An Improved Analysis of Stochastic Gradient Descent with Momentum

Yanli Liu* **Yuan Gao†** **Wotao Yin***
*Department of Mathematics, University of California, Los Angeles
†Department of IEOR, Columbia University
{yanli, wotaoyin}@math.ucla.edu, {gao.yuan}@columbia.edu

Abstract

SGD with momentum (SGDM) has been widely applied in many machine learning tasks, and it is often applied with dynamic stepsizes and momentum weights tuned in a stagewise manner. Despite of its empirical advantage over SGD, the role of momentum is still unclear in general since previous analyses on SGDM either provide worse convergence bounds than those of SGD, or assume Lipschitz or quadratic objectives, which fail to hold in practice. Furthermore, the role of dynamic parameters has not been addressed. In this work, we show that SGDM converges as fast as SGD for smooth objectives under both strongly convex and nonconvex settings. We also prove that multistage strategy is beneficial for SGDM compared to using fixed parameters. Finally, we verify these theoretical claims by numerical experiments.

1 Introduction

Stochastic gradient methods have been a widespread practice in machine learning. They aim to minimize the following empirical risk:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} \ell(x, q_i), \tag{1}$$

where ℓ is a loss function and $\{q_i\}_{i=1}^n$ denotes the training data, x denotes the trainable parameters of the machine learning model, e.g., the weight matrices in a neural network.

In general, stochastic gradient methods can be written as

$$\begin{aligned} m^k &= \beta m^{k-1} + (1-\beta)\tilde{g}^k, \\ x^{k+1} &= x^k - \alpha m^k. \end{aligned} \tag{2}$$

where $\alpha > 0$ is a stepsize, $\beta \in [0, 1)$ is called momentum weight, and $m^0 = 0$. The classical Stochastic Gradient Descent(SGD) method [21] uses $\beta = 0$ and $m^k = \tilde{g}^k$, where $\tilde{g}^k$ is a stochastic gradient of $f(x)$ at x^k. To boost the practical performance, one often applies a momentum weight of $\beta > 0$. and the resulting algorithm is often called SGD with momentum (SGDM). SGDM is very popular for training neural networks with remarkable empirical successes, and has been implemented as the default SGD optimizer in Pytorch [19] and Tensorflow [1][1].

The idea behind SGDM originates from Polyak's heavy-ball method [20] for deterministic optimization. For strongly convex and smooth objectives, heavy-ball method enjoys an accelerated linear

[1]Their implementation of SGDM does not have the $(1-\beta)$ before $\tilde{g}^k$, which gives $m^k = \sum_{i=1}^k \beta^{k-i}\tilde{g}^i$, while $m^k = (1-\beta)\sum_{i=1}^k \beta^{k-i}\tilde{g}^i$ for (2). Therefore, they only differ by a constant scaling.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

convergence rate over gradient descent [7]. However, the theoretical understanding of its stochastic counterpart is far from being complete.

In the case of fixed stepsize and momentum weight, most of the current results only apply to restrictive settings. In [15, 16] and [12], the behavior of SGDM on least square regression is analyzed and linear convergence is established. [9] analyzes the local convergence rate of SGDM for strongly convex and smooth functions, where the initial point x^0 is assumed to be close enough to the minimizer x^*. [25] provides global convergence of SGDM, but only for objectives with *uniformly bounded gradients*, thus excluding many machine learning models such as Ridge regression. Very recently, [26] presents a convergence bound of $\mathcal{O}(\frac{1}{k\alpha} + \frac{\alpha}{1-\beta})$ for general smooth nonconvex objectives[3]. When $\beta = 0$, this recovers the classical convergence bound of $\mathcal{O}(\frac{1}{k\alpha} + \alpha)$ of SGD [4]. However, the size of stationary distribution $\mathcal{O}(\frac{\alpha}{1-\beta})$ is $\frac{1}{1-\beta}$ times larger than that of SGD. This factor is not negligible, especially when large β values such as 0.99 and 0.995 is applied [24]. Therefore, their result does not explain the competitiveness of SGDM compared to SGD. Concurrent to this work, [22] shows that SGDM converges as fast as SGD under convexity and strong convexity, and that it is asymptotically faster than SGD for overparameterized models. Remarkably, their analysis considers a different stepsize and momentum weight schedule from this work, and applies to arbitrary sampling without assuming the bounded variance of the gradient noise.

In deep learning, SGDM is often applied with various parameter tuning rules to achieve efficient training. One of the most widely adopted rules is called "constant and drop", where a constant stepsize is applied for a long period and is dropped by some constant factor to allow for refined training, while the momentum weight is either kept unchanged (usually 0.9) or gradually increasing. We call this strategy Multistage SGDM and summarize it in Algorithm 1. Practically, (multistage) SGDM was successfully applied to training large-scale neural networks [13, 11], and it was found that appropriate parameter tuning leads to superior performance [24]. Since then, (multistage) SGDM has become increasingly popular [23].

At each stage, Multistage SGDM (Algorithm 1) requires three parameters: stepsize, momentum weight, and stage length. In [8] and [10], doubling argument based rules are analyzed for SGD on strongly convex objectives, where the stage length is doubled whenever the stepsize is halved. Recently, certain stepsize schedules are shown to yield faster convergence for SGD on nonconvex objectives satisfying growth conditions [27, 5], and a nearly optimal stepsize schedule is provided for SGD on least square regression [6]. These results consider only the momentum-free case. Another recent work focuses on the asymptotic convergence of SGDM (i.e., without convergence rate) [9], which requires the momentum weights to approach either 0 or 1, and therefore contradicts the common practice in neural network training. In summary, the convergence rate of Multistage SGDM (Algorithm 1) has not been established except for the momentum-free case, and the role of parameters in different stages is unclear.

Algorithm 1 Multistage SGDM

Input: problem data $f(x)$ as in (1), number of stages n, momentum weights $\{\beta_i\}_{i=1}^n \subseteq [0, 1)$, step sizes $\{\alpha_i\}_{i=1}^n$, and stage lengths $\{T_i\}_{i=1}^n$ at n stages, initialization $x^1 \in \mathbb{R}^{\mathrm{d}}$ and $m^0 = 0$, iteration counter $k = 1$.

1: **for** $i = 1, 2, ..., n$ **do**
2: $\quad \alpha \leftarrow \alpha_i, \beta \leftarrow \beta_i$;
3: $\quad$ **for** $j = 1, 2, ..., T_i$ **do**
4: $\quad\quad$ Sample a minibatch ζ^k uniformly from the training data;
5: $\quad\quad$ $\tilde{g}^k \leftarrow \nabla_x l(x^k, \zeta^k)$;
6: $\quad\quad$ $m^k \leftarrow \beta m^{k-1} + (1-\beta)\tilde{g}^k$;
7: $\quad\quad$ $x^{k+1} \leftarrow x^k - \alpha m^k$;
8: $\quad\quad$ $k \leftarrow k + 1$;
9: $\quad$ **end for**
10: **end for**
11: **return** $\tilde{x}$, which is generated by first choosing a stage $l \in \{1, 2, ...n\}$ uniformly at random, and then choosing $\tilde{x} \in \{x^{T_1+\ldots+T_{l-1}+1}, x^{T_1+\ldots+T_{l-1}+2}, ..., x^{T_1+\ldots+T_l}\}$ uniformly at random;

[3]Here k is the number of iterations. Note that in [26], a different but equivalent formulation of SGDM is analyzed; their stepsize γ is effectively $\frac{\alpha}{1-\beta}$ in our setting.

1.1 Our contributions

In this work, we provide new convergence analysis for SGDM and Multistage SGDM that resolve the aforementioned issues. A comparison of our results with prior work can be found in Table 1.

1. We show that for both strongly convex and nonconvex objectives, SGDM (2) enjoys the same convergence bound as SGD. This helps explain the empirical observations that SGDM is at least as fast as SGD [23]. Our analysis relies on a new observation that, the update direction m^k of SGDM (2) has a controllable deviation from the current full gradient $\nabla f(x^k)$, and enjoys a smaller variance. Inspired by this, we construct a new Lyapunov function that properly handles this deviation and exploits an auxiliary sequence to take advantage of the reduced variance.

   Compared to aforementioned previous work, our analysis applies to not only least squares, does not assume uniformly bounded gradient, and improves the convergence bound.

2. For the more popular SGDM in the multistage setting (Algorithm 1), we establish its convergence and demonstrate that the multistage strategy are faster at initial stages. Specifically, we allow larger stepsizes in the first few stages to boost initial performance, and smaller stepsizes in the final stages decrease the size of stationary distribution. Theoretically, we properly redefine the aforementioned auxiliary sequence and Lyapunov function to incorporate the stagewise parameters.

   To the best of our knowledge, this is the first convergence guarantee for SGDM in the multistage setting.

Method	Additional Assumptions	Convergence Bound
SGDM [25]	Bounded gradient	$\mathbb{E}[\\|\nabla f(x_{\text{out}})\\|^2] = \mathcal{O}\left(\frac{1}{k\alpha} + \frac{\alpha\sigma^2}{1-\beta}\right)$
SGDM [26]	-	$\mathbb{E}[\\|\nabla f(x_{\text{out}})\\|^2] = \mathcal{O}\left(\frac{1}{k\alpha} + \frac{\alpha\sigma^2}{1-\beta}\right)$
SGDM (*)	-	$\mathbb{E}[\\|\nabla f(x_{\text{out}})\\|^2] = \mathcal{O}\left(\frac{1}{k\alpha} + \alpha\sigma^2\right)$
SGDM (*)	Strong convexity	$\mathbb{E}[f(x^k) - f^*] = \mathcal{O}\left((1-\alpha\mu)^k + \alpha\sigma^2\right)$
Multistage SGDM(*)	-	$\mathbb{E}[\\|\nabla f(x_{\text{out}})\\|^2] = \mathcal{O}\left(\frac{1}{nA_2} + \frac{1}{n}\sum_{l=1}^{n}\alpha_l\sigma^2\right)$

Table 1: Comparison of our results (*) with prior work under Assumption 1 and additional assumptions. “Bounded gradient" stands for the bounded gradient assumption $\|\nabla f(x)\| \leq G$ for some $G > 0$ and all $x \in \mathbb{R}^{\mathrm{d}}$. This work removes this assumption and improves convergence bounds. Strongly convex setting and multistage setting are also analyzed. We omit the results of [8] and [10] as their analysis only applies to SGD (momentum-free case).

1.2 Other related work

Nesterov's momentum achieves optimal convergence rate in deterministic optimization [18], and has also been combined with SGD for neural network training [24]. Recently, its multistage version has been analyzed for convex or strongly convex objectives [3, 14]. Other forms of momentum for stochastic optimization include PID Control-based methods [2], Accelerated SGD [12], and Quasi-Hyperbolic Momentum [17]. In this work, we restrict ourselves to heavy-ball momentum, which is arguably the most popular form of momentum in current deep learning practice.

2 Notation and Preliminaries

Throughout this paper, we use $\|\cdot\|$ for vector ℓ_2-norm, $\langle\cdot,\cdot\rangle$ stands for dot product. Let g^k denote the full gradient of f at x^k, i.e., $g^k := \nabla f(x^k)$, and $f^* := \min_{x\in\mathbb{R}^{\mathrm{d}}} f(x)$.

Definition 1. *We say that $f : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is $L-$smooth with $L \geq 0$, if it is differentiable and satisfies*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^{\mathrm{d}}\,.$$

We say that $f : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is $\mu-$strongly convex with $\mu \geq 0$, if it satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^{\mathrm{d}}.$$

The following assumption is effective throughout, which is standard in stochastic optimization.

Assumption 1.
1. ***Smoothness:*** *The objective $f(x)$ in (1) is $L-$smooth.*
2. ***Unbiasedness:*** *At each iteration k, $\tilde{g}^k$ satisfies $\mathbb{E}_{\zeta^k}[\tilde{g}^k] = g^k$.*
3. ***Independent samples:*** *the random samples $\{\zeta_k\}_{k=1}^{\infty}$ are independent.*
4. ***Bounded variance:*** *the variance of $\tilde{g}^k$ with respect to ζ^k satisfies $\mathrm{Var}_{\zeta^k}(\tilde{g}^k) = \mathbb{E}_{\zeta^k}[\|\tilde{g}^k - g^k\|^2] \leq \sigma^2$ for some $\sigma^2 > 0$.*

Unless otherwise noted, all the proof in the paper are deferred to the appendix.

3 Key Ingredients of Convergence Theory

In this section, we present some key insights for the analysis of stochastic momentum methods. For simplicity, we first focus on the case of fixed stepsize and momentum weight, and make proper generalizations for Multistage SGDM in App. C.

3.1 A key observation on momentum

In this section, we make the following observation on the role of momentum:

> *With a momentum weight $\beta \in [0, 1)$, the update vector m^k enjoys a reduced "variance" of $(1-\beta)\sigma^2$, while having a controllable deviation from the full gradient g^k in expectation.*

First, without loss of generality, we can take $m^0 = 0$, and express m^k as

$$m^k = (1-\beta)\sum_{i=1}^{k} \beta^{k-i}\tilde{g}^i. \tag{3}$$

m^k is a moving average of the past stochastic gradients, with smaller weights for older ones[1].

we have the following result regarding the "variance" of m^k, which is measured between m^k and its deterministic version $(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i$.

Lemma 1. *Under Assumption 1, the update vector m^k in SGDM (2) satisfies*

$$\mathbb{E}\left[\left\|m^k - (1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i\right\|^2\right] \leq \frac{1-\beta}{1+\beta}(1-\beta^{2k})\sigma^2.$$

Lemma 1 follows directly from the property of the moving average.

On the other hand, $(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i$ is a moving average of all past gradients, which is in contrast to SGD. It seems unclear how far is $(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i$ from the ideal descent direction g^k, which could be unbounded unless stronger assumptions are imposed. Previous analysis such as [25] and [9] make the blanket assumption of bounded ∇f to circumvent this difficulty.

In this work, we provide a different perspective to resolve this issue.

Lemma 2. *Under Assumption 1, we have*

$$\mathbb{E}\left[\left\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right] \leq \sum_{i=1}^{k-1} a_{k,i}\,\mathbb{E}[\|x^{i+1} - x^i\|^2],$$

where

$$a_{k,i} = \frac{L^2\beta^{k-i}}{1-\beta^k}\left(k - i + \frac{\beta}{1-\beta}\right). \tag{4}$$

[1]Note the sum of weights $(1-\beta)\sum_{i=1}^{k}\beta^{k-i} = 1 - \beta^k \to 1$ as $k \to \infty$.

From Lemma 2, we know the deviation of $\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i$ from g^k is controllable sum of past successive iterate differences, in the sense that the coefficients $a_{k,i}$ decays linearly for older ones. This inspires the construction of a new Lyapunov function to handle the deviation brought by the momentum, as we shall see next.

3.2 A new Lyapunov function

Let us construct the following Lyapunov function for SGDM:

$$L^k = \big(f(z^k) - f^\star\big) + \sum_{i=1}^{k-1} c_i \|x^{k+1-i} - x^{k-i}\|^2. \tag{5}$$

In the Lyapunov function (5), $\{c_i\}_{i=1}^{\infty}$ are positive constants to be specified later corresponding to the deviation described in Lemma 2. Since the coefficients in (4) converges linearly to 0 as $k \to \infty$, we can choose $\{c_i\}_{i=1}^{\infty}$ in a diminishing fashion, such that this deviation can be controlled, and L^k defined in (5) is indeed a Lyapunov function under strongly convex and nonconvex settings (see Propositions 1 and 2).

In (5), z^k is an auxiliary sequence defined as

$$z^k = \begin{cases} x^k & k = 1, \\ \frac{1}{1-\beta}x^k - \frac{\beta}{1-\beta}x^{k-1} & k \geq 2. \end{cases} \tag{6}$$

This auxiliary sequence first appeared in the analysis of deterministic heavy ball methods in [7], and later applied in the analysis of SGDM [26, 25]. It enjoys the following property.

Lemma 3. *z^k defined in* (6) *satisfies*

$$z^{k+1} - z^k = -\alpha \tilde{g}^k.$$

Lemma 3 indicates that it is more convenient to analyze z^k than x^k since z^k behaves more like a SGD iterate, although the stochastic gradient $\tilde{g}^k$ is not taken at z^k.

Since the coefficients of the deviation in Lemma 2 converges linearly to 0 as $k \to \infty$, we can choose $\{c_i\}_{i=1}^{\infty}$ in a diminishing fashion, such that this deviation can be controlled. Remarkably, we shall see in Sec. 4 that with $c_1 = \mathcal{O}\left(\frac{L}{1-\beta}\right)$, L^k defined in (5) is indeed a Lyapunov function under strongly convex and nonconvex settings, and that SGDM converges as fast as SGD.

Now, let us turn to the Multistage SGDM (Algorithm 1), which has been very successful in neural network training. However, its convergence still remains unclear except for the momentum-free case. To establish convergence, we require the parameters of Multistage SGDM to satisfy

$$\begin{aligned} \frac{\alpha_i \beta_i}{1-\beta_i} &\equiv A_1, \text{ for } i = 1, 2, ...n. \\ \alpha_i T_i &\equiv A_2, \text{ for } i = 1, 2, ...n. \\ 0 \leq \beta_1 \leq \beta_2 &\leq ... \leq \beta_n < 1. \end{aligned} \tag{7}$$

where α_i, β_i, and T_i are the stepsize, momentum weight, and stage length of ith stage, respectively, and A_1, A_2 are properly chosen constants. In principle, one applies larger stepsizes α_i at the initial stages, which will accelerate initial convergence, and smaller stepsizes for the final stages, which will shrink the size of final stationary distribution. As a result, (7) stipulates that less iterations are required for stages with large stepsizes and more iterations for stages with small stepsizes. Finally, (7) requires the momentum weights to be monotonically increasing, which is consistent with what's done in practice [24]. often, using constant momentum weight also works.

Under the parameter choices in (7), let us define the auxiliary sequence z^k by

$$z^k = x^k - A_1 m^{k-1}. \tag{8}$$

This $\{z^k\}_{k=1}^{\infty}$ sequence reduces to (6) when a constant stepsize and momentum weight are applied. Furthermore, the observations made in Lemmas 1, 2, and 3 can also be generalized (see Lemmas 4, 5, 6, and 7 in App. C). In Sec. 5. we shall see that with (7) and appropriately chosen $\{c_i\}_{i=1}^{\infty}$, L^k in (5) also defines a Lyapunov function in the multistage setting, which in turn leads to the convergence of Multistage SGDM.

