{\mathsf{T}}\left(\nabla_{\theta_i}g(\theta_i)-\nabla_{\theta_i}g(\theta_i,\xi_i))\right)+\frac{c}{2}\epsilon_i^{1+\eta_0}\|v_i\|^2,
\end{aligned} \tag{28}$$

where the first inequality is due to $\epsilon_i\leq\epsilon_{i-1}$ in Assumption 3.2, and the last equality is from (10). Since ξ_i and θ_i are independent, taking the mathematical expectation of (28) and noting that

$$\mathbb{E}\left(\left(\epsilon_i^{1+\eta_0}\alpha v_{i-1}+\epsilon_i^{2+\eta_0}\nabla_{\theta_i}g(\theta_i)\right)^{\mathsf{T}}\left(\nabla_{\theta_i}g(\theta_i)-\nabla_{\theta_i}g(\theta_i,\xi_i))\right)\right)=0,$$

yield

$$\begin{aligned}
&\mathbb{E}\left(\epsilon_{i+1}^{1+\eta_0}g(\theta_{i+1})\right)-\mathbb{E}\left(\epsilon_i^{1+\eta_0}g(\theta_i)\right)\\
\leq&\alpha\epsilon_i^{\eta_0}\,\mathbb{E}(v_i^{\mathsf{T}}v_{i-1})-\epsilon_i^{\eta_0}\,\mathbb{E}\left(\|v_i\|^2\right)+\epsilon_i^{2+\eta_0}\,\mathbb{E}\left(\left\|\nabla_{\theta_i}g(\theta_i)-\nabla_{\theta_i}g(\theta_i,\xi_i)\right\|^2\right)+\frac{c}{2}\epsilon_i^{1+\eta_0}\,\mathbb{E}\left(\|v_i\|^2\right).
\end{aligned} \tag{29}$$

Moreover, it holds that

$$\begin{aligned}
&\mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2\right) \\
=&\mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i, \xi_i)\|^2\right) - \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i)\|^2\right) \\
=&\frac{1}{\epsilon_i}\,\mathbb{E}\left(\|v_i - \alpha v_{i-1}\|^2\right) - \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i)\|^2\right) \\
=&\frac{1}{\epsilon_i}\left(\mathbb{E}\left(\|v_i\|^2\right) + \alpha^2\,\mathbb{E}\left(\|v_{i-1}\|^2\right) - 2\alpha\,\mathbb{E}(v_i^{\mathsf{T}} v_{i-1})\right) - \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i)\|^2\right).
\end{aligned} \tag{30}$$

Combining (29) and (30), we get

$$\begin{aligned}
&\mathbb{E}\left(\epsilon_{i+1}^{1+\eta_0} g(\theta_{i+1})\right) - \mathbb{E}\left(\epsilon_i^{1+\eta_0} g(\theta_i)\right) \\
\leq& -\frac{1}{2}\left(\mathbb{E}\left(\epsilon_i^{1+\eta_0}\|v_i\|^2\right) - \alpha^2\,\mathbb{E}\left(\epsilon_{i-1}^{1+\eta_0}\|v_{i-1}\|^2\right)\right) - \frac{\epsilon_i^{2+\eta_0}}{2}\,\mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i)\|^2\right) + \frac{c}{2}\epsilon_i^{1+\eta_0}\,\mathbb{E}\left(\|v_i\|^2\right) \\
&+ \frac{\epsilon_i^{2+\eta_0}}{2}\,\mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2\right).
\end{aligned} \tag{31}$$

Since $\epsilon_i \to 0$, given any $\delta > 0$, there is an integer $i_0 \geq 0$, such that for $i \geq i_0$, $1 - c\epsilon_i > \alpha^{1-\delta}(\delta > 0)$. Since i_0 is finite, without loss of generality, we assume $i_0 = 0$ for convenience, i.e., $1 - c\epsilon_i > \alpha^{1-\delta}(\delta > 0)(i \geq 1)$. Thus, we have

$$\begin{aligned}
&\mathbb{E}\left(\epsilon_{i+1}^{1+\eta_0} g(\theta_{i+1})\right) - \mathbb{E}\left(\epsilon_i^{1+\eta_0} g(\theta_i)\right) \\
\leq& -\frac{\alpha^{1-\delta}}{2}\left(\mathbb{E}\left(\epsilon_i^{1+\eta_0}\|v_i\|^2\right) - \alpha^{1+\delta}\,\mathbb{E}\left(\epsilon_{i-1}^{1+\eta_0}\|v_{i-1}\|^2\right)\right) + \frac{\epsilon_i^{2+\eta_0}}{2}\,\mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2\right).
\end{aligned} \tag{32}$$

Multiplying both sides of (32) by $Z(i+1)$, and noticing that $Z(i) > Z(i+1)$, we have

$$\begin{aligned}
&\left(Z(i+1)\,\mathbb{E}\left(\epsilon_i^{\eta_0}\|v_i\|^2\right) - \alpha^{1+\delta} Z(i)\,\mathbb{E}\left(\epsilon_{i-1}^{\eta_0}\|v_{i-1}\|^2\right)\right) \\
\leq& -\frac{2}{\alpha^{1-\delta}} Z(i+1)\left(\mathbb{E}\left(\epsilon_{i+1}^{1+\eta_0} g(\theta_{i+1})\right) - \mathbb{E}\left(\epsilon_i^{1+\eta_0} g(\theta_i)\right)\right) + \frac{\epsilon_i^{2+\eta_0}}{\alpha^{1-\delta}} Z(i+1)\,\mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2\right).
\end{aligned}$$

Then (15) is obtained by recursively applying the above inequality.

B.5 Proof of Lemma B.9

From (31), we have

$$\begin{aligned}
&\mathbb{E}\left(\epsilon_{n+1}^{1+2\eta_0} g(\theta_{n+1})\right) - \mathbb{E}\left(\epsilon_1^{1+2\eta_0} g(\theta_1)\right) \\
<& -\frac{1}{2}\sum_{t=1}^{n}\mathbb{E}\left(\epsilon_t^{2\eta_0}\|v_t\|^2\right) + \frac{\alpha^2}{2}\sum_{t=1}^{n}\mathbb{E}\left(\epsilon_{t-1}^{2\eta_0}\|v_{t-1}\|^2\right) - \sum_{t=1}^{n}\frac{\epsilon_t^{2+2\eta_0}}{2}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t)\|^2\right) + \frac{c}{2}\sum_{t=1}^{n}\epsilon_t^{1+2\eta_0}\,\mathbb{E}\left(\|v_t\|^2\right) \\
&+ \sum_{t=1}^{n}\frac{\epsilon_t^{2+2\eta_0}}{2}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2\right).
\end{aligned} \tag{33}$$

It follows from Assumption 3.1 3) and Lemma B.8 that

$$\epsilon_t^{\eta_0}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2\right) < \epsilon_t^{\eta_0}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2\right) MT(\theta_1, v_0) < +\infty.$$

Because of $\sum_{t=1}^{n}\epsilon_t^{2+\eta_0} < +\infty$, there is a scalar $\bar{M} > 0$ such that for $\forall n$

$$\sum_{t=1}^{n}\frac{\epsilon_t^{2+2\eta_0}}{2}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2\right) < MT(\theta_1, v_0)\sum_{t=1}^{n}\frac{\epsilon_t^{2+\eta_0}}{2} < \bar{M} < +\infty.$$

Then it follows from (33) that

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{n}(1 - \alpha^2 - c\epsilon_t)\,\mathbb{E}\left(\epsilon_t^{2\eta_0}\|v_t\|^2\right) \\
\leq& \bar{M} + \epsilon_1^{2\eta_0} g(\theta_1) - \epsilon_{n+1}^{1+2\eta_0}\,\mathbb{E}\left(g(\theta_{n+1})\right) + \frac{\alpha^2}{2}\,\mathbb{E}\left(\epsilon_0^{1+2\eta_0}\|v_0\|^2 - \epsilon_n^{1+2\eta_0}\|v_n\|^2\right) \\
&- \sum_{t=1}^{n}\frac{\epsilon_t^{2+2\eta_0}}{2}\,\mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t)\|^2\right) < K,
\end{aligned}$$

where K is a positive scalar. Since $\epsilon_n \to 0$ when n is large enough, it holds that $\frac{1}{5}(1 - \alpha^2) < 1 - \alpha^2 - c\epsilon_n$. Without loss of generality, assume $\frac{1}{5}(1 - \alpha^2) < 1 - \alpha^2 - c\epsilon_n^2$ for $n \geq 0$, so $\sum_{t=1}^{n}\mathbb{E}\left(\epsilon_t^{2\eta_0}\|v_t\|^2\right) < \frac{10K}{1-\alpha^2} < +\infty$. By Lemma B.5, we obtain $\sum_{t=1}^{n}\epsilon_t^{2\eta_0}\|v_t\|^2 < +\infty$.

B.6 Proof of Lemma B.10

Through Taylor expansion, we derive that

$$\begin{aligned}&\epsilon_{t+1}^{2\eta_0} g(\theta_{t+1}) - \epsilon_t^{2\eta_0} g(\theta_t)\\ =&\epsilon_t^{2\eta_0} \nabla_{\theta_{\xi_t}} g(\theta_{\xi_t})^T (\theta_{t+1} - \theta_t) = -\epsilon_t^{2\eta_0} \nabla_{\theta_t} g(\theta_t)^T v_t + \epsilon_t^{2\eta_0} \Big(\nabla_{\theta_{\xi_t}} g(\theta_{\xi_t}) - \nabla_{\theta_t} g(\theta_t)\Big)^T (\theta_{t+1} - \theta_t).\end{aligned} \tag{34}$$

We first focus on the term $\epsilon_t^{2\eta_0} \nabla_{\theta_t} g(\theta_t)^T v_t$. It holds that

$$\begin{aligned}&\epsilon_t^{2\eta_0} \nabla_{\theta_t} g(\theta_t)^T v_t = \epsilon_t^{2\eta_0} (\nabla_{\theta_t} g(\theta_t))^\mathsf{T} (\alpha v_{t-1} + \epsilon_t \nabla_{\theta_n} g(\theta_t, \xi_t))\\ &= \alpha \epsilon_t^{2\eta_0} (\nabla_{\theta_{t-1}} g(\theta_{t-1}) + \nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_{t-1}} g(\theta_{t-1}))^\mathsf{T} v_{t-1} + \epsilon_t^{1+2\eta_0} \nabla_{\theta_t} g(\theta_t)^\mathsf{T} \nabla_{\theta_t} g(\theta_t, \xi_t)\\ &= \alpha \epsilon_{t-1}^{2\eta_0} \nabla_{\theta_{t-1}} g(\theta_{t-1})^\mathsf{T} v_{t-1} + \alpha \epsilon_{t-1}^{2\eta_0} (\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_{t-1}} g(\theta_{t-1}))^T v_{t-1} + \epsilon_t^{1+2\eta_0} \nabla_{\theta_t} g(\theta_t)^\mathsf{T} \nabla_{\theta_t} g(\theta_t, \xi_t)\\ &+ (\epsilon_t^{2\eta_0} - \epsilon_{t-1}^{2\eta_0}) \nabla_{\theta_{t-1}} g(\theta_{t-1})^\mathsf{T} v_{t-1}.\end{aligned}$$

