Journal of Machine Learning Research 26 (2025) 1-51 Submitted 8/24; Revised 4/25; Published 7/25

Last-iterate Convergence of Shuffling Momentum Gradient Method under the Kurdyka-Łojasiewicz Inequality

Yuqing Liang LIANGYQ337@NENU.EDU.CN
Dongpo Xu* XUDP100@NENU.EDU.CN
Key Laboratory for Applied Statistics of MOE
School of Mathematics and Statistics
Northeast Normal University
Changchun 130024, China

Editor: Nicolas Le Roux

Abstract

Shuffling gradient algorithms are extensively used to solve finite-sum optimization problems in machine learning. However, their theoretical properties still need to be further explored, especially the last-iterate convergence in the non-convex setting. In this paper, we study the last-iterate convergence behavior of shuffling momentum gradient (SMG) method, a shuffling gradient algorithm with momentum. Specifically, we focus on the non-convex scenario and provide theoretical guarantees under arbitrary shuffling strategies. For non-convex objectives, we achieve the convergence of gradient norms at the last-iterate, showing that every accumulation point of the iterative sequence is a stationary point of the non-convex problem. Our analysis also reveals that the function values of the last-iterate converge to a finite value. Additionally, we obtain the asymptotic convergence rates of gradient norms at the minimum-iterate. By employing a uniform without-replacement sampling strategy, we further achieve an improved convergence rate for the minimum-iterate output. Under the Kurdyka-Łojasiewicz (KL) inequality, we establish the challenging strong limit-point convergence results. In particular, we prove that the whole sequence of iterates exhibits convergence to a stationary point of the finite-sum problem. By choosing an appropriate stepsize, we also obtain the corresponding rate of last-iterate convergence, matching available results in the strongly convex setting. Given that the last iteration is typically preferred as the output of the algorithm in applied scenarios, this paper contributes to narrowing the gap between theory and practice.

Keywords: Shuffling momentum gradient method, last-iterate convergence, strong limit-point convergence, non-convex optimization, Kurdyka-Łojasiewicz inequality

1. Introduction

In this paper, we consider the following finite-sum optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d}\left\{F(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}f(\boldsymbol{x};i)\right\}, \tag{1}$$

where $f(\cdot\,;i):\mathbb{R}^d\to\mathbb{R}$ is the loss associated with the i-th training sample and is differentiable for all $i\in[n]$. The problem (1) is also referred to as the empirical risk minimization (ERM) and models many machine learning tasks, such as training deep neural networks (Bottou et al., 2018; Mandic and Chambers, 2001). In particular, the number of losses $f(\cdot\,;i)$, denoted by n, is very large in the context of large-scale machine learning. Due to the computational burden of evaluating the

. * Corresponding Author.

full gradient, deterministic algorithms become less effective for solving problem (1). By leveraging the gradient information of a single sample in each iteration, stochastic gradient descent (SGD) significantly reduces the computation and storage costs, which iterates as follows.

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \nabla f(\boldsymbol{x}_t; i_t), \tag{2}$$

where α_t is the stepsize, $\nabla f(\boldsymbol{x}_t; i_t)$ denotes the stochastic gradient, and i_t is randomly sampled from $[n]$. The original idea of SGD is proposed by Robbins and Monro (1951), and its convergence properties are widely studied in Nguyen et al. (2019), Lei et al. (2020), and Wang and Yuan (2023). To further enhance the performance of SGD, shuffling algorithms employ without-replacement strategy to sample the index i_t for the stochastic gradient. Specifically, the permutation π^t of the set $[n]$ is introduced to implement sampling without replacement, and then the update of shuffling-type gradient method (Nguyen et al., 2021) in the t-th epoch is given by

$$\left[\begin{array}{l} \text{Set } \boldsymbol{x}_0^t = \tilde{\boldsymbol{x}}_{t-1} \text{ and generate a permutation } \pi^t \text{ of } [n] \\ \quad \textbf{For } i = 0, 1, \cdots, n-1 \textbf{ do: } \quad \boldsymbol{x}_{i+1}^t = \boldsymbol{x}_i^t - \alpha_t \nabla f(\boldsymbol{x}_i^t; \pi^t(i+1)) \ , \\ \text{Set } \tilde{\boldsymbol{x}}_t = \boldsymbol{x}_n^t \end{array}\right. \tag{3}$$

where α_t is the stepsize, and $\pi^t(i+1)$ is the $i+1$-th element of π^t. The incorporation of permutation π^t enables shuffling algorithms to better learn all the data information during each epoch. Depending on the specific choice of π^t, shuffling schemes can be categorized into three popular used cases: random reshuffling (RR), shuffle once (SO), and incremental gradient (IG). More specifically, RR randomly generates a new π^t at the beginning of each epoch, SO randomly obtains π^t only once and reuses it in the following epochs, and IG involves applying a deterministic π^t in all epochs. Among these, the superiority of RR over SGD is substantiated through empirical evidence (Bottou, 2009; Kasai, 2018; Koloskova et al., 2024). Meanwhile, shuffling gradient methods are embedded in many machine learning frameworks, including PyTorch and TensorFlow. However, the sampling without-replacement regime leads to the absence of unbiasedness, that is, $\mathbb{E}[\nabla f(\boldsymbol{x}_i^t; \pi^t(i+1))|\boldsymbol{x}_i^t] \neq \nabla F(\boldsymbol{x}_i^t)$, which brings significant challenges to theoretical analyses. Further, more involved proof techniques are required to achieve the convergence of shuffling-based methods. In recent years, a great effort has been devoted to exploring the convergence properties of shuffling gradient methods in various settings (Mishchenko et al., 2020; Gürbüzbalaban et al., 2021; Nguyen et al., 2021; Li et al., 2023). Despite the excellent performance and research advances, there remains a gap between theoretical results and applied scenarios. In practice, optimization algorithms commonly return the last-iterate as their output, whereas most of the current analyses rely on the average and minimum iterations (Ahn et al., 2020; Qin et al., 2023). To this end, we delve into the last-iterate convergence of shuffling momentum gradient (SMG) method proposed by Tran et al. (2021), which is a momentum version of shuffling algorithm and contains shuffling-type gradient in (3) as a special case. It is worth emphasizing that our analysis provides last-iterate convergence guarantees for both gradient norms and iterative sequences in the non-convex setting.

1.1 Related work

Shuffling gradient algorithms have emerged as efficient methods for solving problem (1), with extensive theoretical investigation in different scenarios. For strongly convex problems, Safran and Shamir (2020) provide non-asymptotic lower bounds for both RR and SO. Gürbüzbalaban et al. (2021) analyze the convergence rate of RR, revealing that the q-suffix average-iterate converges at the rate of $\mathcal{O}(1/T)$. For general convex objectives, Nagaraj et al. (2019) study the performance of RR for constraint minimization and deduce the rate of $\mathcal{O}(1/\sqrt{nT})$ under the bounded gradient

assumption. Moreover, Tran et al. (2022) introduce Nesterov's extrapolation technique to propose the NASG method and provide the last-iterate convergence guarantee for function values.

However, convexity may be absent in many practical scenarios, such as neural network training and signal processing. Then many efforts are dedicated to establishing the convergence of shuffling methods in the non-convex setting (Mishchenko et al., 2020; Nguyen et al., 2021; Li et al., 2023). In the theoretical analysis of non-convex objectives, four types of gradient norms are usually employed as the convergence measure of the method, including (i) **minimum-iterate**: $\min_{1\leq t\leq T}\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2$, (ii) **random-iterate**: $\|\nabla F(\tilde{\boldsymbol{x}}_\tau)\|^2$, (iii) **average-iterate**: $(1/T)\sum_{t=1}^{T}\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2$, and (iv) **last-iterate**: $\|\nabla F(\tilde{\boldsymbol{x}}_T)\|^2$, where τ is randomly selected from the set $\{1, 2, \cdots, T\}$, and T is the total number of epochs. In particular, Mishchenko et al. (2020) investigate the rate of minimum-iterate convergence for RR, SO, and IG equipped with a constant stepsize. By leveraging the proximal operator, Mishchenko et al. (2022) propose ProxRR to solve composite optimization problems with a finite-sum structure and study the behavior of minimum-iterate output. Meanwhile, Wang et al. (2023) combine randomly reshuffled strategy with AdaGrad, a popular adaptive gradient algorithm, and achieve the minimum-iterate convergence rate of $\mathcal{O}(\ln\sqrt{T}/\sqrt{T})$. Huang et al. (2023) propose a distributed version of random reshuffling method, named D-RR, and attain the minimum-iterate convergence rate of $\mathcal{O}(1/T^{2/3})$. However, the minimum-iterate output calls for accurate calculation of the full gradients, which is not advisable in modern machine learning. In addition, some studies focus on the convergence behavior of random and average iterations. Specifically, Tran et al. (2021) develop the SMG method, in which the momentum technique and shuffling strategy are integrated. By randomly choosing the output from the whole sequence of iterates according to some probability distributions, they show that the random-iterate is convergent at the rate of $\mathcal{O}(1/T^{2/3})$. Due to the high dimension of the data, the storage cost for generating a random-iterate output is expensive. Under any shuffling strategy, Nguyen et al. (2021) prove that the average-iterate of generic shuffling-type gradient algorithm converges at the rate of $\mathcal{O}(1/T^{2/3})$. By combining the gradient estimator in variance reduction method with different shuffling regimes, Malinovsky et al. (2023) explore the behavior of SVRG with constant stepsize and attain the average-iterate convergence rate of $\mathcal{O}(1/T)$. For IG and SO methods, Koloskova et al. (2024) provide a tight convergence rate for average-iterate and emphasize the advantages of exploiting SO. The sign-based RR method presented by Qin et al. (2023) employs the sign of stochastic gradients to update, and the theoretical behavior of average-iterate output is analyzed. Notably, the average-iterate can be reformulated as sampling the output randomly from the iterative sequence following the uniform distribution. As discussed before, the effort of storing iterative sequences may weaken the application of returning average-iterate output in large-scale problems. In contrast, the last-iterate requires no additional computation or storage, making it a commonly adopted choice of output in practice. Nevertheless, the theoretical guarantees for last-iterate convergence in the non-convex setting are rare. Therefore, there still exists a gap between the theoretical findings of shuffling gradient methods and their empirical success.

For non-convex problems, the convergence analysis of algorithms under the Kurdyka-Łojasiewicz (KL) inequality has also attracted much attention (Bolte et al., 2014; Li and Lin, 2015; Zeng and Yin, 2018), primarily owing to its ease of satisfaction. In particular, the KL inequality is a local geometrical property that generally holds for many objective functions of learning models (Bolte et al., 2014; Ge et al., 2016), thereby serving as a mild assumption in the non-convex setting. Under the KL property, Bolte et al. (2014) establish the convergence of PALM algorithm, showing that every bounded iterative sequence exhibits global convergence to a stationary point. Moreover, Li and Lin (2015) propose the monotone APG-type algorithm and study the convergence property when the objective function is coercive. To further improve the practical performance of APG, they also introduce a non-monotone variant that reduces the calculation cost of each iteration. For consensus optimization problems, Zeng and Yin (2018) analyze the convergence rate of DGD and

Prox-DGD according to different KL exponents. By constructing a separable Lyapunov function with sufficient descent property, Barakat and Bianchi (2020) explore the convergence behavior of adaptive gradient algorithm with bounded adaptive stepsizes in the KL setting. Additionally, Guo et al. (2023a) investigate the global convergence property of linearized proximal ADMM method for KL functions from the perspective of dynamical systems. Yang and Li (2023) present the PPGD algorithm with a new projection operator and obtain the convergence rate of $\mathcal{O}(1/T^2)$ for function values at the last-iterate. It is worth mentioning that standard theoretical analysis under the KL condition critically depends on the sufficient descent property of algorithm, whereas most shuffling gradient algorithms fail to fulfill. More recently, Li et al. (2023) pioneer the extension of classical KL inequality to non-descent algorithms and analyze the convergence of RR within the proposed framework. Specifically, they establish the strong limit-point convergence and convergence rate of $\mathcal{O}(1/T)$ under the bound domain assumption, revealing that the whole sequence of iterates converges to a stationary point. To the best of our knowledge, the results in Li et al. (2023) are the first to achieve the last-iterate convergence of iterative sequence for RR without strong convexity. Further, an interesting and challenging topic is to understand the last-iterate behavior of iterative sequences for other shuffling-based methods in the non-convex KL scenario.

Despite the widespread analysis, the convergence of shuffling gradient algorithms remains not fully understood, especially for the behavior of last-iterate. To address this issue, we perform the last-iterate convergence analysis for SMG method (Tran et al., 2021), which is a shuffling gradient algorithm with momentum and subsumes RR as a special case. It is worth emphasizing that our theoretical results not only cover non-convex and KL settings but also hold for arbitrary shuffling schemes. Specifically, we establish the last-iterate convergence guarantees for gradient norms and iterative sequences under some standard assumptions, and the asymptotic convergence rate of gradient norms at the minimum-iterate is also obtained for non-convex objectives. In addition, we investigate the performance of SMG when the uniform without-replacement sampling strategy is leveraged and obtain an improved minimum-iterate convergence rate. Further, we believe that these findings can effectively narrow the gap between theory and practice, and the analysis framework can be extended to derive last-iterate convergence of other shuffling variants.

1.2 Contribution

This paper explores the last-iterate convergence performance of SMG algorithm, which is a momentum variant of shuffling gradient method and includes RR as a special case. The main contributions are summarized as follows.

- We establish the last-iterate convergence analysis for SMG algorithm with arbitrary shuffling schemes in the non-convex scenario. Given that the last-iterate is typically returned as the output of method in practice rather than performing an average or minimum, our theoretical findings regarding the last-iterate are more in line with actual implementations.

- For non-convex problems, we provide the last-iterate convergence guarantee for gradient sequences, that is, $\lim_{T\to\infty}\|\nabla F(\tilde{\boldsymbol{x}}_T)\| = 0$, which is different from previous results on the minimum and average iterations. Additionally, our analysis also demonstrates that the function values of the last-iterate converge to a finite value, that is, $\lim_{T\to\infty} F(\tilde{\boldsymbol{x}}_T) = \bar{F}$.

- We obtain the asymptotic convergence rate of minimum-iterate in the non-convex scenario, that is, $\min_{1\le t\le T}\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2 = o\big(1/\sum_{t=1}^T \alpha_t\big)$, where α_t is the stepsize. In particular, we show that the asymptotic rate can be arbitrarily close to $o(1/T^{2/3})$ with an appropriate polynomial stepsize. Further, under the uniform without-replacement sampling scheme, we establish the rate of $\mathcal{O}(1/(n^{1/3}T^{2/3}))$ for constant stepsize, which exhibits a superior dependence on n.

- Under the KL inequality (see Definition 1) and the coercivity condition (see Assumption 3(a)), we achieve the well-known strong limit-point convergence results for iterative sequences. In particular, we prove that the whole sequence of iterates $\{\tilde{\boldsymbol{x}}_t\}$ converges to a stationary point of the non-convex problem. By choosing a diminishing stepsize, the rate of $\mathcal{O}(1/T^2)+\mathcal{O}(1/T)$ is attained, matching current results of shuffling methods for strongly convex functions.

Notations

We use $[n]$ to denote the set $\{1, 2, \cdots, n\}$ and the bold letters to represent vectors. For any $a, b \in \mathbb{R}$, we follow the convention that $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Without loss of generality, we also follow the convention that $\sum_{i=k}^{\ell} a_i = 0$ when $k > \ell$. The symbols $\mathcal{O}(\cdot)$ and $o(\cdot)$ are the standard asymptotic notations, that is, if there exists a constant $C > 0$ such that $|a_t/b_t| \leq C$, then $a_t = \mathcal{O}(b_t)$, and if $\lim_{t\to\infty}|a_t/b_t| = 0$, then $a_t = o(b_t)$. Further, $\tilde{\mathcal{O}}(\cdot)$ is used to hide the logarithmic factors in $\mathcal{O}(\cdot)$. $\mathbb{E}_t[\,\cdot\,]$ is the conditional expectation with respect to $\pi^t = \{\pi^t(1), \pi^t(2), \cdots, \pi^t(n)\}$ conditioned on previous iterations, and $\mathbb{E}[\,\cdot\,]$ denotes the total expectation. For clarity, we introduce the sequences Δ_t, Φ_t, and Ψ_t to simplify the theoretical proofs in Sections 3 and 4.

$$\Delta_t = \sum_{i=0}^{n-1} \|\boldsymbol{x}_i^t - \boldsymbol{x}_0^t\|^2, \quad \Phi_t = n^3\alpha_t^2\Big(\beta\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 + \sigma_1(1-\beta)\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \sigma_2(1-\beta)\Big), \tag{4}$$

$$\Psi_t = n^2\alpha_t^2\Big(n\beta\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big] + n(1-\beta)(3+2\sigma_1)\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + 2\sigma_2(1-\beta)\Big), \tag{5}$$

where $\boldsymbol{x}_i^t$ is the iterative sequence generated by Algorithm 1, σ_1 and σ_2 are two finite values defined in Assumption 2, $\alpha_t > 0$ is the stepsize, and $0 \leq \beta < 1$ is the momentum parameter. Moreover, we focus on a single sample trajectory of $\{\tilde{\boldsymbol{x}}_t(\omega)\}$, then we omit the notation (ω) in the convergence analysis for simplicity.

2. Algorithm and Assumptions

This section first presents the pseudo-code of SMG algorithm in Section 2.1, and then gives some standard assumptions adopted in our theoretical analysis in Section 2.2.

2.1 Shuffling momentum gradient algorithm

In this section, we provide the pseudo-code of SMG in Algorithm 1 and briefly discuss the relationship between Algorithm 1 and the original SMG method (Tran et al., 2021) as follows.

Notice that the permutation π^t can be generated in both deterministic and random manners, showing that Algorithm 1 works for arbitrary permutations. As a result, the last-iterate convergence guarantees achieved in this paper are valid for any shuffling schemes, demonstrating the generality of our theoretical analysis. Meanwhile, when choosing the momentum parameter $\beta = 0$, Algorithm 1 degenerates into the generic shuffling-type gradient method studied by Nguyen et al. (2021) and Mishchenko et al. (2020), referred to as RR in Li et al. (2023). Additionally, the algorithmic output and update rule of $\boldsymbol{v}_i^t$ in Algorithm 1 are slightly different from the original SMG method developed by Tran et al. (2021). In particular, we choose the last-iterate $\tilde{\boldsymbol{x}}_T$ as the output rather than the random-iterate used in Tran et al. (2021), and line 6 of Algorithm 1 indicates that

$$\begin{aligned}\boldsymbol{v}_n^t &= \Big(1-\frac{1}{n}\Big)\boldsymbol{v}_{n-1}^t + \frac{1}{n}\boldsymbol{g}_{n-1}^t = \cdots = \Big(1-\frac{1}{n}\Big)\cdots\Big(1-\frac{1}{2}\Big)\boldsymbol{v}_1^t + \frac{1}{n}\boldsymbol{g}_1^t + \cdots + \frac{1}{n}\boldsymbol{g}_{n-1}^t \\ &= \frac{1}{n}\Big(\boldsymbol{v}_1^t + \boldsymbol{g}_1^t + \cdots + \boldsymbol{g}_{n-1}^t\Big) = \frac{1}{n}\Big(\boldsymbol{g}_0^t + \boldsymbol{g}_1^t + \cdots + \boldsymbol{g}_{n-1}^t\Big) = \frac{1}{n}\sum_{i=0}^{n-1}\boldsymbol{g}_i^t. \end{aligned} \tag{6}$$

Algorithm 1 Shuffling Momentum Gradient (SMG)

Require: stepsize $\alpha_t > 0$ and momentum parameter $0 \leq \beta < 1$;

Initialization: $\tilde{\boldsymbol{x}}_0 \in \mathbb{R}^d$ and $\tilde{\boldsymbol{v}}_0 = \mathbf{0}$;

1: **for** $t = 1, 2, \cdots, T$ **do**

2: Set $\boldsymbol{x}_0^t = \tilde{\boldsymbol{x}}_{t-1}$, $\boldsymbol{m}_0^t = \tilde{\boldsymbol{v}}_{t-1}$, and $\boldsymbol{v}_0^t = \mathbf{0}$

3: Generate a deterministic or random permutation π^t of $[n]$

4: **for** $i = 0, 1, \cdots, n-1$ **do**

5: Set $\boldsymbol{g}_i^t = \nabla f(\boldsymbol{x}_i^t; \pi^t(i+1))$

6: Update the inner iterate via

$$\begin{cases} \boldsymbol{m}_{i+1}^t = \beta \boldsymbol{m}_0^t + (1-\beta)\boldsymbol{g}_i^t \\ \boldsymbol{v}_{i+1}^t = \big(1 - \frac{1}{i+1}\big)\boldsymbol{v}_i^t + \frac{1}{i+1}\,\boldsymbol{g}_i^t \\ \boldsymbol{x}_{i+1}^t = \boldsymbol{x}_i^t - \alpha_t \boldsymbol{m}_{i+1}^t \end{cases}$$

7: **end for**

8: Set $\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_n^t$ and $\tilde{\boldsymbol{v}}_t = \boldsymbol{v}_n^t$

9: **end for**

10: **Output**: the last-iterate $\tilde{\boldsymbol{x}}_T$

The iteration of $\boldsymbol{v}_i^t$ in Tran et al. (2021) is $\boldsymbol{v}_{i+1}^t = \boldsymbol{v}_i^t + \frac{1}{n}\,\boldsymbol{g}_i^t$ with $\boldsymbol{v}_0^t = \mathbf{0}$, showing that

$$\boldsymbol{v}_n^t = \boldsymbol{v}_{n-1}^t + \frac{1}{n}\boldsymbol{g}_{n-1}^t = \cdots = \boldsymbol{v}_0^t + \frac{1}{n}\boldsymbol{g}_0^t + \frac{1}{n}\boldsymbol{g}_1^t + \cdots + \frac{1}{n}\boldsymbol{g}_{n-1}^t = \frac{1}{n}\sum_{i=0}^{n-1}\boldsymbol{g}_i^t. \tag{7}$$

Hence, the final expression of $\boldsymbol{v}_n^t$ in equations (6) and (7) remains the same, which is the sum of simple average of stochastic gradients in one epoch. However, the number of losses n is typically unknown in advance, then the pseudo-code presented in this paper is more in line with practical applications. Further, more detailed information regarding the design of momentum in SMG method can be found in Section 2 of Tran et al. (2021).

2.2 Basic assumptions

In this section, we present the assumptions used in our convergence analysis.

Assumption 1 *The function $f(\cdot\,; i)$ is L-smooth, that is, there exists a constant $L > 0$ such that for any $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^d$,*

$$\|\nabla f(\boldsymbol{x}; i) - \nabla f(\boldsymbol{y}; i)\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \forall\, i \in [n].$$

$f(\cdot\,; i)$ is also uniformly lower bounded by f^*, that is, $f(\boldsymbol{x}; i) \geq f^*$ for any $\boldsymbol{x} \in \mathbb{R}^d$ and $i \in [n]$.