4 Convergence of SGDM

In this section, we proceed to establish the convergence of SGDM described in (2). First, by following the idea presented in Sec. 3, we can show that L^k defined in (5) is a Lyapunov function.

Proposition 1. *Let Assumption 1 hold. In* (2)*, let* $\alpha \le \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}}$*. Let* $\{c_i\}_{i=1}^{\infty}$ *in* (5) *be defined by*

$$c_1 = \frac{\frac{\beta+\beta^2}{(1-\beta)^3}L^3\alpha^2}{1-4\alpha^2\frac{\beta+\beta^2}{(1-\beta)^2}L^2}, \qquad c_{i+1} = c_i - \left(4c_1\alpha^2 + \frac{L\alpha^2}{1-\beta}\right)\beta^i(i+\frac{\beta}{1-\beta})L^2 \quad \text{for all } i \ge 1.$$

Then, $c_i > 0$ *for all* $i \ge 1$*, and*

$$\begin{aligned} \mathbb{E}[L^{k+1} - L^k] \le & \left(-\alpha + \frac{3-\beta+\beta^2}{2(1-\beta)}L\alpha^2 + 4c_1\alpha^2\right)\mathbb{E}[\|g^k\|^2] \\ & + \left(\frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2\right). \end{aligned} \tag{9}$$

By telescoping (9), we obtain the stationary convergence of SGDM under nonconvex settings.

Theorem 1. *Let Assumption 1 hold. In* (2)*, let* $\alpha \le \alpha \le \min\{\frac{1-\beta}{L(4-\beta+\beta^2)}, \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}}\}$*. Then,*

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\|g^i\|^2] \le \frac{2\left(f(x^1)-f^*\right)}{k\alpha} + \left(\frac{\beta+3\beta^2}{2(1+\beta)}+1\right)L\alpha\sigma^2 = \mathcal{O}\left(\frac{f(x^1)-f^*}{k\alpha} + L\alpha\sigma^2\right). \tag{10}$$

Now let us turn to the strongly convex setting, for which we have

Proposition 2. *Let Assumption 1 hold. Assume in addition that* f *is* $\mu-$*strongly convex. In* (2)*, let* $\alpha \le \min\{\frac{1-\beta}{5L}, \frac{1-\beta}{L\left(3-\beta+2\beta^2+\frac{48\sqrt{\beta}}{25}\frac{2L+18\mu}{L}\right)}\}$*. Then, there exists positive constants* c_i *for* (5) *such that for all* $k \ge k_0 := \lfloor\frac{\log 0.5}{\log\beta}\rfloor$*, we have*

$$\mathbb{E}[L^{k+1}-L^k] \le -\frac{\alpha\mu}{1+\frac{8\mu}{L}}\mathbb{E}[L^k] + (\frac{1+\beta+\beta^2}{2(1+\beta)}L + \frac{1-\beta}{1+\beta}2c_1)\alpha^2\sigma^2 + \frac{\beta^2+\frac{L\alpha}{2}\frac{\beta^2}{1-\beta}}{(1+\frac{8\mu}{L})(1+\beta)}2\mu\alpha^2\sigma^2.$$

The choices of $\{c_i\}_{i=1}^{\infty}$ is similar to those of Proposition 1 and can be found in App. B.4. With Proposition 2, we immediately have

Theorem 2. *Let Assumption 1 hold and assume in addition that* f *is* $\mu-$*strongly convex. Under the same settings as in Proposition 2, for all* $k \ge k_0 = \lfloor\frac{\log 0.5}{\log\beta}\rfloor$ *we have*

$$\begin{aligned} \mathbb{E}[f(z^k)-f^*] \le & \left(1-\frac{\alpha\mu}{1+\frac{8\mu}{L}}\right)^{k-k_0}\mathbb{E}[L^{k_0}] + \left(1+\frac{8\mu}{L}\right)\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L}{\mu}\alpha\sigma^2 \\ & + \left(1+\frac{8\mu}{L}\right)\left(\frac{1}{1+\beta}\frac{12\sqrt{\beta}}{25}\frac{2L+18\mu}{\mu}\alpha\sigma^2 + \frac{\beta^2+\frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}}\frac{2}{1+\beta}\alpha\sigma^2\right) \\ & = \mathcal{O}\left((1-\alpha\mu)^k + \frac{L}{\mu}\alpha\sigma^2\right). \end{aligned}$$

Corollary 1. *Let Assumption 1 hold and assume in addition that* f *is* $\mu-$*strongly convex. Under the same settings as in Proposition 2, for all* $k \ge k_0 = \lfloor\frac{\log 0.5}{\log\beta}\rfloor$ *we have*

$$\mathbb{E}[f(x^k)-f^*] = \mathcal{O}\left(r^k + \frac{L}{\mu}\alpha\sigma^2\right),$$

where $r = \max\{1-\alpha\mu, \beta\}$.

Remark 1. *1. Under nonconvex settings, the classical convergence bound of SGD is $\mathcal{O}\left(\frac{f(x^1)-f^*}{k\alpha}+L\alpha\sigma^2\right)$ with $\alpha=\mathcal{O}(\frac{1}{L})$ (see, e.g., Theorem 4.8 of [4]). Therefore, Theorem 1 tells us that with $\alpha=\mathcal{O}(\frac{1-\beta}{L})$, SGDM achieves the same convergence bound as SGD.*

2. *In contrast, the radius of the stationary distribution for SGDM in [26] and [25] is $\mathcal{O}(\frac{\alpha\sigma^2}{1-\beta})$, and the latter one also assumes that ∇f is uniformly bounded.*
3. *In Theorem 2 and Corollary 1, the convergence bounds hold for $k \geq k_0 = \lfloor\frac{\log 0.5}{\log\beta}\rfloor$, where k_0 is a mild constant[1]. when $r = 1-\alpha\mu$, the $\mathcal{O}(r^k)$ part in Corollary 1 matches the lower bound established in Proposition 3 of [12].*
4. *The convergence bound of SGD under strong convexity is $\mathcal{O}\left((1-\alpha\mu)^k+\frac{L}{\mu}\alpha\sigma^2\right)$ (see, e.g, Theorem 4.6 of [4]), our result for SGDM in Corollary 1 recovers this when $\beta=0$.*

5 Convergence of Multistage SGDM

In this section, we switch to the Multistage SGDM (Algorithm 1).

Let us first show that when the (7) is applied, we can define the constants c_i properly so that (5) still produces a Lyapunov function.

Proposition 3. *Let Assumption 1 hold. In Algorithm 1, let the parameters satisfy* (7) *with $A_1 = \frac{1}{24\sqrt{2}L}$. In addition, let*

$$\frac{1-\beta_1}{\beta_1} \leq 12\frac{1-\beta_n}{\sqrt{\beta_n+\beta_n^2}}, \qquad c_1 = \frac{\frac{\alpha_1^2}{1-\beta_1}\frac{\beta_n+\beta_n^2}{(1-\beta_n)^2}L^3}{1-4\alpha_1^2\frac{\beta_n+\beta_n^2}{(1-\beta_n)^2}L^2},$$

and for any $i \geq 1$, let

$$c_{i+1} = c_i - \left(4c_1\alpha_1^2 + L\frac{\alpha_1^2}{1-\beta_1}\right)\beta_n^i(i+\frac{\beta_n}{1-\beta_n})L^2.$$

Then, we have $c_i > 0$ for any $i \geq 1$. Furthermore, with z^k defined in (8)*, for any $k \geq 1$, we have*

$$\begin{aligned}
&\mathbb{E}[L^{k+1}-L^k] \\
&\leq \left(-\alpha(k)+\frac{3-\beta(k)+2\beta^2(k)}{2(1-\beta(k))}L\alpha^2(k)+4c_1\alpha^2(k)\right)\mathbb{E}[\|g^k\|^2] \\
&\quad + \left(\beta^2(k)L\alpha^2(k)12\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}\sigma^2+\frac{1}{2}L\alpha^2(k)\sigma^2+4c_1(1-\beta_1)\alpha^2(k)\sigma^2\right).
\end{aligned}$$

where $\alpha(k), \beta(k)$ are the stepsize and momentum weight applied at kth iteration, respectively.

Theorem 3. *Let Assumption 1 hold. Under the same settings as in Proposition 3, let $\beta_1 \geq \frac{1}{2}$ and let A_2 be large enough such that*

$$\beta_i^{2T_i} \leq \frac{1}{2} \quad \text{for} \quad i = 1, 2, ...n.$$

Then, we have

$$\begin{aligned}
\frac{1}{n}\sum_{l=1}^{n}\frac{1}{T_l}\sum_{i=T_1+..+T_{l-1}+1}^{T_1+..+T_l}\mathbb{E}[\|g^i\|^2] &\leq \frac{2(f(x^1)-f^*)}{nA_2}+\frac{1}{n}\sum_{l=1}^{n}\left(24\beta_l^2\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}L+3L\right)\alpha_l\sigma^2 \\
&= \mathcal{O}\left(\frac{f(x^1)-f^*}{nA_2}+\frac{1}{n}\sum_{l=1}^{n}L\alpha_l\sigma^2\right).
\end{aligned} \tag{11}$$

Remark 2. *1. On the left hand side of* (11)*, we have the average of the averaged squared gradient norm of n stages.*

[1]For example, we have $k_0 = 6$ for the popular choice $\beta = 0.9$.

2. *On the right hand side of* (11)*, the first term dominates at initial stages, we can apply large* α_i *for these stages to accelerate initial convergence, and use smaller* α_i *for later stages so that the size of stationary distribution is small. In contrast, (static) SGDM need to use a small stepsize* α *to make the size of stationary distribution with small.*
3. *It is unclear whether the iteration complexity of Multistage SGDM is better than SGDM or not. However, we do observe that Multistage SGDM is faster numerically. We leave the possible improved analysis of Multistage SGDM to future work.*

6 Experiments

In this section, we verify our theoretical claims by numerical experiments. For each combination of algorithm and training task, training is performed with 3 random seeds $1, 2, 3$. Unless otherwise stated, we report the average of losses of the past m batches, where m is the number of batches for the whole dataset. Our implementation is available at GitHub[1]. Additional implementation details can be found in App. E.

6.1 Logistic regression

Setup. The MNIST dataset consists of $n = 60000$ labeled examples of 28×28 gray-scale images of handwritten digits in $K = 10$ classes $0, 1, \ldots, 9$. For all algorithms, we use batch size $s = 64$ (and hence number batches per epoch is $m = 1874$), number of epochs $T = 50$. The regularization parameter is $\lambda = 5 \times 10^{-4}$.

The effect of α in (static) SGDM. By Theorem 2 we know that, with a fixed β, a larger α leads to faster loss decrease to the stationary distribution. However, the size of the stationary distribution is also larger. This is well illustrated in Figure 1. For example, $\alpha = 1.0$ and $\alpha = 0.5$ make losses decrease more rapidly than $\alpha = 0.1$. During later iterations, $\alpha = 0.1$ leads to a lower final loss.

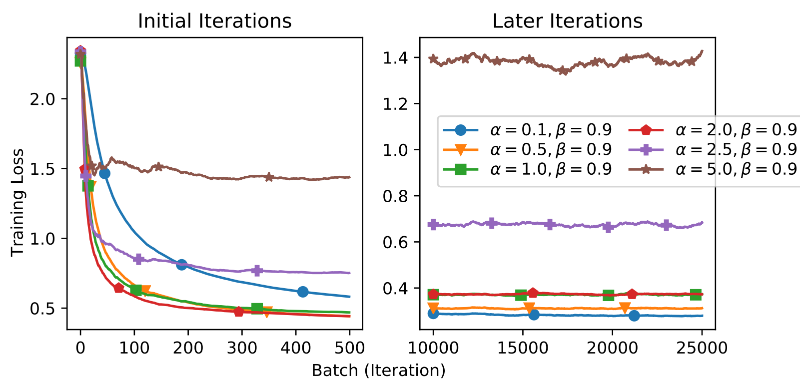

Figure 1: Logistic Regression on the MNIST Dataset using SGDM with fixed (α, β)

Multistage SGDM. We take 3 stages for Multistage SGDM. The parameters are chosen according to (7): $T_1 = 3, T_2 = 6, T_3 = 21$, $\alpha_i = A_2/T_i$, $\beta_i = A_1/(c_2 + \alpha_i)$, where $A_2 = 2.0$ and $A_1 = 1.0$.[1] We compare Multistage SGDM with SGDM with $(\alpha, \beta) = (0.66, 0.9)$ and $(\alpha, \beta) = (0.095, 0.9)$, where $0.66, 0.095$ are the stepsizes of the first and last stage of Multistage SGDM, respectively. The training losses of initial and later iterations are shown in Figure 2.

We can see that SGDM with $(\alpha, \beta) = (0.66, 0.9)$ converges faster initially, but has a higher final loss; while SGDM with $(\alpha, \beta) = (0.095, 0.9)$ behaves the other way. Multistage SGDM takes the advantage of both, as predicted by Theorem 3. The performances of SGDM and Vanilla SGD with the same stepsize are similar.

6.2 Image classification

For the task of training ResNet-18 on the CIFAR-10 dataset, we compare Multistage SGDM, a baseline SGDM, and YellowFin [28], an automatic momentum tuner based on heuristics from

[1] `https://github.com/gao-yuan-hangzhou/improved-analysis-sgdm`

[1] Here, A_1 is not set by its theoretical value $\frac{1}{24L}$, since the dataset is very large and the gradient Lipschitz constant L cannot be computed easily.

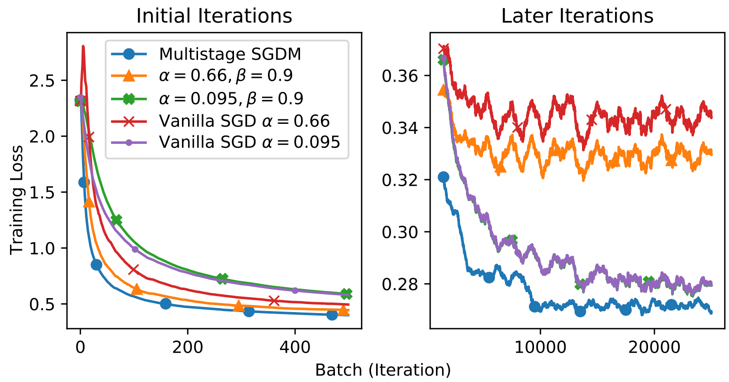

Figure 2: Logistic Regression on the MNIST using Multistage SGDM and SGDM with fixed β

optimizing strongly convex quadratics. The initial learning rate of YellowFin is set to 0.1,[1] and other parameters are set as their default values. All algorithms are run for $T = 50$ epochs and the batch size is fixed as $s = 128$.

For Multistage SGDM, the parameters choices are governed by (7): the stage lengths are $T_1 = 5$, $T_2 = 10$, and $T_3 = 35$. Take $A_1 = 1.0$, $A_2 = 2.0$, set the per-stage stepsizes and momentum weights as $\alpha_i = A_2/T_i$ and $\beta_i = A_1/(A_1 + \alpha_i)$, for stages $i = 1, 2, 3$. For the baseline SGDM, the stepsize schedule of Multistage SGDM is applied, but with a fixed momentum $\beta = 0.9$.

In Figure 3, we present training losses and end-of-epoch validation accuracy of the tested algorithms. We can see that Multistage SGDM performs the best. Baseline SGDM is slightly worse, possibly because of its fixed momentum weight. Finally, Multistage SGDM can reach a test accuracy of 93% around 200 epochs.

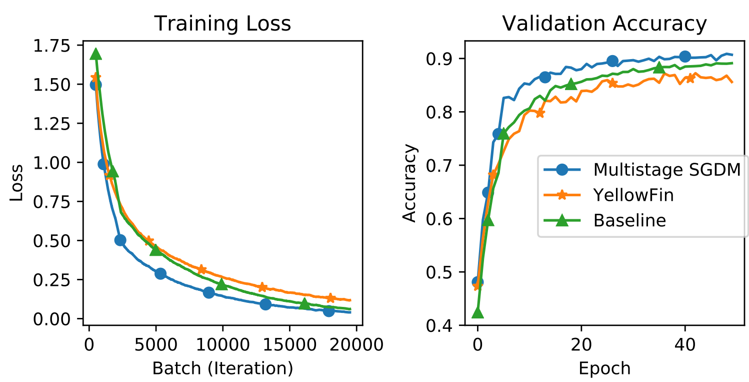

Figure 3: Training ResNet-18 on CIFAR-10

7 Summary and Future Directions

In this work, we provide new theoretical insights into the convergence behavior of SGDM and Multistage SGDM. For SGDM, we show that it is as fast as plain SGD in both nonconvex and strongly convex settings. For the widely adopted multistage SGDM, we establish its convergence and show the advantage of stagewise training.

There are still open problems to be addressed. For example, (a) Is it possible to show that SGDM converges faster than SGD for special objectives such as quadratic ones? (b) Are there more efficient parameter choices than (7) that guarantee even faster convergence?

[1] We have experimented with initial learning rates 0.001 (default), 0.01, 0.1 and 0.5, each repeated 3 times; we found that the choice 0.1 is the best in terms of the final training loss.

Broader Impact

The results of this paper improves the performance of stochastic gradient descent with momentum as well as its multistage version. Our study will also benefit the machine learning community. We do not believe that the results in this work will cause any ethical issue, or put anyone at a disadvantage in our society.

Acknowledgements

Yanli Liu and Wotao Yin were partially supported by ONR Grant N000141712162. Yanli Liu was also supported by UCLA Dissertation Year Fellowship. Yuan Gao was supported by Columbia University School of Engineering and Applied Science.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Wangpeng An, Haoqian Wang, Qingyun Sun, Jun Xu, Qionghai Dai, and Lei Zhang. A pid controller approach for stochastic optimization of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8522–8531, 2018.

[3] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. In *Advances in Neural Information Processing Systems*, pages 8523–8534, 2019.

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[5] Damek Davis, Dmitriy Drusvyatskiy, and Vasileios Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

[6] Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Advances in Neural Information Processing Systems*, pages 14951–14962, 2019.

[7] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315. IEEE, 2015.

[8] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[9] Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 9630–9640, 2019.

[10] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[11] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[12] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] Andrei Kulunchakov and Julien Mairal. A generic acceleration framework for stochastic composite optimization. In *Advances in Neural Information Processing Systems*, pages 12556–12567, 2019.

[15] Nicolas Loizou and Peter Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *arXiv preprint arXiv:1710.10737*, 2017.