By substituting the above equation into (11) and noting $-(\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_{i-1}} g(\theta_{i-1}))^\mathsf{T} v_{i-1} \le \big\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_{i-1}} g(\theta_{i-1})\big\| \|v_{i-1}\| \le c\|v_{t-1}\|^2$, we obtain

$$\begin{aligned}&\epsilon_{t+1}^{2\eta_0} g(\theta_{t+1}) - \epsilon_t^{2\eta_0} g(\theta_t)\\ \le& -\alpha^{t-1} \epsilon_1^{2\eta_0} \nabla_{\theta_1} g(\theta_1)^\mathsf{T} v_1 - \sum_{i=2}^{t} \alpha^{t-i} \epsilon_i^{1+2\eta_0} \nabla_{\theta_i} g(\theta_i)^\mathsf{T} \nabla_{\theta_i} g(\theta_i, \xi_i) + \frac{c}{2} \epsilon_t^{2\eta_0} \|v_t\|^2\\ &- \sum_{i=1}^{t-1} \alpha^{t-i} \epsilon_t^{2\eta_0} (\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_{t-1}} g(\theta_{t-1}))^T v_{t-1}\\ &+ \sum_{i=2}^{t} \alpha^{t-i} (\epsilon_t^{2\eta_0} - \epsilon_{t-1}^{2\eta_0}) \nabla_{\theta_{t-1}} g(\theta_{t-1})^\mathsf{T} v_{t-1}\\ <& -\alpha^{t-1} \epsilon_1^{2\eta_0} \nabla_{\theta_1} g(\theta_1)^\mathsf{T} v_1 - \sum_{i=2}^{t} \alpha^{t-i} \epsilon_i^{1+2\eta_0} \nabla_{\theta_i} g(\theta_i)^\mathsf{T} \nabla_{\theta_i} g(\theta_i, \xi_i) + c' \sum_{i=1}^{t} \alpha^{t-i} \epsilon_i^{2\eta_0} \|v_i\|^2\\ &+ \sum_{i=2}^{t} \alpha^{t-i} (\epsilon_t^{2\eta_0} - \epsilon_{t-1}^{2\eta_0})^2 \|\nabla_{\theta_{t-1}} g(\theta_{t-1})\|^2,\end{aligned} \tag{35}$$

where c' is a constant which does not affect the result. It follows that

$$\begin{aligned}&\epsilon_{n+1}^{2\eta_0} g(\theta_{n+1})\\ =&\epsilon_1^{2\eta_0} g(\theta_1) + \sum_{t=1}^{n} \big(\epsilon_{t+1}^{2\eta_0} g(\theta_{t+1}) - \epsilon_t^{2\eta_0} g(\theta_t)\big)\\ =&\epsilon_1^{2\eta_0} g(\theta_1) - \frac{1-\alpha^n}{1-\alpha} \epsilon_1^{2\eta_0} \nabla_{\theta_1} g(\theta_1)^\mathsf{T} v_1 + \frac{1-\alpha^n}{1-\alpha} \epsilon_1^{1+2\eta_0} \nabla_{\theta_1} g(\theta_1) \nabla_{\theta_1} g(\theta_1, \zeta_1)\\ &- \sum_{t=1}^{n} \frac{1-\alpha^{n-t+1}}{1-\alpha} \epsilon_t^{1+2\eta_0} \nabla_{\theta_t} g(\theta_t) \nabla_{\theta_t} g(\theta_t, \xi_t) + c'' \sum_{i=1}^{t} \epsilon_i^{2\eta_0} \|v_i\|^2\\ &+ c'' \sum_{i=2}^{t} (\epsilon_t^{2\eta_0} - \epsilon_{t-1}^{2\eta_0})^2 \|\nabla_{\theta_{t-1}} g(\theta_{t-1})\|^2,\end{aligned} \tag{36}$$

where c'' is a constant which does not affect the result. Take the mathematical expectation of (36), and notice Assumption 3.1 1), then we have

$$\begin{aligned}\epsilon_{n+1}^{2\eta_0} \mathbb{E}\left(g(\theta_{n+1})\right) \le& \epsilon_1^{2\eta_0} \mathbb{E}\left(g(\theta_1)\right) + \frac{\alpha \epsilon_1^{2\eta_0}}{1-\alpha} \left|\nabla_{\theta_1} g(\theta_1)^\mathsf{T} v_1\right| + \frac{1}{1-\alpha} \epsilon_1^{1+2\eta_0} \mathbb{E}\left(\|\nabla_{\theta_1} g(\theta_1)\|^2\right)\\ &- c''' \sum_{t=1}^{n} (\epsilon_t^{1+2\eta_0} - (\epsilon_t^{2\eta_0} - \epsilon_{t-1}^{2\eta_0})^2) \mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t)\|^2\right) + c'' \sum_{i=1}^{t} \epsilon_i^{2\eta_0} \mathbb{E}(\|v_i\|^2),\end{aligned}$$

where c''' is a constant which does not affect the result. From Lemma B.9, it follows that for some positive constant Q,

$$c'' \sum_{t=1}^{n} \epsilon_i^{2\eta_0} \mathbb{E}\left(\|v_t\|^2\right) < Q.$$

Since $(\epsilon_t^{1+2\eta_0} - (\epsilon_t^{2\eta_0} - \epsilon_{t-1}^{2\eta_0})^2)$ can be controlled by $k_0\epsilon_t^{1+2\eta_0}$, it holds that

$$\sum_{t=1}^{n} \epsilon_t^{1+2\eta_0}\, \mathbb{E}\left(\|\nabla_{\theta_t} g(\theta_t)\|^2\right) < \frac{Q'}{k_0} < +\infty, \tag{37}$$

where Q' is a constant. From Lemma B.5, we have $\sum_{t=1}^{n} \epsilon_t^{1+2\eta_0} \|\nabla_{\theta_t} g(\theta_t)\|^2 < +\infty\ a.s.$.

B.7 Proof of Lemma B.11

It follows from (11) and (12) that

$$\begin{aligned}
&g(\theta_{t+1}) - g(\theta_t) \\
&\leq \alpha^{t-1}\nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 + c\sum_{i=1}^{t} \alpha^{t-i}\|v_i\|^2 - \sum_{i=2}^{t} \alpha^{t-i}\epsilon_i \left\|\nabla_{\theta_i} g(\theta_i)\right\|^2 \\
&+ \sum_{i=2}^{t} \alpha^{t-i}\epsilon_i \nabla_{\theta_i} g(\theta_i)^T (\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)).
\end{aligned}$$

Then in light of (16), we get that

$$\begin{aligned}
&\sum_{i=1}^{t} \alpha^{t-i}\|v_i\|^2 \\
&\leq \sum_{i=1}^{t} \alpha^{t-i}\alpha^{(1+\delta)i}\|v_0\|^2 + \sigma_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 \|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2 \\
&- \sigma_1 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)} \left(\epsilon_{k+1} g(\theta_{k+1}) - \epsilon_k g(\theta_k)\right),
\end{aligned}$$

where σ_0 and σ_1 are two constants not affecting the analysis. Then we make some transformations to obtain

$$\begin{aligned}
&\sigma_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 \|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2 \\
&= \sigma_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2\, \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2 \middle| \mathcal{F}_n\right) + \sigma_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 M_i,
\end{aligned}$$

where

$$M_i = \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2 \middle| \mathcal{F}_n\right) - \|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2.$$

From Assumption 3.1 3), it follows that

$$\begin{aligned}
&\sigma_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2\, \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2 \middle| \mathcal{F}_n\right) \leq \sigma_0 M_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 \|\nabla_{\theta_i} g(\theta_i)\|^2 \\
&+ a' \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2.
\end{aligned}$$

Then we derive that

$$\begin{aligned}
&\sum_{i=1}^{t} \alpha^{t-i}\|v_i\|^2 \\
&\leq \sum_{i=1}^{t} \alpha^{t-i}\alpha^{(1+\delta)i}\|v_0\|^2 + \sigma_0 M_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 \|\nabla_{\theta_i} g(\theta_i)\|^2 \\
&+ a' \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 - \sigma_1 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)} \left(\epsilon_{k+1} g(\theta_{k+1}) - \epsilon_k g(\theta_k)\right).
\end{aligned}$$

It holds that

$$
\begin{aligned}
&g(\theta_{t+1}) - g(\theta_t) \\
&\leq k_0\alpha^{t-1} + k_1\alpha^t + \sigma_0 M_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 \|\nabla_{\theta_i} g(\theta_i)\|^2 \\
&+ a' \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 - \sigma_1 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\big(\epsilon_{k+1} g(\theta_{k+1}) - \epsilon_k g(\theta_k)\big) - \sum_{i=2}^{t} \alpha^{t-i}\epsilon_i \big\|\nabla_{\theta_i} g(\theta_i)\big\|^2 \\
&+ \sum_{i=2}^{t} \alpha^{t-i}\epsilon_i \nabla_{\theta_i} g(\theta_i)^T (\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)).
\end{aligned}
$$

It follows from some transformations that

$$
g(\theta_{t+1}) - g(\theta_t) \leq \frac{a'}{1-\alpha^{1+\delta}}\epsilon_t^{1+2\eta_0} + Q_t, \tag{38}
$$

where

$$
\begin{aligned}
Q_t &= a'\left(\sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 - \frac{1}{1-\alpha^{1+\delta}}\epsilon_t^2\right) + k_0\alpha^{t-1} + k_1\alpha^t \\
&+ \sigma_0 M_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i \|\nabla_{\theta_i} g(\theta_i)\|^2 - \sigma_1 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\big(\epsilon_{k+1} g(\theta_{k+1}) - \epsilon_k g(\theta_k)\big) \\
&+ \sum_{i=2}^{t} \alpha^{t-i}\epsilon_i \nabla_{\theta_i} g(\theta_i)^T (\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)).
\end{aligned}
$$

Next we will prove that $\sum_{t=1}^{+\infty} Q_t$ exists and is finite $a.s.$. First through Lemma B.7, we get

$$
\sum_{t=1}^{+\infty} a'\left(\sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 - \frac{1}{1-\alpha^{1+\delta}}\epsilon_t^2\right) = 0.
$$

Obviously

$$
\sum_{t=1}^{+\infty} \left(k_0\alpha^{t-1} + k_1\alpha^t\right) < +\infty.
$$

For the term $\sigma_0 M_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 \|\nabla_{\theta_i} g(\theta_i)\|^2$, from (B.10) and $0 < \eta_0 < 1/2$, we derive

$$
\begin{aligned}
&\sum_{t=1}^{+\infty} \sigma_0 M_0 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\epsilon_i^2 \|\nabla_{\theta_i} g(\theta_i)\|^2 = \frac{\sigma_0 M_0}{1-\alpha^{1+\delta}} \sum_{t=1}^{+\infty} \epsilon_t^2 \|\nabla_{\theta_t} g(\theta_t)\|^2 \\
&< \frac{\sigma_0 M_0}{1-\alpha^{1+\delta}} \sum_{t=1}^{+\infty} \epsilon_t^{1+2\eta_0} \|\nabla_{\theta_t} g(\theta_t)\|^2 < +\infty \ \ a.s..
\end{aligned}
$$

For the term $-\sigma_1 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\big(\epsilon_{k+1} g(\theta_{k+1}) - \epsilon_k g(\theta_k)\big)$, from (36), it follows that $\epsilon_n^{2\eta_0} g(\theta_n) < \zeta' < +\infty \ \ (\forall\, n > 0)$. Furthermore, it holds that $\epsilon_n g(\theta_n) = \epsilon_n^{1-2\eta_0}\epsilon_n^{2\eta_0} g(\theta_n) \to 0 \ \ a.s.$ and

$$
\sum_{t=1}^{+\infty} -\sigma_1 \sum_{i=1}^{t} \alpha^{(1+\delta)(t-i)}\big(\epsilon_{k+1} g(\theta_{k+1}) - \epsilon_k g(\theta_k)\big) = \frac{\sigma_0 \epsilon_1 g(\theta_1)}{1-\alpha^{1+\delta}}.
$$

For the term $\sum_{i=2}^{t} \alpha^{t-i}\epsilon_i \nabla_{\theta_i} g(\theta_i)^T (\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i))$, we first consider $\sum_{t=1}^{+\infty} \epsilon_t \nabla_{\theta_t} g(\theta_t)^T (\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t))$. From Lemma B.10, it follows that

$$
\begin{aligned}
&\sum_{t=1}^{+\infty} \mathbb{E}\left(\epsilon_t^2 \big(\nabla_{\theta_t} g(\theta_t)^T \big(\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\big)\big)^2 \big|\mathcal{F}_n\right) \\
&\leq a \sum_{t=1}^{+\infty} \mathbb{E}\left(\epsilon_t^2 \big\|\nabla_{\theta_t} g(\theta_t)\big\|^2 \big|\mathcal{F}_n\right) + \hat{M} \sum_{t=1}^{+\infty} \mathbb{E}\left(\epsilon_t^2 \big\|\nabla_{\theta_t} g(\theta_t)\big\|^4 \big|\mathcal{F}_n\right) < +\infty \ \ a.s.
\end{aligned}
$$

This means $\sum_{t=1}^{+\infty} \epsilon_t \nabla_{\theta_t} g(\theta_t)^T (\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t))$ exists and is finite a.s.. From Lemma B.7, it holds that $\sum_{i=2}^{t} \alpha^{t-i}\epsilon_i \nabla_{\theta_i} g(\theta_i)^T (\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i))$ exists and is finite a.s.. Thus, $\sum_{t=1}^{+\infty} Q_t$ exists and is finite a.s..