It should be noted that the L-smoothness of $f(\cdot; i)$ is a standard assumption used in the convergence analysis of various stochastic optimization algorithms, such as SpiderBoost (Wang et al., 2019), generic shuffling-type gradient (Nguyen et al., 2021), RelaySGD (Vogels et al., 2021), D-RR (Huang et al., 2023), and SGDEM (Ramezani-Kebrya et al., 2024). In addition, by combining Assumption 1 with the finite-sum structure of F in problem (1), we can obtain

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\big(\nabla f(\boldsymbol{x}; i) - \nabla f(\boldsymbol{y}; i)\big)\right\| \leq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(\boldsymbol{x}; i) - \nabla f(\boldsymbol{y}; i)\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|,$$

which means that F is also L-smooth. Meanwhile, it follows from the lower boundedness of $f(\cdot\,;i)$ that F is also lower bounded by f^*, that is, $F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x};i) \geq \frac{1}{n}\sum_{i=1}^n f^* = f^*$.

Assumption 2 *There exist two constants σ_1 and $\sigma_2 \geq 0$ such that for any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f(\boldsymbol{x};i)\|^2 \leq \sigma_1 \|\nabla F(\boldsymbol{x})\|^2 + \sigma_2.$$

Note that Assumption 2 is also known as the weak growth condition, which appears frequently in the analysis of finite-sum problems (Pham et al., 2020; Nguyen et al., 2021; Tran et al., 2021). In particular, variance assumption of $\frac{1}{n}\sum_{i=1}^n \|\nabla f(\boldsymbol{x};i) - \nabla F(\boldsymbol{x})\|^2 \leq \theta_1 \|\nabla F(\boldsymbol{x})\|^2 + \theta_2$ (Nguyen et al., 2021; Tran et al., 2021) is equivalent to Assumption 2, differing only in the coefficients. That is,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \|\nabla f(\boldsymbol{x};i) - \nabla F(\boldsymbol{x})\|^2 = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f(\boldsymbol{x};i) - \frac{1}{n}\sum_{i=1}^n \nabla f(\boldsymbol{x};i)\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \|\nabla f(\boldsymbol{x};i)\|^2 - \left\|\frac{1}{n}\sum_{i=1}^n \nabla f(\boldsymbol{x};i)\right\|^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f(\boldsymbol{x};i)\|^2 - \|\nabla F(\boldsymbol{x})\|^2.
\end{aligned}$$

Thus, the variance assumption $\frac{1}{n}\sum_{i=1}^n \|\nabla f(\boldsymbol{x};i) - \nabla F(\boldsymbol{x})\|^2 \leq \theta_1 \|\nabla F(\boldsymbol{x})\|^2 + \theta_2$ implies that Assumption 2 is satisfied with $\sigma_1 = 1 + \theta_1$ and $\sigma_2 = \theta_2$. In addition, if Assumption 2 is satisfied, then $\frac{1}{n}\sum_{i=1}^n \|\nabla f(\boldsymbol{x};i) - \nabla F(\boldsymbol{x})\|^2 \leq \theta_1 \|\nabla F(\boldsymbol{x})\|^2 + \theta_2$ holds with $\theta_1 = \sigma_1 - 1$ and $\theta_2 = \sigma_2$.

3. Convergence and Convergence Rate Analysis for Non-convex Objectives

This section studies the convergence properties of Algorithm 1 for solving non-convex problems, and the corresponding proofs are placed in Appendix B. Specifically, we first establish the recursive estimates for the sequence of function values in Lemma 1, considering both arbitrary shuffling schemes and uniform sampling strategy. Further, the main results on the last-iterate and minimum-iterate are shown in Sections 3.1 and 3.2, respectively.

Lemma 1 *Let Assumptions 1 and 2 hold, if the stepsize $\alpha_t \leq \frac{1}{\sqrt{2}nL}$, then for any $t \geq 2$, we have*

$$\begin{aligned}
F(\boldsymbol{x}_0^{t+1}) \leq\ & F(\boldsymbol{x}_0^t) - \frac{1}{2}\left(\frac{1}{n\alpha_t} - L\right)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 \\
& - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{3nL^2 D_1}{2}\beta^{t-1}\alpha_t + 3L^2\alpha_t \sum_{j=2}^t \beta^{t-j}\Phi_j,
\end{aligned} \tag{8}$$

where $D_1 = \frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$ is a finite value (independent of n), and Φ_j is defined in (4). If additionally assuming that π^t is uniformly sampled at random without replacement from $[n]$, then for any $t \geq 2$, we have

$$\begin{aligned}
\mathbb{E}\big[F(\boldsymbol{x}_0^{t+1})\big] \leq\ & \mathbb{E}\big[F(\boldsymbol{x}_0^t)\big] - \frac{1}{2}\left(\frac{1}{n\alpha_t} - L\right)\mathbb{E}\big[\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2\big] \\
& - \frac{n\alpha_t}{2}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + \frac{3nL^2 D_2}{2}\beta^{t-1}\alpha_t + 3L^2\alpha_t \sum_{j=2}^t \beta^{t-j}\Psi_j,
\end{aligned} \tag{9}$$

where $D_2 = \frac{1}{L^2}(\sigma_1\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^1)\|^2] + \sigma_2)$ is a finite value (independent of n), and Ψ_j is defined in (5).

3.1 Last-iterate convergence of SMG

In this section, we prove the last-iterate convergence of function values in Theorem 1, and the gradient norms are demonstrated to be convergent to zero at the last-iterate in Theorem 2. Before presenting these results, we provide some useful propositions from Bertsekas and Tsitsiklis (2000) and Liu et al. (2022), which are crucial to achieving last-iterate convergence.

Proposition 1 (Lemma 1, Bertsekas and Tsitsiklis 2000) *Let Y_t, W_t, and Z_t be three sequences such that W_t is non-negative for all t. Assume that*

$$Y_{t+1} \leq Y_t - W_t + Z_t, \quad t = 0,\ 1,\ \cdots,$$

and that the series $\sum_{t=1}^{T} Z_t$ converges as $T \to \infty$. Then either $Y_t \to -\infty$ or else Y_t converges to a finite value and $\sum_{t=1}^{\infty} W_t < \infty$.

Proposition 2 (Corollary A.1, Liu et al. 2022) *Let a_t and b_t be two non-negative sequences of real values such that*

$$\sum_{t=1}^{\infty} a_t = \infty, \qquad \sum_{t=1}^{\infty} a_t b_t^2 < \infty, \qquad |b_{t+1} - b_t| \leq \mu a_t$$

for a positive constant μ. Then we have $\lim_{t\to\infty} b_t = 0$.

Note that Propositions 1 and 2 are effective theoretical tools for establishing the last-iterate convergence of function values and gradient norms. In particular, by leveraging Proposition 1 and the recursion estimation in Lemma 1, we deduce the following theorem, revealing that the function values of the last-iterate are convergent when solving non-convex problems.

Theorem 1 *Suppose that Assumptions 1 and 2 hold, and $\tilde{\boldsymbol{x}}_t$ is generated by Algorithm 1. If the stepsize α_t is non-increasing and satisfies $\sum_{t=1}^{\infty} \alpha_t^3 < \infty$ and $\alpha_t \leq \frac{1}{nL\sqrt{K}}$, where $K = \max\left\{2, \frac{9(\sigma_1+\beta)}{1-\beta}\right\}$ is a finite value, then we obtain that $F(\tilde{\boldsymbol{x}}_T)$ converges to a finite point $\bar{F} \in \mathbb{R}$.*

Under the same assumptions as Theorem 1, we show that the squared norm of gradient sequences is upper bounded in the following Corollary 1. This indicates that the sequence $\{\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2\}$ has convergent sub-sequences.

Corollary 1 *Suppose that the conditions in Theorem 1 are satisfied. If $\tilde{\boldsymbol{x}}_t$ and $\boldsymbol{m}_{i+1}^t$ are generated by Algorithm 1, then there exist two constants G and $M \geq 0$ such that for any $t \geq 1$, we have*

$$\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2 \leq G, \qquad \sum_{i=0}^{n-1} \|\boldsymbol{m}_{i+1}^t\|^2 \leq M.$$

By further assuming that $\sum_{t=1}^{\infty} \alpha_t = \infty$, we achieve the convergence of $\|\nabla F(\tilde{\boldsymbol{x}}_t)\|$ and prove that $\|\nabla F(\tilde{\boldsymbol{x}}_t)\|$ converges to zero in Theorem 2. It is worth emphasizing that Proposition 2 serves as the key technique to attain the convergence of gradient norms at the last-iterate for SMG method.

Theorem 2 *Suppose that Assumptions 1 and 2 hold, and $\tilde{\boldsymbol{x}}_t$ is generated by Algorithm 1. If the stepsize α_t is non-increasing and satisfies $\sum_{t=1}^{\infty} \alpha_t = \infty$, $\sum_{t=1}^{\infty} \alpha_t^3 < \infty$, and $\alpha_t \leq \frac{1}{nL\sqrt{K}}$, where $K = \max\left\{2, \frac{9(\sigma_1+\beta)}{1-\beta}\right\}$ is a finite value, then we have*

$$\lim_{T\to\infty} \|\nabla F(\tilde{\boldsymbol{x}}_T)\| = 0,$$

that is, every accumulation point of the sequence $\{\tilde{\boldsymbol{x}}_T\}$ is a stationary point of problem (1).

Remark 1 *To the best of our knowledge, Theorem 2 establishes the first last-iterate convergence of gradient norms for SMG algorithm in the non-convex scenario. In particular, the conclusion of Theorem 2 is slightly stronger than the asymptotic convergence result of* $\liminf_{T\to\infty}\|\nabla F(\tilde{\boldsymbol{x}}_T)\| = 0$ *for generic shuffling-type gradient method (*$\beta = 0$ *in Algorithm 1) obtained in Nguyen et al. (2021). Meanwhile, most previous analyses for shuffling gradient algorithms focus on the behavior of minimum-iterate (Qin et al., 2023; Wang et al., 2023) and average-iterate (Nguyen et al., 2021; Malinovsky et al., 2023) outputs. It should be noted that the last-iterate is more convenient and cheaper to acquire than the minimum and average iterations, so it is usually taken as the output of algorithm in practice. Thus, our theoretical findings on the last-iterate are more relevant to practical applications. In addition, the stepsize conditions in Theorem 2 are often observed in the analysis of different algorithms. Specifically, the stepsize conditions of* $\sum_{t=1}^{\infty}\alpha_t = \infty$ *and* $\sum_{t=1}^{\infty}\alpha_t^3 < \infty$ *are usually employed to achieve the convergence of shuffling methods (Nguyen et al., 2021; Li et al., 2023). The requirement of* $\alpha_t \le 1/(nL\sqrt{K})$ *is commonly used to obtain the convergence of various deterministic and stochastic methods, including GD (Tran-Dinh and van Dijk, 2024), SGD (Nguyen et al., 2019), and SGDM (Liu et al., 2020). Given that the condition* $\sum_{t=1}^{\infty}\alpha_t^3 < \infty$ *implies that* $\lim_{t\to\infty}\alpha_t = 0$*, then* $\alpha_t \le 1/(nL\sqrt{K})$ *is easily satisfied for sufficiently large* t*. Moreover, the polynomial-type stepsizes* $\alpha_t = \gamma/t^p$ *meet the conditions of Theorem 2 when* $p \in (1/3, 1]$*.*

3.2 Minimum-iterate convergence rate of SMG

In this section, we establish the asymptotic convergence rate of gradient norms at the minimum-iterate, as shown in Theorem 3 and Corollary 2. In particular, these results are obtained based on the following proposition from Liu and Yuan (2022). Moreover, we analyze the convergence rate of the minimum-iterate under the uniform without-replacement sampling scheme in Theorem 4.

Proposition 3 (Lemma 2, Liu and Yuan 2022) *Let* $\{X_t\}$ *be a sequence of non-negative real numbers, and* $\{a_t\}$ *be a non-increasing sequence of positive real numbers such that*

$$\sum_{t=1}^{\infty} a_t X_t < \infty, \qquad \sum_{t=2}^{\infty} \frac{a_t}{\sum_{i=1}^{t-1} a_i} = \infty,$$

then we have

$$\min_{1\le i\le t} X_i = o\left(\frac{1}{\sum_{i=1}^{t} a_i}\right).$$

Note that Proposition 3, proposed by Liu and Yuan (2022), serves as an essential tool to achieve the almost sure convergence rates of SGD, SHB, and SNAG. In this paper, we extend the application of Proposition 3 to analyze the rate of shuffling gradient methods, see Theorem 3.

Theorem 3 *Suppose that Assumptions 1 and 2 hold, and* $\tilde{\boldsymbol{x}}_t$ *is generated by Algorithm 1. If the stepsize* α_t *is non-increasing and satisfies* $\sum_{t=1}^{\infty}\alpha_t^3 < \infty$*,* $\sum_{t=2}^{\infty}\frac{\alpha_{t+1}}{\sum_{i=2}^{t}\alpha_i} = \infty$*, and* $\alpha_t \le \frac{1}{nL\sqrt{K}}$*, where* $K = \max\left\{2, \frac{9(\sigma_1+\beta)}{1-\beta}\right\}$ *is a finite value, then we have*

$$\min_{1\le t\le T}\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2 = o\left(\frac{1}{\sum_{t=1}^{T}\alpha_{t+1}}\right).$$

By employing specific decreasing stepsizes, the following corollary obtains the rate of $o(1/T^{2/3-\epsilon})$ for SMG method in the non-convex setting.

Corollary 2 *Suppose that the assumptions in Theorem 3 hold. If choosing the stepsize* $\alpha_t = \frac{\gamma}{t^{1/3+\epsilon}}$ *for any* $\epsilon \in (0, \frac{2}{3})$ *and* $0 < \gamma \leq \frac{1}{nL\sqrt{K}}$, *where* $K = \max\left\{2, \frac{9(\sigma_1+\beta)}{1-\beta}\right\}$ *is a finite value. Then we have*

$$\min_{1\leq t\leq T} \|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2 = o\left(\frac{1}{T^{2/3-\epsilon}}\right).$$

Remark 2 *It should be noted that previous studies on shuffling gradient algorithms primarily focus on the non-asymptotic convergence rates (see, Mishchenko et al., 2020; Nguyen et al., 2021; Tran et al., 2021). More specifically, under the constant stepsize condition, Nguyen et al. (2021) and Tran et al. (2021) achieve the rate of* $\mathcal{O}(1/T^{2/3})$ *for generic shuffling-type gradient algorithm and SMG method, respectively. As shown in Corollary 2, we derive the asymptotic convergence rate of* $o(1/T^{2/3-\epsilon})$ *for SMG method with decreasing stepsizes. In particular, by choosing* $\epsilon \to 0$ *in the stepsize* α_t, *the asymptotic rate can be arbitrarily close to* $o(1/T^{2/3})$.

In particular, when the uniform without-replacement sampling strategy is employed, Theorem 4 establishes an improved convergence rate for the minimum-iterate of SMG.

Theorem 4 *Suppose that Assumptions 1 and 2 are satisfied, and* $\tilde{\boldsymbol{x}}_t$ *is generated by Algorithm 1. If* π^t *is uniformly sampled at random without replacement from* $[n]$, *and the stepsize* $\alpha_t = \frac{\gamma}{n^{2/3}T^{1/3}}$ *for some* $\gamma > 0$, *then for any* $T \geq 2\sqrt{2}n\gamma^3L^3\max\left\{1, \frac{27(\sigma_1+2)^{3/2}}{(1-\beta)^{3/2}}\right\}$, *we have*

$$\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2\right] = \mathcal{O}\left(\frac{1}{n^{1/3}T^{2/3}}\right).$$

Remark 3 *Notice that the number of loss functions* n *is often large in practice, making the investigation of* n*-dependent convergence results important for shuffling gradient algorithms. Specifically, Ahn et al. (2020) establish the* n*-dependent convergence rates for shuffling SGD in both strongly convex and Polyak-Łojasiewicz (PL) settings. For non-convex objectives, Mishchenko et al. (2020) achieve the minimum-iterate convergence rate of* $\mathcal{O}(1/(n^{1/3}T^{2/3}))$ *for RR with a constant stepsize. Additionally, Nguyen et al. (2021) prove that the generic shuffling-type gradient algorithm converges at the rate of* $\mathcal{O}(1/T^{2/3})$ *for any shuffling strategy and at an improved rate of* $\mathcal{O}(1/(n^{1/3}T^{2/3}))$ *under the uniform sampling scheme. In this paper, we attain the convergence rate of* $\mathcal{O}(1/(n^{1/3}T^{2/3}))$ *for the SMG method by leveraging the uniform sampling strategy. It is worth mentioning that the rate obtained in Theorem 4 demonstrates a better dependence on* n *than the* $\mathcal{O}(1/T^{2/3})$ *rate for arbitrary permutations and matches the results in Mishchenko et al. (2020) and Nguyen et al. (2021).*

4. Strong Limit-point Convergence under the Kurdyka-Łojasiewicz Inequality

This section provides strong limit-point convergence results for Algorithm 1 under the KL inequality, and the proof details can be found in Appendix C. Specifically, we provide some standard definitions and assumptions in Section 4.1, and the convergence of iterative sequence $\{\tilde{\boldsymbol{x}}_t\}$ is achieved in Section 4.2. In Section 4.3, we perform the convergence rate analysis under the Łojasiewicz inequality.

4.1 Definitions and assumptions

In this section, we first introduce the definitions of KL inequality and quasi-additivity from Li et al. (2023), followed by some assumptions adopted in the theoretical analysis.

Definition 1 (KL inequality, Li et al. 2023) *The function F is said to satisfy the KL inequality at a point $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\bar{\boldsymbol{x}}$, and a continuous and concave function $\rho : [0, \eta) \to \mathbb{R}^+$ with*

$$\rho \in C^1(0, \eta), \quad \rho(0) = 0, \quad \text{and} \quad \rho'(x) > 0 \quad \forall x \in (0, \eta), \tag{10}$$

such that for all $\boldsymbol{x} \in U \cap \{\boldsymbol{x} \in \mathbb{R}^d : 0 < |F(\boldsymbol{x}) - F(\bar{\boldsymbol{x}})| < \eta\}$, the KL inequality holds, that is,

$$\rho'\big(|F(\boldsymbol{x}) - F(\bar{\boldsymbol{x}})|\big) \cdot \|\nabla F(\boldsymbol{x})\| \geq 1. \tag{11}$$

The KL property in Definition 1 is employed by Li et al. (2023) to establish the strong limit-point convergence results of RR in the non-convex setting. Specifically, the KL inequality in (11) is a local property that only needs to hold within a neighborhood of $\bar{\boldsymbol{x}}$. In addition, Definition 1 exhibits a slight difference from the classical analysis, with an additional absolute value operation on the function values, that is, $|F(\boldsymbol{x}) - F(\bar{\boldsymbol{x}})|$. It is worth highlighting that the KL inequality in Definition 1 is also easily satisfied as that used in the classical analysis. To support this claim, we provide some illustrative examples as follows.

(a) Some of the real polynomial functions, such as $F(x) = x^2 - 2x + 1$ satisfies Definition 1 at the point $\bar{x} = 1 \in \mathbb{R}$. Specifically, there exists $\eta = 1$, a neighborhood $U = (0, 2)$, and a concave function $\rho(z) = 2\sqrt{z}$ with $\rho'(z) = 1/\sqrt{z}$ for any $z > 0$, such that for any $x \in U \cap \{x \in \mathbb{R} : 0 < x^2 - 2x + 1 < 1\} = (0, 1) \cup (1, 2)$, we have

$$\frac{1}{\rho'\big(|F(x) - F(\bar{x})|\big)} = \sqrt{(x-1)^2} = |x - 1| \leq |2x - 2| = |F'(x)|,$$

which indicates that the KL inequality in (11) is valid.

(b) The logistic function $F(x) = \ln(1 + \exp(-x))$ fulfills Definition 1 at the point $\bar{x} = -1 \in \mathbb{R}$. That is, there exist $\eta = 1$, a neighborhood $U = (-2, 0)$, and a concave function $\rho(z) = 2z$ with $\rho'(z) = 2$ for any $z > 0$, such that for all $x \in U \cap \big\{x \in \mathbb{R} : 0 < |\ln(1 + \exp(-x)) - \ln(1 + \exp(1))| < 1\big\} = (-2, -1) \cup (-1, 0)$, we have

$$\frac{1}{\rho'\big(|F(x) - F(\bar{x})|\big)} = \frac{1}{2} \leq \frac{\exp(-x)}{1 + \exp(-x)} = \Big|\frac{-\exp(-x)}{1 + \exp(-x)}\Big| = |F'(x)|,$$

where the inequality follows from $x < 0$ and this means that the KL inequality in (11) holds.

(c) The non-convex PL function $F(x) = x^2 + 3\sin^2(x)$ presented in Karimi et al. (2016) meets Definition 1 at the point $\bar{x} = 0 \in \mathbb{R}$. In particular, there exist $\eta = 1/4 + 3\sin^2(1/2) > 0$, a neighborhood $U = (-1, 1)$, and a concave function $\rho(z) = 8\sqrt{z}$ with $\rho'(z) = 4/\sqrt{z}$ for any $z > 0$, such that for all $x \in U \cap \big\{x \in \mathbb{R} : 0 < x^2 + 3\sin^2(x) < 1/4 + 3\sin^2(1/2)\big\} = (-1/2, 0) \cup (0, 1/2)$, we have

$$\frac{1}{\rho'(|F(x) - F(\bar{x})|)} = \frac{\sqrt{x^2 + 3\sin^2(x)}}{4} \leq |2x + 3\sin(2x)| = |F'(x)|,$$

which implies that the KL inequality in (11) holds.

In particular, due to the non-descent nature of SMG algorithm, the KL property (with absolute value on function values) in Definition 1 is critical for theoretical analysis. However, only the KL inequality in (11) is not sufficient enough to attain the strong limit-point convergence. Therefore, we present an important property of $\rho(x)$ in the following definition.

Definition 2 (Quasi-additivity, Li et al. 2023) *Let $\rho : [0, \eta) \to \mathbb{R}^+$ be a continuous and differentiable function. If there exists a constant $C_\rho > 0$ such that*

$$\frac{1}{\rho'(x+y)} \leq C_\rho \left[\frac{1}{\rho'(x)} + \frac{1}{\rho'(y)} \right], \tag{12}$$

then we say that $\rho(\cdot)$ satisfies the quasi-additivity property.