[16] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

[17] Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and adam for deep learning. In *International Conference on Learning Representations*, 2019.

[18] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[20] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[21] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[22] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. On the convergence of the stochastic heavy ball method. *arXiv preprint arXiv:2006.07867*, 2020.

[23] Ruoyu Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

[24] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[25] Y Yan, T Yang, Z Li, Q Lin, and Y Yang. A unified analysis of stochastic momentum methods for deep learning. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018.

[26] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193, 2019.

[27] Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems*, pages 2604–2614, 2019.

[28] Jian Zhang and Ioannis Mitliagkas. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.

A Proof of Preliminary Lemmas

A.1 Proof of Lemma 1

Since $m^k = (1-\beta)\sum_{i=1}^k \beta^{k-i}\tilde{g}^i$, we have

$$\mathbb{E}\left[\left\|m^k - (1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\right\|^2\right] = (1-\beta)^2\,\mathbb{E}\left[\left\|\sum_{i=1}^k \beta^{k-i}(\tilde{g}^i - g^i)\right\|^2\right].$$

Moreover, since $\zeta^1, \zeta^2, ..., \zeta^k$ are independent random variables (item 3 of Assumption 1), we can write the total expectation as $\mathbb{E} = \mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^k}$, and therefore

$$\begin{aligned}
&\mathbb{E}\left[\left\|m^k - (1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\right\|^2\right] \\
&= (1-\beta)^2\mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^k}\left[\left\|\sum_{i=1}^k \beta^{k-i}(\tilde{g}^i - g^i)\right\|^2\right] \\
&= (1-\beta)^2\mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^k}\left[\sum_{i=1}^k\sum_{j=1}^k \langle \beta^{k-i}(\tilde{g}^i - g^i), \beta^{k-j}(\tilde{g}^j - g^j)\rangle\right].
\end{aligned}$$

By applying $\mathbb{E}_{\zeta^i}[\tilde{g}^i] = g^i$ (item 2 in Assumption 1), we further have for any $i > j$ that

$$\begin{aligned}
&\mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^k}\left[\langle \tilde{g}^i - \mathbb{E}_{\zeta^i}[\tilde{g}^i], \tilde{g}^j - \mathbb{E}_{\zeta^j}[\tilde{g}^j]\rangle\right] \\
&= \mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^i}\left[\langle \tilde{g}^i - \mathbb{E}_{\zeta^i}[\tilde{g}^i], \tilde{g}^j - \mathbb{E}_{\zeta^j}[\tilde{g}^j]\rangle\right] \\
&= \mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^{i-1}}\left[\langle \mathbb{E}_{\zeta^i}[\tilde{g}^i] - \mathbb{E}_{\zeta^i}[\tilde{g}^i], \tilde{g}^j - \mathbb{E}_{\zeta^j}[\tilde{g}^j]\rangle\right] \\
&= 0.
\end{aligned}$$

.

It is straightforward to see that the same conclusion holds for $i < j$.

Finally, we know from the item 4 in Assumption 1 that

$$\begin{aligned}
&\mathbb{E}\left[\left\|m^k - (1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\right\|^2\right] \\
&= (1-\beta)^2\mathbb{E}_{\zeta^1}\mathbb{E}_{\zeta^2}...\mathbb{E}_{\zeta^k}\left[\sum_{i=1}^k \beta^{2(k-i)}\|\tilde{g}^i - \mathbb{E}_{\zeta^i}[\tilde{g}^i]\|^2\right] \\
&\le \frac{1-\beta}{1+\beta}(1-\beta^{2k})\sigma^2.
\end{aligned}$$

A.2 Proof of Lemma 2

We have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right] \\
&= \left(\frac{1-\beta}{1-\beta^k}\right)^2 \sum_{i,j=1}^{k} \mathbb{E}[\langle \beta^{k-i}(g^k-g^i), \beta^{k-j}(g^k-g^j)\rangle] \\
&\le \left(\frac{1-\beta}{1-\beta^k}\right)^2 \sum_{i,j=1}^{k} \beta^{2k-i-j}(\frac{1}{2}\,\mathbb{E}[\|g^k-g^i\|^2] + \frac{1}{2}\,\mathbb{E}[\|g^k-g^j\|^2]) \\
&= \left(\frac{1-\beta}{1-\beta^k}\right)^2 \sum_{i=1}^{k}\left(\sum_{j=1}^{k}\beta^{2k-i-j}\right)\frac{1}{2}\,\mathbb{E}[\|g^k-g^j\|^2] \\
&\quad + \left(\frac{1-\beta}{1-\beta^k}\right)^2 \sum_{j=1}^{k}\left(\sum_{i=1}^{k}\beta^{2k-i-j}\right)\frac{1}{2}\,\mathbb{E}[\|g^k-g^i\|^2] \\
&= \left(\frac{1-\beta}{1-\beta^k}\right)^2 \sum_{i=1}^{k}\frac{\beta^{k-i}(1-\beta^k)}{1-\beta}\,\mathbb{E}[\|g^k-g^i\|^2] \\
&= \frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}\,\mathbb{E}[\|g^k-g^i\|^2],
\end{aligned}
$$

where we have applied Cauchy-Schwarz in the first inequality.

By applying triangle inequality and the smoothness of f (item 1 in Assumption 1), we further have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right] \\
&\le \frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}(k-i)\sum_{j=i}^{k-1}\mathbb{E}[\|g^{j+1}-g^j\|^2] \\
&\le \frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}(k-i)\sum_{j=i}^{k-1}L^2\,\mathbb{E}[\|x^{j+1}-x^j\|^2] \\
&= \frac{1-\beta}{1-\beta^k}\sum_{j=1}^{k-1}\left(\sum_{i=1}^{j}\beta^{k-i}(k-i)\right)L^2\,\mathbb{E}[\|x^{j+1}-x^j\|^2].
\end{aligned}
$$

Therefore, by defining $a'_{k,j} = \frac{1-\beta}{1-\beta^k}L^2\sum_{i=1}^{j}\beta^{k-i}(k-i)$, we get

$$
\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right] \le \sum_{j=1}^{k-1}a'_{k,j}\,\mathbb{E}[\|x^{j+1}-x^j\|^2]. \tag{12}
$$

Furthermore, $a'_{k,j}$ can be calculated as

$$
a'_{k,j} = \frac{L^2\beta^k}{1-\beta^k}\left(-(k-1)-\frac{1}{1-\beta}\right) + \frac{L^2\beta^{k-j}}{1-\beta^k}\left(k-j+\frac{\beta}{1-\beta}\right). \tag{13}
$$

Notice that

$$
a'_{k,j} < a_{k,j} \coloneqq \frac{L^2\beta^{k-j}}{1-\beta^k}\left(k-j+\frac{\beta}{1-\beta}\right). \tag{14}
$$

Combining this with (12), we finally arrive at

$$\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right] \le \sum_{i=1}^{k-1} a_{k,i}\,\mathbb{E}[\|x^{i+1}-x^i\|^2],$$

where

$$a_{k,i} = \frac{L^2\beta^{k-i}}{1-\beta^k}\left(k-i+\frac{\beta}{1-\beta}\right).$$

A.3 Proof of Lemma 3

Let us consider the cases of $k=1$ and $k \ge 2$ separately.

For $k=1$, we have

$$z^2 - z^1 = \frac{1}{1-\beta}x^2 - \frac{\beta}{1-\beta}x^1 - x^1 = \frac{1}{1-\beta}(x^2-x^1) = -\alpha\tilde{g}^1.$$

And for $k \ge 2$, we have

$$\begin{aligned} z^{k+1} - z^k &= \frac{1}{1-\beta}(x^{k+1}-x^k) - \frac{\beta}{1-\beta}(x^k - x^{k-1}) \\ &= \frac{1}{1-\beta}(-\alpha m^k) - \frac{\beta}{1-\beta}(-\alpha m^{k-1}) \\ &= \frac{1}{1-\beta}(-\alpha m^k + \alpha\beta m^{k-1}) \\ &= -\alpha\tilde{g}^k. \end{aligned}$$

B Main Theory for SGDM

B.1 Objective descent

In order to prove Proposition 1, let us first show an auxiliary result.

Proposition 4. *Take Assumption 1. Then, for z^k defined in* (6)*, we have*

$$\begin{aligned} \mathbb{E}[f(z^{k+1})] \le \mathbb{E}[f(z^k)] &+ (-\alpha + \frac{1+\beta^2}{1-\beta}L\alpha^2 + \frac{1}{2}L\alpha^2)\,\mathbb{E}[\|g^k\|^2] \\ &+ (\frac{\beta^2}{2(1+\beta)} + \frac{1}{2})L\alpha^2\sigma^2 + \frac{\beta^2(1-\beta^k)^2L\alpha^2}{1-\beta}\,\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right]. \end{aligned} \tag{15}$$

The smoothness of f yields

$$\begin{aligned} \mathbb{E}_{\zeta^k}[f(z^{k+1})] &\le f(z^k) + \mathbb{E}_{\zeta^k}[\langle\nabla f(z^k), z^{k+1}-z^k\rangle] + \frac{L}{2}\mathbb{E}_{\zeta^k}[\|z^{k+1}-z^k\|^2] \\ &= f(z^k) + \mathbb{E}_{\zeta^k}[\langle\nabla f(z^k), -\alpha\tilde{g}^k\rangle] + \frac{L\alpha^2}{2}\mathbb{E}_{\zeta^k}[\|\tilde{g}^k\|^2], \end{aligned} \tag{16}$$

where we have applied Lemma 3 in the second step.

For the inner product term, we can take full expectation $\mathbb{E} = \mathbb{E}_{\zeta^1}...\mathbb{E}_{\zeta^k}$ to get

$$\mathbb{E}[\langle\nabla f(z^k), -\alpha\tilde{g}^k\rangle] = \mathbb{E}[\langle\nabla f(z^k), -\alpha g^k\rangle],$$

which follows from the fact that z^k is determined by the previous $k-1$ random samples $\zeta^1, \zeta^2, ...\zeta^{k-1}$, which is independent of ζ^k, and $\mathbb{E}_{\zeta^k}[\tilde{g}^k] = g^k$.

So, we can bound

$$\begin{aligned}\mathbb{E}[\langle \nabla f(z^k), -\alpha \tilde{g}^k \rangle] &= \mathbb{E}[\langle \nabla f(z^k) - g^k, -\alpha g^k \rangle] - \alpha\, \mathbb{E}[\|g^k\|^2] \\ &\leq \alpha \frac{\rho_0}{2} L^2\, \mathbb{E}[\|z^k - x^k\|^2] + \alpha \frac{1}{2\rho_0}\, \mathbb{E}[\|g^k\|^2] - \alpha\, \mathbb{E}[\|g^k\|^2],\end{aligned}$$

where $\rho_0 > 0$ can be any positive constant (to be determined later).

Combining (16) and the last inequality, we arrive at

$$\begin{aligned}\mathbb{E}[f(z^{k+1})] \leq \mathbb{E}[f(z^k)] &+ \alpha \frac{\rho_0}{2} L^2\, \mathbb{E}[\|z^k - x^k\|^2] \\ &+ (\alpha \frac{1}{2\rho_0} - \alpha)\, \mathbb{E}[\|g^k\|^2] + \frac{L\alpha^2}{2}\, \mathbb{E}[\|\tilde{g}^k\|^2].\end{aligned}$$

Since $z^k = x^k$ when $k = 1$ and $z^k = \frac{1}{1-\beta} x^k - \frac{\beta}{1-\beta} x^{k-1}$ when $k \geq 2$, it can be verified that $z^k - x^k = -\frac{\beta}{1-\beta} \alpha m^{k-1}$. Consequently,

$$\begin{aligned}\mathbb{E}[f(z^{k+1})] \leq \mathbb{E}[f(z^k)] &+ \alpha^3 \frac{\rho_0}{2} L^2 (\frac{\beta}{1-\beta})^2\, \mathbb{E}[\|m^{k-1}\|^2] \\ &+ (\alpha \frac{1}{2\rho_0} - \alpha)\, \mathbb{E}[\|g^k\|^2] + \frac{L\alpha^2}{2}\, \mathbb{E}[\|\tilde{g}^k\|^2].\end{aligned} \tag{17}$$

On the other hand, from Lemma 1 we know that

$$\begin{aligned}\mathbb{E}[\|m^{k-1}\|^2] &\leq 2\, \mathbb{E}[\|m^{k-1} - (1-\beta) \sum_{i=1}^{k-1} \beta^{k-1-i} g^i\|^2] + 2\, \mathbb{E}[\|(1-\beta) \sum_{i=1}^{k-1} \beta^{k-1-i} g^i\|^2] \\ &\leq 2 \frac{1-\beta}{1+\beta} \sigma^2 + 2\, \mathbb{E}[\|(1-\beta) \sum_{i=1}^{k-1} \beta^{k-1-i} g^i\|^2] \\ \mathbb{E}[\|\frac{1-\beta}{1-\beta^{k-1}} \sum_{i=1}^{k-1} \beta^{k-1-i} g^i\|^2] &\leq 2\, \mathbb{E}[\|g^k\|^2] + 2\, \mathbb{E}[\|\frac{1-\beta}{1-\beta^{k-1}} \sum_{i=1}^{k-1} \beta^{k-1-i} g^i - g^k\|^2], \\ \mathbb{E}[\|\tilde{g}^k\|^2] &\leq \sigma^2 + \mathbb{E}[\|g^k\|^2].\end{aligned} \tag{18}$$

Putting these into (17), we arrive at

$$\begin{aligned}\mathbb{E}[f(z^{k+1})] \leq \mathbb{E}[f(z^k)] &+ \left(-\alpha + \alpha \frac{1}{2\rho_0} + 2\alpha^3 \rho_0 L^2 (\frac{\beta}{1-\beta})^2 (1-\beta^{k-1})^2 + \frac{L\alpha^2}{2} \right) \mathbb{E}[\|g^k\|^2] \\ &+ \left(\alpha^3 \rho_0 L^2 (\frac{\beta}{1-\beta})^2 \frac{1-\beta}{1+\beta} \sigma^2 + \frac{L\alpha^2}{2} \sigma^2 \right) \\ &+ 2\alpha^3 \rho_0 L^2 (\frac{\beta}{1-\beta})^2 (1-\beta^{k-1})^2\, \mathbb{E}[\|\frac{1-\beta}{1-\beta^{k-1}} \sum_{i=1}^{k-1} \beta^{k-1-i} g^i - g^k\|^2].\end{aligned}$$

Substituting

$$\begin{aligned}\mathbb{E}[\|\frac{1-\beta}{1-\beta^k} \sum_{i=1}^{k} \beta^{k-i} g^i - g^k\|^2] &= \mathbb{E}[\|\frac{1}{1-\beta^k} \beta (1-\beta) \sum_{i=1}^{k-1} \beta^{k-1-i} g^i - \frac{1-\beta^{k-1}}{1-\beta^k} \beta g^k\|^2] \\ &= \beta^2 (\frac{1-\beta^{k-1}}{1-\beta^k})^2\, \mathbb{E}[\|\frac{1-\beta}{1-\beta^{k-1}} \sum_{i=1}^{k-1} \beta^{k-1-i} g^i - g^k\|^2]\end{aligned}$$

into the last inequality produces

$$\mathbb{E}[f(z^{k+1})] \le \mathbb{E}[f(z^k)] + \left(-\alpha + \alpha\frac{1}{2\rho_0} + 2\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2(1-\beta^{k-1})^2 + \frac{L\alpha^2}{2}\right)\mathbb{E}[\|g^k\|^2]$$
$$+ \left(\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2\frac{1-\beta}{1+\beta}\sigma^2 + \frac{L\alpha^2}{2}\sigma^2\right)$$
$$+ 2\alpha^3\rho_0 L^2(\frac{1}{1-\beta})^2(1-\beta^k)^2\,\mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2]. \tag{19}$$

Finally, using $1-\beta^{k-1} < 1$ and $\rho_0 = \frac{1-\beta}{2L\alpha}$ gives

$$\mathbb{E}[f(z^{k+1})] \le \mathbb{E}[f(z^k)] + (-\alpha + \frac{1+\beta^2}{1-\beta}L\alpha^2 + \frac{1}{2}L\alpha^2)\,\mathbb{E}[\|g^k\|^2]$$
$$+ (\frac{\beta^2}{2(1+\beta)} + \frac{1}{2})L\alpha^2\sigma^2 + \frac{\beta^2(1-\beta^k)^2 L\alpha^2}{1-\beta}\,\mathbb{E}\left[\left\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right].$$

B.2 Proof of Proposition 1

Recall that L^k is defined as

$$L^k = f(z^k) - f^* + \sum_{i=1}^{k-1} c_i\|x^{k+1-i} - x^{k-i}\|^2,$$

Therefore, by (19) we know that

$$\begin{aligned}\mathbb{E}[L^{k+1} - L^k] \le &\left(-\alpha + \alpha\frac{1}{2\rho_0} + 2\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}\right)\mathbb{E}[\|g^k\|^2] \\ &+ \left(\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2\frac{1-\beta}{1+\beta}\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2\right) \\ &+ \sum_{i=1}^{k-1}(c_{i+1} - c_i)\,\mathbb{E}[\|x^{k+1-i} - x^{k-i}\|^2] \\ &+ c_1\,\mathbb{E}[\|x^{k+1} - x^k\|^2] \\ &+ 2\alpha^3\rho_0 L^2(\frac{1}{1-\beta})^2(1-\beta^k)^2\,\mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2],\end{aligned} \tag{20}$$

where $\rho_0 = \frac{1-\beta}{2L\alpha}$.