C Proofs of Theorems 3.1–3.4 and Proposition 3.2

C.1 Proof of Theorem 3.1

We prove this theorem under two situations, namely, $\sum_{n=1}^{+\infty} \epsilon_n^{1+2\eta_0} < +\infty$ and $\sum_{n=1}^{+\infty} \epsilon_n^{1+2\eta_0} = +\infty$. If $\sum_{n=1}^{+\infty} \epsilon_n^{1+2\eta_0} < +\infty$, we can conclude that $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$. Then our condition degenerates to the classical condition $\sum_{i=1}^{+\infty} \epsilon_n = +\infty$, $\sum_{i=1}^{+\infty} \epsilon_n^2 < +\infty$. The results then follow from Theorem 1 in [7]. Thus, the following proof is all based on the condition that $\sum_{n=1}^{+\infty} \epsilon_n^{1+2\eta_0} = +\infty$. First we will prove that there exists a subsequence $\{\theta_{k_n}\}$, such that $\|\nabla_{\theta_{k_n}} g(\theta_{k_n})\| \to 0 \ a.s.$. We prove it by contradiction. We assume $\exists N(\zeta) > 0$, when $\theta_n > N(\zeta)$, it holds that $\|\nabla_{\theta_t} g(\theta_t)\|^2 > \delta_0^2 > 0$. Then we have

$$\sum_{t=N(\zeta)}^{+\infty} \epsilon_t^{1+2\eta_0} \|\nabla_{\theta_t} g(\theta_t)\|^2 > \delta_0^2 \sum_{t=N(\zeta)}^{+\infty} \epsilon_t^{1+2\eta_0} = +\infty, \tag{39}$$

which contradicts with (37). Thus, there exists a subsequence $\{\theta_{k_n}\}$ satisfying $\|\nabla_{\theta_{k_n}} g(\theta_{k_n})\| \to 0 \ a.s.$. It is easy to find that J_i is a bounded closed set. So $\forall \epsilon > 0$, we can construct an open cover $H_\epsilon^{(i)} = \{U(\theta, \epsilon)\}$ $(\theta \in J_i)$ of J_i. Through the $Heine\breve{}Borel\ theorem$, we can get a finite open subcover $\{U(\theta_k, \epsilon)\ (k = 0, 1, ..., n)$ from $H_\epsilon^{(i)}$. Let $U_\epsilon^{(i)} = \bigcup_{k=0}^{n} U(\theta_k, \epsilon)$, then $U_\epsilon^{(i)}$ is a open set. Under Assumption 3.1, $J = \{\theta | \nabla_\theta g(\theta)\}$ has only finite connected components $J_1, J_2, ..., J_m$. So $\inf_{i \neq j} d(J_i, J_j) = \min_{i \neq j} d(J_i, J_j)$. Let $\delta_0 = \min_{i \neq j} d(J_i, J_j)$. It follows from Lemma B.3 that $\exists \epsilon_0 > 0$, such that when $d(\theta, J_i) < \epsilon_0$, it holds that

$$\|\nabla_\theta g(\theta)\|^2 \leq 2c|g(\theta) - g_i|, \tag{40}$$

where g_i denotes $g(\theta)$ $(\theta \in J_i)$. Let $c = min\{\epsilon_0, \delta_0/4\}$ and construct $U_c^{(1)}, U_c^{(2)}, ..., U_c^{(m)}$. It is obvious that $\forall\, U_c^{(i)}, U_c^{(j)}$ $(i \neq j)$, $d(U_c^{(i)}, U^{(j)}) > \delta_0/2$, and $\|\nabla_\theta g(\theta)\|^2 \leq 2c|g(\theta) - g_i|$ $(\theta \in U_c^{(i)})$. Since J is a bounded set, $\exists N > 0$, such that $J \subset K$ (K is the closure of $U(0, N)$). Then we construct a set $M = K / \bigcup_{i=1}^{m} U_c^{(i)}$. Since $U_c^{(i)}$ is an open set and K is a closed set, we conclude M is a closed set. Since $\|\nabla_\theta g(\theta)\|$ is a continuous function, $\exists \theta_0 \in M$, $\|\nabla_{\theta_0} g(\theta_0)\| = \min_{\theta \in M} \|\nabla_\theta g(\theta)\|$. Let $r = \|\nabla_{\theta_0} g(\theta_0)\| > 0$.

Then we prove that $\forall u > 0, \theta \in K$, $\exists \delta > 0$, if $\|\nabla_\theta g(\theta)\| < \delta$, then $d(\theta, J) < u$ holds. We prove it by contradiction. Assume $\exists u_0 > 0$, $\forall \delta_1 > 0$, $\exists\ \theta_{\delta_1}$, such that $\|\nabla_{\theta_{\delta_1}} g(\theta_{\delta_1})\| < \delta_1$ and $d(\theta, J) \geq u_0$. We choose $\delta_1 = 1, 1/2, 1/3...$, and construct a sequence $\{\theta_{1/n}\}$. It is obvious that $\|\nabla_{\theta_{1/n}} g(\theta_{1/n}) \to 0\|$. Since $\{\theta_{1/n}\}$ is bounded, through the $Accumulation\ point\ theorem$, there exists a convergent subsequence $\{\theta_{1/k_n}\} \subset \{\theta_{1/n}\}$. Let $\theta^{(0)} = \lim_{n \to +\infty} \theta_{1/k_n}$. From the continuity of $d(\theta, J)$ and $\|\nabla_\theta g(\theta)\|$, we get $d(\theta^{(0)}) \geq u_0$ and $\|\nabla_{\theta^{(0)}} g(\theta^{(0)})\| = 0$. It contradicts with the definition of J. So $\forall u > 0, \theta \in K$, $\exists \delta > 0$, $\|\nabla_\theta g(\theta)\| < \delta$, such that $d(\theta, J) < u$. Furthermore, due to the continuity of $g(\theta)$, we can get $\forall \epsilon_1 > 0$, $\exists\ \delta' > 0$, if $d(\theta, J_i) < \delta'$, it holds that $|g(\theta) - g_i| < \epsilon_1$. Combine these two consequences, then we can prove $\forall \epsilon_1 > 0$, $\exists b > 0$, if $\theta \in U_c^{(i)}$ and $\|\nabla_\theta g(\theta)\| < b$, it holds that $\left|g(\theta) - g_i\right| < \epsilon_1$. Through (39)

$$\lim_{n \to +\infty} \left\|\nabla_{\theta_{k_n}} g(\theta_{k_n})\right\|^2 = 0 \ \ a.s.. \tag{41}$$

Next we aim to prove $\lim_{n \to +\infty} \|\nabla_{\theta_n} g(\theta_n)\|^2 = 0$. It is equivalent to prove that $\{\|\nabla_n g(\theta_n)\|^2\}$ has no positive accumulation points, that is to say, $\forall e_0 > 0$, there are only finite values of $\|\nabla_{\theta_n} g(\theta)\|$ larger than e_0. Obviously, we just need to prove $\forall 0 < e_0 < r$, there are only finite values of $\|\nabla_{\theta_n} g(\theta)\|$ larger than e. We prove this by contradiction. We suppose $\exists 0 < e < a$, such that the set $S = \{\|\nabla_{\theta_n} g(\theta_n)\| > e\}$ is an infinite set. Then we let $\epsilon_1 = e/8c$ and $o = min\{b, e/4\}$. Due to (41), there exists a subsequence $\{\theta_{p_n}\}$ of $\{\theta_n\}$ which satisfies $\|\nabla_{\theta_{p_n}} g(\theta_{p_n})\| < o$. We rank S as a subsequence $\{\|\nabla_{m_n} g(\theta_{m_n})\|^2\}$ of $\{\|\nabla_n g(\theta_n)\|^2\}$. Then there is an infinite subsequence $\{\|\nabla_{m_{i_n}} g(\theta_{m_{i_n}})\|^2\}$ of $\{\|\nabla_{m_n} g(\theta_{m_n})\|^2\}$ such that $\forall n \in \mathbb{N}_+$, $\exists l,\ n_{p_n} \in (m_{i_l}, m_{i_{l+1}})$. For convenience, we abbreviate $\{m_{i_n}\}$ as $\{i_n\}$. We construct another infinite sequence $\{q_n\}$ as follows

$$\begin{aligned}
&q_1 = \max\left\{n : p_1 < n < \min\{m_{i_l : m_{i_l} > p_1}\}, \left\|\nabla_{\theta_n} g(\theta_n)\right\| \leq o\right\},\\
&q_2 = \min\left\{n : n > q_1, \left\|\nabla_{\theta_n} g(\theta_n)\right\| > e\right\},\\
&q_{2n-1} = \max\left\{n : \min\{m_{i_l} : m_{i_l} > q_{2n-3}\} < n < \min\{m_l : m_l > \min\{m_{i_l} : m_{i_l} > q_{2n-3}\},\right.\\
&\left.\left\|\nabla_{\theta_n} g(\theta_n)\right\| \leq o\right\},\\
&q_{2n} = \min\left\{n : n > q_{2n-1}, \left\|\nabla_{\theta_n} g(\theta_n)\right\| > e\right\}.
\end{aligned}$$

Now we prove that $\exists N_0$, when $q_{2n} > N_0$, it holds that $e < \left\|\nabla_{\theta_{q_{2n}}} g(\theta_{q_{2n}})\right\| < r$. The left-hand side is obvious (the definition of q_{2n}). For the right-hand side, it holds that $\left\|\nabla_{\theta_{q_{2n}-1}} g(\theta_{q_{2n}-1})\right\| \leq e$. From (36) and Lemma

B.7, it follows that

$$\|\theta_{n+1}-\theta_n\|^2 = \left\|\alpha v_{n-1}+\epsilon_n\nabla_{\theta_n}g(\theta_n,\xi_n)\right\|^2 = \left\|\sum_{i=1}^{n}\alpha^{n-i}\epsilon_i\nabla_{\theta_i}g(\theta_i,\xi_i)\right\|^2$$
$$\leq \frac{1}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon_i^2\left\|\nabla_{\theta_i}g(\theta_i,\xi_i)\right\|^2 \leq \frac{M'}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon_i^2\left\|\nabla_{\theta_i}g(\theta_i)\right\|^2 + \frac{a}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon_i^2$$
$$+\frac{1}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon_i^2\Big(\left\|\nabla_{\theta_i}g(\theta_i,\xi_i)\right\|^2 - \mathbb{E}\left(\|\nabla_{\theta_i}g(\theta_i,\xi_i)\|^2|\mathcal{F}_i\right)\Big) \to 0 \ \ a.s..$$

From Assumption 3.1 2), it holds that $\left|\|\nabla_{\theta_{n+1}}g(\theta_{n+1})\|^2 - \|\nabla_{\theta_n}g(\theta_n)\|^2\right| \leq \left|\|\nabla_{\theta_{n+1}}g(\theta_{n+1})\| - \|\nabla_{\theta_n}g(\theta_n)\|\right|^2 \leq \|\nabla_{\theta_{n+1}}g(\theta_{n+1}) - \nabla_{\theta_n}g(\theta_n)\|^2 \leq c\|\theta_{n+1}-\theta_n\| \to 0 \ \ a.s.$, So $\exists N_0$, when $n > N_0$, it holds that $\left|\|\nabla_{\theta_{n+1}}g(\theta_{n+1})\|^2 - \|\nabla_{\theta_n}g(\theta_n)\|\right| < r-e$. Then we can get that when $q_{2n} > N_0+1$, it holds that $\left\|\nabla_{\theta_{q_{2n}}}g(\theta_{q_n})\right\| \leq \|\nabla_{\theta_{q_{2n}-1}}g(\theta_{q_{2n}-1})\| + \left|\|\nabla_{\theta_{q_{2n}}}g(\theta_{q_{2n}})\| - \|\nabla_{\theta_{q_{2n}-1}}g(\theta_{q_{2n}-1})\|\right| \leq e+r-e = r$. This means that $\theta_{q_{2n}} \in \bigcup_{i=1}^{m}U_\tau^{(i)}$. Then $\exists i_0$, such that $\theta_{q_{2n}} \in U_\tau^{(i_0)}$. Due to $\|\nabla_{\theta_n}g(\theta_n)\| \leq e < r \ \ (n \in [q_{2n-1}, q_{2n}))$, it holds that $\forall k \in [q_{2n-1}, q_{2n})$, $\exists i_k$, such that $\theta_n \in U_\tau^{(i_k)}$. Since $\|\theta_{n+1}-\theta_n\| \to 0 a.s.$, it holds that $i_0 = i_k \ \ (\forall\, j \in [q_{2n-1}, q_{2n}))$. For convenience, we let $i_0 = i_{q_{2n-1}} = ... = i_{q_{2n}-1} = i_{q_{2n}}$. Then we conclude that

$$\|\nabla_\theta g(\theta_n)\|^2 \leq 2c|g(\theta_n) - g_{i_{q_{2n}}}| \ \ (n \in [q_{2n-1}, q_{2n}]).$$