Notice that the function satisfying (10) is also called the desingularizing function, see Definition 41 of Lin et al. (2020). Further, if the desingularizing function ρ is also quasi-additive (see Definition 2), we follow Li et al. (2023) to write it as $\rho \in \mathcal{Q}_\eta$. In particular, for some $c > 0$ and any $0 \leq \theta < 1$, the function $\rho(x) = cx^{1-\theta} \in \mathcal{Q}_\eta$. This is follows from the fact that $\rho(x)$ is concave and

$$\frac{1}{\rho'(x+y)} = \frac{(x+y)^\theta}{c(1-\theta)} \leq \frac{x^\theta}{c(1-\theta)} + \frac{y^\theta}{c(1-\theta)} = \frac{1}{\rho'(x)} + \frac{1}{\rho'(y)}, \tag{13}$$

which indicates that the quasi-additivity in (12) holds with $C_\rho = 1$.

Further, to achieve the strong limit-point convergence of SMG, we present the definition of accumulation points. Specifically, let $\{\tilde{\boldsymbol{x}}_t\}$ be generated by Algorithm 1, then we define the associated set of the accumulation points as follows.

$$\mathcal{C} = \left\{ \bar{\boldsymbol{x}} \in \mathbb{R}^d : \exists \text{ a sub-sequence } \{t_k\} \subseteq \{t\} \text{ such that } \lim_{k \to \infty} \tilde{\boldsymbol{x}}_{t_k} = \bar{\boldsymbol{x}} \right\}. \tag{14}$$

Now, we formalize the main assumptions used to establish the strong convergence results.

Assumption 3 *We assume that the following conditions are satisfied.*

(*a*) *The function F is coercive, that is, F is bounded from below, and $F(\boldsymbol{x}) \to \infty$ when $\|\boldsymbol{x}\| \to \infty$.*

(*b*) *The function F satisfies the KL inequality at each point of the set $\mathcal{C}$, and the corresponding desingularizing function $\rho \in \mathcal{Q}_\eta$ is quasi-additive.*

Note that the coercivity condition in Assumption 3(a) is mild and often appears in the convergence analysis of different optimization methods, such as APG (Li and Lin, 2015), Adam (Barakat and Bianchi, 2020), PIGD (Sun et al., 2021), and PPGD (Yang and Li, 2023). Recently, Josz and Lai (2023) investigate the global stability of first-order methods for coercive functions from the differential inclusion viewpoint. In particular, the objective function in matrix completion (Ge et al., 2016) and the strongly convex functions are naturally coercive. Moreover, the regularization term is typically used in many machine learning tasks to avoid over-fitting (Bottou et al., 2018), which also results in the coercivity of objective functions being easily satisfied. It is worth emphasizing that Assumption 3 does not imply any convexity or strong convexity. Thus, this section explores the convergence performance of iterative sequence for SMG method in the non-convex setting.

Next, we present some useful lemmas for the analysis in the KL setting. Specifically, Lemma 2 is also known as the uniformized KL property (Bolte et al., 2014; Guo et al., 2023b), showing that if F satisfies the KL inequality at each point of $\bar{\boldsymbol{u}}_i \in \Omega$ with respect to the neighborhood U and the function $\rho_i \in \mathcal{Q}_{\eta_i}$, then it also satisfies the KL inequality on all $\bar{\boldsymbol{u}} \in \Omega$ with respect to the set Ω and the associate desingularizing function satisfies $\rho \in \mathcal{Q}_\eta$.

Lemma 2 (Lemma 3.5, Li et al. 2023) *Assume that Ω is a compact set and F is constant on Ω. If F satisfies the KL inequality with a quasi-additive desingularizing function at each point of Ω, then there exist ϵ, $\eta > 0$ and $\rho \in \mathcal{Q}_\eta$ such that for all $\bar{\boldsymbol{u}} \in \Omega$ and*

$$\boldsymbol{u} \in \{\boldsymbol{u} \in \mathbb{R}^d : \operatorname{dist}(\boldsymbol{u}, \Omega) < \epsilon\} \cap \{\boldsymbol{u} \in \mathbb{R}^d : 0 < |F(\boldsymbol{u}) - F(\bar{\boldsymbol{u}})| < \eta\},$$

we have

$$\rho'\big(|F(\boldsymbol{u}) - F(\bar{\boldsymbol{u}})|\big) \cdot \|\nabla F(\boldsymbol{u})\| \geq 1.$$

In particular, uniformized KL property plays an important role in achieving strong limit-point convergence results (Li et al., 2023; Guo et al., 2023b). To leverage the uniformized KL property in the analysis of SMG, we derive several properties of accumulation point set in the following lemma.

Lemma 3 *Suppose that Assumptions 1, 2 and 3(a) hold, and the stepsize conditions in Theorem 2 are satisfied, then we have*

(*a*) *the set $\mathcal{C}$ is non-empty and compact.*

(*b*) *the set $\mathcal{C} \subseteq \{\boldsymbol{x} \in \mathbb{R}^d : \nabla F(\boldsymbol{x}) = \boldsymbol{0}\}$.*

(*c*) *the function F is finite and takes the value as a constant on the set $\mathcal{C}$.*

4.2 Convergence of SMG under the Kurdyka-Łojasiewicz inequality

In this section, we investigate the convergence performance of Algorithm 1 under the KL inequality. Specifically, we achieve the convergence of sequence $\{\tilde{\boldsymbol{x}}_t\}$ generated by Algorithm 1 in Theorem 5, and the specific choices of appropriate stepsize for convergence are given in Corollary 3.

Theorem 5 *Let Assumptions 1, 2 and 3 hold, and $\tilde{\boldsymbol{x}}_t$ be generated by Algorithm 1. Suppose that the stepsize α_t is non-increasing and satisfies*

$$\begin{aligned}
&\sum_{t=1}^{\infty} \alpha_t = \infty, \quad \sum_{t=1}^{\infty} \alpha_t^3 < \infty, \quad \alpha_t \leq \frac{1}{nL\sqrt{K}}, \quad \sum_{t=1}^{\infty} \alpha_t \left[\rho'\left(\beta^{t-1}\right)\right]^{-1} < \infty, \\
&\sum_{t=2}^{\infty} \alpha_t \left[\rho'\left(\beta^t \sum_{j=2}^{t} \beta^{-j} \alpha_j^3\right)\right]^{-1} < \infty, \quad \sum_{t=2}^{\infty} \alpha_t \left[\rho'\left(\sum_{j=t}^{\infty} \alpha_j^3\right)\right]^{-1} < \infty,
\end{aligned} \tag{15}$$

where $K = \max\left\{2, \frac{9(\sigma_1+\beta)}{1-\beta}\right\}$ is a finite value. Then we have

(*a*) *the sequence $\{\tilde{\boldsymbol{x}}_t\}$ has finite length, that is, $\sum_{t=1}^{\infty} \|\tilde{\boldsymbol{x}}_t - \tilde{\boldsymbol{x}}_{t-1}\| < \infty$.*

(*b*) *the last-iterate $\tilde{\boldsymbol{x}}_T$ converges to some stationary point $\boldsymbol{x}^*$ of F.*

Theorem 5 discusses the strong limit-point convergence property of SMG under the KL condition, showing that the last-iterate $\tilde{\boldsymbol{x}}_T$ is convergent and converges to a stationary point of F. Note that the stepsize conditions in Theorem 5 may not be very intuitive, see equation (15). For clarity, we then provide the possible selections of stepsize for convergence in the following corollary.

Corollary 3 *Suppose that Assumptions 1 and 2 hold, and Assumption 3 is satisfied with $\rho(x) = cx^{1-\theta}$ ($c > 0,\ 0 < \theta < 1$). If we run Algorithm 1 with*

$$\alpha_t = \frac{1}{t^p}, \qquad \frac{\theta+1}{3\theta+1} < p \leq 1, \tag{16}$$

then $\{\tilde{\boldsymbol{x}}_t\}$ has finite length and converges to some stationary point $\boldsymbol{x}^*$ of F.

Remark 4 *Note that compared to Theorems 1 and 2, we impose additional conditions on the stepsize α_t to achieve the more challenging strong limit-point convergence in Theorem 5. Nevertheless, the stepsize conditions in* (15) *are easily satisfied when the desingularizing function $\rho(\cdot)$ is given. As shown in Corollary 3, the polynomial-type stepsize in* (16) *fulfills the conditions of Theorem 5, thereby establishing the finite length property and the last-iterate convergence of iterative sequence for SMG method. In particular, the polynomial stepsize in Corollary 3 is often employed to guarantee the theoretical convergence of shuffling gradient algorithms (see, Gürbüzbalaban et al., 2021; Nguyen et al., 2021; Tran et al., 2021; Li et al., 2023).*

4.3 Convergence rate of SMG under the Łojasiewicz inequality

In this section, we explore the convergence rate of the iterative sequence $\{\tilde{\boldsymbol{x}}_t\}$ under the Łojasiewicz inequality (for short, L inequality). Specifically, this section focuses on the KL property with the popular desingularizing function $\rho(x) = cx^{1-\theta}$, where $0 \leq \theta < 1$ and $c > 0$. In this case, the KL inequality in (11) becomes the L inequality, that is, for all $\boldsymbol{x} \in U \cap \big\{\boldsymbol{x} \in \mathbb{R}^d : 0 < |F(\boldsymbol{x}) - F(\bar{\boldsymbol{x}})| < \eta\big\}$, we have

$$\|\nabla F(\boldsymbol{x})\| \geq \frac{1}{c(1-\theta)} |F(\boldsymbol{x}) - F(\bar{\boldsymbol{x}})|^{\theta}, \tag{17}$$

where U is a neighborhood of $\bar{\boldsymbol{x}}$, c is the corresponding KL constant, and $0 \leq \theta < 1$ is referred to as the KL exponent. In particular, the L inequality characterizes the geometric relationship between gradient norms and function values. In order to achieve the rate of the last-iterate convergence for SMG method, we first give the following proposition, and the main result is in Theorem 6.

Proposition 4 *Let e_t, λ_t and γ_t be three non-negative real sequences. Assume that $\lambda_t \leq 1$ and $e_{t+1} \leq (1-\lambda_t)e_t + \gamma_t$, then we have*

$$e_{_{T+1}} \leq \exp\bigg(-\sum_{t=1}^{T} \lambda_t\bigg) e_1 + \sum_{t=1}^{T} \exp\bigg(-\sum_{i=t+1}^{T} \lambda_i\bigg) \gamma_t.$$

By combining the L inequality in (17) and the conclusion of Proposition 4, we can obtain the convergence rate of $\{\tilde{\boldsymbol{x}}_t\}$ in the following theorem.

Theorem 6 *Let Assumptions 1 and 2 be satisfied, and Assumption 3 hold with $\rho(x) = c\sqrt{x}$. Suppose that $\tilde{\boldsymbol{x}}_t$ is generated by Algorithm 1, and the stepsize $\alpha_t = \frac{10c^2}{nt}$. Then for sufficiently large T, we have*

$$\|\tilde{\boldsymbol{x}}_{_T} - \boldsymbol{x}^*\| = \mathcal{O}\bigg(\frac{1}{T^2}\bigg) + \mathcal{O}\bigg(\frac{1}{T}\bigg), \tag{18}$$

where $\boldsymbol{x}^*$ is a stationary point of F, and this means that the squared norm $\|\tilde{\boldsymbol{x}}_{_T} - \boldsymbol{x}^*\|^2 = \mathcal{O}(1/T^2)$.

Remark 5 *For the iterative sequence of the SMG algorithm, Theorem 6 reveals that the last-iterate $\tilde{\boldsymbol{x}}_{_T}$ converges to a stationary point $\boldsymbol{x}^*$ of the non-convex problem* (1) *at the rate of $\mathcal{O}(1/T^2) + \mathcal{O}(1/T)$, which is consistent with the state-of-the-art theoretical results for strongly convex (Nguyen et al., 2021) and KL (Li et al., 2023) functions. For clarity, we compare the convergence result of Theorem 6 with some of the existing works in Table 1. In particular, the PL condition is commonly employed to establish the last-iterate convergence and convergence rate of function values (Lei et al., 2020; Nguyen et al., 2021), which takes the form of*

$$\|\nabla F(\boldsymbol{x})\|^2 \geq 2\mu\big(F(\boldsymbol{x}) - F^*\big), \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^d, \tag{19}$$

where $\mu > 0$ is a constant and $F^* = \inf_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x})$ denotes the lower bound of F. For the finite-sum minimax problems, Das et al. (2022) and Cho and Yun (2023) explore the convergence of AGDA-RR and SGDA-RR algorithms under the PL condition, respectively. More specifically, when π^t is sampled uniformly without replacement from $[n]$, they obtain the rate of $\mathcal{O}(\exp(-T)) + \tilde{\mathcal{O}}(1/(nT^2))$ for function values by employing a constant stepsize that depends on the total number of epochs T. Under the L inequality and the diminishing stepsize condition, we prove that the squared norm of the iterative sequence converges at the rate of $\mathcal{O}(1/T^2)$ for arbitrary shuffling strategies. In particular, when the lower bound of F is achievable and the stationary point $\boldsymbol{x}^*$ is a global minimizer of problem (1)*, we have $F^* = \inf_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) = F(\boldsymbol{x}^*)$. Thus, it follows from Theorem 6 that*

$$F(\tilde{\boldsymbol{x}}_T) - F^* = F(\tilde{\boldsymbol{x}}_T) - F(\boldsymbol{x}^*) \leq \frac{L}{2}\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2 \overset{(18)}{=} \mathcal{O}\left(\frac{1}{T^2}\right),$$

where the first inequality holds because F is L-smooth and $\boldsymbol{x}^*$ is a stationary point of F, that is, $\nabla F(\boldsymbol{x}^*) = \mathbf{0}$. In this case, Theorem 6 can recover the last-iterate convergence rate of function values for any shuffling strategy under the PL condition, as obtained in Nguyen et al. (2021) (see the fifth row of Table 1). In addition, the PL condition in (19) *needs to be satisfied for all $\boldsymbol{x} \in \mathbb{R}^d$, whereas the L inequality only requires* (17) *to be valid within a neighborhood of $\bar{\boldsymbol{x}}$. Hence, the L inequality considered in this paper is easier to be fulfilled. Finally, the last column of Table 1 indicates that the convergence rate can be improved by a factor of n when using the uniform without-replacement sampling scheme. Whether uniform sampling can enhance the rate of shuffling gradient algorithms under the L inequality remains an interesting work for future research.*

Table 1: Comparison of the convergence results in different problem settings

Settings	**Algorithms**	**Stepsize**	**Convergence rates***
strongly convex	RR (Mishchenko et al., 2020)	constant	$\mathbb{E}\big[\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2\big] = \tilde{\mathcal{O}}\big(\frac{1}{nT^2}\big)$
	SMG (Tran et al., 2024)	constant/exponential	$\mathbb{E}\big[\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2\big] = \tilde{\mathcal{O}}\big(\frac{1}{nT^2}\big)$
PL condition (19)	Shuffling-type gradient (Nguyen et al., 2021)	diminishing	$\mathbb{E}\big[F(\tilde{\boldsymbol{x}}_T) - F^*\big] = \tilde{\mathcal{O}}\big(\frac{1}{nT^2}\big)$
		diminishing	$F(\tilde{\boldsymbol{x}}_T) - F^* = \tilde{\mathcal{O}}\big(\frac{1}{T^2}\big)$
L inequality (17)	RR (Li et al., 2023)	diminishing	$\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2 = \mathcal{O}\big(\frac{1}{T^2}\big)$
	SMG (**this paper**)	diminishing	$\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*\|^2 = \mathcal{O}\big(\frac{1}{T^2}\big)$

* where T is the total number of epochs, n is the number of loss functions, $\boldsymbol{x}^*$ is the minimizer in the strongly convex setting and the stationary point under the non-convex L inequality, and $F^* = \inf_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x})$ is the lower bound of the PL function F. In addition, the in-expectation results in the second to fourth rows are obtained under the uniform without-replacement sampling scheme, and the other results hold for any shuffling strategy.

5. Conclusion

This paper has investigated the last-iterate convergence performance of SMG method for solving finite-sum optimization problems. Specifically, we have provided the last-iterate convergence guarantees for non-convex objectives, and the theoretical results hold for arbitrary shuffling strategies. In the non-convex setting, we have demonstrated that the function values of the last-iterate converge to a finite value and the gradient norms of the last-iterate exhibit convergence to zero. Moreover, we have also obtained the asymptotic convergence rate of gradient norms at the minimum-iterate,

which can arbitrarily approach $o(1/T^{2/3})$. By leveraging the uniform without-replacement sampling strategy, the convergence rate of the minimum-iterate has been improved to $\mathcal{O}(1/(n^{1/3}T^{2/3}))$ with an additional factor of $1/n^{1/3}$. Further, when the KL property is satisfied, our analysis has achieved the strong limit-point convergence results for SMG method, providing theoretical insight into the behavior of iterative sequences. In particular, we have proven that the whole sequence of iterates converges to a stationary point of the non-convex problem. Additionally, we have also attained the sublinear rate of $\mathcal{O}(1/T^2)+\mathcal{O}(1/T)$ for SMG method with an appropriate selection of stepsize. From the practical perspective, the last-iterate typically serves as the output of algorithm without additional storage or computation requirements. Therefore, the last-iterate convergence findings presented in this paper can better reflect the actual performance of algorithms, thereby effectively closing the gap between theory and practice.

Acknowledgments

This work was funded in part by the National Natural Science Foundation of China (No. 62176051), in part by the Science and Technology Development Plan Project of Jilin Province, China (No. 20240101370JC), and in part by the Scientific Research Program of Jilin Provincial Department of Education. The authors would like to thank the anonymous reviewers for their constructive and expert comments and suggestions, which have led to important improvements.

Appendix A. Basic Estimates of SMG

This section provides some basic estimations for the SMG algorithm. Specifically, we first present a useful result on the uniform sampling in Lemma 4, which plays an important role in analyzing the convergence of shuffling gradient methods. Next, we establish the upper bounds of $\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2$ and $\sum_{i=0}^{n-1}\mathbb{E}[\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2]$ in Lemma 5. The upper bounds of $\sum_{i=0}^{n-1}\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_i^{t-1}\|^2$ and $\sum_{i=0}^{n-1}\mathbb{E}[\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_i^{t-1}\|^2]$ are then obtained in Lemma 6. Further, we achieve the bounds related to the sequences Φ_t and Ψ_t in Lemma 7. Additionally, the conclusion of Lemma 8 is adopted to achieve the last-iterate convergence rate of iterative sequences in Theorem 6.

Lemma 4 (Lemma 1, Mishchenko et al. 2020) *Let $X_1, X_2, \cdots, X_n$ be n fixed vectors in $\mathbb{R}^d$, and $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$ be their average. Fix any $k\in\{1, 2\cdots, n\}$, let $X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_k}$ be sampled uniformly without replacement from $\{X_1, X_2, \cdots, X_n\}$, and $\bar{X}_\pi=\frac{1}{k}\sum_{i=1}^k X_{\pi_i}$ be their average. Then the sample average and variance are given by*

$$\mathbb{E}\big[\bar{X}_\pi\big]=\bar{X},\qquad \mathbb{E}\big[\|\bar{X}_\pi-\bar{X}\|^2\big]=\frac{n-k}{k(n-1)}\frac{1}{n}\sum_{i=1}^n\|X_i-\bar{X}\|^2.$$

Lemma 5 *Let Assumptions 1 and 2 hold, if the stepsize $\alpha_t\le\frac{1}{\sqrt{2}nL}$, then for any $t\ge 1$, we have*

$$\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2\le n\beta^{t-1}D_1+2\sum_{j=2}^{t}\beta^{t-j}\Phi_j, \tag{20}$$

where $D_1=\frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2+\sigma_2)$ is a finite value (independent of n), and Φ_j is defined in (4). If π^t is uniformly sampled at random without replacement from $[n]$, then for any $t\ge 1$, we have

$$\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2\big]\le n\beta^{t-1}D_2+2\sum_{j=2}^{t}\beta^{t-j}\Psi_j, \tag{21}$$

where $D_2 = \frac{1}{L^2}(\sigma_1\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^1)\|^2] + \sigma_2)$ *is a finite value (independent of* n*), and* Ψ_j *is defined in* (5).