To bound the $c_1\,\mathbb{E}[\|x^{k+1} - x^k\|^2]$ term, we need to following inequalities, which are obtained in a similar way as (18).

$$\begin{aligned}\mathbb{E}[\|m^k\|^2] &\le 2\,\mathbb{E}[\|m^k - (1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i\|^2] + 2\,\mathbb{E}[\|(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i\|^2] \\ &\le 2\frac{1-\beta}{1+\beta}\sigma^2 + 2\,\mathbb{E}[\|(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i\|^2] \\ \mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i\|^2] &\le 2\,\mathbb{E}[\|g^k\|^2] + 2\,\mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2], \\ \mathbb{E}[\|\tilde{g}^k\|^2] &\le \sigma^2 + \mathbb{E}[\|g^k\|^2].\end{aligned} \tag{21}$$

Therefore, $c_1\,\mathbb{E}[\|x^{k+1}-x^k\|^2]$ can be bounded as

$$
\begin{aligned}
c_1\,\mathbb{E}[\|x^{k+1}-x^k\|^2] &= c_1\alpha^2\,\mathbb{E}[\|m^k\|^2] \\
&\le c_1\alpha^2\left(2\frac{1-\beta}{1+\beta}\sigma^2 + 4\,\mathbb{E}[\|g^k\|^2](1-\beta^k)^2\right) \\
&\quad + 4c_1\alpha^2\,\mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2] \\
&< c_1\alpha^2\left(2\frac{1-\beta}{1+\beta}\sigma^2 + 4\,\mathbb{E}[\|g^k\|^2]\right) \\
&\quad + 4c_1\alpha^2(1-\beta^k)^2\,\mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2]
\end{aligned}
$$

Combine this with (20), we obtain

$$
\begin{aligned}
&\mathbb{E}[L^{k+1}-L^k] \\
&\le \left(-\alpha + \alpha\frac{1}{2\rho_0} + 2\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2} + 4c_1\alpha^2\right)\mathbb{E}[\|g^k\|^2] \\
&\quad + \left(\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2\frac{1-\beta}{1+\beta}\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2\right) \\
&\quad + \sum_{i=1}^{k-1}(c_{i+1}-c_i)\,\mathbb{E}[\|x^{k+1-i}-x^{k-i}\|^2] \\
&\quad + 4c_1\alpha^2(1-\beta^k)^2\,\mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2] \\
&\quad + 2\alpha^3\rho_0 L^2(\frac{1}{1-\beta})^2(1-\beta^k)^2\,\mathbb{E}[\|\frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2].
\end{aligned} \tag{22}
$$

In the rest of the proof, let us show that the sum of the last three terms in (22) is non-positive.

First of all, by Lemma 2 we know that

$$
\mathbb{E}\left[\left\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2\right] \le \sum_{i=1}^{k-1}a_{k,i}\,\mathbb{E}[\|x^{i+1}-x^i\|^2],
$$

where

$$
a_{k,i} = \frac{L^2\beta^{k-i}}{1-\beta^k}\left(k-i+\frac{\beta}{1-\beta}\right).
$$

Or equivalently,

$$
\mathbb{E}\left\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\right\|^2 \le \sum_{i=1}^{k-1}a_{k,k-i}\,\mathbb{E}\,\|x^{k+1-i}-x^{k-i}\|^2,
$$

where

$$
a_{k,k-i} = \frac{L^2\beta^i}{1-\beta^k}\left(i+\frac{\beta}{1-\beta}\right).
$$

Therefore, in order to make the sum of the last three terms of (22) to be non-positive, we need to have

$$
c_{i+1} \le c_i - \left(4c_1\alpha^2(1-\beta^k)^2 + 2\alpha^3\rho_0 L^2\frac{(1-\beta^k)^2}{(1-\beta)^2}\right)a_{k,k-i}
$$

for all $i \ge 1$.

Since $1-\beta^k<1$, it suffices to enforce the following for all $i\geq 1$:

$$c_{i+1}=c_i-\left(4c_1\alpha^2+2\alpha^3\rho_0 L^2\frac{1}{(1-\beta)^2}\right)\beta^i(i+\frac{\beta}{1-\beta})L^2. \tag{23}$$

And in order for $c_i>0$ for all $i\geq 1$, we can determine c_1 by

$$c_1=\left(4c_1\alpha^2+2\alpha^3\rho_0 L^2\frac{1}{(1-\beta)^2}\right)\sum_{i=1}^{\infty}\beta^i(i+\frac{\beta}{1-\beta})L^2.$$

Since

$$\sum_{i=1}^{j} i\beta^i=\frac{1}{1-\beta}\left(\frac{\beta(1-\beta^j)}{1-\beta}-j\beta^{j+1}\right),$$

we have $\sum_{i=1}^{\infty} i\beta^i=\frac{\beta}{(1-\beta)^2}$ and

$$c_1=\left(4c_1\alpha^2+2\alpha^3\rho_0 L^2\frac{1}{(1-\beta)^2}\right)\frac{\beta+\beta^2}{(1-\beta)^2}L^2.$$

This stipulates that

$$c_1=\frac{2\alpha^3\rho_0 L^4\frac{\beta+\beta^2}{(1-\beta)^4}}{1-4\alpha^2\frac{\beta+\beta^2}{(1-\beta)^2}L^2}. \tag{24}$$

Notice that $\alpha\leq\frac{1-\beta}{4L\sqrt{\beta+\beta^2}}$ ensures $c_1>0$.

With the choices of c_i in (23) and (24), the sum of the last three terms of (22) is non-positive. Therefore,

$$\begin{aligned}\mathbb{E}[L^{k+1}-L^k]\leq&\left(-\alpha+\alpha\frac{1}{2\rho_0}+2\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2+\frac{L\alpha^2}{2}+4c_1\alpha^2\right)\mathbb{E}[\|g^k\|^2]\\&+\left(\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2\frac{1-\beta}{1+\beta}\sigma^2+\frac{1}{2}L\alpha^2\sigma^2+2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2\right).\end{aligned} \tag{25}$$

Finally, taking

$$\rho_0=\frac{1-\beta}{2L\alpha} \tag{26}$$

in (24), (23), and (25) gives

$$c_1=\frac{\frac{\beta+\beta^2}{(1-\beta)^3}L^3\alpha^2}{1-4\alpha^2\frac{\beta+\beta^2}{(1-\beta)^2}L^2},$$

$$c_{i+1}=c_i-\left(4c_1\alpha^2+\frac{L\alpha^2}{(1-\beta)}\right)\beta^i(i+\frac{\beta}{1-\beta})L^2,$$

$$\begin{aligned}\mathbb{E}[L^{k+1}-L^k]\leq&\left(-\alpha+\frac{3-\beta+2\beta^2}{2(1-\beta)}L\alpha^2+4c_1\alpha^2\right)\mathbb{E}[\|g^k\|^2]\\&+\left(\frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2+\frac{1}{2}L\alpha^2\sigma^2+2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2\right).\end{aligned}$$

B.3 Proof of Theorem 1

From (25) we know that

$$\mathbb{E}[L^{k+1}-L^k]\leq -R_1\,\mathbb{E}[\|g^k\|^2]+R_2, \tag{27}$$

where

$$R_1 = \alpha - \alpha\frac{1}{2\rho_0} - 2\alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2 - \frac{L\alpha^2}{2} - 4c_1\alpha^2 \tag{28}$$

$$R_2 = \alpha^3\rho_0 L^2(\frac{\beta}{1-\beta})^2\frac{1-\beta}{1+\beta}\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2, \tag{29}$$

and $\rho_0 = \frac{1-\beta}{2L\alpha}$.

This immediately tells us that

$$L^1 \geq \mathbb{E}[L^1 - L^{k+1}] \geq R_1\sum_{i=1}^{k}\mathbb{E}[\|g^i\|^2] - kR_2,$$

and therefore

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\|g^k\|^2] \leq \frac{L^1}{kR_1} + \frac{R_2}{R_1}. \tag{30}$$

In the rest the proof, we will bound R_1 and R_2 appropriately.

First, let us show that $R_1 \geq \frac{\alpha}{2}$ when $\rho_0 = \frac{1-\beta}{2L\alpha}$ as in (26) and $\alpha \leq \min\{\frac{1-\beta}{L(4-\beta+\beta^2)}, \frac{1-\beta}{2L\sqrt{\beta+\beta^2}}\}$.

From (24) we know that

$$c_1 = \frac{2\alpha^3\rho_0 L^4\frac{\beta+\beta^2}{(1-\beta)^4}}{1-4\alpha^2\frac{\beta+\beta^2}{(1-\beta)^2}L^2}.$$

Since $\alpha \leq \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}}$, we have

$$4\alpha^2\frac{\beta+\beta^2}{(1-\beta)^2}L^2 \leq \frac{1}{2} \tag{31}$$

and

$$c_1 \leq 4\alpha^3\rho_0 L^4\frac{\beta+\beta^2}{(1-\beta)^4} \leq \frac{1}{4}\alpha\rho_0\frac{L^2}{(1-\beta)^2}. \tag{32}$$

Therefore, in order to ensure $R_1 \geq \frac{\alpha}{2}$ where R_1 is defined in (28), it suffices to have

$$\alpha\frac{1}{2\rho_0} + 2\alpha\rho_0 L^2(\frac{\beta}{1-\beta})^2\alpha^2 + \frac{L\alpha^2}{2} + \alpha\rho_0 L^2\frac{1}{(1-\beta)^2}\alpha^2 \leq \frac{\alpha}{2}. \tag{33}$$

Applying $\rho_0 = \frac{1-\beta}{2L\alpha}$ yields

$$\begin{aligned}
&\alpha\frac{1}{2\rho_0} + 2\alpha\rho_0 L^2(\frac{\beta}{1-\beta})^2\alpha^2 + \frac{L\alpha^2}{2} + \alpha\rho_0 L^2\frac{1}{(1-\beta)^2}\alpha^2\\
&= \frac{L\alpha^2}{1-\beta} + \alpha^2 L\frac{\beta^2}{1-\beta} + \frac{1}{2}\alpha^2 L\frac{1}{1-\beta} + \frac{L\alpha^2}{2}\\
&= L\alpha^2\left(\frac{1}{1-\beta} + \frac{\beta^2}{1-\beta} + \frac{1}{2}\frac{1}{1-\beta} + \frac{1}{2}\right)\\
&= L\alpha^2\frac{4-\beta+2\beta^2}{2(1-\beta)}\\
&\leq \frac{\alpha}{2},
\end{aligned}$$

where we have applied $\alpha \leq \frac{1-\beta}{L(4-\beta+2\beta^2)}$ in the last step.

Therefore, (33) is true and

$$R_1 \geq \frac{\alpha}{2}. \tag{34}$$

Now let us turn to R_2. By (32) we know that

$$R_2 = \alpha\rho_0 L^2 (\frac{\beta}{1-\beta})^2 \alpha^2 \frac{1-\beta}{1+\beta}\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 2c_1 \frac{1-\beta}{1+\beta}\alpha^2\sigma^2$$
$$\leq \alpha\rho_0 L^2 (\frac{\beta}{1-\beta})^2 \alpha^2 \frac{1-\beta}{1+\beta}\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 8\alpha^3\rho_0 L^4 \frac{\beta+\beta^2}{(1-\beta)^4}\frac{1-\beta}{1+\beta}\alpha^2\sigma^2.$$

Since $\rho_0 = \frac{1-\beta}{2L\alpha}$, we have

$$R_2 \leq \frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + \frac{4\beta}{(1-\beta)^2}L^3\alpha^4\sigma^2.$$

By applying $\alpha \leq \min\{\frac{1-\beta}{L(4-\beta+\beta^2)}, \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}}\} \leq \frac{1-\beta}{3.75L} < \frac{1-\beta}{4L}$, we further have

$$R_2 \leq \frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + \frac{\beta}{4}L\alpha^2\sigma^2. \tag{35}$$

By putting (34) and (35) into (30), we finally obtain

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\|g^i\|^2] \leq \frac{2\left(f(x^1)-f^*\right)}{k\alpha} + \left(\frac{\beta+3\beta^2}{2(1+\beta)}+1\right)L\alpha\sigma^2$$
$$= \mathcal{O}\left(\frac{f(x^1)-f^*}{k\alpha}\right) + \mathcal{O}\left(L\alpha\sigma^2\right).$$

B.4 Proof of Proposition 2

In order to prove Proposition 2, we will set

$$c_1 = \left(\frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}}\frac{\beta}{1-\beta}\right)\left(\frac{2L^3\alpha^2}{1-\beta} + \frac{18L^2\mu\alpha^2}{(1-\beta)(1+\frac{8\mu}{L})}\right),$$

$$c_{i+1} - c_i + A_3 2L^2\beta^{k-i}\left(k-i+\frac{\beta}{1-\beta}\right) = A_1 c_i, \qquad \forall i \geq 1,$$

Take $\rho_0 = \frac{1-\beta}{2L\alpha}$ in (22), we have

$$\begin{aligned}\mathbb{E}[L^{k+1}-L^k] \leq &\left(-\alpha + \frac{3-\beta+2\beta^2}{2(1-\beta)}L\alpha^2 + 4c_1\alpha^2\right)\mathbb{E}[\|g^k\|^2] \\ &+ \left(\frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2\right) \\ &+ \sum_{i=1}^{k-1}(c_{i+1}-c_i)\,\mathbb{E}[\|x^{k+1-i}-x^{k-i}\|^2] \\ &+ \left(4c_1\alpha^2 + \frac{L\alpha^2}{(1-\beta)}\right)(1-\beta^k)^2\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2].\end{aligned} \tag{36}$$

Let us first derive a lower bound of the first term on the right hand side of (36).

From the strong convexity of f we have

$$\mathbb{E}[\|g^k\|^2] = \mathbb{E}[\|\nabla f(x^k)\|^2] \geq 2\mu\,\mathbb{E}[f(x^k)-f^\star], \tag{37}$$

where $f^\star = \min_{x\in\mathbb{R}^d} f(x)$. On the other hand, for $\mathbb{E}[f(x^k)]$ we have

$$\begin{aligned}
\mathbb{E}[f(z^k)] &\le \mathbb{E}[f(x^k)] + \mathbb{E}[\langle g^k, z^k - x^k\rangle] + \frac{L}{2}\,\mathbb{E}[\|z^k - x^k\|^2] \\
&= \mathbb{E}[f(x^k)] + \mathbb{E}[\langle g^k - \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i + \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i, -\frac{\alpha\beta}{1-\beta} m^{k-1}\rangle] \\
&\quad + \frac{L}{2}\,\mathbb{E}[\|\frac{\alpha\beta}{1-\beta} m^{k-1}\|^2] \\
&\le \mathbb{E}[f(x^k)] + \alpha\frac{\rho}{2}\,\mathbb{E}[\|g^k - \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2] + \frac{\alpha}{2\rho}\,\mathbb{E}[\|\frac{\beta}{1-\beta} m^{k-1}\|^2] \\
&\quad + \mathbb{E}[\langle \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i, -\frac{\alpha\beta}{1-\beta} m^{k-1}\rangle] + \frac{L}{2}\,\mathbb{E}[\|\frac{\alpha\beta}{1-\beta} m^{k-1}\|^2] \\
&\le \mathbb{E}[f(x^k)] + \alpha\frac{\rho}{2}\,\mathbb{E}[\|g^k - \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2] \\
&\quad + \left(\frac{\alpha}{2\rho}(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2\right)\mathbb{E}[\|m^{k-1}\|^2] \\
&\quad + \alpha\frac{\beta}{1-\beta}\left(\frac{\rho_1}{2}\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2] + \frac{1}{2\rho_1}\,\mathbb{E}[\|m^{k-1}\|^2]\right) \\
&= \mathbb{E}[f(x^k)] + \alpha\frac{\rho}{2}\,\mathbb{E}[\|g^k - \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2] \\
&\quad + \left(\frac{\alpha}{2\rho}(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 + \alpha\frac{\beta}{1-\beta}\frac{1}{2\rho_1}\right)\mathbb{E}[\|m^{k-1}\|^2] \\
&\quad + \alpha\frac{\beta}{1-\beta}\frac{\rho_1}{2}\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2],
\end{aligned}$$

where ρ, $\rho_1 > 0$ are to be determined later.

Combining this with (37) gives

$$\begin{aligned}
\mathbb{E}[\|g^k\|^2] \ge 2\mu\Bigg(&\mathbb{E}[f(z^k)] - f^\star - \alpha\frac{\rho}{2}\,\mathbb{E}[\|g^k - \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2] \\
&- \left(\frac{\alpha}{2\rho}(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 + \alpha\frac{\beta}{1-\beta}\frac{1}{2\rho_1}\right)\mathbb{E}[\|m^{k-1}\|^2] \\
&- \alpha\frac{\beta}{1-\beta}\frac{\rho_1}{2}\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2], \Bigg).
\end{aligned} \tag{38}$$

On the other hand, we have from (18) that

$$\begin{aligned}
\mathbb{E}[\|m^{k-1}\|^2] &\le 2\frac{1-\beta}{1+\beta}\sigma^2 + 2(1-\beta^{k-1})^2\left(2\,\mathbb{E}[\|g^k\|^2] + 2\,\mathbb{E}[\|\frac{1}{1-\beta^{k-1}}(1-\beta)\sum_{i=1}^{k-1}\beta^{k-1-i} g^i - g^k\|^2]\right) \\
&= 2\frac{1-\beta}{1+\beta}\sigma^2 \\
&\quad + 2(1-\beta^{k-1})^2\left(2\,\mathbb{E}[\|g^k\|^2] + 2\frac{1}{\beta^2}(\frac{1-\beta^k}{1-\beta^{k-1}})^2\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i - g^k\|^2]\right),
\end{aligned}$$

and that

$$\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i\|^2] \le 2\,\mathbb{E}[\|g^k\|^2] + 2\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i} g^i - g^k\|^2].$$

Putting these two inequalities into (38) and rearranging gives

$$
\begin{aligned}
&\Bigg[1 + 2\mu\Bigg(\Big(\frac{\alpha}{2\rho}(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 + \alpha\frac{\beta}{1-\beta}\frac{1}{2\rho_1} \Big) 4(1-\beta^{k-1})^2 \\
&\qquad\qquad\qquad + \alpha\frac{\beta}{1-\beta}\rho_1 \Bigg)\Bigg] \mathbb{E}[\|g^k\|^2] \\
&\ge 2\mu\Bigg[\mathbb{E}[f(z^k)] - f^\star - \alpha\frac{\rho}{2}\,\mathbb{E}[\|g^k - \frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i}g^i\|^2] \\
&\quad - \Big(\frac{\alpha}{2\rho}(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 + \alpha\frac{\beta}{1-\beta}\frac{1}{2\rho_1} \Big) \\
&\qquad \times \Bigg(2\frac{1-\beta}{1+\beta}\sigma^2 + (1-\beta^{k-1})^2 4\frac{1}{\beta^2}(\frac{1-\beta^k}{1-\beta^{k-1}})^2\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i}g^i - g^k\|^2] \Bigg) \\
&\quad - \alpha\frac{\beta}{1-\beta}\frac{\rho_1}{2}2\,\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i}g^i - g^k\|^2 \Bigg] \\
&= 2\mu\Bigg[\mathbb{E}[f(z^k)] - f^\star \\
&\quad - \Big(\frac{\alpha}{2\rho}(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 + \alpha\frac{\beta}{1-\beta}\frac{1}{2\rho_1} \Big) 2\frac{1-\beta}{1+\beta}\sigma^2 \\
&\quad - \Bigg(\alpha\frac{\rho}{2} + \Big(\frac{\alpha}{2\rho}(\frac{\beta}{1-\beta})^2 + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 + \alpha\frac{\beta}{1-\beta}\frac{1}{2\rho_1} \Big) 4(1-\beta^k)^2\frac{1}{\beta^2} + \alpha\frac{\beta}{1-\beta}\rho_1 \Bigg) \\
&\qquad \times \mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i}g^i - g^k\|^2 \Bigg].
\end{aligned}
$$