Due to the locally sign-preserving property, it follows that

$$\|\nabla_\theta g(\theta_n)\|^2 \leq 2c(g(\theta_n) - g_{i_{q_{2n}}}) \ \ (g(\theta_n) \geq g_{i_{q_{2n}}}) \ \ or$$
$$\|\nabla_\theta g(\theta_n)\|^2 \leq -2c(g(\theta_n) - g_{i_{q_{2n}}}) \ \ (g(\theta_n) \leq g_{i_{q_{2n}}}) \ \ (n \in [q_{2n-1}, q_{2n}]).$$

Since it is the same to study the two cases, we just show how to prove the first case. We derive that

$$e-o < \left\|\nabla_{\theta_{q_{2n}}}g(\theta_{q_{2n}})\right\|^2 - \left\|\nabla_{\theta_{q_{2n-1}}}g(\theta_{q_{2n-1}})\right\|^2 < 2c\big(g(\theta_{q_{2n}}) - g_{i_{q_{2n}}}\big) - \left\|\nabla_{\theta_{q_{2n-1}}}g(\theta_{q_{2n-1}})\right\|^2$$
$$= \left(2c\sum_{i=0}^{q_{2n}-q_{2n-1}-1} g(\theta_{q_{2n-1}+i+1}) - g(\theta_{q_{2n-1}+i})\right) + 2c\Big(g(\theta_{q_{2n-1}}) - g_{i_{q_{2n}}}\Big) - \left\|\nabla_{\theta_{q_{2n-1}}}g(\theta_{q_{2n-1}})\right\|^2.$$

From (38), we obtain

$$g(\theta_{q_{2n-1}+i+1}) - g(\theta_{q_{2n-1}+i}) \leq \frac{a'}{1-\alpha^{1+\delta}}\epsilon_{q_{2n-1}+i}^2 + Q_{q_{2n-1}+i}$$
$$< \frac{a'}{1-\alpha^{1+\delta}}\epsilon_{q_{2n-1}+i}^{1+2\eta_0} + Q_{q_{2n-1}+i}.$$

So it holds that

$$e-o < \sum_{i=0}^{q_{2n}-q_{2n-1}-1} L\epsilon_{q_{2n-1}+i}^{1+2\eta_0} + \sum_{i=0}^{q_{2n}-q_{2n-1}-1} Q_{q_{2n-1}+i}$$
$$+ 2c\Big(g(\theta_{q_{2n-1}}) - g_{i_{2n-1}}\Big) - \left\|\nabla_{\theta_{q_{2n-1}}}g(\theta_{q_{2n-1}})\right\|^2, \tag{42}$$

where

$$L = \frac{a'}{1-\alpha^{1+\delta}}.$$

Due to $\|\nabla_{\theta_{q_{2n-1}}}g(\theta_{q_{2n-1}})\|^2 < o < b$, it follows that $g(\theta_{q_{2n-1}}) - g_{i_{2n-1}} < e/8c$. Substitute it into (43), then we get

$$\sum_{i=0}^{q_{2n}-q_{2n-1}-1} \epsilon_{q_{2n-1}+i}^{1+2\eta_0} > \frac{e}{2L} - \sum_{i=0}^{q_{2n}-q_{2n-1}-1} Q_{q_{2n-1}+i}. \tag{43}$$

From Lemma B.11, it follows that $\sum_{n=1}^{+\infty} Q_n$ is finite almost surely. Thus, we attain

$$\sum_{i=0}^{q_{2n}-q_{2n-1}-1} Q_{q_{2n-1}+i} \to 0 \ a.s.$$

by the $Cauchy's\ test\ for\ convergence$. From $\epsilon_{q_{2n-1}+i}^{1+2\eta_0} \to 0 \ \ a.s.$, it follows that

$$\sum_{i=1}^{q_{2n}-q_{2n-1}-1} \epsilon_{q_{2n-1}+i}^{1+2\eta_0} > \frac{e}{2L} - \epsilon_{q_{2n-1}+i}^{1+2\eta_0} - \sum_{i=0}^{q_{2n}-q_{2n-1}-1} Q_{q_{2n-1}+i} \to \frac{e}{2L} \ \ a.s.. \tag{44}$$

Thus, it holds that

$$\sum_{n=1}^{+\infty}\left(\sum_{i=1}^{q_{2n}-q_{2n-1}-1}\epsilon_{q_{2n-1}+i}^{1+2\eta_0}\right)=+\infty\ \ a.s.. \tag{45}$$

On the other hand, it holds that $\|\nabla_{\theta_{q_{2n-1}+i}}g(\theta_{q_{2n-1}+i})\|>o\ \ (i>0)$. Together with (37), we get

$$\begin{aligned}\sum_{n=1}^{+\infty}\left(\sum_{i=1}^{q_{2n}-q_{2n-1}-1}\epsilon_{q_{2n-1}+i}^{1+2\eta_0}\right)&<\frac{1}{o}\sum_{n=1}^{+\infty}\left(\sum_{i=1}^{q_{2n}-q_{2n-1}-1}\left\|\nabla_{\theta_{q_{2n-1}+i}}g(\theta_{q_{2n-1}+i})\right\|^2\epsilon_{q_{2n-1}+i}^{1+2\eta_0}\right)\\&<\frac{1}{o}\sum_{n=3}^{n}\left\|\nabla_{\theta_n}g(\theta_n)\right\|^2\epsilon_{q_{2n-1}+i}^{1+2\eta_0}<+\infty\ \ a.s..\end{aligned} \tag{46}$$

It contradicts with (45), so we get that $\|\nabla_{\theta_n}g(\theta_n)\|\to 0\ \ a.s.$. Under Assumption 3.1 1), it is safe to conclude that there exists a connected component J^* of J such that $\lim_{n\to\infty}d(\theta_n,J^*)=0$.

C.2 Proof of Theorem 3.2

First we construct an event

$$\hat{B}_n=\{there\ exists\ at\ least\ one\ point\ n_0\in[0,n],\ satisfing\ \|\nabla_{\theta_{n_0}}g(\theta_{n_0})\|^2\le k_0\epsilon_{n_0}^2\},$$

where k_0 is an undetermined coefficient Then we define another events

$$B_i=\{\|\nabla_{\theta_i}g(\theta_i)\|^2<k_0\epsilon_i^2\}$$

,

$$B_{i,j}=\{\|\nabla_{\theta_i}g(\theta_i)\|^2\le k_0\epsilon_i^2\ and\ for\ all\ n_0\in(i,j]\ \|\nabla_{\theta_n}g(\theta_n)\|^2>k_0\epsilon_n^2\}.$$

Specially, we denote $B_{n,n}=B_n$ Absolutely, there is

$$\hat{B}_n=\bigcup_{i=1}^{n}B_{i,n}.$$

Let the characteristic function of B_n be $\hat{I}_n$, $B_{i,j}$ be $\hat{I}_{i,j}$. Through Theorem 3.1, we know θ_n must converge to some stationary point, so $g(\theta_n)\to\sum_{k=1}^m I_kg_k\ \ a.s.$ and $g_k\ \ (k=1,2,...,m)$ must be a local minimum. We denote $\hat{g}=g-\sum_{k=1}^m I_kg_k$. Then we get $\hat{g}\to 0\ a.s.$ and $\liminf_{n\to+\infty}\operatorname{sgn}(\hat{g}(\theta_n))>0\ a.s.$. There is

$$\begin{aligned}&\hat{I}_{i,n+1}\hat{g}(\theta_{n+1})-\hat{I}_{i,n}\hat{g}(\theta_n)\le\mu\alpha^{n-i}\hat{I}_{i,i+1}\|\nabla_{\theta_{i+1}}g(\theta_{i+1})\|^2-\sum_{j=i+1}^{n}\alpha^{n-j}\epsilon_j\hat{I}_{i,j}\nabla_{\theta_j}g(\theta_j)^T\nabla_{\theta_j}g(\theta_j,\xi_j)\\&+c''\sum_{j=i+1}^{n}\alpha^{n-j}\epsilon_j^2\hat{I}_{i,j}\|\nabla_{\theta_j}g(\theta_j)-\nabla_{\theta_j}g(\theta_j,\xi_j)\|^2.\end{aligned}$$

where $\mu>0$ is a constant. We define $\hat{S}=\{g(\theta)<\hat{T}\}$, $\hat{I}_{i,j}^{(a)}=\{some\ t\in(i,j)\ satisfies\ \|\nabla_{\theta_t}g(\theta_t)\|^2>q_t\epsilon_t^2\ and\ \|\nabla_{\theta_j}g(\theta_j)\|^2>a\}$. We denote $k_0=2ac''$. Then exists $a>0$, after adjusting $\{q_t\}$, making $\mathbb{E}(\hat{I}_{i,j}\|\nabla_{\theta_j}g(\theta_j)\|^2)\ge\mathbb{E}(\hat{I}_{i,j}^{(a)}\|\nabla_{\theta_j}g(\theta_j)\|^2)$ and $1/2\,\mathbb{E}(\hat{I}_{i,j})\le\mathbb{E}(\hat{I}_{i,j}^{(a)})$. Then through Assumption 3.1 (3), i.e., $E_{\xi_n}\left(\left\|\nabla_\theta g(\theta,\xi_n)\right\|^2\right)\le M'\left\|\nabla_\theta g(\theta)\right\|^2+a'$. We take the mathematical expectation on above equation to obtain

$$\begin{aligned}&\mathbb{E}\left(\hat{I}_{i,n+1}\hat{g}(\theta_{n+1})\right)-\mathbb{E}\left(\hat{I}_{i,n}\hat{g}(\theta_n)\right)\le\mu\alpha^{n-i}\,\mathbb{E}\left(\hat{I}_{i,i+1}\|\nabla_{\theta_{i+1}}g(\theta_{i+1})\|^2\right)\\&-\sum_{j=i+1}^{n}\alpha^{n-j}\epsilon_j\,\mathbb{E}\left(\hat{I}_{i,j}\|\nabla_{\theta_j}g(\theta_j)\|^2\right)\\&+c''\sum_{j=i+1}^{n}\alpha^{n-j}\epsilon_j^2\left(M'+\frac{a'}{2k_0a}\right)\mathbb{E}\left(\hat{I}_{i,j}\|\nabla_{\theta_j}g(\theta_j)\|^2\right)\\&=\mu\alpha^{n-i}\,\mathbb{E}\left(\hat{I}_{i,i+1}\|\nabla_{\theta_{i+1}}g(\theta_{i+1})\|^2\right)\\&-\sum_{j=i+1}^{n}\alpha^{n-j}\left(\epsilon_j-c''\epsilon_j^2\left(M'+\frac{a'}{2k_0a}\right)\right)\mathbb{E}\left(\hat{I}_{i,j}\|\nabla_{\theta_j}g(\theta_j)\|^2\right).\end{aligned}$$

We construct a variable $F_{i,n+1}$ as follows

$$F_{i,n+1}=\sum_{t=i+1}^{n+1}\left(\frac{1}{2-\alpha}\right)^{n+1-t}\mathbb{E}\left(\hat{I}_{i,t}\hat{g}(\theta_t)\right).$$