Proof First, by the iteration that $\boldsymbol{x}_{i+1}^t = \boldsymbol{x}_i^t - \alpha_t \boldsymbol{m}_{i+1}^t$ in Algorithm 1, we have

$$\boldsymbol{x}_r^t - \boldsymbol{x}_s^t = \begin{cases} \alpha_t \sum_{i=r}^{s-1} \boldsymbol{m}_{i+1}^t, & r \le s. \\ -\alpha_t \sum_{i=s}^{r-1} \boldsymbol{m}_{i+1}^t, & r > s. \end{cases} \tag{22}$$

Thus, it follows from $\boldsymbol{m}_{i+1}^t = \beta \boldsymbol{m}_0^t + (1-\beta)\boldsymbol{g}_i^t$ that

$$\begin{aligned} \|\boldsymbol{x}_r^t - \boldsymbol{x}_s^t\|^2 = \alpha_t^2 \Big\| \sum_{i=r\wedge s}^{r\vee s-1} \boldsymbol{m}_{i+1}^t \Big\|^2 &= \alpha_t^2 \Big\| \beta(r\vee s - r\wedge s)\boldsymbol{m}_0^t + (1-\beta)\sum_{i=r\wedge s}^{r\vee s-1} \boldsymbol{g}_i^t \Big\|^2 \\ &\le \beta\alpha_t^2 (r\vee s - r\wedge s)^2 \|\boldsymbol{m}_0^t\|^2 + (1-\beta)\alpha_t^2 \Big\| \sum_{i=r\wedge s}^{r\vee s-1} \boldsymbol{g}_i^t \Big\|^2, \end{aligned} \tag{23}$$

where the last inequality holds by the Jensen inequality with respect to the convex function $\|\cdot\|^2$. Upon let's set $r = j \ge 0$ and $s = 0$, then $r \vee s = j$, $r \wedge s = 0$ and equation (23) shows

$$\|\boldsymbol{x}_j^t - \boldsymbol{x}_0^t\|^2 \le \beta\alpha_t^2 j^2 \|\boldsymbol{m}_0^t\|^2 + (1-\beta)\alpha_t^2 \Big\| \sum_{i=0}^{j-1} \boldsymbol{g}_i^t \Big\|^2. \tag{24}$$

It follows from the update of $\boldsymbol{v}_i^t$ in Algorithm 1 that for any $t \ge 2$, we have

$$\boldsymbol{m}_0^t = \tilde{\boldsymbol{v}}_{t-1} = \boldsymbol{v}_n^{t-1} \overset{(6)}{=} \frac{1}{n}\sum_{i=0}^{n-1} \boldsymbol{g}_i^{t-1}. \tag{25}$$

Then we can estimate the first term in (24) as follows. Since π^{t-1} is a permutation of $[n]$, then

$$\begin{aligned} \|\boldsymbol{m}_0^t\|^2 \overset{(25)}{=} \Big\| \frac{1}{n}\sum_{i=0}^{n-1} \boldsymbol{g}_i^{t-1} \Big\|^2 &= \Big\| \frac{1}{n}\sum_{i=0}^{n-1} \nabla f(\boldsymbol{x}_i^{t-1}; \pi^{t-1}(i+1)) \Big\|^2 \\ &= \Big\| \frac{1}{n}\sum_{i=0}^{n-1} \Big(\nabla f(\boldsymbol{x}_i^{t-1}; \pi^{t-1}(i+1)) - \nabla f(\boldsymbol{x}_0^{t-1}; \pi^{t-1}(i+1)) \Big) + \nabla F(\boldsymbol{x}_0^{t-1}) \Big\|^2 \\ &\le 2\Big\| \frac{1}{n}\sum_{i=0}^{n-1} \Big(\nabla f(\boldsymbol{x}_i^{t-1}; \pi^{t-1}(i+1)) - \nabla f(\boldsymbol{x}_0^{t-1}; \pi^{t-1}(i+1)) \Big) \Big\|^2 + 2\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 \\ &\le \frac{2}{n}\sum_{i=0}^{n-1} \|\nabla f(\boldsymbol{x}_i^{t-1}; \pi^{t-1}(i+1)) - \nabla f(\boldsymbol{x}_0^{t-1}; \pi^{t-1}(i+1))\|^2 + 2\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 \\ &\le \frac{2L^2}{n}\sum_{i=0}^{n-1} \|\boldsymbol{x}_i^{t-1} - \boldsymbol{x}_0^{t-1}\|^2 + 2\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2, \end{aligned} \tag{26}$$

where the first inequality follows from $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \le 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$, the second inequality holds by $\|\frac{1}{n}\sum_{i=0}^{n-1} \boldsymbol{a}_i\|^2 \le \frac{1}{n}\sum_{i=0}^{n-1}\|\boldsymbol{a}_i\|^2$, and the last inequality is due to the L-smoothness assumption, that is, $\|\nabla f(\boldsymbol{x}; i) - \nabla f(\boldsymbol{y}; i)\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$. Now, we continue to estimate the second term in (24). For any $0 \le k \le l-1 \le n-1$, we can obtain

$$\Big\| \sum_{i=k}^{l-1} \boldsymbol{g}_i^t \Big\|^2 = \Big\| \sum_{i=k}^{l-1} \Big(\nabla f(\boldsymbol{x}_i^t; \pi^t(i+1)) - \nabla f(\boldsymbol{x}_0^t; \pi^t(i+1)) \Big) + \sum_{i=k}^{l-1} \nabla f(\boldsymbol{x}_0^t; \pi^t(i+1)) \Big\|^2$$

$$
\begin{aligned}
&\leq 2\Big\|\sum_{i=k}^{l-1}\Big(\nabla f(\boldsymbol{x}_i^t;\pi^t(i+1))-\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\Big)\Big\|^2+2\Big\|\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\Big\|^2\\
&\leq 2(l-k)\sum_{i=k}^{l-1}\|\nabla f(\boldsymbol{x}_i^t;\pi^t(i+1))-\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\|^2+2(l-k)\sum_{i=k}^{l-1}\|\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\|^2\\
&\leq 2L^2(l-k)\sum_{i=k}^{l-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2+2(l-k)\sum_{i=k}^{l-1}\|\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\|^2\\
&\leq 2L^2(l-k)\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2+2n(l-k)\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f(\boldsymbol{x}_0^t;i+1)\|^2\\
&\leq 2L^2(l-k)\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2+2\sigma_1 n(l-k)\|\nabla F(\boldsymbol{x}_0^t)\|^2+2\sigma_2 n(l-k),
\end{aligned}\tag{27}
$$

where the second inequality holds by $\|\sum_{i=k}^{l-1}\boldsymbol{a}_i\|^2\leq(l-k)\sum_{i=k}^{l-1}\|\boldsymbol{a}_i\|^2$, the third inequality is from Assumption 1, the fourth inequality is due to $0\leq k\leq l-1\leq n-1$, that is, $\sum_{i=k}^{l-1}\|\boldsymbol{a}\|^2\leq\sum_{i=0}^{n-1}\|\boldsymbol{a}\|^2$, and the last inequality is by Assumption 2. Upon setting $k=0$ and $l=j$ in (27), we can get

$$
\Big\|\sum_{i=0}^{j-1}\boldsymbol{g}_i^t\Big\|^2\leq 2L^2 j\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2+2\sigma_1 nj\|\nabla F(\boldsymbol{x}_0^t)\|^2+2\sigma_2 nj.\tag{28}
$$

By substituting (26) and (28) back into (24), we have

$$
\begin{aligned}
\|\boldsymbol{x}_0^t-\boldsymbol{x}_j^t\|^2\leq{}&\frac{2L^2\beta\alpha_t^2 j^2}{n}\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2+2\beta\alpha_t^2 j^2\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\\
&+\ 2L^2(1-\beta)\alpha_t^2 j\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2+2\sigma_1(1-\beta)n\alpha_t^2 j\|\nabla F(\boldsymbol{x}_0^t)\|^2+2\sigma_2(1-\beta)n\alpha_t^2 j.
\end{aligned}
$$

Upon summing up the above equation over j from 0 to $n-1$, we arrive at

$$
\begin{aligned}
\sum_{j=0}^{n-1}\|\boldsymbol{x}_0^t-\boldsymbol{x}_j^t\|^2\leq{}&\frac{2L^2\beta\alpha_t^2}{n}\sum_{j=0}^{n-1}j^2\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2+2\beta\alpha_t^2\sum_{j=0}^{n-1}j^2\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2+2L^2(1-\beta)\\
&\times\alpha_t^2\sum_{j=0}^{n-1}j\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2+2\sigma_1(1-\beta)n\alpha_t^2\sum_{j=0}^{n-1}j\|\nabla F(\boldsymbol{x}_0^t)\|^2+2\sigma_2(1-\beta)n\alpha_t^2\sum_{j=0}^{n-1}j\\
\leq{}&n^2L^2\beta\alpha_t^2\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2+n^2L^2(1-\beta)\alpha_t^2\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2\\
&+\underbrace{n^3\alpha_t^2\Big(\beta\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2+\sigma_1(1-\beta)\|\nabla F(\boldsymbol{x}_0^t)\|^2+\sigma_2(1-\beta)\Big)}_{\Phi_t},
\end{aligned}\tag{29}
$$

where the last inequality is from $\sum_{j=0}^{n-1}j^2\leq\frac{n^3}{2}$ and $\sum_{j=0}^{n-1}j\leq\frac{n^2}{2}$. Upon setting $\Delta_t=\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2$ and $\Phi_t=n^3\alpha_t^2\big(\beta\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2+\sigma_1(1-\beta)\|\nabla F(\boldsymbol{x}_0^t)\|^2+\sigma_2(1-\beta)\big)$, then it follows that

$$
\Delta_t\leq n^2L^2\beta\alpha_t^2\Delta_{t-1}+n^2L^2(1-\beta)\alpha_t^2\Delta_t+\Phi_t
$$

$$\leq n^2L^2\beta\alpha_t^2\Delta_{t-1} + n^2L^2\alpha_t^2\Delta_t + \Phi_t \leq \frac{\beta}{2}\Delta_{t-1} + \frac{1}{2}\Delta_t + \Phi_t,$$

where the second inequality follows from $1-\beta \leq 1$, and the last inequality holds by the stepsize condition of $\alpha_t \leq \frac{1}{\sqrt{2}nL}$. Finally, rearranging the above equation, we have

$$\begin{aligned}\Delta_t &\leq \beta\Delta_{t-1} + 2\Phi_t \leq \beta(\beta\Delta_{t-2} + 2\Phi_{t-1}) + 2\Phi_t = \beta^2\Delta_{t-2} + 2\big(\beta\Phi_{t-1} + \Phi_t\big)\\ &\leq \beta^{t-1}\Delta_1 + 2\big(\beta^{t-2}\Phi_2 + \cdots + \beta\Phi_{t-1} + \Phi_t\big) = \beta^{t-1}\Delta_1 + 2\sum_{j=2}^{t}\beta^{t-j}\Phi_j.\end{aligned}$$

Using similar technique to (29) and $\boldsymbol{m}_0^1 = \mathbf{0}$, we can get that $\Delta_1 \leq \frac{1}{2}\Delta_1 + n^3\alpha_1^2(1-\beta)(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$, which implies that $\Delta_1 \leq 2n^3\alpha_1^2(1-\beta)(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2) \leq \frac{n}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$. Hence, the conclusion (20) is obtained by substituting the bound of Δ_1 in the above equation.

Next, if π^t is sampled uniformly without replacement from $[n]$, then for any $0 \leq k \leq l-1 \leq n-1$,

$$\begin{aligned}&\mathbb{E}_t\left[\left\|\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\right\|^2\right] \;=\; (l-k)^2\mathbb{E}_t\left[\left\|\frac{1}{l-k}\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\right\|^2\right]\\
&=\; (l-k)^2\mathbb{E}_t\left[\left\|\frac{1}{l-k}\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1)) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1) + \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1)\right\|^2\right]\\
&=\; (l-k)^2\mathbb{E}_t\left[\left\|\frac{1}{l-k}\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1)) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1)\right\|^2\right] + (l-k)^2\|\nabla F(\boldsymbol{x}_0^t)\|^2\\
&+\; 2(l-k)^2\left\langle\mathbb{E}_t\left[\frac{1}{l-k}\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\right] - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1),\; \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1)\right\rangle\\
&\overset{(31)}{=}\; (l-k)^2\frac{n-(l-k)}{(l-k)(n-1)}\frac{1}{n}\sum_{i=0}^{n-1}\left\|\nabla f(\boldsymbol{x}_0^t;i+1) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1)\right\|^2 + (l-k)^2\|\nabla F(\boldsymbol{x}_0^t)\|^2\\
&\leq\; (l-k)\frac{1}{n}\sum_{i=0}^{n-1}\left\|\nabla f(\boldsymbol{x}_0^t;i+1) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1)\right\|^2 + (l-k)^2\|\nabla F(\boldsymbol{x}_0^t)\|^2\\
&\leq\; 2(l-k)\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f(\boldsymbol{x}_0^t;i+1)\|^2 + 2(l-k)\|\nabla F(\boldsymbol{x}_0^t)\|^2 + (l-k)^2\|\nabla F(\boldsymbol{x}_0^t)\|^2\\
&\leq\; (l-k)^2\|\nabla F(\boldsymbol{x}_0^t)\|^2 + 2(1+\sigma_1)(l-k)\|\nabla F(\boldsymbol{x}_0^t)\|^2 + 2\sigma_2(l-k), \qquad (30)\end{aligned}$$

where the last equality holds because π^t is uniformly sampled without replacement from $[n]$, thus Lemma 4 indicates that

$$\begin{aligned}&\mathbb{E}_t\left[\frac{1}{l-k}\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\right] = \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1),\\
&\mathbb{E}_t\left[\left\|\frac{1}{l-k}\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1)) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1)\right\|^2\right]\\
&= \frac{n-(l-k)}{(l-k)(n-1)}\frac{1}{n}\sum_{i=0}^{n-1}\left\|\nabla f(\boldsymbol{x}_0^t;i+1) - \frac{1}{n}\sum_{i=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;i+1)\right\|^2. \qquad (31)\end{aligned}$$

Moreover, in (30), the first inequality holds by $k \leq l-1$, that is, $n-(l-k) \leq n-1$, the second inequality follows from $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$, and the last inequality is due to Assumption 2. Further, similar to the proof of (27), we can obtain

$$
\begin{aligned}
\mathbb{E}\left[\Big\|\sum_{i=k}^{l-1}\boldsymbol{g}_i^t\Big\|^2\right] \;&\leq\; 2\mathbb{E}\left[\Big\|\sum_{i=k}^{l-1}\nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\Big\|^2\right] \\
&+\; 2\mathbb{E}\left[\Big\|\sum_{i=k}^{l-1}\Big(\nabla f(\boldsymbol{x}_i^t;\pi^t(i+1)) - \nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\Big)\Big\|^2\right] \\
&\overset{(30)}{\leq}\; 2(l-k)^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + 4(1+\sigma_1)(l-k)\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] \\
&+\; 4\sigma_2(l-k) + 2(l-k)\sum_{i=0}^{n-1}\mathbb{E}\big[\|\nabla f(\boldsymbol{x}_i^t;\pi^t(i+1)) - \nabla f(\boldsymbol{x}_0^t;\pi^t(i+1))\|^2\big] \\
&\leq\; 2(l-k)^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + 4(1+\sigma_1)(l-k)\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] \\
&+\; 4\sigma_2(l-k) + 2L^2(l-k)\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2\big], \qquad (32)
\end{aligned}
$$

where the second inequality is from (30) with total expectation and $\|\sum_{i=k}^{l-1}\boldsymbol{a}_i\|^2 \leq (l-k)\sum_{i=0}^{n-1}\|\boldsymbol{a}_i\|^2$, and the last inequality holds by Assumption 1. Then it follows from (24) that

$$
\begin{aligned}
&\mathbb{E}\big[\|\boldsymbol{x}_0^t-\boldsymbol{x}_j^t\|^2\big] \overset{(24)}{\leq} \beta\alpha_t^2 j^2\mathbb{E}\big[\|\boldsymbol{m}_0^t\|^2\big] + (1-\beta)\alpha_t^2\mathbb{E}\left[\Big\|\sum_{i=0}^{j-1}\boldsymbol{g}_i^t\Big\|^2\right] \\
&\leq \beta\alpha_t^2 j^2\bigg(\frac{2L^2}{n}\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2\big] + 2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big]\bigg) + (1-\beta)\alpha_t^2 \\
&\times\bigg(2j^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + 4(1+\sigma_1)j\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + 4\sigma_2 j + 2L^2 j\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2\big]\bigg) \\
&\leq \frac{2\beta L^2\alpha_t^2 j^2}{n}\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2\big] + 2\beta\alpha_t^2 j^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big] + 2(1-\beta)\alpha_t^2 j^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] \\
&+ 4(1-\beta)(1+\sigma_1)\alpha_t^2 j^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + 4\sigma_2(1-\beta)\alpha_t^2 j + 2L^2\alpha_t^2 j\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2\big],
\end{aligned}
$$

where the second inequality holds by the bound of $\|\boldsymbol{m}_0^t\|^2$ in (26) and (32) with $l=j$ and $k=0$, and the last inequality is due to $1-\beta \leq 1$ and $j \leq j^2$. Similar to the proof of (29), summing the above equation over j from 0 to $n-1$ and using $\sum_{j=0}^{n-1} j \leq \frac{n^2}{2}$ and $\sum_{j=0}^{n-1} j^2 \leq \frac{n^3}{2}$, we arrive at

$$
\begin{aligned}
\sum_{j=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_0^t-\boldsymbol{x}_j^t\|^2\big] &\leq n^2L^2\beta\alpha_t^2\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2\big] + n^2L^2\alpha_t^2\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^t-\boldsymbol{x}_0^t\|^2\big] \\
&+ \underbrace{n^2\alpha_t^2\Big(n\beta\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big] + n(1-\beta)(3+2\sigma_1)\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + 2\sigma_2(1-\beta)\Big)}_{\Psi_t}.
\end{aligned}
$$

By introducing $\Psi_t = n^2\alpha_t^2\big(n\beta\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2] + n(1-\beta)(3+2\sigma_1)\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^t)\|^2] + 2\sigma_2(1-\beta)\big)$ and using $\Delta_t = \sum_{i=0}^{n-1}\|\boldsymbol{x}_0^t - \boldsymbol{x}_i^t\|^2$, the above equation implies that

$$\mathbb{E}[\Delta_t] \leq n^2L^2\beta\alpha_t^2\mathbb{E}[\Delta_{t-1}] + n^2L^2\alpha_t^2\mathbb{E}[\Delta_t] + \Psi_t \leq \frac{\beta}{2}\mathbb{E}[\Delta_{t-1}] + \frac{1}{2}\mathbb{E}[\Delta_t] + \Psi_t,$$

where the last inequality holds by $\alpha_t \leq \frac{1}{\sqrt{2n}L}$. Upon rearranging the above equation, we have

$$\mathbb{E}[\Delta_t] \leq \beta\mathbb{E}[\Delta_{t-1}] + 2\Psi_t \leq \cdots \leq \beta^{t-1}\mathbb{E}[\Delta_1] + 2\sum_{j=2}^{t}\beta^{t-j}\Psi_j.$$

Finally, recalling that $\Delta_1 \leq \frac{n}{L^2}\big(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2\big)$ in the last part of the proof of (20), then taking the total expectation, we have $\mathbb{E}[\Delta_1] \leq \frac{n}{L^2}\big(\sigma_1\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^1)\|^2] + \sigma_2\big)$. Hence, we obtain the conclusion (21) by plugging the bound of $\mathbb{E}[\Delta_1]$ in the above equation. This completes the proof. ■

Lemma 6 *Let Assumptions 1 and 2 hold, if the stepsize $\alpha_t \leq \frac{1}{\sqrt{2n}L}$, then for any $t \geq 2$, we have*

$$\sum_{i=0}^{n-1}\|\boldsymbol{x}_n^{t-1} - \boldsymbol{x}_i^{t-1}\|^2 \leq 2n\beta^{t-2}D_1 + 4\sum_{j=2}^{t-1}\beta^{t-1-j}\Phi_j, \tag{33}$$

where $D_1 = \frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$ is a finite value (independent of n), and Φ_j is defined in (4). If π^t is uniformly sampled at random without replacement from $[n]$, then for any $t \geq 2$, we have

$$\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_n^{t-1} - \boldsymbol{x}_i^{t-1}\|^2\big] \leq 2n\beta^{t-2}D_2 + 4\sum_{j=2}^{t-1}\beta^{t-1-j}\Psi_j, \tag{34}$$

where $D_2 = \frac{1}{L^2}(\sigma_1\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^1)\|^2] + \sigma_2)$ is a finite value (independent of n), and Ψ_j is defined in (5).

Proof First, setting $r = n$ and $s = j \leq n$, we know that $r \wedge s = j$ and $r \vee s = n$, and (23) becomes

$$\|\boldsymbol{x}_n^{t-1} - \boldsymbol{x}_j^{t-1}\|^2 \leq \beta\alpha_{t-1}^2(n-j)^2\|\boldsymbol{m}_0^{t-1}\|^2 + (1-\beta)\alpha_{t-1}^2\bigg\|\sum_{i=j}^{n-1}\boldsymbol{g}_i^{t-1}\bigg\|^2. \tag{35}$$

For the first term in (35), the equation (26) indicates that for any $t \geq 3$,

$$\|\boldsymbol{m}_0^{t-1}\|^2 \leq \frac{2L^2}{n}\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-2} - \boldsymbol{x}_0^{t-2}\|^2 + 2\|\nabla F(\boldsymbol{x}_0^{t-2})\|^2. \tag{36}$$

For the second term in (35), by choosing $k = j$ and $l = n$ in (27), we arrive at

$$\bigg\|\sum_{i=j}^{n-1}\boldsymbol{g}_i^{t-1}\bigg\|^2 \leq 2L^2(n-j)\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1} - \boldsymbol{x}_0^{t-1}\|^2 + 2\sigma_1 n(n-j)\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 + 2\sigma_2 n(n-j). \tag{37}$$

Then we substitute (36) and (37) in (35) to obtain

$$\|\boldsymbol{x}_n^{t-1} - \boldsymbol{x}_j^{t-1}\|^2 \leq \frac{2L^2\beta\alpha_{t-1}^2}{n}(n-j)^2\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-2} - \boldsymbol{x}_0^{t-2}\|^2$$

$$+ 2L^2(1-\beta)(n-j)\alpha_{t-1}^2 \sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2 + 2\beta\alpha_{t-1}^2(n-j)^2\|\nabla F(\boldsymbol{x}_0^{t-2})\|^2$$
$$+ 2\sigma_1(1-\beta)\alpha_{t-1}^2 n(n-j)\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 + 2\sigma_2(1-\beta)\alpha_{t-1}^2 n(n-j).$$

Upon summing the above equation over j from 0 to $n-1$, we have

$$\begin{aligned}
\sum_{j=0}^{n-1}\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_j^{t-1}\|^2 \le{}& \frac{2L^2\beta\alpha_{t-1}^2}{n}\sum_{j=0}^{n-1}(n-j)^2\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-2}-\boldsymbol{x}_0^{t-2}\|^2 + 2\beta\alpha_{t-1}^2\sum_{j=0}^{n-1}(n-j)^2 \\
&\times \|\nabla F(\boldsymbol{x}_0^{t-2})\|^2 + 2L^2(1-\beta)\alpha_{t-1}^2\sum_{j=0}^{n-1}(n-j)\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2 + 2\sigma_1 \\
&\times (1-\beta)\alpha_{t-1}^2 n\sum_{j=0}^{n-1}(n-j)\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 + 2\sigma_2(1-\beta)\alpha_{t-1}^2 n\sum_{j=0}^{n-1}(n-j) \\
\le{}& 2n^2L^2\beta\alpha_{t-1}^2\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-2}-\boldsymbol{x}_0^{t-2}\|^2 + 2n^3\beta\alpha_{t-1}^2\|\nabla F(\boldsymbol{x}_0^{t-2})\|^2 + 2n^2L^2(1-\beta)\alpha_{t-1}^2 \\
&\times \sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2 + 2n^3\sigma_1(1-\beta)\alpha_{t-1}^2\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 + 2n^3\sigma_2(1-\beta)\alpha_{t-1}^2 \\
\le{}& 2n^2L^2\beta\alpha_{t-1}^2\underbrace{\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-2}-\boldsymbol{x}_0^{t-2}\|^2}_{\Delta_{t-2}} + 2n^2L^2\alpha_{t-1}^2\underbrace{\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2}_{\Delta_{t-1}} \\
&+ \underbrace{2n^3\alpha_{t-1}^2\Big(\beta\|\nabla F(\boldsymbol{x}_0^{t-2})\|^2 + \sigma_1(1-\beta)\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 + \sigma_2(1-\beta)\Big)}_{\Phi_{t-1}}, \qquad (38)
\end{aligned}$$

where the second inequality holds because $\sum_{j=0}^{n-1}(n-j)^2 \le n^3$ and $\sum_{j=0}^{n-1}(n-j) \le n^2$, and the last inequality is by $\beta \ge 0$. It follows from the definitions of Δ_t and Φ_t that for any $t \ge 3$,

$$\begin{aligned}
\sum_{i=0}^{n-1}\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_i^{t-1}\|^2 \overset{(38)}{\le}{}& 2n^2L^2\beta\alpha_{t-1}^2\Delta_{t-2} + 2n^2L^2\alpha_{t-1}^2\Delta_{t-1} + 2\Phi_{t-1} \\
\le{}& \beta\Delta_{t-2} + \Delta_{t-1} + 2\Phi_{t-1} \le 2n\beta^{t-2}D_1 + 4\sum_{j=2}^{t-1}\beta^{t-1-j}\Phi_j,
\end{aligned}$$

where the second inequality follows from $\alpha_t \le \frac{1}{\sqrt{2}nL}$, the last inequality holds by (20) in Lemma 5, that is, $\Delta_t \le n\beta^{t-1}D_1 + 2\sum_{j=2}^{t}\beta^{t-j}\Phi_j$ is valid for any $t \ge 1$, and $D_1 = \frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$ is a finite value. Similar to the proof of (38), we can apply $\boldsymbol{m}_0^1 = \boldsymbol{0}$ to obtain that $\sum_{i=0}^{n-1}\|\boldsymbol{x}_n^1 - \boldsymbol{x}_i^1\|^2 \le 2n^2L^2\alpha_1^2\Delta_1 + 2n^3\alpha_1^2(1-\beta)(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2+\sigma_2) \le \Delta_1 + \frac{n}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2+\sigma_2) \le nD_1 + nD_1 = 2nD_1$. Therefore, by combining the bound of $\sum_{i=0}^{n-1}\|\boldsymbol{x}_n^1 - \boldsymbol{x}_i^1\|^2$ with the above equation, we can conclude that (33) holds for any $t \ge 2$.