Taking $\rho = \frac{1}{1-\beta}$ and $\rho_1 = \frac{1}{\beta}$ gives

$$
\begin{aligned}
&\Bigg[1 + 2\mu\Bigg(\Big(\alpha\frac{\beta^2}{1-\beta} + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 \Big) 4(1-\beta^{k-1})^2 + \alpha\frac{1}{1-\beta} \Bigg)\Bigg] \mathbb{E}[\|g^k\|^2] \\
&\ge 2\mu\Bigg[\mathbb{E}[f(z^k)] - f^\star \\
&\quad - \Big(\alpha\frac{\beta^2}{1-\beta} + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 \Big) 2\frac{1-\beta}{1+\beta}\sigma^2 \\
&\quad - \Bigg(\alpha\frac{1}{2(1-\beta)} + \Big(\alpha\frac{\beta^2}{1-\beta} + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2 \Big) 4(1-\beta^k)^2\frac{1}{\beta^2} + \alpha\frac{1}{1-\beta} \Bigg) \\
&\qquad \times \mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^k \beta^{k-i}g^i - g^k\|^2 \Bigg].
\end{aligned} \tag{39}
$$

Since

$$
\alpha \le \frac{1-\beta}{5L}, \tag{40}
$$

(39) gives

$$
\begin{aligned}
&\left(1+8\frac{\mu}{L}\right)\mathbb{E}[\|g^k\|^2] \\
&\geq 2\mu\Bigg[\mathbb{E}[f(z^k)]-f^\star \\
&-\left(\alpha\frac{\beta^2}{1-\beta}+\frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2\right)2\frac{1-\beta}{1+\beta}\sigma^2 \\
&-\left(\alpha\frac{1}{2(1-\beta)}+\left(\alpha\frac{\beta^2}{1-\beta}+\frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2\right)4(1-\beta^k)^2\frac{1}{\beta^2}+\alpha\frac{1}{1-\beta}\right) \\
&\quad\times\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i-g^k\|^2\Bigg].
\end{aligned}
\tag{41}
$$

Since $\alpha\leq\frac{1-\beta}{5L}$, we have that

$$
\begin{aligned}
c_1&=\left(\frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2}+\frac{\sqrt{\beta}}{1-\sqrt{\beta}}\frac{\beta}{1-\beta}\right)\left(\frac{2L^3\alpha^2}{1-\beta}+\frac{18L^2\mu\alpha^2}{(1-\beta)(1+\frac{8\mu}{L})}\right) \\
&\leq\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}+\frac{2\sqrt{\beta}}{1-\beta}\frac{\beta}{1-\beta}\right)\left(\frac{2L(1-\beta)}{25}+\frac{18\mu(1-\beta)}{25(1+\frac{8\mu}{L})}\right) \\
&\leq\frac{6\sqrt{\beta}}{25(1-\beta)}\left(2L+\frac{18\mu}{1+\frac{8\mu}{L}}\right) \\
&\leq\frac{6\sqrt{\beta}}{25(1-\beta)}\left(2L+18\mu\right)
\end{aligned}
\tag{42}
$$

Therefore, by $\alpha\leq\frac{1-\beta}{L\left(3-\beta+2\beta^2+\frac{48\sqrt{\beta}}{25}\frac{2L+18\mu}{L}\right)}$ we have

$$
\begin{aligned}
&-\alpha+\frac{3-\beta+2\beta^2}{2(1-\beta)}L\alpha^2+4c_1\alpha^2 \\
&=-\frac{\alpha}{2}-\frac{\alpha}{2}+\frac{3-\beta+2\beta^2}{2(1-\beta)}L\alpha^2+\frac{24\sqrt{\beta}}{25(1-\beta)}\left(2L+18\mu\right)\alpha^2 \\
&\leq-\frac{\alpha}{2}.
\end{aligned}
\tag{43}
$$

Combine (43) with (36), we have

$$
\begin{aligned}
\mathbb{E}[L^{k+1}-L^k]\leq&-\frac{\alpha}{2}\mathbb{E}[\|g^k\|^2]+\left(\frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2+\frac{1}{2}L\alpha^2\sigma^2+2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2\right) \\
&+\sum_{i=1}^{k-1}(c_{i+1}-c_i)\mathbb{E}[\|x^{k+1-i}-x^{k-i}\|^2] \\
&+\left(4c_1\alpha^2+\frac{L\alpha^2}{(1-\beta)}\right)(1-\beta^k)^2\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i-g^k\|^2].
\end{aligned}
\tag{44}
$$

By combining (44) with (41), we further obtain

$$
\begin{aligned}
\mathbb{E}[L^{k+1}-L^k]\leq&B_1\mathbb{E}[f(z^k)-f^\star]+B_2 \\
&+B_3\mathbb{E}[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i-g^k\|^2]+\sum_{i=1}^{k-1}(c_{i+1}-c_i)\mathbb{E}[\|x^{k+1-i}-x^{k-i}\|^2],
\end{aligned}
\tag{45}
$$

where

$$
\begin{aligned}
B_1 &= -\frac{\alpha}{2}\frac{2\mu}{1+\frac{8\mu}{L}}, \\
B_2 &= \frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2 \\
&\quad + \frac{\alpha}{2}\frac{2\mu\left(\alpha\frac{\beta^2}{1-\beta} + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2\right)2\frac{1-\beta}{1+\beta}\sigma^2}{1+\frac{8\mu}{L}}, \\
B_3 &= 4c_1\alpha^2 + \frac{L\alpha^2}{(1-\beta)} \\
&\quad + \frac{\alpha}{2}\frac{2\mu\left(\alpha\frac{1}{2(1-\beta)} + \left(\alpha\frac{\beta^2}{1-\beta} + \frac{L\alpha^2}{2}(\frac{\beta}{1-\beta})^2\right)4\frac{1}{\beta^2} + \alpha\frac{1}{1-\beta}\right)}{1+\frac{8\mu}{L}}.
\end{aligned} \tag{46}
$$

From Lemma 2 we know that

$$
\mathbb{E}\left[\|\frac{1}{1-\beta^k}(1-\beta)\sum_{i=1}^{k}\beta^{k-i}g^i - g^k\|^2\right] \le \sum_{i=1}^{k-1} a_{k,i}\,\mathbb{E}[\|x^{i+1}-x^i\|^2],
$$

where

$$
a_{k,i} = \frac{L^2\beta^{k-i}}{1-\beta^k}\left(k-i+\frac{\beta}{1-\beta}\right). \tag{47}
$$

Putting this into (45) yields

$$
\begin{aligned}
\mathbb{E}[L^{k+1}-L^k] \le B_1\,\mathbb{E}[f(z^k)-f^\star] + B_2 \\
+ \sum_{i=1}^{k-1}(c_{i+1}-c_i+B_3a_{k,k-i})\,\mathbb{E}[\|x^{k+1-i}-x^{k-i}\|^2.
\end{aligned} \tag{48}
$$

In the rest of the proof, we will show that if the constants c_i are chosen such that

$$
c_1 = \left(\frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}}\frac{\beta}{1-\beta}\right)\left(\frac{4L^3\alpha^2}{1-\beta} + \frac{30L^2\mu\alpha^2}{(1-\beta)(1+\frac{8\mu}{L})}\right), \tag{49}
$$

and

$$
c_{i+1}-c_i+B_3 2L^2\beta^{k-i}\left(k-i+\frac{\beta}{1-\beta}\right) = B_1c_i, \qquad \forall i\ge 1. \tag{50}
$$

Then, we have $c_i>0$ for all $i\ge 1$ and

$$
c_{i+1}-c_i+B_3a_{k,k-i} \le B_1c_i, \qquad \forall i\ge 1. \tag{51}
$$

And therefore, we will have the desired result:

$$
\begin{aligned}
\mathbb{E}[L^{k+1}-L^k] &\le B_1\,\mathbb{E}[f(z^k)-f^\star] + B_2 + B_1\sum_{i=1}^{k-1}c_i\,\mathbb{E}[\|x^{k+1-i}-x^{k-i}\|^2 \\
&= -\frac{\alpha\mu}{1+\frac{8\mu}{L}}\,\mathbb{E}[L^k] + \frac{\beta^2}{2(1+\beta)}L\alpha^2\sigma^2 + \frac{1}{2}L\alpha^2\sigma^2 + 2c_1\frac{1-\beta}{1+\beta}\alpha^2\sigma^2 \\
&\quad + \frac{\beta^2+\frac{L\alpha}{2}\frac{\beta^2}{1-\beta}}{1+\frac{8\mu}{L}}\frac{2}{1+\beta}\mu\alpha^2\sigma^2.
\end{aligned}
$$

First of all. by $k\ge\frac{\log 0.5}{\log\beta}$, we know that $\beta^k\le\frac{1}{2}$, and (47) gives

$$
a_{k,k-i} \le 2L^2\beta^i\left(i+\frac{\beta}{1-\beta}\right).
$$

Therefore, in order for (51) to hold, it suffices to set

$$c_{i+1} - c_i + B_3 2L^2 \beta^{k-i} \left(k - i + \frac{\beta}{1-\beta} \right) = B_1 c_i \qquad \forall i \geq 1.$$

This is exactly (50).

On the other hand, (50) is also equivalent to

$$\frac{c_{i+1}}{(1+B_1)^{i+1}} - \frac{c_i}{(1+B_1)^i} = -\frac{2L^2 B_3}{(1+B_1)^{i+1}} \beta^i \left(i + \frac{\beta}{1-\beta} \right), \qquad \forall i \geq 1.$$

Therefore, in order to have $c_i > 0$ for all $i \geq 1$, we can set

$$c_1 \geq 2L^2 B_3 \sum_{i=1}^{\infty} \left(\frac{\beta}{1+B_1} \right)^i \left(i + \frac{\beta}{1-\beta} \right). \tag{52}$$

Since $\beta \leq \sqrt{\beta} \leq 1 + B_1 = 1 - \alpha\mu \frac{1}{1+\frac{8\mu}{L}}$ and

$$\sum_{i=1}^{j} i q^i = \frac{1}{1-q} \left(\frac{q(1-q^j)}{1-q} - j q^{j+1} \right),$$

for any $q \in (0,1)$, (52) is equivalent to

$$c_1 \geq 2L^2 B_3 \left(\frac{\frac{\beta}{1+B_1}}{(1 - \frac{\beta}{1+B_1})^2} + \frac{\frac{\beta}{1+B_1}}{1 - \frac{\beta}{1+B_1}} \frac{\beta}{1-\beta} \right). \tag{53}$$

Recall from (46) that

$$\begin{aligned} B_3 &= 4c_1\alpha^2 + \frac{L\alpha^2}{(1-\beta)} \\ &\quad + \frac{\alpha}{2} \frac{2\mu \left(\alpha \frac{1}{2(1-\beta)} + \left(\alpha \frac{\beta^2}{1-\beta} + \frac{L\alpha^2}{2} (\frac{\beta}{1-\beta})^2 \right) 4 \frac{1}{\beta^2} + \alpha \frac{1}{1-\beta} \right)}{1 + \frac{8\mu}{L}} \\ &= \left(4c_1\alpha^2 + \frac{L\alpha^2}{(1-\beta)} \right) + \frac{\alpha}{2} \frac{2\mu \left(\alpha \frac{11}{2(1-\beta)} + 2L\alpha^2 (\frac{1}{1-\beta})^2 \right)}{1 + \frac{8\mu}{L}}. \end{aligned}$$

Since $\alpha \leq \frac{1-\beta}{L}$, we further have

$$B_3 \leq \left(4c_1\alpha^2 + \frac{L\alpha^2}{(1-\beta)} \right) + \frac{\alpha}{2} \frac{2\mu \left(\alpha \frac{15}{2(1-\beta)} \right)}{1 + \frac{8\mu}{L}}.$$

Since $B_1 = -\frac{\alpha\mu}{1+\frac{8\mu}{L}}$ and $\alpha \leq \frac{1-\beta}{5L}$, it can be verified that $\frac{\beta}{1+B_1} \leq \sqrt{\beta}$ for all $\beta \in [0,1)$ and $\mu \leq L$. Therefore,

$$\frac{\frac{\beta}{1+B_1}}{(1 - \frac{\beta}{1+B_1})^2} + \frac{\frac{\beta}{1+B_1}}{1 - \frac{\beta}{1+B_1}} \frac{\beta}{1-\beta} \leq \frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta}.$$

As a result, in order to have (53), it suffices to set

$$c_1 \geq 2L^2 \left(\frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right) \left(4c_1\alpha^2 + \frac{L\alpha^2}{(1-\beta)} + \frac{\alpha}{2} \frac{2\mu \left(\alpha \frac{15}{2(1-\beta)} \right)}{1 + \frac{8\mu}{L}} \right), \tag{54}$$

Since $\alpha \leq \frac{1-\beta}{5L}$, we have

$$1 - 8\alpha^2 L^2 \left(\frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right) \geq \frac{1}{2},$$

(54) in turn just requires

$$c_1 = \left(\frac{\sqrt{\beta}}{(1-\sqrt{\beta})^2} + \frac{\sqrt{\beta}}{1-\sqrt{\beta}} \frac{\beta}{1-\beta} \right) \left(\frac{4L^3\alpha^2}{1-\beta} + \frac{30L^2\mu\alpha^2}{(1-\beta)(1+\frac{8\mu}{L})} \right),$$

which is exactly our choice of c_1 as in (49).

B.5 Proof of Theorem 2

From Proposition 2 we know that for all $k \geq k_0 = \lfloor \frac{\log 0.5}{\log \beta} \rfloor$,

$$\mathbb{E}[L^{k+1} - L^k] \leq -\frac{\alpha\mu}{1+\frac{8\mu}{L}} \mathbb{E}[L^k] + \frac{1+\beta+\beta^2}{2(1+\beta)} L\alpha^2\sigma^2 + \frac{1-\beta}{1+\beta} 2c_1\alpha^2\sigma^2 + \frac{\beta^2 + \frac{L\alpha}{2}\frac{\beta^2}{1-\beta}}{(1+\frac{8\mu}{L})(1+\beta)} 2\mu\alpha^2\sigma^2.$$

Rearranging gives

$$\begin{aligned}
\mathbb{E}[L^{k+1}] &\leq \left(1 - \frac{\alpha\mu}{1+\frac{8\mu}{L}}\right) \mathbb{E}[L^k] + \frac{1+\beta+\beta^2}{2(1+\beta)} L\alpha^2\sigma^2 + \frac{1-\beta}{1+\beta} 2c_1\alpha^2\sigma^2 \\
&\quad + \frac{\beta^2 + \frac{L\alpha}{2}\frac{\beta^2}{1-\beta}}{(1+\frac{8\mu}{L})(1+\beta)} 2\mu\alpha^2\sigma^2 \\
&\leq \left(1 - \frac{\alpha\mu}{1+\frac{8\mu}{L}}\right) \mathbb{E}[L^k] + \frac{1+\beta+\beta^2}{2(1+\beta)} L\alpha^2\sigma^2 + \frac{1-\beta}{1+\beta} 2c_1\alpha^2\sigma^2 \\
&\quad + \frac{\beta^2 + \frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}} \frac{2}{1+\beta} \mu\alpha^2\sigma^2,
\end{aligned}$$

where we have applied $\alpha \leq \frac{1-\beta}{5L}$ in the last step. Therefore,

$$\begin{aligned}
&\mathbb{E}[L^{k+1}] - \frac{1}{\frac{\alpha\mu}{1+\frac{8\mu}{L}}} \left(\frac{1+\beta+\beta^2}{2(1+\beta)} L\alpha^2\sigma^2 + \frac{1-\beta}{1+\beta} 2c_1\alpha^2\sigma^2 + \frac{\beta^2 + \frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}} \frac{2}{1+\beta}\mu\alpha^2\sigma^2 \right) \\
&\leq \left(1 - \frac{\alpha\mu}{1+\frac{8\mu}{L}}\right) \\
&\quad \times \left(\mathbb{E}[L^k] - \frac{1}{\frac{\alpha\mu}{1+\frac{8\mu}{L}}} \left(\frac{1+\beta+\beta^2}{2(1+\beta)} L\alpha^2\sigma^2 + \frac{1-\beta}{1+\beta} 2c_1\alpha^2\sigma^2 + \frac{\beta^2 + \frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}} \frac{2}{1+\beta}\mu\alpha^2\sigma^2 \right)\right).
\end{aligned}$$

This immediately yields

$$\begin{aligned}
&\mathbb{E}[L^k] \\
&\leq \left(1 - \frac{\alpha\mu}{1+\frac{8\mu}{L}}\right)^{k-k_0} \\
&\quad \times \left(\mathbb{E}[L^{k_0}] - \frac{1}{\frac{\alpha\mu}{1+\frac{8\mu}{L}}} \left(\frac{1+\beta+\beta^2}{2(1+\beta)} L\alpha^2\sigma^2 + \frac{1-\beta}{1+\beta} 2c_1\alpha^2\sigma^2 + \frac{\beta^2 + \frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}} \frac{2}{1+\beta}\mu\alpha^2\sigma^2 \right)\right) \\
&\quad + \frac{1}{\frac{\alpha\mu}{1+\frac{8\mu}{L}}} \left(\frac{1+\beta+\beta^2}{2(1+\beta)} L\alpha^2\sigma^2 + \frac{1-\beta}{1+\beta} 2c_1\alpha^2\sigma^2 + \frac{\beta^2 + \frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}} \frac{2}{1+\beta}\mu\alpha^2\sigma^2 \right) \\
&\leq \left(1 - \frac{\alpha\mu}{1+\frac{8\mu}{L}}\right)^{k-k_0} \mathbb{E}[L^{k_0}] \\
&\quad + \left(1 + \frac{8\mu}{L}\right) \left(\frac{1+\beta+\beta^2}{4(1+\beta)} \frac{L}{\mu} \alpha\sigma^2 + \frac{1-\beta}{1+\beta} \frac{2c_1}{\mu} \alpha\sigma^2 + \frac{\beta^2 + \frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}} \frac{2}{1+\beta} \alpha\sigma^2 \right).
\end{aligned}$$

By $c_i \geq 0$ for all $i \geq 1$ and (42), we conclude that

$$
\begin{aligned}
&\mathbb{E}[f(z^k) - f^*] \\
&\leq \left(1 - \frac{\alpha\mu}{1 + \frac{8\mu}{L}}\right)^{k-k_0} \mathbb{E}[L^{k_0}] \\
&\quad + \left(1 + \frac{8\mu}{L}\right)\left(\frac{1+\beta+\beta^2}{2(1+\beta)}\frac{L}{\mu}\alpha\sigma^2 + \frac{1}{1+\beta}\frac{12\sqrt{\beta}}{25}\frac{2L+18\mu}{\mu}\alpha\sigma^2 + \frac{\beta^2 + \frac{L\alpha}{10}\beta^2}{1+\frac{8\mu}{L}}\frac{2}{1+\beta}\alpha\sigma^2\right) \\
&= \mathcal{O}\left((1-\alpha\mu)^{k-k_0} + \frac{L}{\mu}\alpha\sigma^2\right).
\end{aligned}
$$

B.6 Proof of Corollary 1

In fact, by (6) we can express x^k as a convex combination of $\{z^i\}_{i=1}^k$:

$$
x^k = (1-\beta)\sum_{i=2}^{k}\beta^{k-i}z^i + \beta^{k-1}z^1.
$$

The desired result follows directly from the convexity of f and Theorem 2.