So we can get

$$F_{i,n+1} - F_{i,n}$$
$$\leq \sigma_0 \left(\frac{1}{2-\alpha}\right)^{n-i} \mathbb{E}\left(\hat{I}_{i,i+1}\|\nabla_{\theta_{i+1}} g(\theta_{i+1})\|^2\right) - \frac{2}{(1-\alpha)^2} \sum_{j=i+1}^{n} \left(1 - \frac{a'c''}{2k_0 a}\right)\left(\frac{1}{2-\alpha}\right)^{n-j} \epsilon_j \, \mathbb{E}\left(\hat{I}_{i,j}\|g(\theta_j)\|^2\right).$$

Through Assumption 3.3, we know when $\theta \in S = \{\|\nabla_\theta g(\theta)\|^2/(g(\theta) - g^*) < \delta_0\}$, there is $g(\theta) < \hat{T}$. We denote $k_0 = a'c''/a$. Then when $\theta_j \in \hat{S}$, there is $\mathbb{E}_{\theta_j \in \hat{S}}\left(\hat{I}_{i,j}\|\nabla_{\theta_{i,j}} g(\theta_{i,j})\|^2\right) = \mathbb{E}_{\theta_j \in \hat{S}}\left(\hat{I}^{(a)}_{i,j}\|\nabla_{\theta_{i,j}} g(\theta_{i,j})\|^2\right) > (a/2\hat{T})\,\mathbb{E}_{\theta_j \in \hat{S}}\left(\hat{I}_{i,j} g(\theta_j)\right)$, and when $\theta_j \in \mathbb{R}^N/S$, there is $\mathbb{E}_{\theta_j \in \mathbb{R}^N/S}\left(\hat{I}_{i,j}\|\nabla_{\theta_j} g(\theta_j)\|^2\right) > 2\delta_0\,\mathbb{E}_{\theta_j \in \mathbb{R}^N/S}\left(\hat{I}_{i,j} g(\theta_j)\right)$. We assign $\hat{c} = \min\{2\delta_0, a/2\hat{T}\}$. Then we can get that

$$F_{i,n} \leq K_0 F_{i,i+1} e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=i+1}^{n}\epsilon_k},$$

where K_0 is a constant. Regarding the value of $F_{i,i+1}$, it holds that

$$F_{i,i+1} \leq \mathbb{E}\left(\hat{I}_{i,i+1}|\hat{g}(\theta_{i+1})|\right) \leq \mathbb{E}(\hat{I}_i|\hat{g}(\theta_{i+1})|),$$
$$\hat{I}_i|\hat{g}(\theta_{i+1}) - \hat{g}(\theta_i)| = \hat{I}_i|\nabla_i g(\theta_i)^T v_i| = \hat{I}_i|\nabla_i g(\theta_i)^T(\alpha v_{i-1} + \epsilon_i \nabla_i g(\theta_i, \xi_i))| \leq O(\epsilon_i^2).$$

As a result,

$$F_{i,n} = O\left(\epsilon_i^2 e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=i+1}^{n}\epsilon_k}\right).$$

Then we get

$$\sum_{i=1}^{n} \mathbb{E}(\hat{I}_{i,n}\|\nabla_{\theta_n} g(\theta_n)\|^2) = O\left(\sum_{i=1}^{n} \epsilon_i^2 e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=i+1}^{n}\epsilon_k}\right).$$

For the adverse events of $\hat{B}_n$,

$$\Omega/\hat{B}_n = \{there\ is\ no\ point\ n_0 \in [0,n],\ satisfying\ \|\nabla_{\theta_{n_0}} g(\theta_{n_0})\|^2 \leq k_0 \epsilon_{n_0}^2\},$$

we use the same function, getting

$$\mathbb{E}\left(I_{\Omega/\hat{B}_n}\|\nabla_{\theta_n} g(\theta_n)\|^2\right) = O\left(e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=1}^{n}\epsilon_k}\right).$$

Finally, we get that

$$\mathbb{E}\left(\|\nabla_{\theta_n} g(\theta_n)\|^2\right) = O\left(\sum_{i=1}^{n} \epsilon_i^2 e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=i}^{n}\epsilon_k}\right) = O\left(e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=1}^{n}\epsilon_k} \sum_{i=1}^{n} \epsilon_i^2 e^{\frac{2\hat{c}}{(1-\alpha)^2}\sum_{k=1}^{i}\epsilon_k}\right).$$

C.3 Proof of Theorem 3.3

Proof. First we have

$$\begin{aligned} v_n &= \alpha v_{n-1} + \epsilon_n \nabla_{\theta_n} g(\theta_n, \xi_n) \\ \theta_{n+1} &= \theta_n - v_n. \end{aligned} \tag{47}$$

Then we calculate that

$$g(\theta_{n+1}) \leq g(\theta_n) - \nabla_{\theta_n} g(\theta_n)^T v_n + \frac{c}{2}\|v_n\|^2.$$

First, we focus on the term $\nabla_{\theta_n} g(\theta_n)^T v_n$, which satisfies

$$\begin{aligned} &- \nabla_{\theta_n} g(\theta_n)^T v_n = -\nabla_{\theta_n} g(\theta_n)^T(\alpha v_{n-1} + \epsilon_n \nabla_{\theta_n} g(\theta_n, \xi_n)) \\ &\leq -\alpha \nabla_{\theta_{n-1}} g(\theta_{n-1})^T v_{n-1} + c\alpha\|v_{n-1}\|^2 - \epsilon_n \nabla_{\theta_n} g(\theta_n)^T \nabla_{\theta_n} g(\theta_n, \xi_n)). \end{aligned} \tag{48}$$

We multiple $I_n^{(a)}$ on both sides of (48), leading to

$$\begin{aligned} &- I_n^{(a)} \nabla_{\theta_n} g(\theta_n)^T v_n \\ &\leq -\alpha I_{n-1}^{(a)} \nabla_{\theta_{n-1}} g(\theta_{n-1})^T v_{n-1} + \alpha\left(I_{n-1}^{(a)} - I_n^{(a)}\right) \nabla_{\theta_{n-1}} g(\theta_{n-1})^T v_{n-1} + c\alpha I_{n-1}^{(a)}\|v_{n-1}\|^2 \\ &- \epsilon_n I_n^{(a)} \nabla_{\theta_n} g(\theta_n)^T \nabla_{\theta_n} g(\theta_n, \xi_n)). \end{aligned} \tag{49}$$

For convenience, we denote $\hat{u}_{n-1} = \left(I_{n-1}^{(a)} - I_n^{(a)}\right) \nabla_{\theta_{n-1}} g(\theta_{n-1})^T v_{n-1}$, and then we iterate (49), getting

$$\begin{aligned} &- I_n^{(a)} \nabla_{\theta_n} g(\theta_n)^T v_n \\ &\leq -\alpha^{n-1} I_1^{(a)} \nabla_{\theta_1} g(\theta_1)^T v_1 + \sum_{t=1}^{n-1} \alpha^{n-t} c_t + \sum_{t=1}^{n-1} c\alpha^{n-t} I_t^{(a)}\|v_t\|^2 \\ &- \sum_{t=1}^{n-1} \epsilon_{t+1} \alpha^{n-t-1} I_{t+1}^{(a)} \nabla_{\theta_{t+1}} g(\theta_{t+1})^T \nabla_{\theta_{t+1}} g(\theta_{t+1}, \xi_{t+1}). \end{aligned} \tag{50}$$

We multiple $I_n^{(a)}$ on both sides of (48) and note $I_n^{(a)} \geq I_n^{(a)}$, then it holds that

$$I_{n+1}^{(a)} g(\theta_{n+1}) \leq I_n^{(a)} g(\theta_n) - I_n^{(a)} \nabla_{\theta_n} g(\theta_n)^T v_n + \frac{c}{2} I_n^{(a)} \|v_n\|^2. \tag{51}$$

Then we substitute (50) into (51) to obtain

$$\begin{aligned} I_{n+1}^{(a)} g(\theta_{n+1}) \leq & I_n^{(a)} g(\theta_n) - \alpha^{n-1} I_1^{(a)} \nabla_{\theta_1} g(\theta_1)^T v_1 + \sum_{t=1}^{n-1} \alpha^{n-t} c_t + \sum_{t=1}^{n-1} c\alpha^{n-t} I_t^{(a)} \|v_t\|^2 \\ & - \sum_{t=1}^{n-1} \epsilon_{t+1} \alpha^{n-t-1} I_{t+1}^{(a)} \nabla_{\theta_{t+1}} g(\theta_{t+1})^T \nabla_{\theta_{t+1}} g(\theta_{t+1}, \xi_{t+1}) + \frac{c}{2} I_n^{(a)} \|v_n\|^2. \end{aligned} \tag{52}$$

Then we use the $am - gm\ inequality$ on the term $\sum_{t=1}^{n-1} \alpha^{n-t} c_{t-1}$ to attain

$$\begin{aligned} \sum_{t=1}^{n-1} \alpha^{n-t} c_t = & \sum_{t=1}^{n-1} \alpha^{n-t} \big(I_t^{(a)} - I_{t+1}^{(a)}\big) \nabla_{\theta_t} g(\theta_t)^T v_t \leq \sum_{t=1}^{n-1} \frac{\alpha^{n-t}}{2} \big(I_t^{(a)} - I_{t+1}^{(a)}\big) I_t^{(a)} \big\|\nabla_{\theta_t} g(\theta_t)\big\|^2 \\ & + \sum_{t=1}^{n-1} \frac{\alpha^{n-t}}{2} \big(I_t^{(a)} - I_{t+1}^{(a)}\big) I_t^{(a)} \|v_t\|^2. \end{aligned} \tag{53}$$

We substitute (53) into (52), note $I_n^{(a)} - I_{n+1}^{(a)} = o(\epsilon_n)\ \ a.e.$ (due to $\sum_{n=1}^{+\infty} I_n^{(a)} - I_{n+1}^{(a)} < +\infty,\ \sum_{n=1}^{+\infty} \epsilon_n = +\infty$), and make the mathematical expectation, then we derive

$$\begin{aligned} \mathbb{E}(I_{n+1}^{(a)} g(\theta_{n+1})) \leq & \mathbb{E}(I_n^{(a)} g(\theta_n)) - \alpha^{n-1} \big| I_1^{(a)} \nabla_{\theta_1} g(\theta_1)^T v_1 \big| + 2 \sum_{t=1}^{n} c\alpha^{n-t}\, \mathbb{E}(I_t^{(a)} \|v_t\|^2) \\ & - \sum_{t=1}^{n} (\epsilon_t - o(\epsilon_t)) \alpha^{n-t}\, \mathbb{E}(I_t^{(a)} \|\nabla_{\theta_t} g(\theta_t)\|^2). \end{aligned} \tag{54}$$

On the other hand, we have

$$\begin{aligned} & \mathbb{E}\left(\epsilon_{n+1} I_{n+1}^{(a)} g(\theta_{n+1})\right) - \mathbb{E}\left(\epsilon_n I_n^{(a)} g(\theta_n)\right) \\ \leq & -\frac{1}{2}\left(\mathbb{E}\left(I_n^{(a)} \|v_n\|^2\right) - \alpha^2\, \mathbb{E}\left(I_{n-1}^{(a)} \|v_{n-1}\|^2\right)\right) - \frac{\epsilon_n^2}{2}\, \mathbb{E}\left(I_n^{(a)} \|\nabla_{\theta_n} g(\theta_n)\|^2\right) + \frac{c}{2} \epsilon_n\, \mathbb{E}\left(I_n^{(a)} \|v_n\|^2\right) \\ & + \frac{\epsilon_n^2}{2}\, \mathbb{E}\left(I_n^{(a)} \|\nabla_{\theta_n} g(\theta_n) - \nabla_{\theta_n} g(\theta_n, \xi_n)\|^2\right). \end{aligned} \tag{55}$$