Next, taking the total expectation on both sides of (35), we have

$$\mathbb{E}\big[\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_j^{t-1}\|^2\big] \le \beta\alpha_{t-1}^2(n-j)^2\mathbb{E}\big[\|\boldsymbol{m}_0^{t-1}\|^2\big] + (1-\beta)\alpha_{t-1}^2\mathbb{E}\left[\Big\|\sum_{i=j}^{n-1}\boldsymbol{g}_i^{t-1}\Big\|^2\right]$$

$$\leq \beta\alpha_{t-1}^2(n-j)^2\bigg(\frac{2L^2}{n}\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-2}-\boldsymbol{x}_0^{t-2}\|^2\big]+2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-2})\|^2\big]\bigg)+(1-\beta)\alpha_{t-1}^2\bigg(4(1+\sigma_1)(n-j)$$
$$\times\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big]+2(n-j)^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big]+4\sigma_2(n-j)+2L^2(n-j)\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2\big]\bigg)$$
$$\leq \frac{2L^2\beta\alpha_{t-1}^2}{n}(n-j)^2\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-2}-\boldsymbol{x}_0^{t-2}\|^2\big]+2\beta\alpha_{t-1}^2(n-j)^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-2})\|^2\big]+4(1-\beta)$$
$$\times(1+\sigma_1)\alpha_{t-1}^2(n-j)^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big]+2(1-\beta)\alpha_{t-1}^2(n-j)^2\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big]$$
$$+2L^2\alpha_{t-1}^2(n-j)\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2\big]+4\sigma_2(1-\beta)\alpha_{t-1}^2(n-j),$$

where the second inequality is from the bound of $\|\boldsymbol{m}_0^{t-1}\|^2$ in (36) and (32) with $k=j$ and $l=n$, and the last inequality holds by $n-j\leq(n-j)^2$ and $\beta\geq 0$. Similar to the proof of (38), we sum the above equation over j from 0 to $n-1$, and then use $\sum_{j=0}^{n-1}j\leq n^2$ and $\sum_{j=0}^{n-1}j^2\leq n^3$ to get

$$\sum_{j=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_j^{t-1}\|^2\big]\leq 2n^2L^2\beta\alpha_{t-1}^2\underbrace{\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-2}-\boldsymbol{x}_0^{t-2}\|^2\big]}_{\mathbb{E}[\Delta_{t-2}]}+2n^2L^2\alpha_{t-1}^2\underbrace{\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_i^{t-1}-\boldsymbol{x}_0^{t-1}\|^2\big]}_{\mathbb{E}[\Delta_{t-1}]}$$
$$+\underbrace{2n^2\alpha_{t-1}^2\Big(n\beta\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-2})\|^2\big]+n(1-\beta)(3+2\sigma_1)\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2\big]+2\sigma_2(1-\beta)\Big)}_{\Psi_{t-1}}.$$

By the conclusion (21) of Lemma 5, that is, $\mathbb{E}[\Delta_t]\leq n\beta^{t-1}D_2+2\sum_{j=2}^{t}\beta^{t-j}\Psi_j$ holds for any $t\geq 1$, we can get that for any $t\geq 3$,

$$\sum_{j=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_j^{t-1}\|^2\big]\leq 2n^2L^2\beta\alpha_{t-1}^2\mathbb{E}[\Delta_{t-2}]+2n^2L^2\alpha_{t-1}^2\mathbb{E}[\Delta_{t-1}]+2\Psi_{t-1}$$
$$\leq\beta\mathbb{E}[\Delta_{t-2}]+\mathbb{E}[\Delta_{t-1}]+2\Psi_{t-1}\leq 2n\beta^{t-2}D_2+4\sum_{j=2}^{t-1}\beta^{t-1-j}\Psi_j,$$

where the second inequality holds by $\alpha_t\leq\frac{1}{\sqrt{2}nL}$ and $D_2=\frac{1}{L^2}(\sigma_1\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^1)\|^2]+\sigma_2)$ is a finite value. Recalling that $\sum_{i=0}^{n-1}\|\boldsymbol{x}_n^1-\boldsymbol{x}_i^1\|^2\leq\Delta_1+\frac{n}{L^2}\big(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2+\sigma_2\big)$ in the last part of the proof of (33), we know that $\sum_{i=0}^{n-1}\mathbb{E}[\|\boldsymbol{x}_n^1-\boldsymbol{x}_i^1\|^2]\leq\mathbb{E}[\Delta_1]+\frac{n}{L^2}(\sigma_1\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^1)\|^2]+\sigma_2)\leq nD_2+nD_2=2nD_2$. Finally, combining the bound of $\sum_{i=0}^{n-1}\mathbb{E}[\|\boldsymbol{x}_n^1-\boldsymbol{x}_i^1\|^2]$ and the above equation, we can obtain that (34) holds for any $t\geq 2$. This completes the proof. ■

Lemma 7 *Assume that $\alpha_t>0$ is non-increasing, then the following holds.*

$$\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j\leq\frac{n^3\beta\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^1)\|^2+\frac{n^3(\sigma_1+\beta)}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^j)\|^2+\frac{\sigma_2n^3}{1-\beta}\sum_{j=2}^{T}\alpha_j^3,\tag{39}$$

$$\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j\leq\frac{n^3\beta\alpha_2^3}{1-\beta}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^1)\|^2\big]+\frac{2n^3(\sigma_1+2)}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^j)\|^2\big]+\frac{2\sigma_2n^2}{1-\beta}\sum_{j=2}^{T}\alpha_j^3,\tag{40}$$

where Φ_j and Ψ_j are defined in (4) *and* (5)*, respectively.*

Proof From the definition of Φ_j in (4), we have

$$\begin{aligned}\sum_{j=2}^{t}\beta^{t-j}\Phi_j &= \sum_{j=2}^{t}\beta^{t-j}\left(n^3\alpha_j^2\Big(\beta\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2+\sigma_1(1-\beta)\|\nabla F(\boldsymbol{x}_0^{j})\|^2+\sigma_2(1-\beta)\Big)\right)\\ &\le n^3\beta\underbrace{\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2}_{A_t}+\sigma_1 n^3\underbrace{\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2\|\nabla F(\boldsymbol{x}_0^{j})\|^2}_{B_t}+\sigma_2 n^3\underbrace{\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2}_{C_t},\end{aligned}\tag{41}$$

where the last inequality holds by $0<1-\beta\le 1$. Then it follows from the non-increasing of α_t that $\alpha_t\le\alpha_j$ for all $j\le t$. Thus, for the first term A_t in the RHS of (41), we have

$$\begin{aligned}\sum_{t=2}^{T}\alpha_t A_t &= \sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2 \le \sum_{t=2}^{T}\sum_{j=2}^{t}\beta^{t-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2\\ &= \sum_{j=2}^{T}\sum_{t=j}^{T}\beta^{t-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2 = \sum_{j=2}^{T}\Big(\sum_{t=j}^{T}\beta^t\Big)\beta^{-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2\\ &\overset{(i)}{\le}\sum_{j=2}^{T}\Big(\sum_{t=j}^{\infty}\beta^t\Big)\beta^{-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2 \overset{(ii)}{\le}\frac{1}{1-\beta}\sum_{j=2}^{T}\beta^j\,\beta^{-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2\\ &= \frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2 \le \frac{\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^{1})\|^2+\frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j})\|^2,\end{aligned}\tag{42}$$

where (i) and (ii) hold by $0\le\beta<1$ and the last inequality is from the non-increasing of α_t. For the second term B_t in the RHS of (41), we use the same approach to get

$$\begin{aligned}\sum_{t=2}^{T}\alpha_t B_t &= \sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2\|\nabla F(\boldsymbol{x}_0^{j})\|^2 \le \sum_{t=2}^{T}\sum_{j=2}^{t}\beta^{t-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j})\|^2\\ &= \sum_{j=2}^{T}\sum_{t=j}^{T}\beta^{t-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j})\|^2 \le \sum_{j=2}^{T}\Big(\sum_{t=j}^{\infty}\beta^t\Big)\beta^{-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j})\|^2\\ &\le \frac{1}{1-\beta}\sum_{j=2}^{T}\beta^j\beta^{-j}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j})\|^2 = \frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^{j})\|^2.\end{aligned}\tag{43}$$

Then we calculate the last term C_t in the RHS of (41) in the same way as follows.

$$\begin{aligned}\sum_{t=2}^{T}\alpha_t C_t &= \sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2 \le \sum_{t=2}^{T}\sum_{j=2}^{t}\beta^{t-j}\alpha_j^3 = \sum_{j=2}^{T}\sum_{t=j}^{T}\beta^{t-j}\alpha_j^3\\ &\le \sum_{j=2}^{T}\Big(\sum_{t=j}^{\infty}\beta^t\Big)\beta^{-j}\alpha_j^3 \le \frac{1}{1-\beta}\sum_{j=2}^{T}\beta^j\beta^{-j}\alpha_j^3 = \frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3.\end{aligned}\tag{44}$$

By combining the equations (42), (43) and (44), we finally have

$$\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j \overset{(41)}{\le} n^3\beta\sum_{t=2}^{T}\alpha_t A_t+\sigma_1 n^3\sum_{t=2}^{T}\alpha_t B_t+\sigma_2 n^3\sum_{t=2}^{T}\alpha_t C_t$$

$$
\begin{aligned}
&\leq n^3\beta\bigg(\frac{\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^1)\|^2+\frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^j)\|^2\bigg)\\
&+\sigma_1 n^3\bigg(\frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^j)\|^2\bigg)+\sigma_2 n^3\bigg(\frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\bigg)\\
&=\frac{n^3\beta\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^1)\|^2+\frac{n^3(\sigma_1+\beta)}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^j)\|^2+\frac{\sigma_2 n^3}{1-\beta}\sum_{j=2}^{T}\alpha_j^3,
\end{aligned}
$$

which implies that (39) is valid. For the conclusion in (40), we leverage similar proof technique to the one used in the above equation, together with the definition of Ψ_t in (5), to estimate the bound of $\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j$. Given that (39) and (40) follow a similar proof approach, we omit the proof details for (40) here. This completes the proof. ■

Lemma 8 *Assume that $0\leq\beta<1$, then we have*

$$
\begin{aligned}
&(a)\ \sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{\beta^{i-1}}{i}}\leq\frac{1}{\sqrt{2(1-\beta)}}\frac{1}{(1-\sqrt{\beta})^2},\\
&(b)\ \sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}}\leq\frac{1}{\sqrt{1-\beta}}\frac{4\sqrt{T}}{1-\sqrt{\beta}}+\frac{\sqrt{2}(\ln T+T)}{\sqrt{1-\beta}}+C_6,
\end{aligned}
$$

where $C_6=\frac{1}{\sqrt{1-\beta}}\frac{4}{(1-\sqrt{\beta})^2}+\frac{\sqrt{2}}{\sqrt{1-\beta}}$ is a finite value.

Proof First, it follows from $i\geq t\geq k\geq 2$ that $\frac{1}{i}\leq\frac{1}{2}$, thus we have

$$
\begin{aligned}
\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{\beta^{i-1}}{i}}&\leq\frac{1}{\sqrt{2}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\beta^{i-1}}\overset{(i)}{\leq}\frac{1}{\sqrt{2}}\frac{1}{\sqrt{1-\beta}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t-1}{2}}}{t}\\
&\overset{(ii)}{\leq}\frac{1}{\sqrt{2}}\frac{1}{\sqrt{1-\beta}}\sum_{k=2}^{T}\sum_{t=k}^{\infty}\beta^{\frac{t-1}{2}}\leq\frac{1}{\sqrt{2}}\frac{1}{\sqrt{1-\beta}}\frac{1}{1-\sqrt{\beta}}\sum_{k=2}^{T}\beta^{\frac{k-1}{2}}\\
&\leq\frac{1}{\sqrt{2}}\frac{1}{\sqrt{1-\beta}}\frac{1}{1-\sqrt{\beta}}\sum_{k=2}^{\infty}\beta^{\frac{k-1}{2}}\leq\frac{1}{\sqrt{2}}\frac{1}{\sqrt{1-\beta}}\frac{1}{(1-\sqrt{\beta})^2},
\end{aligned}
$$

where (i) follows from $0\leq\beta<1$, (ii) holds by $t\geq k$, that is, $\frac{1}{t}\leq\frac{1}{k}$, the last three inequalities are due to $0\leq\sqrt{\beta}<1$. Next, applying the fact that $\sum_{j=2}^{i}a_j\leq\sum_{j=2}^{t}a_j+\sum_{j=t}^{i}a_j$ for any $a_j\geq 0$, we can obtain

$$
\begin{aligned}
\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}&\leq\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{t}\frac{\beta^{i-j}}{j^2}+\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=t}^{i}\frac{\beta^{i-j}}{j^2}\\
&\leq\sum_{i=t}^{\infty}\sum_{j=2}^{t}\frac{\beta^{i-j}}{j^3}+\sum_{i=t}^{\infty}\sum_{j=t}^{i}\frac{\beta^{i-j}}{j^3}=\sum_{i=t}^{\infty}\beta^i\sum_{j=2}^{t}\frac{\beta^{-j}}{j^3}+\sum_{j=t}^{\infty}\sum_{i=j}^{\infty}\frac{\beta^{i-j}}{j^3}
\end{aligned}
$$

$$\leq \frac{\beta^t}{1-\beta}\sum_{j=2}^{t}\frac{\beta^{-j}}{j^3}+\sum_{j=t}^{\infty}\frac{\beta^{-j}}{j^3}\Big(\sum_{i=j}^{\infty}\beta^i\Big) \leq \frac{\beta^t}{1-\beta}\sum_{j=2}^{t}\frac{\beta^{-j}}{j^3}+\frac{1}{1-\beta}\sum_{j=t}^{\infty}\frac{1}{j^3},$$

where the second inequality holds by $i \geq j$, and the last two inequalities follow from $0 \leq \beta < 1$. Then it follows that

$$\begin{aligned}
\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}} &\leq \sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\frac{\beta^t}{1-\beta}\sum_{j=2}^{t}\frac{\beta^{-j}}{j^3}+\frac{1}{1-\beta}\sum_{j=t}^{\infty}\frac{1}{j^3}} \\
&\overset{(i)}{\leq} \sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\left(\sqrt{\frac{\beta^t}{1-\beta}\sum_{j=2}^{t}\frac{\beta^{-j}}{j^3}}+\sqrt{\frac{1}{1-\beta}\sum_{j=t}^{\infty}\frac{1}{j^3}}\right) \\
&= \frac{1}{\sqrt{1-\beta}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\sqrt{\sum_{j=2}^{t}\frac{\beta^{-j}}{j^3}}+\frac{1}{\sqrt{1-\beta}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{j=t}^{\infty}\frac{1}{j^3}} \\
&\overset{(46)}{\leq} \frac{1}{\sqrt{1-\beta}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\sqrt{\sum_{j=2}^{t}\frac{\beta^{-j}}{j^3}}+\frac{1}{\sqrt{1-\beta}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\frac{1}{2(t-1)^2}} \\
&\overset{(ii)}{\leq} \frac{1}{\sqrt{1-\beta}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\sum_{j=2}^{t}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}+\frac{1}{\sqrt{2(1-\beta)}}\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\frac{1}{(t-1)} \\
&\leq \frac{1}{\sqrt{1-\beta}}\underbrace{\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\sum_{j=2}^{t}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}}_{U_t}+\frac{1}{\sqrt{2(1-\beta)}}\underbrace{\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{(t-1)^2}}_{V_t},
\end{aligned} \tag{45}$$

where inequalities (i) and (ii) hold by $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$, and the third inequality follows from the integral test inequality, that is,

$$\sum_{j=t}^{\infty}\frac{1}{j^3} \leq \int_{t-1}^{\infty}\frac{1}{x^3}\,dx = -\frac{1}{2x^2}\Big|_{t-1}^{\infty} \leq \frac{1}{2(t-1)^2}. \tag{46}$$

The last inequality in (45) is due to the fact that $\frac{1}{t} \leq \frac{1}{t-1}$ for any $t \geq 2$. Upon we estimate U_t and V_t as follows. For U_t, we have

$$\begin{aligned}
U_t &= \sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\sum_{j=2}^{t}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}} \overset{(i)}{\leq} \sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\Big(\sum_{j=2}^{k}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}+\sum_{j=k}^{t}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}\Big) \\
&= \sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\sum_{j=2}^{k}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}+\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{\beta^{\frac{t}{2}}}{t}\sum_{j=k}^{t}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}} \\
&\overset{(ii)}{\leq} \sum_{k=2}^{T}\sum_{t=k}^{\infty}\beta^{\frac{t}{2}}\sum_{j=2}^{k}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}+\sum_{k=2}^{T}\sum_{t=k}^{\infty}\beta^{\frac{t}{2}}\sum_{j=k}^{t}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}} \\
&\leq \frac{1}{1-\sqrt{\beta}}\sum_{k=2}^{T}\beta^{\frac{k}{2}}\sum_{j=2}^{k}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}+\sum_{k=2}^{T}\sum_{j=k}^{\infty}\Big(\sum_{t=j}^{\infty}\beta^{\frac{t}{2}}\Big)\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}} \\
&\leq \frac{1}{1-\sqrt{\beta}}\sum_{j=2}^{T}\frac{\beta^{-\frac{j}{2}}}{j^{\frac{3}{2}}}\Big(\sum_{k=j}^{\infty}\beta^{\frac{k}{2}}\Big)+\frac{1}{1-\sqrt{\beta}}\sum_{k=2}^{T}\sum_{j=k}^{\infty}\frac{1}{j^{\frac{3}{2}}}
\end{aligned}$$

$$\leq \frac{1}{(1-\sqrt{\beta})^2}\sum_{j=2}^{T}\frac{1}{j^{\frac{3}{2}}}+\frac{1}{1-\sqrt{\beta}}\sum_{k=2}^{T}\sum_{j=k}^{\infty}\frac{1}{j^{\frac{3}{2}}},$$

where (i) holds because $\sum_{j=2}^{t}a_j\leq\sum_{j=2}^{k}a_j+\sum_{j=k}^{t}a_j$ is valid for any $a_j\geq 0$, (ii) is due to $t\geq k$, that is, $\frac{1}{t}\leq\frac{1}{k}$, and the last three inequalities follows from $0\leq\beta<1$. Upon using the integral test inequality, we can get that for any $t\geq 2$,

$$\sum_{j=t}^{\infty}\frac{1}{j^{\frac{3}{2}}}\leq\int_{t-1}^{\infty}\frac{1}{x^{\frac{3}{2}}}\,dx=-\frac{2}{\sqrt{x}}\Big|_{t-1}^{\infty}\leq\frac{2}{\sqrt{t-1}}. \tag{47}$$

Thus, we continue to estimate U_t as follows.

$$\begin{aligned}U_t &\overset{(47)}{\leq} \frac{1}{(1-\sqrt{\beta})^2}\sum_{j=2}^{\infty}\frac{1}{j^{\frac{3}{2}}}+\frac{1}{1-\sqrt{\beta}}\sum_{k=2}^{T}\frac{2}{\sqrt{k-1}}\overset{(47)}{\leq}\frac{2}{(1-\sqrt{\beta})^2}+\frac{2}{1-\sqrt{\beta}}\sum_{k=2}^{T}\frac{1}{\sqrt{k-1}}\\ &\leq \frac{2}{(1-\sqrt{\beta})^2}+\frac{2}{1-\sqrt{\beta}}\Big(1+\sum_{k=2}^{T}\frac{1}{\sqrt{k}}\Big)\leq\frac{2}{(1-\sqrt{\beta})^2}+\frac{2\big(1+2\sqrt{T}\big)}{1-\sqrt{\beta}},\end{aligned} \tag{48}$$

where the third inequality holds because $\sum_{k=2}^{T}\frac{1}{\sqrt{k-1}}=\sum_{k=1}^{T-1}\frac{1}{\sqrt{k}}\leq 1+\sum_{k=2}^{T}\frac{1}{\sqrt{k}}$, and the last inequality follows from the integral test inequality, that is,

$$\sum_{k=2}^{T}\frac{1}{\sqrt{k}}\leq\int_{1}^{T}\frac{1}{\sqrt{x}}\,dx=2\sqrt{x}\,\Big|_{1}^{T}=2\sqrt{T}-2\leq 2\sqrt{T}.$$

Then for V_t, we can get

$$\begin{aligned}V_t&=\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{(t-1)^2}=\sum_{k=2}^{T}\frac{k}{(k-1)^2}+\sum_{k=2}^{T}k\sum_{t=k+1}^{\infty}\frac{1}{(t-1)^2}\\ &\leq 2\sum_{k=2}^{T}\frac{1}{k-1}+\sum_{k=2}^{T}\frac{k}{k-1}\leq 2\big(1+\ln T\big)+2T,\end{aligned} \tag{49}$$

where the first inequality follows from $k\geq 2$, thus $k\leq 2(k-1)$ and

$$\sum_{t=k+1}^{\infty}\frac{1}{(t-1)^2}\leq\int_{k}^{\infty}\frac{1}{(x-1)^2}\,dx=-\frac{1}{x-1}\Big|_{k}^{\infty}\leq\frac{1}{k-1}.$$

The last inequality in (49) holds by $\sum_{k=2}^{T}\frac{k}{k-1}\leq 2\sum_{k=2}^{T}1=2(T-1)\leq 2T$ and

$$\sum_{k=2}^{T}\frac{1}{k-1}\leq 1+\sum_{k=2}^{T}\frac{1}{k}\leq 1+\int_{1}^{T}\frac{1}{x}\,dx=1+\ln\,x\,\Big|_{1}^{T}=1+\ln T.$$