C Generalizations of Lemmas 1, 2, and 3 for Multistage SGDM

In order to establish the convergence of Multistage SGDM(Algorithm 1), we need to generalize the Lemmas 1 and 2 , which play a key role in the convergence of SGDM in (2).

C.1 Generalization of Lemma 1 for Multistage SGDM

Lemma 4. *Under the assumptions of Theorem 3, the variance of update vector m^k in Algorithm 1 satisfies*

$$
\frac{1}{1-\beta_1}\mathbb{E}[\|m^k - \sum_{i=1}^{k} b_{k,i}g^i\|^2] \leq 2\sigma^2,
$$

where $b_{k,i} = \big(1-\beta(i)\big)\prod_{j=i+1}^{k}\beta(j)$.

Proof. To begin with, let us express m^k by the past stochastic gradients:

$$
\begin{aligned}
m^k &= \beta(k)m^{k-1} + \big(1-\beta(k)\big)\tilde{g}^k \\
&= \beta(k)\beta(k-1)m^{k-2} + \beta(k)\big(1-\beta(k-1)\big)\tilde{g}^{k-1} \\
&\qquad\qquad + \cdots + \big(1-\beta(k)\big)\tilde{g}^k \\
&= \ldots \\
&= \prod_{i=1}^{k}\beta(i)m^0 + \prod_{i=2}^{k}\beta(i)\big(1-\beta(1)\big)\tilde{g}^1 \\
&\qquad\qquad + \cdots + \big(1-\beta(k)\big)\tilde{g}^k \\
&= \sum_{i=1}^{k} b_{k,i}\tilde{g}^i,
\end{aligned}
\tag{55}
$$

where we have applied $m^0 = 0$ and defined

$$
b_{k,i} = \big(1-\beta(i)\big)\prod_{j=i+1}^{k}\beta(j) \tag{56}
$$

in the last step.

It can be verified that the sum of weights is

$$\sum_{i=1}^{k} b_{k,i} = 1 - \prod_{i=1}^{k} \beta(i). \tag{57}$$

As a result, by applying Assumption 1 we have

$$\mathbb{E}[\|m^k - \sum_{i=1}^{k} b_{k,i} g^i\|^2] = \mathbb{E}[\|\sum_{i=1}^{k} b_{k,i}(\tilde{g}^i - g^i)\|^2] \leq \sum_{i=1}^{k} b_{k,i}^2 \sigma^2.$$

Note that by setting $k = T_1 + \cdots + T_{n_k} + r_k$, we have

$$b_{k,i} = \begin{cases} \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} ... \beta_2^{T_2}(1-\beta_1)\beta_1^{T_1-i}, 1 \leq i \leq T_1, \\ \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} ... \beta_3^{T_3}(1-\beta_2)\beta_1^{T_1+T_2-i}, T_1 + 1 \leq i \leq T_1 + T_2, \\ ..... \\ (1-\beta_{n_k+1})\beta_1^{T_1+\cdots+T_{n_k}+r_k-i}, \sum_{l=1}^{n_k} T_l + 1 \leq i \leq \sum_{l=1}^{n_k} T_l + r_k. \end{cases}$$

Therefore,

$$\begin{aligned} \mathbb{E}[\|m^k - \sum_{i=1}^{k} b_{k,i} g^i\|^2] \leq & (\beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} ... \beta_2^{T_2})^2 \frac{1-\beta_1}{1+\beta_1}(1-\beta_1^{2T_1})\sigma^2 \\ & + (\beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} ... \beta_3^{T_3})^2 \frac{1-\beta_2}{1+\beta_2}(1-\beta_2^{2T_2})\sigma^2 \\ & + \ldots \\ & + (\beta_{n_k+1}^{r_k})^2 \frac{1-\beta_{n_k}}{1+\beta_{n_k}}(1-\beta_{n_k}^{2T_{n_k}})\sigma^2 \\ & + \frac{1-\beta_{n_k+1}}{1+\beta_{n_k+1}}(1-\beta_{n_k+1}^{2r_k})\sigma^2. \end{aligned}$$

Since for any $l \in [1, n]$, we have

$$\begin{aligned} (\beta_l^{T_l})^2 &\leq \frac{1}{2}, \\ 1 - \beta_l &\leq 1 - \beta_1, \\ 1 + \beta_l &\geq \frac{3}{2}, \\ 1 - \beta_l^{2T_l} &< 1. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{1}{1-\beta_1}\mathbb{E}[\|m^k - \sum_{i=1}^{k} b_{k,i} g^i\|^2] \leq & (\beta_{n_k+1}^{r_k})^2 (\frac{1}{2})^{n_k-1} \frac{2}{3} \cdot \frac{1-\beta_1}{1-\beta_1}\sigma^2 \\ & + (\beta_{n_k+1}^{r_k})^2 (\frac{1}{2})^{n_k-2} \frac{2}{3} \cdot \frac{1-\beta_2}{1-\beta_1}\sigma^2 \\ & + \ldots \\ & + (\beta_{n_k+1}^{r_k})^2 (\frac{1}{2})^{0} \frac{2}{3} \cdot \frac{1-\beta_{n_k}}{1-\beta_1}\sigma^2 \\ & + \frac{2}{3} \cdot \frac{1-\beta_{n_k+1}}{1-\beta_1}\sigma^2 \\ \leq & 2\sigma^2. \end{aligned}$$

□

Lemma 5. *Under the assumptions of Theorem 3, the update vector m^{k-1} in Algorithm 1 satisfies*

$$\frac{1}{1-\beta(k)}\mathbb{E}[\|m^{k-1}-\sum_{i=1}^{k-1}b_{k-1,i}g^i\|^2] \leq 24\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}\sigma^2.$$

Proof. By setting $k-1 = T_1+\cdots+T_{n_{k-1}}+r_{k-1}$, we have

$$\begin{aligned}\mathbb{E}[\|m^{k-1}-\sum_{i=1}^{k-1}b_{k-1,i}g^i\|^2] \leq{}& (\beta_{n_{k-1}+1}^{r_{k-1}}\beta_{n_{k-1}}^{T_{n_{k-1}}}...\beta_2^{T_2})^2\frac{1-\beta_1}{1+\beta_1}(1-\beta_1^{2T_1})\sigma^2\\ &+(\beta_{n_{k-1}+1}^{r_{k-1}}\beta_{n_{k-1}}^{T_{n_{k-1}}}...\beta_3^{T_3})^2\frac{1-\beta_2}{1+\beta_2}(1-\beta_2^{2T_2})\sigma^2\\ &+\dots\\ &+(\beta_{n_{k-1}+1}^{r_{k-1}})^2\frac{1-\beta_{n_{k-1}}}{1+\beta_{n_{k-1}}}(1-\beta_{n_{k-1}}^{2T_{n_{k-1}}})\sigma^2\\ &+\frac{1-\beta_{n_{k-1}+1}}{1+\beta_{n_{k-1}+1}}(1-\beta_{n_{k-1}+1}^{2r_{k-1}})\sigma^2.\end{aligned}$$

Similar as before, we have

$$\begin{aligned}\frac{1}{1-\beta(k)}\mathbb{E}[\|m^{k-1}-\sum_{i=1}^{k}b_{k-1,i}g^i\|^2] \leq{}& (\beta_{n_{k-1}+1}^{r_{k-1}})^2(\frac{1}{2})^{n_{k-1}-1}\frac{2}{3}\cdot\frac{1-\beta_1}{1-\beta(k)}\sigma^2\\ &+(\beta_{n_{k-1}+1}^{r_{k-1}})^2(\frac{1}{2})^{n_{k-1}-2}\frac{2}{3}\cdot\frac{1-\beta_1}{1-\beta(k)}\sigma^2\\ &+\dots\\ &+(\beta_{n_{k-1}+1}^{r_{k-1}})^2(\frac{1}{2})^{0}\frac{2}{3}\cdot\frac{1-\beta_1}{1-\beta(k)}\sigma^2\\ &+\frac{2}{3}\cdot\frac{1-\beta_1}{1-\beta(k)}\sigma^2\\ \leq{}& 2\frac{1-\beta_1}{1-\beta(k)}\sigma^2.\end{aligned}$$

Finally, by applying

$$\frac{1-\beta_1}{1-\beta_n} \leq 12\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}},$$

we arrive at

$$\frac{1}{1-\beta(k)}\mathbb{E}[\|m^{k-1}-\sum_{i=1}^{k}b_{k-1,i}g^i\|^2] \leq 24\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}\sigma^2.$$

□

C.2 Generalization of Lemma 3 for Multistage SGDM

Lemma 6. *z^k defined in* (8) *satisfies*

$$z^{k+1}-z^k = -\alpha(k)\tilde{g}^k,$$

where $\alpha(k)$ is the stepsize applied at the kth step.

Proof. Recall that the auxiliary sequence z^k is defined by

$$z^k = x^k - A_1 m^{k-1},$$

where $A_1 \equiv \frac{\alpha_i \beta_i}{1-\beta_i}$ and α_i, β_i are the stepsize and momentum weight at the ith stage, respectively. Therefore, we also have

$$A_1 \equiv \frac{\alpha(k)\beta(k)}{1-\beta(k)},$$

where $\alpha(k), \beta(k)$ are the stepsize and momentum weight applied at the kth step. Using this, we obtain

$$\begin{aligned}
z^{k+1} - z^k &= x^{k+1} - x^k - A_1(m^k - m^{k-1}) \\
&= -\alpha(k)m^k - A_1(1-\beta(k))(\tilde{g}^k - m^{k-1}) \\
&= -\alpha(k)m^k - \alpha(k)\beta(k)(\tilde{g}^k - m^{k-1}) \\
&= \alpha(k)(\beta(k)m^{k-1} - m^k) - \alpha(k)\beta(k)\tilde{g}^k \\
&= -\alpha(k)\tilde{g}^k.
\end{aligned}$$

□

C.3 Generalization of Lemma 2 for Multistage SGDM

Lemma 7. *In Multistage SGDM(Algorithm 1), assume that the momentum weights at n stages satisfy $\beta_1 \le \beta_2 \le ... \le \beta_n$. Then, we have*

$$\mathbb{E}\left[\left\|\frac{1}{1-\prod_{i=1}^k \beta(i)} \sum_{i=1}^k b_{k,i} g^i - g^k\right\|^2\right] \le \sum_{i=1}^{k-1} a_{k,i}\, \mathbb{E}[\|x^{j+1} - x^j\|^2],$$

where $b_{k,i} = \big(1-\beta(i)\big)\prod_{j=i+1}^k \beta(j)$ and $\beta(i)$ is the momentum weight applied at the ith iteration, and

$$a_{k,i} = \frac{L^2 \beta^{k-i}(k)}{1-\prod_{i=1}^k \beta(i)} \left(k - i + \frac{\beta(k)}{1-\beta(k)}\right). \tag{58}$$

Proof. By By (55), (56) and (57), we can compute that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{1-\prod_{i=1}^k \beta(i)} \sum_{i=1}^k b_{k,i} g^i - g^k\right\|^2\right] \\
&= \mathbb{E}\,\|\frac{1}{1-\prod_{j=1}^k \beta(j)} \sum_{i=1}^k b_{k,i}(g^i - g^k)\|^2 \\
&= \left(\frac{1}{1-\prod_{j=1}^k \beta(j)}\right)^2 \sum_{i,j=1}^k b_{k,i} b_{k,j}\, \mathbb{E}\langle (g^k - g^i), (g^k - g^j)\rangle \\
&\le \left(\frac{1}{1-\prod_{j=1}^k \beta(j)}\right)^2 \sum_{i,j=1}^k b_{k,i} b_{k,j} (\frac{1}{2}\,\mathbb{E}\,\|g^k - g^i\|^2 \\
&\qquad\qquad + \frac{1}{2}\,\mathbb{E}\,\|g^k - g^j\|^2) \\
&= \left(\frac{1}{1-\prod_{j=1}^k \beta(j)}\right) \sum_{j=1}^k b_{k,j}\, \mathbb{E}\,\|g^k - g^j\|^2 \\
&\le \left(\frac{1}{1-\prod_{j=1}^k \beta(j)}\right) \sum_{j=1}^k b_{k,j}(k-j) \sum_{i=j}^{k-1} L^2\, \mathbb{E}\,\|x^{i+1} - x^i\|^2,
\end{aligned}$$

where we have used (57) in the first and third equality, and Cauchy-Schwarz in the first inequality. In the last inequality, we have applied the triangle inequality and the $L-$smoothness of f.

Consequently, we have

$$
\begin{aligned}
&\mathbb{E}\left\|\frac{1}{1-\prod_{j=1}^{k}\beta(j)}\sum_{i=1}^{k} b_{k,i}g^i - g^k\right\|^2 \\
&\leq \left(\frac{1}{1-\prod_{j=1}^{k}\beta(j)}\right)\sum_{j=1}^{k} b_{k,j}(k-j)\sum_{i=j}^{k-1} L^2\,\mathbb{E}\,\|x^{i+1}-x^i\|^2 \\
&= \left(\frac{1}{1-\prod_{j=1}^{k}\beta(j)}\right)\sum_{i=1}^{k-1}\sum_{j=1}^{i} b_{k,j}(k-j)L^2\,\mathbb{E}\,\|x^{i+1}-x^i\|^2 \\
&= \sum_{i=1}^{k-1} d_{k,i}\,\mathbb{E}[\|x^{i+1}-x^i\|^2],
\end{aligned}
$$

where in the last step we have defined

$$
d_{k,i} = \left(\frac{L^2}{1-\prod_{j=1}^{k}\beta(j)}\right)\sum_{j=1}^{i}(k-j)b_{k,j}. \tag{59}
$$

In the Proposition 5 below, we shall see that $d_{k,i} \leq a_{k,i}$ for all $i \leq k-1$, where $a_{k,i}$ is defined in (58). Therefore,

$$
\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\sum_{i=1}^{k} b_{k,i}g^i - g^k\|^2] \leq \sum_{i=1}^{k-1} a_{k,i}\,\mathbb{E}[\|x^{j+1}-x^j\|^2],
$$

and the proof will be complete.

□

Proposition 5. *$d_{k,i}$ defined in* (59) *and $a_{k,i}$ defined in* (58) *satisfy*

$$
d_{k,i} \leq a_{k,i} \quad \textit{for all} \quad i \leq k-1.
$$

Proof. We aim to show that $d_{k,i} \leq a_{k,i}$ for all $i \leq k-1$. Or equivalently, $d_{k,j} \leq a_{k,j}$ for all $j \leq k-1$.

In order to show $d_{k,j} \leq a_{k,j}$, we just need to show that

$$
\sum_{i=1}^{j}(k-i)b_{k,i} \leq \beta^{k-j}(k)\left(k-j+\frac{\beta(k)}{1-\beta(k)}\right), \tag{60}
$$

where

$$
b_{k,i} = \big(1-\beta(i)\big)\prod_{j=i+1}^{k}\beta(j).
$$

Let $k = T_1 + T_2 + \cdots + T_{n_k} + r_k$, where $0 \leq n_k \leq n-1$. If $n_k < n-1$, then $0 \leq r_k \leq T_{n_k+1}-1$. If $n_k = n-1$, then $0 \leq r_k \leq T_{n_k+1} = T_n$.

Since $j \leq k-1$, we have $j = T_1 + \cdots + T_{n_j} + r_j$, where $0 \leq n_j \leq n_k$.