Substitute (55) into (54) yields

$$\begin{aligned} & \mathbb{E}(I_{n+1}^{(a)} g(\theta_{n+1})) - \mathbb{E}(I_n^{(a)} g(\theta_n)) \leq \mu \alpha^{n-1} \|\nabla_{\theta_1} g(\theta_1)\|^2 - \sum_{i=1}^{n} \alpha^{n-i} (\epsilon_i - o(\epsilon_i))\, \mathbb{E}\left(I_i^{(a)} \|\nabla_{\theta_i} g(\theta_i)\|^2\right) \\ & + c'' \sum_{i=1}^{n} \alpha^{n-i} \epsilon_i^2\, \mathbb{E}\left(I_i^{(a)} \|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2\right), . \end{aligned} \tag{56}$$

where c'' and μ are two constants which do not affect results Then we note when $\|\nabla_{\theta_n} g(\theta_n)\|^2 > a$, there is

$$\mathbb{E}(\|\nabla_{\theta_n} g(\theta_n) - \nabla_{\theta_n} g(\theta_n, \xi_n)\|^2) \leq \left(M' - 1 + \frac{a'}{a}\right) \mathbb{E}\left(\|\nabla_{\theta_n} g(\theta_n)\|^2\right) \leq \hat{M}\, \mathbb{E}\left(\|\nabla_{\theta_n} g(\theta_n)\|^2\right),$$

where $\hat{M}$ is a constant which does not affect results. So we get

$$\sum_{i=1}^{n} \alpha^{n-i} \epsilon_i^2\, \mathbb{E}\left(I_i^{(a)} \|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2\right) \leq \hat{M} \sum_{i=1}^{n} \alpha^{n-i} \epsilon_i^2\, \mathbb{E}\left(I_i^{(a)} \|\nabla_{\theta_i} g(\theta_i)\|^2\right). \tag{57}$$

Substitute (57) into (56) and note $c'' \hat{M} \epsilon_n^2 = o(\epsilon_n)$, then it follows that

$$\mathbb{E}(I_{n+1}^{(a)} g(\theta_{n+1})) - \mathbb{E}(I_n^{(a)} g(\theta_n)) \leq \mu \alpha^{n-i} \|\nabla_{\theta_1} g(\theta_1)\|^2 - \sum_{i=1}^{n} \alpha^{n-i} (\epsilon_i - o(\epsilon_i))\, \mathbb{E}\left(I_i^{(a)} \|\nabla_{\theta_i} g(\theta_i)\|^2\right). \tag{58}$$

We denote

$$\hat{F}_n^{(a)} = \sum_{i=1}^{n} \left(\frac{1}{2-\alpha}\right)^{n-i} \mathbb{E}\left(I_n^{(a)} g(\theta_n)\right) \tag{59}$$

For convenience, we let

$$\hat{G}_n^{(a)} = \frac{2}{(1-\alpha)^2}\sum_{i=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-i}(\epsilon_i + o(\epsilon_i))\,\mathbb{E}\left(I_i^{(a)}\|\nabla_{\theta_n} g(\theta_n)\|^2\right).$$

Through (54), we get

$$\hat{F}_{n+1}^{(a)} - \hat{F}_n^{(a)} \le \hat{\mu}\left(\frac{1}{2-\alpha}\right)^n - \hat{G}_n^{(a)},$$

so there is

$$\begin{aligned}\hat{F}_{n+1}^{(a)} &\le \hat{F}_n^{(a)}\left(1-\frac{\hat{G}_n^{(a)}}{\hat{F}_n^{(a)}}\right) + \hat{\mu}\left(\frac{1}{2-\alpha}\right)^n \\ &\le \hat{F}_1^{(a)}\prod_{i=1}^{n}\left(1-\frac{\hat{G}_i^{(a)}}{\hat{F}_i^{(a)}}\right) + \sum_{t=1}^{n}\hat{\mu}\left(\frac{1}{2-\alpha}\right)^n\prod_{i=n-t}^{n}\left(1-\frac{\hat{G}_i^{(a)}}{\hat{F}_i^{(a)}}\right) \le \hat{q}_0\prod_{i=1}^{n}\left(1-\frac{\hat{G}_i^{(a)}}{\hat{F}_i^{(a)}}\right),\end{aligned} \tag{60}$$

where $\hat{q}_0$ is a constant. We focus on $\frac{\hat{G}_i^{(a)}}{\hat{F}_i^{(a)}}$. Using $O'stolz\ theorem$ yields

$$\begin{aligned}\liminf_{i\to+\infty}\frac{\hat{G}_i^{(a)}}{\epsilon_i\hat{F}_i^{(a)}} &= \liminf_{i\to+\infty}\frac{2}{(1-\alpha)^2}\frac{\sum_{t=1}^{i}(\frac{1}{2-\alpha})^{i-t}\,\mathbb{E}(I_t^{(a)}\|\nabla_{\theta_t} g(\theta_t)\|^2)}{\sum_{t=1}^{i}(\frac{1}{2-\alpha})^{i-t}\,\mathbb{E}(I_t^{(a)} g(\theta_t))} \\ &\ge \liminf_{i\to+\infty}\frac{2}{(1-\alpha)^2}\frac{\mathbb{E}(I_i^{(a)}\|\nabla_{\theta_i} g(\theta_i)\|^2)}{\mathbb{E}(I_t^{(a)} g(\theta_i))}.\end{aligned}$$

Through Assumption 3.3, we know when $\theta \in S = \{\|\nabla_\theta g(\theta)\|^2 < \delta g(\theta)\}$, there is $g(\theta) < \hat{T}$. We define $\hat{S} = \{g(\theta) < 2\hat{T}\}$. Then when $\theta_i \in \hat{S}$, there is $I_i^{(a)}\|\nabla_{\theta_i} g(\theta_i)\|^2 > (a/2\hat{T})I_i^{(a)} g(\theta_i)$, and when $\theta_i \in \mathbb{R}^N/S$, there is $I_i^{(a)}\|\nabla_{\theta_i} g(\theta_i)\|^2 > 2\delta_0 I_i^{(a)} g(\theta_i)$. We assign $\hat{c} = \min\{2\delta_0, a/2\hat{T}\}$, then it holds that

$$\liminf_{i\to+\infty}\frac{2}{(1-\alpha)^2}\frac{\mathbb{E}(I_i^{(a)}\|\nabla_{\theta_i} g(\theta_i)\|^2)}{\mathbb{E}(I_t^{(a)} g(\theta_i))} \ge \frac{2\hat{c}}{(1-\alpha)^2}. \tag{61}$$

Thus, it follows that

$$\mathbb{E}(I_n^{(a)} g(\theta_n)) \le \hat{F}_{n+1}^{(a)} = O\left(e^{-\frac{2\hat{c}}{(1-\alpha)^2}\sum_{i=1}^{n}\epsilon_n}\right). \tag{62}$$

□

C.4 Proof of Proposition 3.1

We use the inverse inequality

$$\|\nabla_x g(x) - \nabla_y g(y)\| \ge -c\|x-y\|,$$

to get the inverse inequation

$$\begin{aligned}I_{n+1}^{(a)} g(\theta_{n+1}) - I_n^{(a)} g(\theta_n) &= \mu_0\alpha^{n-1}\|\nabla_{\theta_1} g(\theta_1)\|^2 - \sum_{i=1}^{n}\alpha^{n-i}(\epsilon_i + o(\epsilon_i))I_i^{(a)}\|\nabla_{\theta_i} g(\theta_i)\|^2 \\ &+ \zeta_n - (I_n^{(a)} - I_{n+1}^{(a)})g(\theta_{n+1}),\end{aligned} \tag{63}$$

where $\mathbb{E}(\zeta_n) = 0$ and k'' is a constant which does not affect the result. We denote

$$\hat{F}_n^{(a)} = \sum_{i=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-i} I_i^{(a)} g(\theta_i). \tag{64}$$

For convenience, we let

$$\begin{aligned}\hat{H}_n^{(a)} &= \frac{2}{(1-\alpha)^2}\sum_{i=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-i}(\epsilon_i + o(\epsilon_i))\,\mathbb{E}\left(I_i^{(a)}\|\nabla_{\theta_i} g(\theta_i)\|^2\right), \\ \hat{G}_n^{(a)} &= \sum_{t=1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-t}(I_n^a - I_{n+1}^{(a)})g(\theta_{n+1}).\end{aligned}$$

Through (63), we get

$$\hat{F}_{n+1}^{(a)} - \hat{F}_n^{(a)} = \hat{\mu}_0\left(\frac{1}{2-\alpha}\right)^n + \hat{\zeta}_n - \hat{H}_n^{(a)} - \hat{G}_n^{(a)},$$

which leads to

$$\hat{F}_{n+1}^{(a)} = \hat{F}_n^{(a)}\left(1 - \frac{\hat{H}_n^{(a)} + \hat{G}_n^{(a)}}{\hat{F}_n^{(a)}}\right) + \hat{\zeta}_n + \hat{\mu}_0\left(\frac{1}{2-\alpha}\right)^n.$$

Then we calculate

$$\prod_{i=1}^{n}\left(1 - \frac{\hat{H}_i^{(a)} + \hat{G}_i^{(a)}}{\hat{F}_i^{(a)}}\right) \geq \exp\left\{-\hat{k}_0 \sum_{i=1}^{n} \frac{\hat{H}_i^{(a)} + \hat{G}_i^{(a)}}{\hat{F}_i^{(a)}}\right\}.$$

Then it holds that

$$\sum_{t=1}^{n} \frac{\hat{G}_t^{(a)}}{\hat{F}_t^{(a)}} = \sum_{t=1}^{n} \frac{\sum_{i=1}^{t}\left(\frac{1}{2-\alpha}\right)^{t-i}(I_i^{(a)} - I_{i+1}^{(a)})g(\theta_{i+1})}{\sum_{i=1}^{t}\left(\frac{1}{2-\alpha}\right)^{t-i} I_i^{(a)} g(\theta_i)} \leq \hat{v}_0 \sum_{t=1}^{n} \frac{I_t^{(a)}(I_t^{(a)} - I_{t+1}^{(a)})}{I_t^{(a)}} \leq I_1^{(a)}\hat{v}_0 \leq \hat{v}_0 < +\infty.$$

The second inequation above is due to the *cauchy inequality*.

$$\sum_{t=1}^{n} \frac{\sum_{i=1}^{t}\left(\frac{1}{2-\alpha}\right)^{t-i}(I_i^{(a)} - I_{i+1}^{(a)})g(\theta_{i+1})}{\sum_{i=1}^{t}\left(\frac{1}{2-\alpha}\right)^{t-i} I_i^{(a)} g(\theta_i)} \leq \sum_{t=1}^{n} \frac{\sum_{i=1}^{t}\left(\frac{1}{2-\alpha}\right)^{t-i}(I_i^{(a)} - I_{i+1}^{(a)})}{\sum_{i=1}^{t}\left(\frac{1}{2-\alpha}\right)^{t-i} I_i^{(a)}} \leq$$

As a result, we have

$$\prod_{i=1}^{n}\left(1 - \frac{\hat{H}_i^{(a)} + \hat{G}_i^{(a)}}{\hat{F}_i^{(a)}}\right) \geq e^{-\hat{v}_0} \exp\left\{-\hat{k}_0 \sum_{i=1}^{n} \frac{\hat{H}_i^{(a)}}{\hat{F}_i^{(a)}}\right\}.$$

We focus on $\frac{\hat{H}_i^{(a)}}{\hat{F}_i^{(a)}}$. Using $O'stolz\ theorem$ and Lemma B.2 yields

$$\limsup_{i\to+\infty} \frac{\hat{H}_i^{(a)}}{\epsilon_i \hat{F}_i^{(a)}} \leq \limsup_{i\to+\infty} \frac{2}{(1-\alpha)^2} \frac{\sum_{t=1}^{i} \epsilon_t (\frac{1}{2-\alpha})_t^{i-t(a)} g(\theta_t)}{\epsilon_i \sum_{t=1}^{i} (\frac{1}{2-\alpha})^{i-t} I_t^{(a)} g(\theta_t)} \leq \frac{4c}{(1-\alpha)^2}.$$

Thus, by recursively using the above inequality and making the mathematical expectation, due to $\mathbb{E}(\zeta_n) = 0$, it follows that

$$\mathbb{E}(\hat{F}_{n+1}^{(a)}) \geq \hat{l_0} e^{-\frac{4c}{(1-\alpha)^2}\sum_{i=1}^{n}\epsilon_i}.$$

Then through Theorem 3.3, we obtain that $g(\theta_n)$ is convergent a.s.. Thus, there exist $\hat{k_0} > 0,\ \hat{k_1} > 0$, such that

$$P(\tau_n^{(a)} \geq n) \geq \hat{k}'\, \mathbb{E}(I_n^{(a)} g(\theta_n)) \geq \hat{k}_0 e^{-\frac{4c}{(1-\alpha)^2}\sum_{i=1}^{n}\epsilon_n}. \tag{65}$$

C.5 Proof of Theorem 3.4.