Finally, substituting (48) and (49) in (45), we have

$$\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}}\overset{(45)}{\leq}\frac{1}{\sqrt{1-\beta}}\,U_t+\frac{1}{\sqrt{2(1-\beta)}}\,V_t$$

$$\leq \frac{1}{\sqrt{1-\beta}}\left(\frac{2}{(1-\sqrt{\beta})^2}+\frac{2\big(1+2\sqrt{T}\big)}{1-\sqrt{\beta}}\right)+\frac{2\big(1+\ln T\big)+2T}{\sqrt{2(1-\beta)}}$$
$$\leq \frac{1}{\sqrt{1-\beta}}\frac{4}{(1-\sqrt{\beta})^2}+\frac{1}{\sqrt{1-\beta}}\frac{4\sqrt{T}}{1-\sqrt{\beta}}+\frac{\sqrt{2}}{\sqrt{1-\beta}}+\frac{\sqrt{2}(\ln T+T)}{\sqrt{1-\beta}}.$$

This completes the proof. ■

Appendix B. Proofs for Section 3

B.1 Proof of Lemma 1

Proof First, by combining the L-smoothness of $f(\cdot;i)$ and the finite-sum structure of F in (1), we know that F is also L-smooth. Thus, the application of descent lemma (Bertsekas, 1999) gives

$$\begin{aligned}
F(\boldsymbol{x}_0^{t+1}) &\leq F(\boldsymbol{x}_0^t)+\langle\nabla F(\boldsymbol{x}_0^t),\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\rangle+\frac{L}{2}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2\\
&= F(\boldsymbol{x}_0^t)+\frac{n\alpha_t}{2}\left\|\nabla F(\boldsymbol{x}_0^t)-\frac{1}{n\alpha_t}\Big(\boldsymbol{x}_0^t-\boldsymbol{x}_0^{t+1}\Big)\right\|^2\\
&- \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2-\frac{1}{2n\alpha_t}\left\|\boldsymbol{x}_0^t-\boldsymbol{x}_0^{t+1}\right\|^2+\frac{L}{2}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2\\
&\overset{(22)}{=} F(\boldsymbol{x}_0^t)-\frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2\\
&- \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2+\frac{n\alpha_t}{2}\underbrace{\left\|\nabla F(\boldsymbol{x}_0^t)-\frac{1}{n}\sum_{j=0}^{n-1}\boldsymbol{m}_{j+1}^t\right\|^2}_{\spadesuit},
\end{aligned}\tag{50}$$

where the first equality holds because $-\langle\boldsymbol{a},\boldsymbol{b}\rangle=\frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|^2-\frac{1}{2}\|\boldsymbol{a}\|^2-\frac{1}{2}\|\boldsymbol{b}\|^2$ is valid with $\boldsymbol{a}=\sqrt{n\alpha_t}\,\nabla F(\boldsymbol{x}_0^t)$ and $\boldsymbol{b}=\frac{1}{\sqrt{n\alpha_t}}(\boldsymbol{x}_0^t-\boldsymbol{x}_0^{t+1})$, and the last equality follows from $\boldsymbol{x}_0^{t+1}=\tilde{\boldsymbol{x}}_t=\boldsymbol{x}_n^t$ and equation (22) with $r=0$ and $s=n$. From the iterate of $\boldsymbol{m}_{j+1}^t=\beta\boldsymbol{m}_0^t+(1-\beta)\boldsymbol{g}_j^t$, we estimate the last term in (50) as follows. For any $t\geq 2$, we have

$$\begin{aligned}
\spadesuit &= \left\|\nabla F(\boldsymbol{x}_0^t)-\frac{1}{n}\sum_{j=0}^{n-1}\boldsymbol{m}_{j+1}^t\right\|^2=\left\|\nabla F(\boldsymbol{x}_0^t)-\frac{1}{n}\sum_{j=0}^{n-1}\big(\beta\boldsymbol{m}_0^t+(1-\beta)\boldsymbol{g}_j^t\big)\right\|^2\\
&= \left\|\nabla F(\boldsymbol{x}_0^t)-\Big(\beta\boldsymbol{m}_0^t+\frac{1-\beta}{n}\sum_{j=0}^{n-1}\boldsymbol{g}_j^t\Big)\right\|^2\overset{(25)}{=}\left\|\nabla F(\boldsymbol{x}_0^t)-\Big(\frac{\beta}{n}\sum_{j=0}^{n-1}\boldsymbol{g}_j^{t-1}+\frac{1-\beta}{n}\sum_{j=0}^{n-1}\boldsymbol{g}_j^t\Big)\right\|\\
&\overset{(i)}{=}\left\|\frac{\beta}{n}\sum_{j=0}^{n-1}\Big(\nabla f(\boldsymbol{x}_0^t;\pi^{t-1}(j+1))-\boldsymbol{g}_j^{t-1}\Big)+\frac{1-\beta}{n}\sum_{j=0}^{n-1}\Big(\nabla f(\boldsymbol{x}_0^t;\pi^t(j+1))-\boldsymbol{g}_j^t\Big)\right\|^2\\
&\leq \beta\underbrace{\left\|\frac{1}{n}\sum_{j=0}^{n-1}\Big(\nabla f(\boldsymbol{x}_0^t;\pi^{t-1}(j+1))-\boldsymbol{g}_j^{t-1}\Big)\right\|^2}_{\mathrm{I}_t}+(1-\beta)\underbrace{\left\|\frac{1}{n}\sum_{j=0}^{n-1}\Big(\nabla f(\boldsymbol{x}_0^t;\pi^t(j+1))-\boldsymbol{g}_j^t\Big)\right\|^2}_{\mathrm{II}_t},
\end{aligned}\tag{51}$$

where (i) is due to the fact that π^{t-1} and π^t are the permutations of $[n]$, that is,

$$\nabla F(\boldsymbol{x}_0^t)=\beta\,\nabla F(\boldsymbol{x}_0^t)+(1-\beta)\nabla F(\boldsymbol{x}_0^t)$$

$$= \frac{\beta}{n}\sum_{j=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;\pi^{t-1}(j+1)) + \frac{(1-\beta)}{n}\sum_{j=0}^{n-1}\nabla f(\boldsymbol{x}_0^t;\pi^{t}(j+1)).$$

The last inequality in (51) is from $\|\beta\boldsymbol{a}+(1-\beta)\boldsymbol{b}\|^2 \le \beta\|\boldsymbol{a}\|^2+(1-\beta)\|\boldsymbol{b}\|^2$. Further, we estimate I_t and II_t in (51) as follows. Using the Jensen inequality of $\left\|\frac{1}{n}\sum_{i=0}^{n-1}\boldsymbol{a}_i\right\|^2 \le \frac{1}{n}\sum_{i=0}^{n-1}\|\boldsymbol{a}_i\|^2$, we have

$$\begin{aligned}\mathrm{I}_t &= \left\|\frac{1}{n}\sum_{j=0}^{n-1}\Big(\nabla f(\boldsymbol{x}_0^t;\pi^{t-1}(j+1)) - \boldsymbol{g}_j^{t-1}\Big)\right\|^2 \le \frac{1}{n}\sum_{j=0}^{n-1}\left\|\nabla f(\boldsymbol{x}_0^t;\pi^{t-1}(j+1)) - \boldsymbol{g}_j^{t-1}\right\|^2\\ &= \frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(\boldsymbol{x}_0^t;\pi^{t-1}(j+1)) - \nabla f(\boldsymbol{x}_j^{t-1};\pi^{t-1}(j+1))\|^2 \le \frac{L^2}{n}\sum_{j=0}^{n-1}\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_j^{t-1}\|^2, \qquad (52)\end{aligned}$$

where the last inequality follows from $\boldsymbol{x}_0^t = \tilde{\boldsymbol{x}}_{t-1} = \boldsymbol{x}_n^{t-1}$ and Assumption 1, that is, $\|\nabla f(\boldsymbol{x};i) - \nabla f(\boldsymbol{y};i)\| \le L\|\boldsymbol{x}-\boldsymbol{y}\|$. Then for II_t in (51), we have

$$\begin{aligned}\mathrm{II}_t &= \left\|\frac{1}{n}\sum_{j=0}^{n-1}\Big(\nabla f(\boldsymbol{x}_0^t;\pi^{t}(j+1)) - \boldsymbol{g}_j^{t}\Big)\right\|^2 \le \frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(\boldsymbol{x}_0^t;\pi^{t}(j+1)) - \boldsymbol{g}_j^{t}\|^2\\ &= \frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(\boldsymbol{x}_0^t;\pi^{t}(j+1)) - \nabla f(\boldsymbol{x}_j^{t};\pi^{t}(j+1))\|^2 \le \frac{L^2}{n}\sum_{j=0}^{n-1}\|\boldsymbol{x}_0^t-\boldsymbol{x}_j^t\|^2. \qquad (53)\end{aligned}$$

Upon substituting (52) and (53) in (51), we arrive at

$$\begin{aligned}\spadesuit &\overset{(51)}{\le} \beta\,\mathrm{I}_t + (1-\beta)\,\mathrm{II}_t \;\le\; \beta\left(\frac{L^2}{n}\sum_{j=0}^{n-1}\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_j^{t-1}\|^2\right) + (1-\beta)\left(\frac{L^2}{n}\sum_{j=0}^{n-1}\|\boldsymbol{x}_0^{t}-\boldsymbol{x}_j^{t}\|^2\right)\\ &\le \frac{\beta L^2}{n}\left(2n\beta^{t-2}D_1 + 4\sum_{j=2}^{t-1}\beta^{t-1-j}\Phi_j\right) + \frac{(1-\beta)L^2}{n}\left(n\beta^{t-1}D_1 + 2\sum_{j=2}^{t}\beta^{t-j}\Phi_j\right)\\ &\le 3\beta^{t-1}L^2D_1 + \frac{4L^2}{n}\sum_{j=2}^{t-1}\beta^{t-j}\Phi_j + \frac{2L^2}{n}\sum_{j=2}^{t}\beta^{t-j}\Phi_j \;\le\; 3\beta^{t-1}L^2D_1 + \frac{6L^2}{n}\sum_{j=2}^{t}\beta^{t-j}\Phi_j, \qquad (54)\end{aligned}$$

where $D_1 = \frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2+\sigma_2)$ is a finite value and Φ_j is defined in (4), the third inequality is from (20) in Lemma 5 and (33) in Lemma 6, and the last two inequalities holds by $1-\beta \le 1$ and $\Phi_t \ge 0$, respectively. Finally, we substitute (54) in (50) to get

$$\begin{aligned}&F(\boldsymbol{x}_0^{t+1}) \overset{(50)}{\le} F(\boldsymbol{x}_0^t) - \frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2 - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{n\alpha_t}{2}\spadesuit\\ &\le F(\boldsymbol{x}_0^t) - \frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2 - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{n\alpha_t}{2}\left(3\beta^{t-1}L^2D_1 + \frac{6L^2}{n}\sum_{j=2}^{t}\beta^{t-j}\Phi_j\right)\\ &= F(\boldsymbol{x}_0^t) - \frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2 - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{3nL^2D_1}{2}\beta^{t-1}\alpha_t + 3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j.\end{aligned}$$

Next, if π^t is uniformly sampled at random without replacement from $[n]$, then taking the total expectation on both sides of (50), we have

$$\mathbb{E}\big[F(\boldsymbol{x}_0^{t+1})\big] \le \mathbb{E}\big[F(\boldsymbol{x}_0^t)\big] - \frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\mathbb{E}\big[\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2\big] - \frac{n\alpha_t}{2}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + \frac{n\alpha_t}{2}\mathbb{E}[\spadesuit].$$

Similar to the proof of (54), we leverage (21) in Lemma 5 and (34) in Lemma 6 to get

$$\begin{aligned}\mathbb{E}[\spadesuit] &\leq \beta\bigg(\frac{L^2}{n}\sum_{j=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_n^{t-1}-\boldsymbol{x}_j^{t-1}\|^2\big]\bigg)+(1-\beta)\bigg(\frac{L^2}{n}\sum_{j=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_0^{t}-\boldsymbol{x}_j^{t}\|^2\big]\bigg)\\ &\leq \frac{\beta L^2}{n}\bigg(2n\beta^{t-2}D_2+4\sum_{j=2}^{t-1}\beta^{t-1-j}\Psi_j\bigg)+\frac{(1-\beta)L^2}{n}\bigg(n\beta^{t-1}D_2+2\sum_{j=2}^{t}\beta^{t-j}\Psi_j\bigg)\\ &\leq 3\beta^{t-1}L^2D_2+\frac{4L^2}{n}\sum_{j=2}^{t-1}\beta^{t-j}\Psi_j+\frac{2L^2}{n}\sum_{j=2}^{t}\beta^{t-j}\Psi_j\leq 3\beta^{t-1}L^2D_2+\frac{6L^2}{n}\sum_{j=2}^{t}\beta^{t-j}\Psi_j.\end{aligned}$$

Thus, it follows that

$$\begin{aligned}\mathbb{E}\big[F(\boldsymbol{x}_0^{t+1})\big] &\leq \mathbb{E}\big[F(\boldsymbol{x}_0^{t})\big]-\frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\mathbb{E}\big[\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^{t}\|^2\big]\\ &\quad-\frac{n\alpha_t}{2}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t})\|^2\big]+\frac{n\alpha_t}{2}\bigg(3\beta^{t-1}L^2D_2+\frac{6L^2}{n}\sum_{j=2}^{t}\beta^{t-j}\Psi_j\bigg)\\ &= \mathbb{E}\big[F(\boldsymbol{x}_0^{t})\big]-\frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\mathbb{E}\big[\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^{t}\|^2\big]\\ &\quad-\frac{n\alpha_t}{2}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t})\|^2\big]+\frac{3nL^2D_2}{2}\beta^{t-1}\alpha_t+3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j,\end{aligned}$$

This completes the proof. ■

B.2 Proof of Theorem 1

Proof We begin the proof with the conclusion (8) of Lemma 1, that is,

$$\begin{aligned}F(\boldsymbol{x}_0^{t+1}) &\leq F(\boldsymbol{x}_0^{t})-\frac{1}{2}\left(\frac{1}{n\alpha_t}-L\right)\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^{t}\|^2\\ &\quad-\frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^{t})\|^2+\frac{3nL^2D_1}{2}\beta^{t-1}\alpha_t+3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j\\ &\leq F(\boldsymbol{x}_0^{t})-\frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^{t})\|^2+\frac{3nL^2D_1}{2}\beta^{t-1}\alpha_t+3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j,\end{aligned}\tag{55}$$

where $D_1=\frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2+\sigma_2)$ is a finite value, Φ_j is defined in (4), and the last inequality holds by $\alpha_t\leq\frac{1}{\sqrt{2}nL}\leq\frac{1}{nL}$. Upon rearranging (55) and summing over t from 2 to T, we arrive at

$$\begin{aligned}\sum_{t=2}^{T}\frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^{t})\|^2 &\leq \sum_{t=2}^{T}\Big(F(\boldsymbol{x}_0^{t})-F(\boldsymbol{x}_0^{t+1})\Big)+\frac{3nL^2D_1}{2}\sum_{t=2}^{T}\beta^{t-1}\alpha_t+3L^2\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j\\ &= F(\boldsymbol{x}_0^{2})-F(\boldsymbol{x}_0^{T+1})+\frac{3nL^2D_1}{2}\sum_{t=2}^{T}\beta^{t-1}\alpha_t+3L^2\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j\end{aligned}$$

$$\leq F(\boldsymbol{x}_0^2) - f^* + \frac{3nL^2D_1}{2}\sum_{t=2}^{\infty}\beta^{t-1}\alpha_t + 3L^2\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j$$
$$\leq F(\boldsymbol{x}_0^2) - f^* + \frac{3\beta LD_1}{2(1-\beta)} + 3L^2\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j$$
$$\leq 2\big(F(\boldsymbol{x}_0^1) - f^*\big) + \frac{(1+2\beta)LD_1}{2(1-\beta)} + 3L^2\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j,$$

where the second inequality holds by Assumption 1, that is, $F(\boldsymbol{x}) \geq f^*$, the third inequality is due to $0 \leq \beta < 1$ and $\alpha_t \leq \frac{1}{\sqrt{2}nL} \leq \frac{1}{nL}$, and the last inequality follows from

$$\begin{aligned}
F(\boldsymbol{x}_0^2) &\overset{(50)}{\leq} F(\boldsymbol{x}_0^1) - \frac{1}{2}\left(\frac{1}{n\alpha_1} - L\right)\|\boldsymbol{x}_0^2 - \boldsymbol{x}_0^1\|^2 - \frac{n\alpha_1}{2}\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \frac{n\alpha_1}{2}\left\|\nabla F(\boldsymbol{x}_0^1) - \frac{1}{n}\sum_{j=0}^{n-1}\boldsymbol{m}_{j+1}^1\right\|^2 \\
&\leq F(\boldsymbol{x}_0^1) + \frac{1}{2L}\left\|\beta\nabla F(\boldsymbol{x}_0^1) + (1-\beta)\nabla F(\boldsymbol{x}_0^1) - \frac{1-\beta}{n}\sum_{j=0}^{n-1}\boldsymbol{g}_j^1\right\|^2 \\
&= F(\boldsymbol{x}_0^1) + \frac{1}{2L}\left\|\beta\nabla F(\boldsymbol{x}_0^1) + \frac{1-\beta}{n}\sum_{i=0}^{n-1}\big(\nabla f(\boldsymbol{x}_0^1;\pi^1(i+1)) - \nabla f(\boldsymbol{x}_i^1;\pi^1(i+1))\big)\right\|^2 \\
&\leq F(\boldsymbol{x}_0^1) + \frac{1}{2L}\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \frac{L}{2n}\sum_{i=0}^{n-1}\|\boldsymbol{x}_0^1 - \boldsymbol{x}_i^1\|^2 \\
&\leq 2F(\boldsymbol{x}_0^1) - f^* + \frac{L}{2n}\sum_{i=0}^{n-1}\|\boldsymbol{x}_0^1 - \boldsymbol{x}_i^1\|^2 \overset{(20)}{\leq} 2F(\boldsymbol{x}_0^1) - f^* + \frac{LD_1}{2},
\end{aligned} \tag{56}$$

where the second inequality holds because $\alpha_t \leq \frac{1}{\sqrt{2}nL} \leq \frac{1}{nL}$ and $\boldsymbol{m}_{j+1}^1 = (1-\beta)\boldsymbol{g}_j^1$, the third inequality follows from the Jensen inequality and $0 \leq \beta < 1$, the fourth inequality is due to the L-smoothness of F, that is, $\|\nabla F(\boldsymbol{x})\|^2 \leq 2L\big(F(\boldsymbol{x}) - f^*\big)$, and the last inequality holds by (20) in Lemma 5, that is, $\sum_{i=0}^{n-1}\|\boldsymbol{x}_0^1 - \boldsymbol{x}_i^1\|^2 \leq nD_1$. Further, the bound of $\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j$ established in Lemma 7 gives

$$\begin{aligned}
\frac{n}{2}\sum_{t=2}^{T}\alpha_t\|\nabla F(\boldsymbol{x}_0^t)\|^2 &\leq 2\big(F(\boldsymbol{x}_0^1) - f^*\big) + \frac{(1+2\beta)LD_1}{2(1-\beta)} + 3L^2\bigg(\frac{n^3\beta\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^1)\|^2 \\
&\quad + \frac{n^3(\sigma_1+\beta)}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^j)\|^2 + \frac{\sigma_2 n^3}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\bigg) \\
&\leq \underbrace{2\big(F(\boldsymbol{x}_0^1) - f^*\big) + \frac{(1+2\beta)LD_1}{2(1-\beta)} + \frac{3\beta}{2(1-\beta)L}\|\nabla F(\boldsymbol{x}_0^1)\|^2}_{C_1} \\
&\quad + \frac{3L^2n^3(\sigma_1+\beta)}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^j)\|^2 + \frac{3L^2\sigma_2 n^3}{1-\beta}\sum_{j=2}^{T}\alpha_j^3 \\
&\leq C_1 + \frac{n}{3}\sum_{t=2}^{T}\alpha_t\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{3L^2\sigma_2 n^3}{1-\beta}\sum_{t=2}^{T}\alpha_t^3,
\end{aligned} \tag{57}$$

where we introduce the constant $C_1 = 2\big(F(\boldsymbol{x}_0^1) - f^*\big) + \frac{3LD_1}{2(1-\beta)} + \frac{3\beta}{2(1-\beta)L}\|\nabla F(\boldsymbol{x}_0^1)\|^2$ to simplify the proof, the second inequality is from $\alpha_t \leq \frac{1}{\sqrt{2}nL}$, and the last inequality holds by $\alpha_t \leq \frac{\sqrt{1-\beta}}{3Ln\sqrt{\sigma_1+\beta}}$. From the stepsize condition of $\sum_{t=1}^{\infty}\alpha_t^3 < \infty$, we know that there exists a constant $C_2 > 0$ such that $\frac{\sigma_2 n^3}{1-\beta}\sum_{t=2}^{T}\alpha_t^3 \leq C_2$. Thus, it follows that

$$\frac{n}{6}\sum_{t=2}^{T}\alpha_t\|\nabla F(\boldsymbol{x}_0^t)\|^2 \overset{(57)}{\leq} C_1 + \frac{3L^2\sigma_2 n^3}{1-\beta}\sum_{t=2}^{T}\alpha_t^3 \leq C_1 + 3L^2C_2. \tag{58}$$

Upon recalling the equation (55) and using $\frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 \geq 0$, we have

$$F(\boldsymbol{x}_0^{t+1}) \leq F(\boldsymbol{x}_0^t) + \underbrace{\frac{3nL^2D_1}{2}\beta^{t-1}\alpha_t + 3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j}_{Z_t}. \tag{59}$$