Now, let us compute the left hand side of (60) explicitly.

$$
\begin{aligned}
&\sum_{i=1}^{j}(k-i)b_{k,i} \\
&= \left(\sum_{i=1}^{T_1} + \sum_{i=T_1+1}^{T_1+T_2} + \cdots + \sum_{i=T_1+\cdots+T_{n_j}+1}^{T_1+\cdots+T_{n_j}+r_j}\right)(k-i)b_{k,i}.
\end{aligned}
$$

Notice that

$$b_{k,i} = \begin{cases} \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} \cdots \beta_2^{T_2} (1-\beta_1) \beta_1^{T_1 - i}, & 1 \le i \le T_1, \\ \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} \cdots \beta_3^{T_3} (1-\beta_2) \beta_1^{T_1+T_2-i}, & \\ \qquad\qquad T_1 + 1 \le i \le T_1 + T_2, & \\ \ldots\ldots & \\ (1-\beta_{n_k+1}) \beta_1^{T_1+\cdots+T_{n_k}+r_k-i}, & \\ \qquad\qquad \sum_{l=1}^{n_k} T_l + 1 \le i \le \sum_{l=1}^{n_k} T_l + r_k. & \end{cases}$$

As a result, we have

$$\begin{aligned} & \sum_{i=1}^{j} (k-i) b_{k,i} \\ & = \left(\sum_{i=1}^{T_1} + \sum_{i=T_1+1}^{T_1+T_2} + \cdots + \sum_{i=T_1+\cdots+T_{n_j}+1}^{T_1+\cdots+T_{n_j}+r_j} \right) (k-i) b_{k,i} \\ & \le \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} \cdots \beta_2^{T_2} (1-\beta_1) \sum_{i=1}^{T_1} \beta_1^{T_1-i} (k-i) \\ & \quad + \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} \cdots \beta_3^{T_3} (1-\beta_2) \sum_{i=T_1+1}^{T_1+T_2} \beta_2^{T_1+T_2-i} (k-i) \\ & \quad + \ldots \\ & \quad + \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} ... \beta_{n_j+1}^{T_{n_j+1}} (1-\beta_{n_j}) \\ & \qquad \sum_{i=T_1+\cdots+T_{n_j-1}+1}^{T_1+\cdots+T_{n_j}} \beta_{n_l}^{T_1+\cdots+T_{n_j}-i} (k-i) \\ & \quad + \beta_{n_k+1}^{r_k} \beta_{n_k}^{T_{n_k}} ... \beta_{n_j+2}^{T_{n_j}+2} (1-\beta_{n_j+1}) \\ & \qquad \sum_{i=T_1+\cdots+T_{n_j}+1}^{T_1+\cdots+T_{n_j}+r_j} \beta_{n_j+1}^{T_1+\cdots+T_{n_j}+r_j-i} (k-i), \end{aligned} \tag{61}$$

where we have applied $r_j \le T_{n_j+1}$ if $n_j < n_k$ and $r_j \le r_k$ if $n_j = n_k$ in the last term. Since

$$\begin{aligned} \sum_{i=1}^{l} \beta^{k-i} (k-i) = \beta^k \left(-\frac{k-1}{1-\beta} - \frac{1}{(1-\beta)^2} \right) \\ + \beta^{k-l} \left(\frac{k-l}{1-\beta} + \frac{\beta}{(1-\beta)^2} \right). \end{aligned}$$

we have

$$
\begin{aligned}
\sum_{i=1}^{T_1} \beta_1^{T_1-i}(k-i) &= \beta_1^{T_1-k} \sum_{i=1}^{T_1} \beta_1^{k-i}(k-i) \\
&= \beta_1^{T_1} \left(-\frac{k-1}{1-\beta_1} - \frac{1}{(1-\beta_1)^2} \right) \\
&\quad + \left(\frac{k-T_1}{1-\beta_1} + \frac{\beta_1}{(1-\beta_1)^2} \right), \\
\sum_{i=T_1+1}^{T_1+T_2} \beta_2^{T_1+T_2-i}(k-i) &= \sum_{i=1}^{T_2} \beta_2^{T_2-i}(k-T_1-i) \\
&= \beta_2^{T_1+T_2-k} \sum_{i=1}^{T_2} \beta_2^{k-T_1-i}(k-T_1-i) \\
&= \beta_2^{T_2} \left(-\frac{k-T_1-1}{1-\beta_2} - \frac{1}{(1-\beta_2)^2} \right) \\
&\quad + \left(\frac{k-T_1-T_2}{1-\beta_2} + \frac{\beta_2}{(1-\beta_2)^2} \right).
\end{aligned}
$$

And that in general

$$
\begin{aligned}
&\sum_{i=T_1+\cdots+T_{n_j}+1}^{T_1+\cdots+T_{n_j}+r_j} \beta_{n_j+1}^{T_1+\cdots+T_{n_j}+r_j-i}(k-i) \\
&= \sum_{i=1}^{r_j} \beta_{n_j+1}^{r_j-i}(k-T_1-\cdots-T_{n_j}-i) \\
&= \beta_{n_j+1}^{T_1+\cdots+T_{n_j}+r_j-k} \sum_{i=1}^{r_j} \beta_{n_j+1}^{k-T_1-\cdots-T_{n_j}-i}(k-T_1-\cdots-T_{n_j}-i) \\
&= \beta_{n_j+1}^{r_j} \left(-\frac{k-T_1-\cdots-T_{n_j}-1}{1-\beta_{n_j+1}} - \frac{1}{(1-\beta_{n_j+1})^2} \right) \\
&\quad + \left(\frac{k-T_1-\cdots-T_{n_j}-r_j}{1-\beta_{n_j+1}} + \frac{\beta_{n_j+1}}{(1-\beta_{n_j+1})^2} \right).
\end{aligned}
$$

By applying these equalities on (61), we have

$$
\begin{aligned}
&\sum_{i=1}^{j}(k-i)b_{k,i}\\
&=\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_2^{T_2}\left(\beta_1^{T_1}\left(-(k-1)-\frac{1}{1-\beta_1}\right)+\left((k-T_1)+\frac{\beta_1}{1-\beta_1}\right)\right)\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_3^{T_3}\left(\beta_2^{T_2}\left(-(k-T_1-1)-\frac{1}{1-\beta_2}\right)\right.\\
&\qquad\qquad\qquad\qquad\left.+\left((k-T_1-T_2)+\frac{\beta_2}{1-\beta_2}\right)\right)\\
&\quad+\dots\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+1}^{T_{n_j+1}}\\
&\qquad\left(\beta_{n_j}^{T_{n_j}}\left(-(k-T_1-\dots-T_{n_j-1}-1)-\frac{1}{1-\beta_{n_j}}\right)\right.\\
&\qquad\qquad\left.+\left((k-T_1-\dots-T_{n_j})+\frac{\beta_{n_j}}{1-\beta_{n_j}}\right)\right)\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+2}^{T_{n_j+2}}\\
&\qquad\left(\beta_{n_j+1}^{r_j}\left(-(k-T_1-\dots-T_{n_j}-1)-\frac{1}{1-\beta_{n_j+1}}\right)\right.\\
&\qquad\qquad\left.+\left((k-T_1-T_{n_j}-r_j)+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right)\right).
\end{aligned}
$$

This yields

$$
\begin{aligned}
\sum_{i=1}^{j}(k-i)b_{k,i}&=\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_2^{T_2}\beta_1^{T_1}\left(-(k-1)-\frac{1}{1-\beta_1}\right)\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_2^{T_2}\left(\frac{\beta_1}{1-\beta_1}+1-\frac{1}{1-\beta_2}\right)\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_3^{T_3}\left(\frac{\beta_2}{1-\beta_2}+1-\frac{1}{1-\beta_3}\right)\\
&\quad+\dots\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j}^{T_{n_j}}\left(\frac{\beta_{n_j-1}}{1-\beta_{n_j-1}}+1-\frac{1}{1-\beta_{n_j}}\right)\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+1}^{T_{n_j+1}}\left(k-T_1-\dots-T_{n_j}+\frac{\beta_{n_j}}{1-\beta_{n_j}}\right)\\
&\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+2}^{T_{n_j+2}}\\
&\qquad\cdot\left(\beta_{n_j+1}^{r_j}\left(-(k-T_1-\dots-T_{n_j}-1)-\frac{1}{1-\beta_{n_j+1}}\right)\right.\\
&\qquad\qquad\left.+\left((k-T_1-T_{n_j}-r_j)+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right)\right).
\end{aligned}
$$

On the right hand side, the first n_j terms are non-positive since $\beta_1 \leq \beta_2 \leq ... \leq \beta_n$. Therefore,

$$\begin{aligned}\sum_{i=1}^{j}(k-i)b_{k,i} \leq \beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+1}^{T_{n_j+1}}\left(k-T_1-\cdots-T_{n_j}+\frac{\beta_{n_j}}{1-\beta_{n_j}}\right)\\ +\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+2}^{T_{n_j+2}}\\ \left(\beta_{n_j+1}^{r_j}\left(-(k-T_1-\cdots-T_{n_j}-1)-\frac{1}{1-\beta_{n_j+1}}\right)\right.\\ \left.+\left((k-T_1-T_{n_j}-r_j)+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right)\right).\end{aligned}$$

By applying $\beta_{n_j+1}^{r_j} \geq \beta_{n_j+1}^{T_{n_j}+1}$ and $k-T_1-\cdots-T_{n_j}-1 = k-(j-r_j)-1 \geq 0$ (since $j \leq k-1$), we arrive at

$$\begin{aligned}\sum_{i=1}^{j}(k-i)b_{k,i} &\leq \beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+1}^{T_{n_j+1}}\left(k-T_1-\cdots-T_{n_j}+\frac{\beta_{n_j}}{1-\beta_{n_j}}\right)\\ &\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+2}^{T_{n_j+2}}\\ &\qquad\left(\beta_{n_j+1}^{T_{n_j}+1}\left(-(k-T_1-\cdots-T_{n_j}-1)-\frac{1}{1-\beta_{n_j+1}}\right)\right.\\ &\qquad\left.+\left((k-T_1-T_{n_j}-r_j)+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right)\right)\\ &\leq \beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+1}^{T_{n_j+1}}\left(\frac{\beta_{n_j}}{1-\beta_{n_j}}+1-\frac{1}{1-\beta_{n_j+1}}\right)\\ &\quad+\beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+2}^{T_{n_j+2}}\left(k-T_1-\cdots-T_{n_j}-r_j+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right)\\ &\leq \beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+2}^{T_{n_j+2}}\left(k-T_1-\cdots-T_{n_j}-r_j+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right)\\ &= \beta_{n_k+1}^{r_k}\beta_{n_k}^{T_{n_k}}...\beta_{n_j+2}^{T_{n_j+2}}\left(k-j+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right).\end{aligned}$$

Now let us consider two cases: $r_k > 0$ and $r_k = 0$.

1. $r_k > 0$.

   In this case, we apply $\beta_1 \leq ... \leq \beta_n$ to get

$$\begin{aligned}&\sum_{i=1}^{j}(k-i)b_{k,i}\\ &\leq \beta_{n_k+1}^{r_k+T_{n_k}+\cdots+T_{n_j+2}}\left(k-j+\frac{\beta_{n_k+1}}{1-\beta_{n_k+1}}\right).\end{aligned}$$

   Notice that

$$\begin{aligned}r_k+T_{n_k}+\cdots+T_{n_j+2} &= (T_1+\cdots+T_{n_k}+r_k)\\ &\quad-(T_1+\cdots+T_{n_j}+T_{n_j+1})\\ &\leq (T_1+\cdots+T_{n_k}+r_k)\\ &\quad-(T_1+\cdots+T_{n_j}+r_j)\\ &= k-j.\end{aligned}$$

This tells us that

$$\sum_{i=1}^{j}(k-i)b_{k,i} \le \beta_{n_k+1}^{k-j}\left(k-j+\frac{\beta_{n_k+1}}{1-\beta_{n_k+1}}\right).$$

Since $r_k > 0$, iteration k is at the $(n_k+1)-$th stage, we have $\beta(k) = \beta_{n_k+1}$, and the above inequality is exactly what we want to show in (60).

2. $r_k = 0$

In this case, we apply $\beta_1 \le ... \le \beta_n$ to get

$$\sum_{i=1}^{j}(k-i)b_{k,i} \le \beta_{n_k}^{T_{n_k}+\cdots+T_{n_j+2}}\left(k-j+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right).$$

Notice that

$$\begin{aligned} r_k + T_{n_k} + \cdots + T_{n_j+2} &= (T_1 + \cdots + T_{n_k} + r_k) \\ &\quad - (T_1 + \cdots + T_{n_j} + T_{n_j+1}) \\ &\le (T_1 + \cdots + T_{n_k} + r_k) \\ &\quad - (T_1 + \cdots + T_{n_j} + r_j) \\ &= k - j. \end{aligned}$$

This tells us that

$$\sum_{i=1}^{j}(k-i)b_{k,i} \le \beta_{n_k}^{k-j}\left(k-j+\frac{\beta_{n_j+1}}{1-\beta_{n_j+1}}\right),$$

Since $r_k = 0$, we have $\beta(k) = \beta_{n_k}$ and by $j \le k-1$ we deduce that $n_j \le n_k - 1$ (Otherwise $j = T_1 + \cdots + T_{n_j} + r_j = T_1 + \cdots + T_{n_k} + r_j \ge T_1 + \cdots + T_{n_k} = k$). Therefore, we have

$$\sum_{i=1}^{j}(k-i)b_{k,i} \le \beta_{n_k}^{k-j}\left(k-j+\frac{\beta_{n_k}}{1-\beta_{n_k}}\right),$$

which is exactly what we want to show in (60).

□

D Main Theory for Multistage SGDM

In this section, we prove the main convergence theory of Multistage SGDM.

D.1 Proof of Proposition 3

Proposition 3 is a generalization of Propositions 4 and 1 to the multistage case. Therefore, its proof is similar to those of Propositions 4 and 1.

First of all, by the smoothness of f we have

$$\begin{aligned} \mathbb{E}[f(z^{k+1})] &\le \mathbb{E}[f(z^k)] + \mathbb{E}\langle \nabla f(z^k), z^{k+1} - z^k\rangle + \frac{L}{2}\,\mathbb{E}\,\|z^{k+1} - z^k\|^2 \\ &= \mathbb{E}[f(z^k)] + \mathbb{E}\langle \nabla f(z^k), -\alpha(k)\tilde{g}^k\rangle + \frac{L\alpha^2(k)}{2}\,\mathbb{E}\,\|\tilde{g}^k\|^2, \end{aligned} \tag{62}$$

where we have applied Lemma 6 in the second step. Note that $\alpha(k)$ is the stepsize applied at the $k-$th iteration.

For the inner product term, we have

$$\mathbb{E}\langle \nabla f(z^k), -\alpha(k)\tilde{g}^k\rangle = \mathbb{E}\langle \nabla f(z^k), -\alpha(k)g^k\rangle,$$

which follows from the fact that z^k is determined by the previous $k-1$ random samples $\zeta^1, \zeta^2, ...\zeta^{k-1}$, which is independent of ζ^k, and $\mathbb{E}_{\zeta^k}[\tilde{g}^k] = g^k$.

As a result, we can write

$$\begin{aligned}\mathbb{E}\langle \nabla f(z^k), -\alpha(k)\tilde{g}^k\rangle &= \mathbb{E}\langle \nabla f(z^k) - g^k, -\alpha(k)g^k\rangle - \alpha(k)\,\mathbb{E}\,\|g^k\|^2 \\ &\leq \alpha(k)\frac{\rho_{0,k}}{2}L^2\,\mathbb{E}[\|z^k - x^k\|^2] + \alpha(k)\frac{1}{2\rho_{0,k}}\,\mathbb{E}[\|g^k\|^2] - \alpha(k)\,\mathbb{E}[\|g^k\|^2],\end{aligned} \tag{63}$$

where $\rho_{0,k} > 0$ can be any positive constant.

Combining (62) and (63) gives

$$\begin{aligned}\mathbb{E}[f(z^{k+1})] \leq{}& \mathbb{E}[f(z^k)] + \alpha(k)\frac{\rho_{0,k}}{2}L^2\,\mathbb{E}[\|z^k - x^k\|^2] \\ &+ \big(\alpha(k)\frac{1}{2\rho_{0,k}} - \alpha(k)\big)\,\mathbb{E}[\|g^k\|^2] + \frac{L\alpha^2(k)}{2}\,\mathbb{E}[\|\tilde{g}^k\|^2]\end{aligned}$$

By (8) we know that $z^k - x^k = -A_1 m^{k-1}$, which leads to

$$\begin{aligned}\mathbb{E}[f(z^{k+1})] \leq{}& \mathbb{E}[f(z^k)] + \alpha(k)\frac{\rho_{0,k}}{2}L^2 A_1^2\,\mathbb{E}[\|m^{k-1}\|^2] \\ &+ \big(\alpha(k)\frac{1}{2\rho_{0,k}} - \alpha(k)\big)\,\mathbb{E}[\|g^k\|^2] + \frac{L\alpha^2(k)}{2}(\sigma^2 + \mathbb{E}[\|g^k\|^2]).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[L^{k+1} - L^k] \leq{}& \alpha(k)\frac{\rho_{0,k}}{2}L^2 A_1^2\,\mathbb{E}[\|m^{k-1}\|^2] \\ &+ \left(\alpha(k)\frac{1}{2\rho_{0,k}} - \alpha(k) + \frac{L\alpha^2(k)}{2}\right)\mathbb{E}[\|g^k\|^2] + \frac{L\alpha^2(k)}{2}\sigma^2 \\ &+ c_1\alpha^2(k)\,\mathbb{E}[\|m^k\|^2] \\ &+ \sum_{i=1}^{k-1}(c_{i+1} - c_i)\,\mathbb{E}[\|x^{k+1-i} - x^{k-i}\|^2] \\ \leq{}& \alpha(k)\frac{\rho_{0,k}}{2}L^2 A_1^2\left(2\mathbb{E}[\|m^{k-1} - \sum_{i=1}^{k-1} b_{k-1,i}g^i\|^2] + 2\mathbb{E}[\|\sum_{i=1}^{k-1} b_{k-1,i}g^i\|^2]\right) \\ &+ \left(\alpha(k)\frac{1}{2\rho_{0,k}} - \alpha(k) + \frac{L\alpha^2(k)}{2}\right)\mathbb{E}[\|g^k\|^2] + \frac{L\alpha^2(k)}{2}\sigma^2 \\ &+ c_1\alpha^2(k)\left(2\mathbb{E}[\|m^k - \sum_{i=1}^{k} b_{k,i}g^i\|^2] + 2\mathbb{E}[\|\sum_{i=1}^{k} b_{k,i}g^i\|^2]\right) \\ &+ \sum_{i=1}^{k-1}(c_{i+1} - c_i)\,\mathbb{E}[\|x^{k+1-i} - x^{k-i}\|^2].\end{aligned} \tag{64}$$

On the other hand, we know that

$$\begin{aligned}\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^k \beta(i)}\sum_{i=1}^k b_{k,i}g^i\|^2] &= \mathbb{E}[\|\frac{1}{1-\prod_{i=1}^k \beta(i)}\sum_{i=1}^k b_{k,i}g^i - g^k + g^k\|^2] \\ &\leq 2\,\mathbb{E}[\|g^k\|^2] + 2\,\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^k \beta(i)}\sum_{i=1}^k b_{k,i}g^i - g^k\|^2].\end{aligned} \tag{65}$$