From (31) with $\epsilon_n \equiv \epsilon$, it follows that

$$\begin{aligned}
&\mathbb{E}\left(\epsilon g(\theta_{n+1})\right) - \mathbb{E}\left(\epsilon g(\theta_n)\right) \\
\leq &-\frac{1}{2}\left(\mathbb{E}\left(\|v_n\|^2\right) - \alpha^2\,\mathbb{E}\left(\|v_{n-1}\|^2\right)\right) - \frac{\epsilon^2}{2}\,\mathbb{E}\left(\|\nabla_{\theta_n} g(\theta_n)\|^2\right) + \frac{c}{2}\epsilon\,\mathbb{E}\left(\|v_n\|^2\right) \\
&+ \frac{\epsilon^2}{2}\,\mathbb{E}\left(\|\nabla_{\theta_n} g(\theta_n) - \nabla_{\theta_n} g(\theta_n, \xi_n)\|^2\right).
\end{aligned} \tag{66}$$

Let the event $A_n^{(\varphi)} = \left\{\|\nabla_{\theta_1} g(\theta_1)\|^2 > \varphi,\ \|\nabla_{\theta_2} g(\theta_2)\|^2 > \varphi, \ldots,\ \|\nabla_{\theta_n} g(\theta_n)\|^2 > \varphi\right\}$ and the characteristic function be $I_n^{(\varphi)}$, then we attain

$$\sum_{i=1}^{n} \mathbb{E}(I_i^{(\varphi)}\|v_i\|^2) \leq \hat{c} + \hat{\sigma_0} \sum_{i=1}^{n} \epsilon^2\, \mathbb{E}(I_i^{(\varphi)}\|\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_i} g(\theta_i, \xi_i)\|^2). \tag{67}$$

Using Taylor expansion, we derive that

$$\begin{aligned}
&I_{t+1}^{(\varphi)} g(\theta_{t+1}) - I_t^{(\varphi)} g(\theta_t) \\
=&I_t^{(\varphi)} \nabla_{\theta_{\xi_t}} g(\theta_{\xi_t})^T (\theta_{t+1} - \theta_t) = -I_t^{(\varphi)} \nabla_{\theta_t} g(\theta_t)^T v_t \\
&+ I_t^{(\varphi)}\left(\nabla_{\theta_{\xi_t}} g(\theta_{\xi_t}) - \nabla_{\theta_t} g(\theta_t)\right)^T (\theta_{t+1} - \theta_t).
\end{aligned} \tag{68}$$

Considering the term $\nabla_{\theta_t} g(\theta_t)^T v_t$, it holds that

$$\begin{aligned}
&I_t^{(\varphi)} \nabla_{\theta_t} g(\theta_t)^T v_t \\
&= I_{t-1}^{(\varphi)} \alpha \nabla_{\theta_{t-1}} g(\theta_{t-1})^{\mathsf{T}} v_{t-1} + \alpha I_{t-1}^{(\varphi)} (\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_{t-1}} g(\theta_{t-1}))^T v_{t-1} \\
&+ \epsilon I_{t-1}^{(\varphi)} \nabla_{\theta_t} g(\theta_t)^{\mathsf{T}} \nabla_{\theta_t} g(\theta_t, \xi_t) + (I_t^{(\varphi)} - I_{t-1}^{(\varphi)}) \nabla_{\theta_t} g(\theta_t)^T v_t.
\end{aligned}$$

Note $(I_t^{(\varphi)} - I_{t-1}^{(\varphi)})I_{t-1}^{(\varphi)} = I_t^{(\varphi)} - I_{t-1}^{(\varphi)}$ and $I_t^{(\varphi)} - I_{t-1}^{(\varphi)} \to 0 \ \ a.s.$. Then by substituting the above equation into (11) and noting $-(\nabla_{\theta_{(i)}} g(\theta_{(i)}) - \nabla_{\theta_{i-1}} g(\theta_{i-1}))^{\mathsf{T}} v_{i-1} \leq \left\|\nabla_{\theta_{(i)}} g(\theta_{(i)}) - \nabla_{\theta_{i-1}} g(\theta_{i-1})\right\| \|v_{i-1}\| \leq c\|v_{t-1}\|^2$, we obtain

$$\begin{aligned} & I_{t+1}^{(\varphi)} g(\theta_{t+1}) - I_t^{(\varphi)} g(\theta_t) \\ \leq & -\alpha^{t-1} I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 - \sum_{i=2}^{t} \alpha^{t-i} \epsilon I_{i-1}^{(\varphi)} \nabla_{\theta_i} g(\theta_i)^{\mathsf{T}} \nabla_{\theta_i} g(\theta_i, \xi_i) + \frac{c}{2}\|v_t\|^2 \\ & - \sum_{i=1}^{t-1} \alpha^{t-i} I_{i-1}^{(\varphi)} (\nabla_{\theta_i} g(\theta_i) - \nabla_{\theta_{i-1}} g(\theta_{i-1}))^T v_{i-1}, \end{aligned} \tag{69}$$

where $c > 0$ is a constant. It follows that

$$\begin{aligned} & I_{t+1}^{(\varphi)} g(\theta_{n+1}) \\ = & I_1^{(\varphi)} g(\theta_1) + \sum_{t=1}^{n} \left(I_{t+1}^{(\varphi)} g(\theta_{t+1}) - I_t^{(\varphi)} g(\theta_t) \right) \\ = & I_1^{(\varphi)} g(\theta_1) - \frac{1-\alpha^n}{1-\alpha} I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 + \frac{1-\alpha^n}{1-\alpha} \epsilon I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1) \nabla_{\theta_1} g(\theta_1, \xi_1) \\ & - \sum_{t=1}^{n} \frac{1-\alpha^{n-t+1}}{1-\alpha} \epsilon I_{t-1}^{(\varphi)} \nabla_{\theta_t} g(\theta_t) \nabla_{\theta_t} g(\theta_t, \xi_t) + c'' \sum_{i=1}^{t} \|v_i\|^2, \end{aligned} \tag{70}$$

where c'' is a constant which does not affect the result. Substituting (67) into (70) yields

$$\begin{aligned} & I_{t+1}^{(\varphi)} g(\theta_{n+1}) \\ \leq & I_1^{(\varphi)} g(\theta_1) - \frac{1-\alpha^n}{1-\alpha} I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 + \frac{1-\alpha^n}{1-\alpha} \epsilon I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1) \nabla_{\theta_1} g(\theta_1, \xi_1) \\ & - \sum_{t=1}^{n} \frac{1-\alpha^{n-t+1}}{1-\alpha} \epsilon I_{t-1}^{(\varphi)} \nabla_{\theta_t} g(\theta_t) \nabla_{\theta_t} g(\theta_t, \xi_t) + c''\hat{c} + c''\hat{\sigma}_0 \sum_{t=1}^{n} \epsilon^2 I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2. \end{aligned} \tag{71}$$

Then we multiply $I_n^{(\varphi)}$ on the both sides of (71) and notice $I_n^{(\varphi)} \geq I_{n+1}^{(\varphi)}$ to derive that

$$\begin{aligned} & I_{n+1}^{(\varphi)} g(\theta_{n+1}) \\ \leq & I_1^{(\varphi)} g(\theta_1) - \frac{1-\alpha^n}{1-\alpha} I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 + \frac{1-\alpha^n}{1-\alpha} \epsilon I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1) \nabla_{\theta_1} g(\theta_1, \xi_1) \\ & - \sum_{t=1}^{n} \frac{1-\alpha^{n-t+1}}{1-\alpha} \epsilon I_{t-1}^{(\varphi)} \nabla_{\theta_t} g(\theta_t) \nabla_{\theta_t} g(\theta_t, \xi_t) + c''\hat{c} + c''\hat{\sigma}_0 \sum_{t=1}^{n} \epsilon^2 I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2. \end{aligned} \tag{72}$$

Then we take the expectation on (71) and obtain

$$\begin{aligned} & \mathbb{E}\left(I_{n+1}^{(\varphi)} g(\theta_{n+1}) \right) \\ \leq & \mathbb{E}\left(I_1^{(\varphi)} g(\theta_1) \right) - \frac{1-\alpha^n}{1-\alpha} \mathbb{E}\left(I_1^{(\varphi)} \nabla_{\theta_1} g(\theta_1)^{\mathsf{T}} v_1 \right) + \frac{1-\alpha^n}{1-\alpha} \epsilon \mathbb{E}\left(\nabla_{\theta_1} g(\theta_1) \nabla_{\theta_1} g(\theta_1, \xi_1) \right) \\ & - \sum_{t=1}^{n} \frac{1-\alpha^{n-t+1}}{1-\alpha} \epsilon \mathbb{E}\left(I_{t-1}^{(\varphi)} \nabla_{\theta_t} g(\theta_t) \nabla_{\theta_t} g(\theta_t, \xi_t) \right) + c''\hat{c} + c''\hat{\sigma}_0 \sum_{t=1}^{n} \epsilon^2 \mathbb{E}\left(I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2 \right). \end{aligned}$$

Under Assumption 3.1 2), it holds that

$$\begin{aligned} & c''\hat{\sigma}_0 \sum_{t=1}^{n} \epsilon^2 \mathbb{E}\left(I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2 \right) = c''\hat{\sigma}_0 \sum_{t=1}^{n} \epsilon^2 \mathbb{E}\left(\mathbb{E}\left(I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t) - \nabla_{\theta_t} g(\theta_t, \xi_t)\|^2 \middle| \mathcal{F}_t \right) \right) \\ \leq & c''\hat{\sigma}_0 \epsilon^2 \left(M_0 + \frac{a}{\varphi} \right) \sum_{t=1}^{n} \mathbb{E}\left(I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t)\|^2 \right). \end{aligned}$$

It follows that

$$\mathbb{E}\left(I_{n+1}^{(\varphi)} g(\theta_{n+1}) \right) \leq \hat{K} - \sum_{t=1}^{n} \left(\epsilon - c''\hat{\sigma}_0 \epsilon^2 \left(M_0 + \frac{a}{\varphi} \right) \right) \mathbb{E}\left(I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t)\|^2 \right).$$

Let $\mu_0 = 1/c''\hat{\sigma}_0 \left(M_0 + \frac{a}{\varphi} \right)$. Then when $\epsilon < \mu_0$, it holds that $\left(\epsilon - c''\hat{\sigma}_0 \epsilon^2 \left(M_0 + \frac{a}{\varphi} \right) \right) > 0$. We derive that

$$\sum_{t=1}^{n} \varphi \mathbb{E}\left(I_t^{(\varphi)} \right) < \sum_{t=1}^{n} \mathbb{E}\left(I_t^{(\varphi)} \|\nabla_{\theta_t} g(\theta_t)\|^2 \right) < \frac{\hat{K}}{\epsilon - c''\hat{\sigma}_0 \epsilon^2 \left(M_0 + \frac{a}{\varphi} \right)} < +\infty. \tag{73}$$

In addition, we have

$$\lim_{n\to+\infty} P(A_n^{(\varphi)}) = \lim_{n\to+\infty} \mathbb{E}\left(I_t^{(\varphi)}\right) = 0.$$

Due to $A_1^{(\varphi)} \supset A_2^{(\varphi)} \supset ... \supset A_n^{(\varphi)}$, we get

$$P(\lim_{n\to+\infty} A_n^{(\varphi)}) = \lim_{n\to+\infty} P(A_n^{(\varphi)}) = 0.$$

Through the arbitrariness of ϕ, we attain the conclusion.

C.6 Proof of Proposition 3.2.