Then from the conclusion of Lemma 7, we know that

$$\begin{aligned}
\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j &\leq \frac{n^3\beta\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \frac{n^3(\sigma_1+\beta)}{1-\beta}\sum_{j=2}^{T}\alpha_j^3\|\nabla F(\boldsymbol{x}_0^j)\|^2 + \frac{\sigma_2 n^3}{1-\beta}\sum_{j=2}^{T}\alpha_j^3 \\
&\leq \frac{n^3\beta\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \frac{n^3(\sigma_1+\beta)\alpha_1^2}{1-\beta}\sum_{j=2}^{T}\alpha_j\|\nabla F(\boldsymbol{x}_0^j)\|^2 + \frac{\sigma_2 n^3}{1-\beta}\sum_{j=2}^{T}\alpha_j^3 \\
&\overset{(58)}{\leq} \frac{n^3\beta\alpha_2^3}{1-\beta}\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \frac{n^3(\sigma_1+\beta)\alpha_1^2}{1-\beta}\,\frac{6}{n}\big(C_1 + 3L^2C_2\big) + C_2,
\end{aligned}$$

where the second inequality holds by the non-increasing of α_t, that is, $\alpha_t \leq \alpha_1$. This implies that the series $\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j$ is upper bounded. Thus, it follows that

$$\begin{aligned}
\sum_{t=2}^{\infty} Z_t &\overset{(59)}{=} \frac{3nL^2D_1}{2}\sum_{t=2}^{\infty}\beta^{t-1}\alpha_t + 3L^2\sum_{t=2}^{\infty}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j \\
&\leq \frac{3\beta LD_1}{2(1-\beta)} + 3L^2\sum_{t=2}^{\infty}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j < \infty,
\end{aligned} \tag{60}$$

where the first inequality follows from $\beta < 1$ and $\alpha_t \leq \frac{1}{\sqrt{2}nL} \leq \frac{1}{nL}$. Finally, by combining equations (60) and (59), we have

$$F(\boldsymbol{x}_0^{t+1}) \overset{(59)}{\leq} F(\boldsymbol{x}_0^t) + Z_t, \qquad \sum_{t=2}^{\infty} Z_t \overset{(60)}{<} \infty.$$

Since F is lower bounded, then Proposition 1 shows that there exists a finite point $\bar{F} \in \mathbb{R}$ such that

$$\lim_{T\to\infty} F(\tilde{\boldsymbol{x}}_T) = \lim_{T\to\infty} F(\boldsymbol{x}_0^{T+1}) = \bar{F}.$$

This completes the proof. ■

B.3 Proof of Corollary 1

Proof First, it follows from Theorem 1 that $F(\tilde{\boldsymbol{x}}_t)$ is convergent, then it is also bounded for all t. Since F is L-smooth, we use Lemma 4 of Li and Orabona (2019) to get $\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2 \leq 2L(F(\tilde{\boldsymbol{x}}_t)-f^*)$. Thus, $\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2$ is also bounded for all t, that is, there exists a constant $G \geq 0$ such that

$$\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2 \leq G.$$

Further, by using the conclusion (20) of Lemma 5, we have

$$\begin{aligned}
\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t - \boldsymbol{x}_0^t\|^2 &\leq n\beta^{t-1}D_1 + 2\sum_{j=2}^{t}\beta^{t-j}\Phi_j \\
&\overset{(4)}{\leq} nD_1 + 2\sum_{j=2}^{t}\beta^{t-j}\Big(n^3\alpha_j^2\Big(\beta\|\nabla F(\boldsymbol{x}_0^{j-1})\|^2 + \sigma_1(1-\beta)\|\nabla F(\boldsymbol{x}_0^j)\|^2 + \sigma_2(1-\beta)\Big)\Big) \\
&\leq nD_1 + 2n^3\Big(\beta G + \sigma_1(1-\beta)G + \sigma_2(1-\beta)\Big)\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2 \\
&\leq nD_1 + 2n^3\alpha_1^2\Big(\beta G + \sigma_1(1-\beta)G + \sigma_2(1-\beta)\Big)\sum_{j=2}^{t}\beta^{t-j} \\
&\leq nD_1 + \frac{2n^3\alpha_1^2}{1-\beta}\Big(\beta G + \sigma_1(1-\beta)G + \sigma_2(1-\beta)\Big) := C_3, \qquad (61)
\end{aligned}$$

where $D_1 = \frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$ is a finite value, the second inequality holds by $0 \leq \beta < 1$, and the fourth inequality holds by $\alpha_t \leq \alpha_1$. Upon applying the L-smoothness of $f(\cdot\,; i)$, we have

$$\begin{aligned}
\|\boldsymbol{g}_i^t\|^2 &= \|\nabla f(\boldsymbol{x}_i^t; \pi^t(i+1))\|^2 = \|\nabla f(\boldsymbol{x}_i^t; \pi^t(i+1)) - \nabla f(\boldsymbol{x}_0^t; \pi^t(i+1)) + \nabla f(\boldsymbol{x}_0^t; \pi^t(i+1))\|^2 \\
&\leq 2\|\nabla f(\boldsymbol{x}_i^t; \pi^t(i+1)) - \nabla f(\boldsymbol{x}_0^t; \pi^t(i+1))\|^2 + 2\|\nabla f(\boldsymbol{x}_0^t; \pi^t(i+1))\|^2 \\
&\leq 2L^2\|\boldsymbol{x}_i^t - \boldsymbol{x}_0^t\|^2 + 4L\big(f(\boldsymbol{x}_0^t; \pi^t(i+1)) - f^*\big), \qquad (62)
\end{aligned}$$

where the first inequality follows from $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$. Then since F has the finite-sum structure, we obtain that for any $j \in [n]$,

$$\begin{aligned}
F(\boldsymbol{x}) &\overset{(1)}{=} \frac{1}{n}\Big(f(\boldsymbol{x}; 1) + \cdots + f(\boldsymbol{x}; j-1) + f(\boldsymbol{x}; j) + f(\boldsymbol{x}; j+1) + \cdots + f(\boldsymbol{x}; n)\Big) \\
&\geq \frac{1}{n}\Big(f^* + \cdots + f^* + f(\boldsymbol{x}; j) + f^* + \cdots + f^*\Big) = \frac{n-1}{n}f^* + \frac{1}{n}f(\boldsymbol{x}; j),
\end{aligned}$$

where the last inequality holds by $f(\boldsymbol{x}; i) \geq f^*$ in Assumption 1. This indicates that $f(\boldsymbol{x}; j) - f^* \leq n\big(F(\boldsymbol{x}) - f^*\big)$. Recalling that $F(\boldsymbol{x}_0^t)$ is convergent in Theorem 1, then it is also bounded. Hence, $f(\boldsymbol{x}_0^t; j) - f^*$ is also bounded for all j, that is, there exists a constant $C_4 \geq 0$ such that

$$f(\boldsymbol{x}_0^t; j) - f^* \leq C_4, \quad \forall\, j \in [n]. \qquad (63)$$

By the fact that $\pi^{t-1}(i+1) \in [n]$, we substitute (63) in equation (62) to get

$$\sum_{i=0}^{n-1}\|\boldsymbol{g}_i^t\|^2 \overset{(62)}{\leq} 2L^2\sum_{i=0}^{n-1}\|\boldsymbol{x}_i^t - \boldsymbol{x}_0^t\|^2 + 4L\sum_{i=0}^{n-1}\big(f(\boldsymbol{x}_0^t; \pi^t(i+1)) - f^*\big)$$

$$\overset{(63)}{\leq} 2L^2 \sum_{i=0}^{n-1} \|\boldsymbol{x}_i^t - \boldsymbol{x}_0^t\|^2 + 4nLC_4 \overset{(61)}{\leq} 2L^2 C_3 + 4nLC_4 = 2L(LC_3 + 2nC_4). \tag{64}$$

From the iterate of $\boldsymbol{m}_{i+1}^t$ in Algorithm 1, we arrive at

$$\begin{aligned}
\sum_{i=0}^{n-1} \|\boldsymbol{m}_{i+1}^t\|^2 &= \sum_{i=0}^{n-1} \|\beta \boldsymbol{m}_0^t + (1-\beta)\boldsymbol{g}_i^t\|^2 \leq n\beta \|\boldsymbol{m}_0^t\|^2 + (1-\beta) \sum_{i=0}^{n-1} \|\boldsymbol{g}_i^t\|^2 \\
&\overset{(25)}{=} n\beta \left\| \frac{1}{n} \sum_{i=0}^{n-1} \boldsymbol{g}_i^{t-1} \right\|^2 + (1-\beta) \sum_{i=0}^{n-1} \|\boldsymbol{g}_i^t\|^2 \leq \beta \sum_{i=0}^{n-1} \|\boldsymbol{g}_i^{t-1}\|^2 + (1-\beta) \sum_{i=0}^{n-1} \|\boldsymbol{g}_i^t\|^2 \\
&\overset{(64)}{\leq} 2\beta L(LC_3 + 2nC_4) + 2(1-\beta)L(LC_3 + 2nC_4) = 2L(LC_3 + 2nC_4) := M,
\end{aligned}$$

where the first two inequality holds by the Jensen inequality. This completes the proof. ■

B.4 Proof of Theorem 2

Proof Since Assumptions 1, 2 and the stepsize condition of $\sum_{t=1}^{\infty} \alpha_t^3 < \infty$ are satisfied, then the equation (58) in the proof of Theorem 1 still holds. That is, $\frac{n}{6} \sum_{t=2}^{T} \alpha_t \|\nabla F(\boldsymbol{x}_0^t)\|^2 \leq C_1 + 3L^2 C_2$ is valid with two finite values C_1 and C_2, which indicates that $\sum_{t=1}^{\infty} \alpha_t \|\nabla F(\boldsymbol{x}_0^t)\|^2 < \infty$. By setting $a_t = \alpha_t$ and $b_t = \|\nabla F(\boldsymbol{x}_0^t)\|$, then it follows that $\sum_{t=1}^{\infty} a_t b_t^2 < \infty$. Thus, combining the condition of $\sum_{t=1}^{\infty} \alpha_t = \infty$ and Proposition 2, we estimate $|b_{t+1} - b_t|$ as follows.

$$|b_{t+1} - b_t| = |\|\nabla F(\boldsymbol{x}_0^{t+1})\| - \|\nabla F(\boldsymbol{x}_0^t)\|| \leq \|\nabla F(\boldsymbol{x}_0^{t+1}) - \nabla F(\boldsymbol{x}_0^t)\| \leq L\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|, \tag{65}$$

where the last inequality holds by the L-smoothness of F. Then by the iterate of $\boldsymbol{x}_0^t$, we have

$$\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 = \|\boldsymbol{x}_n^t - \boldsymbol{x}_0^t\|^2 \overset{(22)}{=} \alpha_t^2 \left\| \sum_{i=0}^{n-1} \boldsymbol{m}_{i+1}^t \right\|^2 \leq n\alpha_t^2 \sum_{i=0}^{n-1} \|\boldsymbol{m}_{i+1}^t\|^2 \leq nM\alpha_t^2,$$

where the first equality is due to $\boldsymbol{x}_0^{t+1} = \tilde{\boldsymbol{x}}_t = \boldsymbol{x}_n^t$, the first inequality holds by the Jensen inequality, and the last inequality follows from Corollary 1, that is, $\sum_{i=0}^{n-1} \|\boldsymbol{m}_{i+1}^t\|^2 \leq M$. Thus, it implies that $\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| \leq \sqrt{nM}\alpha_t$. Finally, recalling (65) and using the definition of $a_t = \alpha_t$, we have

$$|b_{t+1} - b_t| \overset{(65)}{\leq} L\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| \leq L\sqrt{nM}\alpha_t = L\sqrt{nM}a_t,$$

which implies that the condition of $|b_{t+1} - b_t| \leq \mu a_t$ is satisfied with $\mu = L\sqrt{nM}$. Therefore, we can apply Proposition 2 to obtain

$$\lim_{T\to\infty} \|\nabla F(\tilde{\boldsymbol{x}}_T)\| = \lim_{T\to\infty} \|\nabla F(\boldsymbol{x}_0^{T+1})\| = \lim_{T\to\infty} b_{T+1} = 0.$$

This completes the proof. ■

B.5 Proof of Theorem 3

Proof Given that Assumptions 1, 2 and the stepsize condition of $\sum_{t=1}^{\infty}\alpha_t^3<\infty$ are satisfied, the equation (58) in the proof of Theorem 1 still holds, that is, $\frac{n}{6}\sum_{t=2}^{T}\alpha_t\|\nabla F(\boldsymbol{x}_0^t)\|^2\leq C_1+3L^2C_2$, where C_1 and C_2 are two finite values. Upon setting $a_t=\alpha_{t+1}$ and $X_t=\|\nabla F(\boldsymbol{x}_0^{t+1})\|^2$, we combine (58) and the stepsize condition of $\sum_{t=2}^{\infty}\frac{\alpha_{t+1}}{\sum_{i=2}^{t}\alpha_i}=\infty$ to obtain

$$\sum_{t=1}^{\infty}a_tX_t=\sum_{t=1}^{\infty}\alpha_{t+1}\|\nabla F(\boldsymbol{x}_0^{t+1})\|^2\overset{(58)}{<}\infty,\qquad\sum_{t=2}^{\infty}\frac{a_t}{\sum_{i=1}^{t-1}a_i}=\sum_{t=2}^{\infty}\frac{\alpha_{t+1}}{\sum_{i=2}^{t}\alpha_i}=\infty.$$

Since α_t is non-increasing, then the application of Proposition 3 gives

$$\min_{1\leq t\leq T}\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2=\min_{1\leq t\leq T}\|\nabla F(\boldsymbol{x}_0^{t+1})\|^2=o\left(\frac{1}{\sum_{t=1}^{T}\alpha_{t+1}}\right).$$

This completes the proof. ■

B.6 Proof of Corollary 2

Proof By the conditions of $\alpha_t=\frac{\gamma}{t^{1/3+\epsilon}}$ and $\gamma\leq\frac{1}{nL\sqrt{K}}$, we know that $\alpha_t\leq\frac{1}{nL\sqrt{K}}$ and

$$\sum_{i=2}^{t}\alpha_i=\sum_{i=2}^{t}\frac{\gamma}{i^{1/3+\epsilon}}\leq\int_1^t\frac{\gamma}{x^{1/3+\epsilon}}\,dx=\frac{\gamma x^{2/3-\epsilon}}{\frac{2}{3}-\epsilon}\bigg|_1^t=\frac{\gamma(t^{2/3-\epsilon}-1)}{\frac{2}{3}-\epsilon}\leq\frac{\gamma t^{2/3-\epsilon}}{\frac{2}{3}-\epsilon}.$$

Thus, it follows that

$$\sum_{t=2}^{\infty}\frac{\alpha_{t+1}}{\sum_{i=2}^{t}\alpha_i}\geq\sum_{t=2}^{\infty}\frac{\gamma}{(t+1)^{1/3+\epsilon}}\frac{\frac{2}{3}-\epsilon}{\gamma t^{2/3-\epsilon}}=\sum_{t=2}^{\infty}\frac{1}{(t+1)^{1/3+\epsilon}}\frac{\frac{2}{3}-\epsilon}{t^{2/3-\epsilon}}\geq\sum_{t=2}^{\infty}\frac{\frac{2}{3}-\epsilon}{t+1}=\infty,$$

where the last inequality holds because $\epsilon<\frac{2}{3}$. By the condition that $\epsilon>0$, we can get

$$\sum_{t=1}^{\infty}\alpha_t^3=\sum_{t=1}^{\infty}\frac{\gamma^3}{t^{1+3\epsilon}}<\infty.$$

Further, we know that the stepsize conditions in Theorem 3 are satisfied. Thus, it follows from the conclusion of Theorem 3 that

$$\min_{1\leq t\leq T}\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2=o\left(\frac{1}{\sum_{t=1}^{T}\alpha_{t+1}}\right)=o\left(\frac{1}{T^{\frac{2}{3}-\epsilon}}\right),$$

where the last equality follows from the definition of little-o and

$$\sum_{t=1}^{T}\alpha_{t+1}=\sum_{t=1}^{T}\frac{\gamma}{(t+1)^{1/3+\epsilon}}\geq\int_1^{T+1}\frac{\gamma}{(x+1)^{1/3+\epsilon}}\,dx=\frac{\gamma\big((T+2)^{\frac{2}{3}-\epsilon}-2^{\frac{2}{3}-\epsilon}\big)}{\frac{2}{3}-\epsilon}\geq\frac{\gamma(T+2)^{\frac{2}{3}-\epsilon}}{3},$$

where the last inequality holds by the concavity of $x^{2/3-\epsilon}$ and $T\geq 1$. This completes the proof. ■

B.7 Proof of Theorem 4

Proof It follows from $T \geq 2\sqrt{2}n\gamma^3L^3$ that $\alpha_t = \frac{\gamma}{n^{2/3}T^{1/3}} \leq \frac{1}{\sqrt{2}nL}$. Then by the condition that π^t is sampled uniformly without replacement from $[n]$ and the stepsize $\alpha_t \leq \frac{1}{\sqrt{2}nL}$, we begin the proof with (9) in Lemma 1, that is,

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}_0^{t+1})] &\overset{(9)}{\leq} \mathbb{E}[F(\boldsymbol{x}_0^t)] - \frac{1}{2}\Big(\frac{1}{n\alpha_t} - L\Big)\mathbb{E}\big[\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2\big] \\ &\quad - \frac{n\alpha_t}{2}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + \frac{3nL^2D_2}{2}\beta^{t-1}\alpha_t + 3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j \\ &\leq \mathbb{E}[F(\boldsymbol{x}_0^t)] - \frac{n\alpha_t}{2}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] + \frac{3LD_2}{2}\beta^{t-1} + 3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j,\end{aligned}$$

where the second inequality holds by $T \geq 2\sqrt{2}n\gamma^3L^3 \geq n\gamma^3L^3$, that is, $\alpha_t = \frac{1}{n^{2/3}T^{1/3}} \leq \frac{1}{nL}$. Upon rearranging the above equation and summing over t from 2 to $T+1$, we have

$$\begin{aligned}\frac{n}{2}\sum_{t=2}^{T+1}\alpha_t\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] &\leq \sum_{t=2}^{T+1}\Big(\mathbb{E}[F(\boldsymbol{x}_0^t)] - \mathbb{E}[F(\boldsymbol{x}_0^{t+1})]\Big) + \frac{3LD_2}{2}\sum_{t=2}^{T+1}\beta^{t-1} + 3L^2\sum_{t=2}^{T+1}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j \\ &\leq \mathbb{E}[F(\boldsymbol{x}_0^2)] - f^* + \frac{3LD_2}{2}\sum_{t=2}^{\infty}\beta^{t-1} + 3L^2\sum_{t=2}^{T+1}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j \\ &\leq \mathbb{E}[F(\boldsymbol{x}_0^2)] - f^* + \frac{3\beta LD_2}{2(1-\beta)} + 3L^2\sum_{t=2}^{T+1}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j \\ &\leq 2\big(\mathbb{E}\big[F(\boldsymbol{x}_0^1)\big] - f^*\big) + \frac{(1+2\beta)LD_2}{2(1-\beta)} + 3L^2\sum_{t=2}^{T+1}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j,\end{aligned}$$

where the second inequality holds by the fact that $F(\boldsymbol{x}) \geq f^*$ in Assumption 1, the third inequality is due to $0 \leq \beta < 1$, and the last inequality follows from (56) and (21) in Lemma 5, that is,

$$\mathbb{E}\big[F(\boldsymbol{x}_0^2)\big] \overset{(56)}{\leq} 2\mathbb{E}\big[F(\boldsymbol{x}_0^1)\big] - f^* + \frac{L}{2n}\sum_{i=0}^{n-1}\mathbb{E}\big[\|\boldsymbol{x}_0^1 - \boldsymbol{x}_i^1\|^2\big] \overset{(21)}{\leq} 2\mathbb{E}\big[F(\boldsymbol{x}_0^1)\big] - f^* + \frac{LD_2}{2},$$

where $D_2 = \frac{1}{L^2}(\sigma_1\mathbb{E}[\|\nabla F(\boldsymbol{x}_0^1)\|^2] + \sigma_2)$ is a finite value (independent of n). Upon leveraging the bound of $\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Psi_j$ in (40) of Lemma 7, we have

$$\begin{aligned}\frac{n}{2}\sum_{t=2}^{T+1}\alpha_t\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big] &\leq 2\big(\mathbb{E}\big[F(\boldsymbol{x}_0^1)\big] - f^*\big) + \frac{(1+2\beta)LD_2}{2(1-\beta)} + 3L^2\bigg(\frac{n^3\beta\alpha_2^3}{1-\beta}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^1)\|^2\big] \\ &\quad + \frac{2n^3(\sigma_1+2)}{1-\beta}\sum_{j=2}^{T+1}\alpha_j^3\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^j)\|^2\big] + \frac{2\sigma_2n^2}{1-\beta}\sum_{j=2}^{T+1}\alpha_j^3\bigg) \\ &\leq \underbrace{2\big(\mathbb{E}\big[F(\boldsymbol{x}_0^1)\big] - f^*\big) + \frac{(1+2\beta)LD_2}{2(1-\beta)} + \frac{3\beta}{2(1-\beta)L}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^1)\|^2\big]}_{C_9}\end{aligned}$$

$$+\frac{6L^2n^3(\sigma_1+2)}{1-\beta}\sum_{t=2}^{T+1}\alpha_t^3\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big]+\frac{6L^2\sigma_2n^2}{1-\beta}\sum_{t=2}^{T+1}\alpha_t^3$$
$$\leq C_9+\frac{n}{3}\sum_{t=2}^{T+1}\alpha_t\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^t)\|^2\big]+\frac{6L^2\sigma_2n^2}{1-\beta}\sum_{t=2}^{T+1}\alpha_t^3,$$

where the second inequality follows from $T\geq 2\sqrt{2}n\gamma^3L^3$, that is, $\alpha_t\leq\frac{1}{\sqrt{2}nL}$, and the last inequality holds because $T\geq\frac{54\sqrt{2}n\gamma^3L^3(\sigma_1+2)^{3/2}}{(1-\beta)^{3/2}}$, that is, $\alpha_t\leq\frac{\sqrt{1-\beta}}{3\sqrt{2}nL\sqrt{\sigma_1+2}}$. Upon rearranging the above equation, we arrive at

$$\sum_{t=1}^{T}\alpha_{t+1}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t+1})\|^2\big]=\sum_{t=2}^{T+1}\alpha_t\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t})\|^2\big]\leq\frac{6C_9}{n}+\frac{36L^2\sigma_2n}{1-\beta}\sum_{t=2}^{T+1}\alpha_t^3. \tag{66}$$

Thus, by the fact that $\boldsymbol{x}_0^t=\tilde{\boldsymbol{x}}_{t-1}$ and using $\alpha_t=\frac{\gamma}{n^{2/3}T^{1/3}}$, we finally obtain

$$\begin{aligned}\min_{1\leq t\leq T}\mathbb{E}\big[\|\nabla F(\tilde{\boldsymbol{x}}_t)\|^2\big] &= \min_{1\leq t\leq T}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t+1})\|^2\big]\leq\frac{\sum_{t=1}^{T}\alpha_{t+1}\mathbb{E}\big[\|\nabla F(\boldsymbol{x}_0^{t+1})\|^2\big]}{\sum_{t=1}^{T}\alpha_{t+1}}\\ &\overset{(66)}{\leq}\frac{6C_9}{\gamma}\frac{1}{n^{1/3}T^{2/3}}+\frac{36L^2\gamma^2\sigma_2}{1-\beta}\frac{1}{n^{1/3}T^{2/3}}=\mathcal{O}\left(\frac{1}{n^{1/3}T^{2/3}}\right).\end{aligned}$$

Notably, since D_2 is a constant independent of n, then C_9 is also a finite value that is independent of n, thus the last equality in the above equation holds. This completes the proof. ■

Appendix C. Proofs for Section 4

C.1 Proof of Lemma 3

Proof (a) Since the conditions $\sum_{t=1}^{\infty}\alpha_t^3<\infty$ and $\alpha_t\leq\frac{1}{nL\sqrt{K}}$ are satisfied, then the conclusion of Theorem 1 still holds, that is, $F(\tilde{\boldsymbol{x}}_t)$ is convergent. Then the sequence $\{F(\tilde{\boldsymbol{x}}_t)\}$ is also bounded, that is, there exists $C>0$ such that $F(\tilde{\boldsymbol{x}}_t)\leq C$ for all t. By the condition that F is coercive, we know that all the level sets of F are bounded. Therefore, the iterative sequence $\{\tilde{\boldsymbol{x}}_t\}\subseteq\{\boldsymbol{x}:F(\boldsymbol{x})\leq C\}$ is also bounded. This indicates that $\{\tilde{\boldsymbol{x}}_t\}$ has a convergent sub-sequence, and the convergent sub-sequence is also bounded, thus the accumulation points set $\mathcal{C}$ is non-empty and compact.