Furthermore,

$$
\begin{aligned}
&\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\sum_{i=1}^{k}b_{k,i}g^i - g^k\|^2] \\
&= \mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\beta(k)\sum_{i=1}^{k-1}b_{k-1,i}g^i + \frac{1-\beta(k)}{1-\prod_{i=1}^{k}\beta(i)}g^k - g^k\|^2] \\
&= \mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\beta(k)\sum_{i=1}^{k-1}b_{k-1,i}g^i - \beta(k)\frac{1-\prod_{i=1}^{k-1}\beta(i)}{1-\prod_{i=1}^{k}\beta(i)}g^k\|^2] \\
&= \beta^2(k)\left(\frac{1-\prod_{i=1}^{k-1}\beta(i)}{1-\prod_{i=1}^{k}\beta(i)}\right)^2 \mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k-1}\beta(i)}\sum_{i=1}^{k-1}b_{k-1,i}g^i - g^k\|^2].
\end{aligned}
\tag{66}
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k-1}\beta(i)}\sum_{i=1}^{k-1}b_{k-1,i}g^i\|^2] \\
&= \mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k-1}\beta(i)}\sum_{i=1}^{k-1}b_{k-1,i}g^i - g^k + g^k\|^2] \\
&\le 2\,\mathbb{E}[\|g^k\|^2] + 2\,\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k-1}\beta(i)}\sum_{i=1}^{k-1}b_{k-1,i}g^i - g^k\|^2] \\
&= 2\,\mathbb{E}[\|g^k\|^2] + 2\frac{1}{\beta^2(k)}\left(\frac{1-\prod_{i=1}^{k}\beta(i)}{1-\prod_{i=1}^{k-1}\beta(i)}\right)^2 \mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\sum_{i=1}^{k}b_{k,i}g^i - g^k\|^2].
\end{aligned}
\tag{67}
$$

Plugging (65) and (67) into (64) gives us

$$
\begin{aligned}
&\mathbb{E}[L^{k+1} - L^k] \\
&\le \left(-\alpha(k) + \alpha(k)\frac{1}{2\rho_{0,k}} + 2\alpha(k)\rho_{0,k}L^2A_1^2 + \frac{L\alpha^2(k)}{2} + 4c_1\alpha^2(k)\right)\mathbb{E}[\|g^k\|^2] \\
&\quad + \left(\alpha(k)\rho_{0,k}L^2A_1^2\mathbb{E}[\|m^{k-1} - \sum_{i=1}^{k-1}b_{k-1,i}g^i\|^2]) + \frac{1}{2}L\alpha^2(k)\sigma^2 + 2c_1\alpha^2(k)\mathbb{E}[\|m^k - \sum_{i=1}^{k}b_{k,i}g^i\|^2]\right) \\
&\quad + \sum_{i=1}^{k-1}(c_{i+1} - c_i)\,\mathbb{E}[\|x^{k+1-i} - x^{k-i}\|^2] \\
&\quad + 2\alpha(k)\rho_{0,k}L^2A_1^2\frac{1}{\beta^2(k)}\left(1-\prod_{i=1}^{k}\beta(i)\right)^2\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\sum_{i=1}^{k}b_{k,i}g^i - g^k\|^2] \\
&\quad + 4c_1\alpha^2(k)\left(1-\prod_{i=1}^{k}\beta(i)\right)^2\mathbb{E}[\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\sum_{i=1}^{k}b_{k,i}g^i - g^k\|^2]
\end{aligned}
\tag{68}
$$

In the rest of the proof, we will show that the sum of the last three terms in (68) is non-positive. First, by Lemma 7 we know that

$$
\mathbb{E}\,\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\sum_{i=1}^{k}b_{k,i}g^i - g^k\|^2 \le \sum_{i=1}^{k-1}a_{k,i}\,\mathbb{E}\,\|x^{i+1} - x^i\|^2,
$$

where

$$
a_{k,i} = \frac{L^2\beta^{k-i}(k)}{1-\prod_{i=1}^{k}\beta(i)}\left(k - i + \frac{\beta(k)}{1-\beta(k)}\right).
$$

Or equivalently,

$$\mathbb{E}\,\|\frac{1}{1-\prod_{i=1}^{k}\beta(i)}\sum_{i=1}^{k} b_{k,i}g^i - g^k\|^2 \le \sum_{i=1}^{k-1} a_{k,k-i}\,\mathbb{E}\,\|x^{k+1-i}-x^{k-i}\|^2,$$

where

$$a_{k,k-i} = \frac{L^2\beta^i(k)}{1-\prod_{i=1}^{k}\beta(i)}\left(i+\frac{\beta(k)}{1-\beta(k)}\right).$$

Therefore, in order to make the sum of the last three terms of (68) to be non-positive, we need to enforce that

$$c_{i+1} \le c_i - \left(4c_1\alpha^2(k)\big(1-\prod_{i=1}^{k}\beta(i)\big)^2 + 2\alpha(k)\rho_{0,k}L^2A_1^2\frac{1}{\beta^2(k)}\big(1-\prod_{i=1}^{k}\beta(i)\big)^2\right)a_{k,k-i}$$

for all $i \ge 1$ and $k \ge 1$.

Since $1-\prod_{i=1}^{k}\beta(i) < 1$, $\beta_1 \le \beta(k) \le \beta_n$, and $\alpha_1 \le \alpha(k) \le \alpha_n$, we need to enforce the following for all $i \ge 1$:

$$c_{i+1} \le c_i - \left(4c_1\alpha_1^2 + 2\alpha(k)\rho_{0,k}L^2A_1^2\frac{1}{\beta_1^2}\right)\beta_n^i(i+\frac{\beta_n}{1-\beta_n})L^2.$$

Recall that $\frac{\alpha_i\beta_i}{1-\beta_i} \equiv A_1$ for all n stages $i = 1, 2, ..., n$. This gives us

$$c_{i+1} \le c_i - \left(4c_1\alpha_1^2 + 2\alpha(k)\rho_{0,k}L^2\frac{\alpha_1^2}{(1-\beta_1)^2}\right)\beta_n^i(i+\frac{\beta_n}{1-\beta_n})L^2.$$

Let us also set

$$\rho_{0,k} = \frac{1-\beta(k)}{2L\alpha(k)}. \tag{69}$$

Then, we need to enforce

$$c_{i+1} \le c_i - \left(4c_1\alpha_1^2 + 2\frac{1-\beta(k)}{2}L\frac{\alpha_1^2}{(1-\beta_1)^2}\right)\beta_n^i(i+\frac{\beta_n}{1-\beta_n})L^2.$$

Since $\beta_1 \le \beta(k)$, it suffices to enforce that

$$c_{i+1} = c_i - \left(4c_1\alpha_1^2 + L\frac{\alpha_1^2}{(1-\beta_1)}\right)\beta_n^i(i+\frac{\beta_n}{1-\beta_n})L^2. \tag{70}$$

Note that the equalities in (70) does not depend on k. In order for $c_i > 0$ for all $i \ge 1$, we can determine c_1 by

$$c_1 = \left(4c_1\alpha_1^2 + L\frac{\alpha_1^2}{(1-\beta_1)}\right)\sum_{i=1}^{\infty}\beta_n^i(i+\frac{\beta_n}{1-\beta_n})L^2.$$

Since

$$\sum_{i=1}^{j} i\beta_n^i = \frac{1}{1-\beta_n}\left(\frac{\beta_n(1-\beta_n^j)}{1-\beta_n} - j\beta_n^{j+1}\right),$$

we have $\sum_{i=1}^{\infty} i\beta_n^i = \frac{\beta_n}{(1-\beta_n)^2}$ and

$$c_1 = \left(4c_1\alpha_1^2 + L\frac{\alpha_1^2}{(1-\beta_1)}\right)\frac{\beta_n+\beta_n^2}{(1-\beta_n)^2}L^2.$$

This stipulates that

$$c_1 = \frac{\frac{\alpha_1^2}{(1-\beta_1)}\frac{\beta_n+\beta_n^2}{(1-\beta_n)^2}L^3}{1-4\alpha_1^2\frac{\beta_n+\beta_n^2}{(1-\beta_n)^2}L^2}. \tag{71}$$

Notice that $A_1 = \frac{1}{24\sqrt{2}L}$ and $\frac{1-\beta_1}{\beta_1} \leq 12\frac{1-\beta_n}{\sqrt{\beta_n+\beta_n^2}}$ ensures

$$4L^2\alpha_1^2\frac{\beta_n+\beta_n^2}{(1-\beta_n)^2} \leq \frac{1}{2}$$

and therefore

$$0 < c_1 \leq 2\frac{\alpha_1^2}{(1-\beta_1)}\frac{\beta_n+\beta_n^2}{(1-\beta_n)^2}L^3 \leq \frac{L}{4(1-\beta_1)}. \tag{72}$$

With the choices of c_i in (70) and (71), the sum of the last three terms of (68) is non-positive. Therefore,

$$\begin{aligned}
&\mathbb{E}[L^{k+1}-L^k] \\
&\leq \Big(-\alpha(k)+\alpha(k)\frac{1}{2\rho_{0,k}}+2\alpha(k)\rho_{0,k}L^2A_1^2+\frac{L\alpha^2(k)}{2}+4c_1\alpha^2(k)\Big)\,\mathbb{E}[\|g^k\|^2] \\
&\quad+\Big(\alpha(k)\rho_{0,k}L^2A_1^2\mathbb{E}[\|m^{k-1}-\sum_{i=1}^{k-1}b_{k-1,i}g^i\|^2]+\frac{1}{2}L\alpha^2(k)\sigma^2+2c_1\alpha^2(k)\mathbb{E}[\|m^k-\sum_{i=1}^{k}b_{k,i}g^i\|^2]\Big).
\end{aligned} \tag{73}$$

Taking $\rho_{0,k} = \frac{1-\beta(k)}{2L\alpha(k)}$ in (73) gives

$$\begin{aligned}
&\mathbb{E}[L^{k+1}-L^k] \\
&\leq \Big(-\alpha(k)+\frac{3-\beta(k)+2\beta^2(k)}{2(1-\beta(k))}L\alpha^2(k)+4c_1\alpha^2(k)\Big)\,\mathbb{E}[\|g^k\|^2] \\
&\quad+\Big(\frac{\beta^2(k)}{2(1-\beta(k))}L\alpha^2(k)\mathbb{E}[\|m^{k-1}-\sum_{i=1}^{k-1}b_{k-1,i}g^i\|^2]+\frac{1}{2}L\alpha^2(k)\sigma^2+2c_1\alpha^2(k)\mathbb{E}[\|m^k-\sum_{i=1}^{k}b_{k,i}g^i\|^2]\Big).
\end{aligned}$$

Finally, by applying Lemma 4 and Lemma 5, we arrive at

$$\begin{aligned}
&\mathbb{E}[L^{k+1}-L^k] \\
&\leq \Big(-\alpha(k)+\frac{3-\beta(k)+2\beta^2(k)}{2(1-\beta(k))}L\alpha^2(k)+4c_1\alpha^2(k)\Big)\,\mathbb{E}[\|g^k\|^2] \\
&\quad+\Big(\frac{\beta^2(k)}{2}L\alpha^2(k)24\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}\sigma^2+\frac{1}{2}L\alpha^2(k)\sigma^2+4c_1(1-\beta_1)\alpha^2(k)\sigma^2\Big).
\end{aligned}$$

D.2 Proof of Theorem 3

From (73) we know that

$$\mathbb{E}[L^{k+1}-L^k] \leq -R_{1,k}\,\mathbb{E}[\|g^k\|^2]+R_{2,k}, \tag{74}$$

where

$$R_{1,k} = \alpha(k)-\alpha(k)\frac{1}{2\rho_{0,k}}-2\alpha(k)\rho_{0,k}L^2A_1^2-\frac{L\alpha^2(k)}{2}-4c_1\alpha^2(k) \tag{75}$$

$$R_{2,k} = \alpha(k)\rho_{0,k}L^2A_1^2\mathbb{E}[\|m^{k-1}-\sum_{i=1}^{k-1}b_{k-1,i}g^i\|^2]+\frac{1}{2}L\alpha^2(k)\sigma^2+2c_1\alpha^2(k)\mathbb{E}[\|m^k-\sum_{i=1}^{k}b_{k,i}g^i\|^2]. \tag{76}$$

This immediately tells us that

$$L^1 \geq \mathbb{E}[L^1-L^{k+1}] \geq \sum_{i=1}^{k}R_{1,i}\,\mathbb{E}[\|g^i\|^2]-\sum_{i=1}^{k}R_{2,i}, \tag{77}$$

In the rest the proof, we will bound $R_{1,i}$ and $R_{2,i}$ appropriately.

First, let us show that $R_{1,i} \geq \frac{\alpha(i)}{2}$ under $\rho_{0,i} = \frac{1-\beta(i)}{2L\alpha(i)}$ as in (69) and $\alpha(i) = \frac{A_1(1-\beta(i))}{\beta(i)} = \frac{1-\beta(i)}{24\sqrt{2}L\beta(i)}$.

From (72) we know that

$$c_1 \leq \frac{L}{4(1-\beta_1)}.$$

Therefore, in order for $R_{1,i} \geq \frac{\alpha(i)}{2}$, it suffices to have

$$\alpha(i)\frac{1}{2\rho_{0,i}} + 2\alpha(i)\rho_{0,i}L^2A_1^2 + \frac{L\alpha^2(i)}{2} + 4\frac{L}{4(1-\beta_1)}\alpha^2(i) \leq \frac{\alpha(i)}{2}. \tag{78}$$

By $\beta(i) \geq \beta_1 \geq \frac{1}{2}$ we know that

$$\alpha(i) = \frac{1-\beta(i)}{24\sqrt{2}L\beta(i)} \leq \frac{1}{2L}.$$

Therefore, $\frac{L\alpha^2(i)}{2} \leq \frac{\alpha(i)}{4}$. Furthermore, $\rho_{0,i} = \frac{1-\beta(i)}{2L\alpha(i)}$ yields

$$\begin{aligned}
&\alpha(i)\frac{1}{2\rho_{0,i}} + 2\alpha(i)\rho_{0,i}L^2A_1^2 + 4\frac{L}{4(1-\beta_1)}\alpha^2(i)\\
&= \frac{L\alpha^2(i)}{1-\beta(i)} + \frac{\beta^2(i)L\alpha^2(i)}{\big(1-\beta(i)\big)} + \frac{L}{(1-\beta_1)}\alpha^2(i)\\
&\leq \frac{\alpha(i)}{12} + \frac{\alpha(i)}{12} + \frac{\alpha(i)}{12}\\
&= \frac{\alpha(i)}{4},
\end{aligned}$$

where in the inequality above, we have applied

$$\begin{aligned}
\alpha(i) &= \frac{1-\beta(i)}{24\sqrt{2}L\beta(i)} \leq \frac{1-\beta(i)}{24L\frac{1}{2}} \leq \frac{1-\beta(i)}{12L},\\
\alpha(i) &= \frac{1-\beta(i)}{24\sqrt{2}L\beta(i)} \leq \frac{1-\beta(i)}{12L\beta^2(i)},\\
\alpha(i) &= \frac{1-\beta(i)}{24\sqrt{2}L\beta(i)} \leq \frac{1-\beta_1}{24L\beta_1} \leq \frac{1-\beta_1}{12L}.
\end{aligned}$$

Therefore, (78) is true and

$$R_{1,i} \geq \frac{\alpha(i)}{2}. \tag{79}$$

Now let us turn to $R_{2,i}$. By (76) and (72) we know that

$$\begin{aligned}
R_{2,i} &= \alpha(k)\rho_{0,i}L^2A_1^2\mathbb{E}[\|m^{i-1} - \sum_{j=1}^{i-1} b_{i-1,j}g^j\|^2] + \frac{1}{2}L\alpha^2(i)\sigma^2 + 2c_1\alpha^2(i)\mathbb{E}[\|m^i - \sum_{j=1}^{i} b_{i,j}g^j\|^2].\\
&\leq \alpha(k)\rho_{0,i}L^2A_1^2\mathbb{E}[\|m^{i-1} - \sum_{j=1}^{i-1} b_{i-1,j}g^j\|^2] + \frac{1}{2}L\alpha^2(i)\sigma^2 + \frac{L}{2(1-\beta_1)}\alpha^2(i)\mathbb{E}[\|m^i - \sum_{j=1}^{i} b_{i,j}g^j\|^2].
\end{aligned}$$

Since $\rho_{0,i} = \frac{1-\beta(i)}{2L\alpha(i)}$ and $\frac{\alpha(i)\beta(i)}{1-\beta(i)} \equiv A_1$, we have

$$R_{2,i} \leq \frac{1}{2}L\alpha^2(i)\beta^2(i)\frac{1}{1-\beta(i)}\mathbb{E}[\|m^{i-1} - \sum_{j=1}^{i-1} b_{i-1,i}g^j\|^2] + \frac{1}{2}L\alpha^2(i)\sigma^2 + \frac{L}{2(1-\beta_1)}\alpha^2(i)\mathbb{E}[\|m^i - \sum_{j=1}^{i} b_{i,j}g^j\|^2].$$

By applying Lemmas 4 and 5, we further have

$$R_{2,i} \leq 12L\alpha^2(i)\beta^2(i)\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}\sigma^2 + \frac{3}{2}L\alpha^2(i)\sigma^2. \tag{80}$$

By putting (79) and (80) into (77) with $k = T_1 + T_2 + \cdots + T_n$, we obtain

$$\sum_{l=1}^{n}\frac{\alpha_l}{2}\sum_{i=T_1+\cdots+T_{l-1}+1}^{T_1+\cdots+T_l}\mathbb{E}[\|g^i\|^2] \leq L^1 + \sum_{l=1}^{n}T_l\left(12L\alpha_l^2\beta_l^2\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}\sigma^2 + \frac{3}{2}L\alpha_l^2\sigma^2\right).$$

Dividing both sides by $\frac{1}{2}nA_2 \equiv \frac{1}{2}n\alpha_l T_l$ gives

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{n}\frac{1}{T_l}\sum_{i=T_1+\cdots+T_{l-1}+1}^{T_1+\cdots+T_l}\mathbb{E}[\|g^i\|^2] \\
&\leq \frac{2\left(f(x^1)-f^*\right)}{nA_2} + \frac{1}{n}\sum_{l=1}^{n}\left(24\beta_l^2\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}L\alpha_l\sigma^2 + 3L\alpha_l\sigma^2\right) \\
&= \mathcal{O}\left(\frac{f(x^1)-f^*}{nA_2}\right) + \mathcal{O}(\frac{1}{n}\sum_{l=1}^{n}L\alpha_l\sigma^2).
\end{aligned}$$

E Details of computational infrastructure

All experiments were performed on a computing server with Intel(R) Core(TM) i9-9940X CPU @ 3.30GHz and NVidia GeForce RTX 2080 P8. The weights of the neural networks are initialized by the default, random initialization routines.