From (73), it follows that $\mathbb{E}\left(I_n^{(\varphi)}\|\nabla_{\theta_n} g(\theta_n)\|^2\right) \to 0$. So for any $\varphi > 0$, there exists n_0, when $n > n_0$, it holds that

$$\mathbb{E}\left(I_n^{(\varphi)}\|\nabla_{\theta_n} g(\theta_n)\|^2\right) < \varphi. \tag{74}$$

Then we define another event First we construct an event

$$\hat{B}_n = \{there\ exists\ at\ least\ one\ point\ n_0 \in [0, n],\ satisfing\ g(\theta_{n_0}) - g^* \leq \varphi\},$$

where k_0 is an undetermined coefficient Then we define another events

$$B_i = \{g(\theta_i) - g^* < \varphi\}$$

,

$$B_{i,j} = \{g(\theta_i) - g^* \leq \varphi\ and\ for\ all\ n_0 \in (i, j]\ g(\theta_n) - g^* > \varphi\}.$$

Specially, we denote $B_{n,n} = B_n$. There is

$$\hat{B}_n = \bigcup_{i=1}^{n} B_{i,n}.$$

Let the characteristic function of B_n be $\hat{I}_n$, $B_{i,j}$ be $\hat{I}_{i,j}$. It is obvious that

$$\mathbb{E}\left(\hat{I}_{n,n}(g(\theta_n) - g^*)\right) \leq \varphi\,\mathbb{E}(\hat{I}_{n,n}) \leq \varphi. \tag{75}$$

Then from (71) we can get that

$$\begin{aligned}
&\hat{I}_{i,n+1}\hat{g}(\theta_{n+1}) - \hat{I}_{i,n}\hat{g}(\theta_n) \leq \mu\alpha^{n-i}\hat{I}_{i,i+1}\|\nabla_{\theta_{i+1}} g(\theta_{i+1})\|^2 - \sum_{j=i+1}^{n} \alpha^{n-j}\epsilon\hat{I}_{i,j}\nabla_{\theta_j} g(\theta_j)^T\nabla_{\theta_j} g(\theta_j, \xi_j) \\
&+ c''\sum_{j=i+1}^{n} \alpha^{n-j}\epsilon^2\hat{I}_{i,j}\|\nabla_{\theta_j} g(\theta_j) - \nabla_{\theta_j} g(\theta_j, \xi_j)\|^2,
\end{aligned} \tag{76}$$

where $\mu > 0$ is a constant. Then we denote $\hat{g} = g - \varphi$. We take the mathematical expectation on (76) to obtain

$$\begin{aligned}
&\mathbb{E}\left(\hat{I}_{i,n+1}\hat{g}(\theta_{n+1})\right) - \mathbb{E}\left(\hat{I}_{i,n}\hat{g}(\theta_n)\right) \leq \mu\alpha^{n-i}\,\mathbb{E}\left(\|\nabla_{\theta_{i+1}} g(\theta_{i+1})\|^2\right) \\
&- \sum_{j=i+1}^{n} \alpha^{n-j}\epsilon\,\mathbb{E}\left(\hat{I}_{i,j}\|\nabla_{\theta_j} g(\theta_j)\|^2\right) \\
&+ c''\sum_{j=i+1}^{n} \alpha^{n-j}\epsilon^2\left(M' + \frac{a}{\varphi}\right)\mathbb{E}\left(\hat{I}_{i,j}\|\nabla_{\theta_j} g(\theta_j)\|^2\right) \\
&= \mu\alpha^{n-i}\,\mathbb{E}\left(\hat{I}_{i,i+1}\|\nabla_{\theta_{i+1}} g(\theta_{i+1})\|^2\right) \\
&- \left(\epsilon - c''\epsilon^2\left(M' + \frac{a}{\varphi}\right)\right)\sum_{j=i+1}^{n} \alpha^{n-j}\,\mathbb{E}\left(\hat{I}_{i,j}\|\nabla_{\theta_j} g(\theta_j)\|^2\right).
\end{aligned}$$

We construct a variable $F_{i,n+1}$ as follows

$$F_{i,n+1} = \sum_{t=i+1}^{n+1}\left(\frac{1}{2-\alpha}\right)^{n+1-t}\mathbb{E}\left(\hat{I}_{i,t}\hat{g}(\theta_t)\right).$$

So we can get

$$F_{i,n+1} - F_{i,n} \leq \sigma_0\left(\frac{1}{2-\alpha}\right)^{n-i}\mathbb{E}\left(\hat{I}_{i,i+1}\|\nabla_{\theta_{i+1}} g(\theta_{i+1})\|^2\right) - \digamma_0\sum_{t=i+1}^{n}\left(\frac{1}{2-\alpha}\right)^{n-t}\epsilon\,\mathbb{E}\left(\hat{I}_{i,t}\|\hat{g}(\theta_t)\|^2\right),$$

where σ_0 and $\digamma_0$ are two constants not dependent on ϵ or φ. We let

$$Q_{i,n} := \digamma_0 \sum_{t=i+1}^{n} \big(\frac{1}{2-\alpha}\big)^{n-t} \epsilon\, \mathbb{E}\left(\hat{I}_{i,t} \|\hat{g}(\theta_t)\|^2\right).$$

Then we get

$$F_{i,n+1} \leq F_{i,n}\left(1 - \frac{Q_{i,n}}{F_{i,n}}\right) + \sigma_0 \left(\frac{1}{2-\alpha}\right)^{n-i} \mathbb{E}\left(\hat{I}_{i,i+1} \|\nabla_{\theta_{i+1}} g(\theta_{i+1})\|^2\right).$$

Then we get $Q_{i,n}/F_{i,n} > \epsilon \digamma_0/\varphi$. So we can attain

$$F_{i,n+1} \leq F_{i,i+1}(1-\epsilon\varphi\digamma_0)^{n-i-1} + y_0^{n-i-1+}\sigma_0\, \mathbb{E}\left(\hat{I}_{i,i+1} \|\nabla_{\theta_{i+1}} g(\theta_{i+1})\|^2\right), \tag{77}$$

where $y_0 = \max\{1-\epsilon\varphi\digamma_0, 1/2-\alpha\}$. Then we first prove $\mathbb{E}(g(\theta_n))$ is uniformly bounded. We have

$$\begin{aligned}
&\mathbb{E}\left(\hat{g}(\theta_{i+1})\right) - \mathbb{E}\left(\hat{g}(\theta_i)\right) = -\mathbb{E}\left(\nabla_{\theta_i} g(\theta_i)^T v_i\right) = -\mathbb{E}\left(\nabla_{\theta_i} g(\theta_i)^T (\alpha v_{i-1} + \epsilon \nabla_{\theta_i} g(\theta_i, \xi_i))\right) \\
&\leq \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i)\| \|v_{i-1}\|\right) + \epsilon\, \mathbb{E}\left(\|\nabla_{\theta_i} g(\theta_i))\|^2\right) \leq \sqrt{\varphi}\|v_{i-1}\| + \epsilon\varphi < \hat{m},
\end{aligned}$$

where $\hat{m}$ is a constant. Then we get

$$\mathbb{E}\left(g(\theta_n)\right) \leq \varphi + \mathbb{E}\left(\hat{g}(\theta_n)\right) \leq \varphi + \hat{m}\sum_{i=0}^{+\infty}(1-\epsilon\varphi\digamma_0)^i + \hat{m}\sum_{i=0}^{+\infty} y_0^i < \frac{\hat{K}}{\epsilon\varphi} < +\infty. \tag{78}$$

Now we have proved $\mathbb{E}\left(g(\theta_n)\right)$ uniformly bounded. Then we aim to prove $\mathbb{E}\left(g(\theta_n)\right)$ can become arbitrarily small if n sufficient large and ϵ sufficient small. First We have

$$\begin{aligned}
&\mathbb{E}\|v_n\|^2 = \mathbb{E}\|\theta_{n+1} - \theta_n\|^2 = \left\|\alpha v_{n-1} + \epsilon \nabla_{\theta_n} g(\theta_n, \xi_n)\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{n} \alpha^{n-i}\epsilon \nabla_{\theta_i} g(\theta_i, \xi_i)\right\|^2 \\
&\leq \frac{1}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon^2\, \mathbb{E}\left\|\nabla_{\theta_i} g(\theta_i, \xi_i)\right\|^2 \leq \frac{M'}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon^2\, \mathbb{E}\left\|\nabla_{\theta_i} g(\theta_i)\right\|^2 + \frac{a}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon^2 \leq \hat{p}_0\frac{\epsilon}{\varphi},
\end{aligned}$$

where $\hat{p}_0$ is a constant. Then we get that

$$\begin{aligned}
&\mathbb{E}\left(\hat{g}(\theta_{i+1})\right) - \mathbb{E}\left(\hat{g}(\theta_i)\right) = -\mathbb{E}\left(\nabla_{\theta_i} g(\theta_i)^T v_i\right) = -\mathbb{E}\left(\nabla_{\theta_i} g(\theta_i)^T (\alpha v_{i-1} + \epsilon \nabla_{\theta_i} g(\theta_i, \xi_i))\right) \\
&\leq \sum_{i=1}^{n}\alpha^{n-i}\|v_i\|^2 \leq \hat{p}_1\frac{\epsilon}{\varphi}.
\end{aligned}$$

Since $\mathbb{E}\left(\hat{g}(\theta_i)\right) \leq 0$, we get $\mathbb{E}\left(\hat{g}(\theta_{i+1})\right) \leq \hat{p}_1\epsilon/\varphi$. Finally we substituted it into (77), getting

$$\mathbb{E}\left\|\nabla_{\theta_n} g(\theta_n)\right\|^2 \leq \mathbb{E}\left(g(\theta_n)\right) \leq \varphi + \mathbb{E}\left(\hat{g}(\theta_n)\right) \leq \varphi + \hat{p}_1\frac{\epsilon}{\varphi}\sum_{i=0}^{+\infty}(1-\epsilon\varphi\digamma_0)^i + \hat{p}_1\frac{\epsilon}{\varphi}\sum_{i=0}^{+\infty} y_0^i < \frac{\hat{p}_2}{\varphi},$$

where $\hat{p}_2$ is a constant. Then we can get

$$\begin{aligned}
&\mathbb{E}\|v_n\|^2 = \mathbb{E}\|\theta_{n+1} - \theta_n\|^2 = \left\|\alpha v_{n-1} + \epsilon \nabla_{\theta_n} g(\theta_n, \xi_n)\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{n} \alpha^{n-i}\epsilon \nabla_{\theta_i} g(\theta_i, \xi_i)\right\|^2 \\
&\leq \frac{1}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon^2\, \mathbb{E}\left\|\nabla_{\theta_i} g(\theta_i, \xi_i)\right\|^2 \leq \frac{M'}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon^2\, \mathbb{E}\left\|\nabla_{\theta_i} g(\theta_i)\right\|^2 + \frac{a}{1-\alpha}\sum_{i=1}^{n}\alpha^{n-i}\epsilon^2 \leq \hat{p}_0'\frac{\epsilon^2}{\varphi},
\end{aligned}$$

where $\hat{p}_0'$ is a constant. Then we get that

$$\begin{aligned}
&\mathbb{E}\left(\hat{g}(\theta_{i+1})\right) - \mathbb{E}\left(\hat{g}(\theta_i)\right) = -\mathbb{E}\left(\nabla_{\theta_i} g(\theta_i)^T v_i\right) = -\mathbb{E}\left(\nabla_{\theta_i} g(\theta_i)^T (\alpha v_{i-1} + \epsilon \nabla_{\theta_i} g(\theta_i, \xi_i))\right) \\
&\leq \sum_{i=1}^{n}\alpha^{n-i}\|v_i\|^2 \leq \hat{p}_1'\frac{\epsilon^2}{\varphi},
\end{aligned}$$

where $\hat{p}_1'$ is a constant. Then we can get

$$\mathbb{E}\left\|\nabla_{\theta_n} g(\theta_n)\right\|^2 \leq \mathbb{E}\left(g(\theta_n)\right) \leq \varphi + \mathbb{E}\left(\hat{g}(\theta_n)\right) \leq \varphi + \hat{p}_1'\frac{\epsilon^2}{\varphi}\sum_{i=0}^{+\infty}(1-\epsilon\varphi\digamma_0)^i + \hat{p}_1\frac{\epsilon}{\varphi}\sum_{i=0}^{+\infty} y_0^i < \frac{\hat{p}_2'\epsilon}{\varphi},$$

where $\hat{p}_2'$ is a constant. Now we can find that provided ϵ small enough, we can get for any small bound φ_0, $\mathbb{E}\left\|\nabla_{\theta_n} g(\theta_n)\right\|^2 < \varphi_0$ (when n is large enough).