(b) Assume that there exists a sub-sequence $\{\tilde{\boldsymbol{x}}_{t_k}\}\subseteq\{\tilde{\boldsymbol{x}}_t\}$ such that $\lim_{k\to\infty}\tilde{\boldsymbol{x}}_{t_k}=\bar{\boldsymbol{x}}\in\mathcal{C}$, but $\nabla F(\bar{\boldsymbol{x}})\neq\mathbf{0}$. Then by the continuity of ∇F and $\|\cdot\|$, and $\lim_{t\to\infty}\|\nabla F(\tilde{\boldsymbol{x}}_t)\|=0$ obtained in Theorem 2, we have $\lim_{k\to\infty}\|\nabla F(\tilde{\boldsymbol{x}}_{t_k})\|=\|\nabla F(\bar{\boldsymbol{x}})\|=0$, which is in contradiction to $\nabla F(\bar{\boldsymbol{x}})\neq\mathbf{0}$. Thus, if $\bar{\boldsymbol{x}}\in\mathcal{C}$, then we can get that $\nabla F(\bar{\boldsymbol{x}})=\mathbf{0}$, that is, $\mathcal{C}\subseteq\{\boldsymbol{x}\in\mathbb{R}^d:\nabla F(\boldsymbol{x})=\mathbf{0}\}$.

(c) For any $\bar{\boldsymbol{x}}\in\mathcal{C}$, there exists a sub-sequence $\{\tilde{\boldsymbol{x}}_{t_k}\}\subseteq\{\tilde{\boldsymbol{x}}_t\}$ such that $\lim_{k\to\infty}\tilde{\boldsymbol{x}}_{t_k}=\bar{\boldsymbol{x}}$. By the continuity of F, we have $\lim_{k\to\infty}F(\tilde{\boldsymbol{x}}_{t_k})=F(\bar{\boldsymbol{x}})$. From Theorem 1, we can obtain that $\lim_{t\to\infty}F(\tilde{\boldsymbol{x}}_t)=\bar{F}$. Further, by the uniqueness of the limitation, we obtain that $F(\bar{\boldsymbol{x}})=\bar{F}\in\mathbb{R}$. Thus, for any $\bar{\boldsymbol{x}}\in\mathcal{C}$, we can derive that $F(\bar{\boldsymbol{x}})\in\mathbb{R}$ is finite and takes the value as a constant of $\bar{F}$. This completes the proof. ■

C.2 Proof of Theorem 5

Proof Since Assumptions 1 and 2 are satisfied, then we begin with (8) in Lemma 1, that is,

$$\begin{aligned}
F(\boldsymbol{x}_0^{t+1}) &\leq F(\boldsymbol{x}_0^t) - \frac{1}{2}\left(\frac{1}{n\alpha_t} - L\right)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 \\
&\quad - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{3nL^2D_1}{2}\beta^{t-1}\alpha_t + 3L^2\alpha_t\sum_{j=2}^{t}\beta^{t-j}\Phi_j \\
&\leq F(\boldsymbol{x}_0^t) - \frac{1}{2}\left(\frac{1}{n\alpha_t} - L\right)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 \\
&\quad - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{3nL^2D_1}{2}\beta^{t-1}\alpha_t + 3L^2R\alpha_t\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2, \qquad (67)
\end{aligned}$$

where $D_1 = \frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$ and $R = n^3(\beta G + \sigma_1(1-\beta)G + \sigma_2(1-\beta))$ are finite values, and the last inequality holds by the definition of Φ_t and $\|\nabla F(\boldsymbol{x}_0^t)\|^2 \leq G$ in Corollary 1, that is,

$$\begin{aligned}
\Phi_t &\overset{(4)}{=} n^3\alpha_t^2\big(\beta\|\nabla F(\boldsymbol{x}_0^{t-1})\|^2 + \sigma_1(1-\beta)\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \sigma_2(1-\beta)\big) \\
&\leq n^3\big(\beta G + \sigma_1(1-\beta)G + \sigma_2(1-\beta)\big)\alpha_t^2 := R\alpha_t^2.
\end{aligned}$$

Upon let's set

$$u_t = \frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i + 3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2. \qquad (68)$$

By the condition that α_t is non-increasing, we have $\sum_{t=1}^{T}\beta^{t-1}\alpha_t \leq \alpha_1\sum_{t=1}^{\infty}\beta^{t-1} \leq \frac{\alpha_1}{1-\beta}$ and

$$\sum_{t=2}^{T}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2 \leq \sum_{t=2}^{T}\sum_{j=2}^{t}\beta^{t-j}\alpha_j^3 = \sum_{j=2}^{T}\sum_{t=j}^{T}\beta^{t-j}\alpha_j^3 \leq \sum_{j=2}^{T}\left(\sum_{t=j}^{\infty}\beta^t\right)\beta^{-j}\alpha_j^3 \leq \frac{1}{1-\beta}\sum_{j=2}^{T}\alpha_j^3. \qquad (69)$$

Thus, it follows that $\sum_{t=1}^{\infty}\beta^{t-1}\alpha_t < \infty$ and $\lim_{t\to\infty}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i = 0$. Moreover, since $\sum_{t=1}^{\infty}\alpha_t^3 < \infty$, then $\sum_{t=2}^{\infty}\alpha_t\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2 < \infty$ and $\lim_{t\to\infty}\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2 = 0$. Hence, it follows from the definition of u_t in (68) that $\lim_{t\to\infty}u_t = 0$. Now, adding u_{t+1} to both sides of (67), we have

$$\begin{aligned}
F(\boldsymbol{x}_0^{t+1}) + u_{t+1} &\leq F(\boldsymbol{x}_0^t) - \frac{1}{2}\Big(\frac{1}{n\alpha_t} - L\Big)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 + \frac{3nL^2D_1}{2}\beta^{t-1}\alpha_t \\
&\quad + 3L^2R\alpha_t\sum_{j=2}^{t}\beta^{t-j}\alpha_j^2 + \left(\frac{3nL^2D_1}{2}\sum_{i=t+1}^{\infty}\beta^{i-1}\alpha_i + 3L^2R\sum_{i=t+1}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2\right) \\
&= F(\boldsymbol{x}_0^t) - \frac{1}{2}\Big(\frac{1}{n\alpha_t} - L\Big)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 \\
&\quad + \frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i + 3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2 \\
&= F(\boldsymbol{x}_0^t) + u_t - \frac{1}{2}\Big(\frac{1}{n\alpha_t} - L\Big)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2. \qquad (70)
\end{aligned}$$

From the stepsize condition of $\alpha_t \leq \frac{1}{\sqrt{2}nL} < \frac{1}{nL}$, we can obtain that $\frac{1}{n\alpha_t} - L > 0$. Then equation (70) indicates that $F(\boldsymbol{x}_0^{t+1}) + u_{t+1} < F(\boldsymbol{x}_0^t) + u_t$, thus the sequence $\{F(\boldsymbol{x}_0^t) + u_t\}$ is decreasing.

Further, since Assumptions 1, 2 and 3 are satisfied, then the conclusions of Lemma 3 still hold, that is, $\mathcal{C}$ is compact and F is constant on $\mathcal{C}$. Upon combining the KL inequality at each point of $\mathcal{C}$ in Assumption 3, we can get that the conditions of Lemma 2 are satisfied. Thus, there exist ϵ, $\eta > 0$ and $\rho \in \mathcal{Q}_\eta$ such that for all $\bar{\boldsymbol{x}} \in \Omega$ and $\boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{R}^d : \operatorname{dist}(\boldsymbol{x}, \mathcal{C}) < \epsilon\} \cap \{\boldsymbol{x} \in \mathbb{R}^d : 0 < |F(\boldsymbol{x}) - \bar{F}| < \eta\}$, we have $\rho'\big(|F(\boldsymbol{x}) - \bar{F}|\big) \cdot \|\nabla F(\boldsymbol{x})\| \geq 1$, where we use the fact that $F(\bar{\boldsymbol{x}}) = \bar{F}$. From the definition of $\mathcal{C}$ in (14), we have that $\lim_{t\to\infty} \operatorname{dist}(\boldsymbol{x}_0^t, \mathcal{C}) = \lim_{t\to\infty} \operatorname{dist}(\tilde{\boldsymbol{x}}_{t-1}, \mathcal{C}) = 0$. Thus, there exists $t_1 > 0$, such that $\operatorname{dist}(\boldsymbol{x}_0^t, \mathcal{C}) < \epsilon$ for any $t \geq t_1$. In addition, from the conclusion of Theorem 1, we know that $\lim_{t\to\infty} F(\boldsymbol{x}_0^t) = \bar{F}$, then there exists $t_2 > 0$ such that $0 < |F(\boldsymbol{x}_0^t) - \bar{F}| < \eta$ for any $t \geq t_2$. Upon let's take $t_3 = \max\{t_1, t_2\}$, then for any $t \geq t_3$, we have $\operatorname{dist}(\boldsymbol{x}_0^t, \mathcal{C}) < \epsilon$ and $0 < |F(\boldsymbol{x}_0^t) - \bar{F}| < \eta$, which means that $\boldsymbol{x}_0^t \in \{\boldsymbol{x} \in \mathbb{R}^d : \operatorname{dist}(\boldsymbol{x}, \mathcal{C}) < \epsilon\} \cap \{\boldsymbol{x} \in \mathbb{R}^d : 0 < |F(\boldsymbol{x}) - \bar{F}| < \eta\}$. Thus, it follows that

$$\rho'\big(|F(\boldsymbol{x}_0^t) - \bar{F}|\big) \cdot \|\nabla F(\boldsymbol{x}_0^t)\| \geq 1. \tag{71}$$

By combining $\lim_{t\to\infty} F(\boldsymbol{x}_0^t) = \bar{F}$ obtained in Theorem 1 and $\lim_{t\to\infty} u_t = 0$ achieved in (69), we know that $\lim_{t\to\infty} \big(F(\boldsymbol{x}_0^t) + u_t\big) = \lim_{t\to\infty} F(\boldsymbol{x}_0^t) + \lim_{t\to\infty} u_t = \bar{F}$. Upon since $\{F(\boldsymbol{x}_0^t) + u_t\}$ is decreasing in (70), then $F(\boldsymbol{x}_0^t) + u_t - \bar{F} > 0$, thus $\rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big)$ is well defined. So, applying the concavity of ρ, we have

$$\begin{aligned}
&\rho\big(F(\boldsymbol{x}_0^{t+1}) - \bar{F} + u_{t+1}\big) \\
&\leq \rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) + \rho'\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big)\big(F(\boldsymbol{x}_0^{t+1}) - \bar{F} + u_{t+1} - \big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big)\big) \\
&= \rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) + \rho'\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big)\big(F(\boldsymbol{x}_0^{t+1}) + u_{t+1} - \big(F(\boldsymbol{x}_0^t) + u_t\big)\big) \\
&\overset{(70)}{\leq} \rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) + \rho'\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big)\left(-\frac{1}{2}\left(\frac{1}{n\alpha_t} - L\right)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 - \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2\right) \\
&\overset{(73)}{\leq} \rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) - \rho'\big(|F(\boldsymbol{x}_0^t) - \bar{F}| + u_t\big)\left(\frac{1}{2}\left(\frac{1}{n\alpha_t} - L\right)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 + \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2\right) \\
&\overset{(74)}{\leq} \rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) - \frac{1}{C_\rho}\frac{\frac{1}{2}\big(\frac{1}{n\alpha_t} - L\big)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 + \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2}{\big[\rho'(|F(\boldsymbol{x}_0^t) - \bar{F}|)\big]^{-1} + \big[\rho'(u_t)\big]^{-1}} \\
&\overset{(71)}{\leq} \rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) - \frac{1}{C_\rho}\frac{\frac{1}{2}\big(\frac{1}{n\alpha_t} - L\big)\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 + \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2}{\|\nabla F(\boldsymbol{x}_0^t)\| + \big[\rho'(u_t)\big]^{-1}} \\
&\leq \rho\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) - \frac{1}{C_\rho}\frac{\frac{1}{10n\alpha_t}\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 + \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2}{\|\nabla F(\boldsymbol{x}_0^t)\| + \big[\rho'(u_t)\big]^{-1}},
\end{aligned} \tag{72}$$

where the third inequality holds by $\alpha_t \leq \frac{1}{\sqrt{2}nL}$ and $\rho'(x)$ is non-increasing, that is, $\frac{1}{n\alpha_t} - L \geq 0$ and

$$F(\boldsymbol{x}_0^t) - \bar{F} + u_t \leq |F(\boldsymbol{x}_0^t) - \bar{F}| + u_t \implies \rho'\big(F(\boldsymbol{x}_0^t) - \bar{F} + u_t\big) \geq \rho'\big(|F(\boldsymbol{x}_0^t) - \bar{F}| + u_t\big). \tag{73}$$

The fourth inequality in (72) is due to the condition that $\rho \in \mathcal{Q}_\eta$, that is,

$$\begin{aligned}
&\frac{1}{\rho'\big(|F(\boldsymbol{x}_0^t) - \bar{F}| + u_t\big)} \leq C_\rho\left[\frac{1}{\rho'(|F(\boldsymbol{x}_0^t) - \bar{F}|)} + \frac{1}{\rho'(u_t)}\right] \\
&\qquad\qquad \implies \rho'\big(|F(\boldsymbol{x}_0^t) - \bar{F}| + u_t\big) \geq \frac{1}{C_\rho}\frac{1}{\big[\rho'(|F(\boldsymbol{x}_0^t) - \bar{F}|)\big]^{-1} + \big[\rho'(u_t)\big]^{-1}}.
\end{aligned} \tag{74}$$

The last inequality in (72) follows from $\rho'(x) > 0$ and the stepsize condition of $\alpha_t \leq \frac{1}{\sqrt{2}nL}$, that is, $1 - n\alpha_t L \geq \frac{\sqrt{2}-1}{\sqrt{2}} \geq \frac{1}{5}$. Upon setting $\delta_t = \rho(F(\boldsymbol{x}_0^t) - \bar{F} + u_t)$ and rearranging (72), we arrive at

$$\frac{1}{10n\alpha_t}\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\|^2 + \frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2 \leq C_\rho\big(\delta_t - \delta_{t+1}\big)\Big(\|\nabla F(\boldsymbol{x}_0^t)\| + \big[\rho'(u_t)\big]^{-1}\Big)$$

$$\leq \frac{2C_\rho^2\big(\delta_t-\delta_{t+1}\big)^2}{n\alpha_t}+\frac{n\alpha_t}{8}\Big(\|\nabla F(\boldsymbol{x}_0^t)\|+\big[\rho'(u_t)\big]^{-1}\Big)^2, \quad (75)$$

where the last inequality follows from the fact that $ab\leq\frac{1}{2}(a^2+b^2)$ is valid with $a=\frac{2C_\rho}{\sqrt{n\alpha_t}}(\delta_t-\delta_{t+1})$ and $b=\frac{\sqrt{n\alpha_t}}{2}(\|\nabla F(\boldsymbol{x}_0^t)\|+[\rho'(u_t)]^{-1})$.

Next, the application of $a+b\leq\sqrt{2(a^2+b^2)}$ gives

$$\begin{aligned}\frac{1}{\sqrt{10n\alpha_t}}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|+\frac{\sqrt{n\alpha_t}}{\sqrt{2}}\|\nabla F(\boldsymbol{x}_0^t)\| &\leq \sqrt{2\left(\frac{1}{10n\alpha_t}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|^2+\frac{n\alpha_t}{2}\|\nabla F(\boldsymbol{x}_0^t)\|^2\right)}\\ &\overset{(75)}{\leq} \sqrt{\frac{4C_\rho^2\big(\delta_t-\delta_{t+1}\big)^2}{n\alpha_t}+\frac{n\alpha_t}{4}\Big(\|\nabla F(\boldsymbol{x}_0^t)\|+\big[\rho'(u_t)\big]^{-1}\Big)^2}\\ &\leq \frac{2C_\rho\big(\delta_t-\delta_{t+1}\big)}{\sqrt{n\alpha_t}}+\frac{\sqrt{n\alpha_t}}{2}\|\nabla F(\boldsymbol{x}_0^t)\|+\frac{\sqrt{n\alpha_t}}{2}\big[\rho'(u_t)\big]^{-1}, \quad (76)\end{aligned}$$

where the last inequality holds by $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ and $\delta_{t+1}\leq\delta_t$ in (72). Thus, it follows that

$$\begin{aligned}\frac{1}{\sqrt{10n\alpha_t}}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\| &\leq \frac{2C_\rho\big(\delta_t-\delta_{t+1}\big)}{\sqrt{n\alpha_t}}+\sqrt{n\alpha_t}\Big(\frac{1}{2}-\frac{1}{\sqrt{2}}\Big)\|\nabla F(\boldsymbol{x}_0^t)\|+\frac{\sqrt{n\alpha_t}}{2}\big[\rho'(u_t)\big]^{-1}\\ &\leq \frac{2C_\rho\big(\delta_t-\delta_{t+1}\big)}{\sqrt{n\alpha_t}}+\frac{\sqrt{n\alpha_t}}{2}\big[\rho'(u_t)\big]^{-1}.\end{aligned}$$

Upon multiplying both sides of the above equation by $\sqrt{10n\alpha_t}$, we arrive at

$$\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|\leq 2\sqrt{10}C_\rho\big(\delta_t-\delta_{t+1}\big)+\frac{\sqrt{10}n\alpha_t}{2}\big[\rho'(u_t)\big]^{-1}.$$

Then summing the above equation over t from t_3 to T, we have

$$\begin{aligned}\sum_{t=t_3}^{T}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\| &\leq 2\sqrt{10}C_\rho\sum_{t=t_3}^{T}\big(\delta_t-\delta_{t+1}\big)+\frac{\sqrt{10}n}{2}\sum_{t=t_3}^{T}\alpha_t\big[\rho'(u_t)\big]^{-1}\\ &= 2\sqrt{10}C_\rho\big(\delta_{t_3}-\delta_{T+1}\big)+\frac{\sqrt{10}n}{2}\sum_{t=t_3}^{T}\alpha_t\big[\rho'(u_t)\big]^{-1}.\end{aligned}$$

It follows from the definition of δ_t in (75) that $\delta_t=\rho\big(F(\boldsymbol{x}_0^t)-\bar{F}+u_t\big)$. Upon combining the fact that $\lim_{t\to\infty}F(\boldsymbol{x}_0^t)=\bar{F}$ in Theorem 1, $\lim_{t\to\infty}u_t=0$ in (69), and the continuity of ρ, we have

$$\lim_{t\to\infty}\delta_t=\lim_{t\to\infty}\rho\big(F(\boldsymbol{x}_0^t)-\bar{F}+u_t\big)=\rho(0)=0, \quad (77)$$

where the last equality follows from the definition of $\rho(\cdot)$ in (10). Then we can get that δ_t is bounded, that is, there exists $C_5>0$ such that $|\delta_t|\leq C_5$, thus

$$\begin{aligned}\sum_{t=t_3}^{T}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\| &\leq 2\sqrt{10}C_\rho\big(|\delta_{t_3}|+|\delta_{T+1}|\big)+\frac{\sqrt{10}n}{2}\sum_{t=t_3}^{T}\alpha_t\big[\rho'(u_t)\big]^{-1}\\ &\leq 4\sqrt{10}C_\rho C_5+\frac{\sqrt{10}n}{2}\sum_{t=t_3}^{T}\alpha_t\big[\rho'(u_t)\big]^{-1}. \quad (78)\end{aligned}$$

Now, we estimate the last term in (78) as follows. Recalling the definition of u_t in (68), we have

$$\sum_{t=t_3}^{T}\alpha_t\big[\rho'(u_t)\big]^{-1}=\sum_{t=t_3}^{T}\alpha_t\bigg[\rho'\bigg(\frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i+3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2\bigg)\bigg]^{-1}$$
$$\leq C_\rho\sum_{t=t_3}^{T}\alpha_t\Bigg(\underbrace{\bigg[\rho'\bigg(\frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i\bigg)\bigg]^{-1}}_{\mathrm{I}_t}+\underbrace{\bigg[\rho'\bigg(3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2\bigg)\bigg]^{-1}}_{\mathrm{II}_t}\Bigg),\tag{79}$$

where the last inequality follows from the quasi-additivity of $\rho(\cdot)$ in (12). Then by the non-increasing of α_t and $0\leq\beta<1$, we obtain that $\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i\leq\alpha_1\sum_{i=t}^{\infty}\beta^{i-1}\leq\frac{\alpha_1\beta^{t-1}}{1-\beta}$. Upon since $\rho(\cdot)$ is concave, then $\rho'(\cdot)$ is non-increasing and $[\rho'(\cdot)]^{-1}$ is non-decreasing, thus

$$\mathrm{I}_t\overset{(79)}{=}\bigg[\rho'\bigg(\frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i\bigg)\bigg]^{-1}\leq\bigg[\rho'\bigg(\frac{3nL^2D_1\alpha_1\beta^{t-1}}{2(1-\beta)}\bigg)\bigg]^{-1}.\tag{80}$$

On the other hand, the non-increasing of α_t, that is, $\alpha_i\leq\alpha_j$ for $i\geq j$, gives

$$\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2\leq\sum_{i=t}^{\infty}\sum_{j=2}^{i}\beta^{i-j}\alpha_j^3\leq\sum_{i=t}^{\infty}\sum_{j=2}^{t}\beta^{i-j}\alpha_j^3+\sum_{i=t}^{\infty}\sum_{j=t}^{i}\beta^{i-j}\alpha_j^3$$
$$=\sum_{i=t}^{\infty}\beta^i\sum_{j=2}^{t}\beta^{-j}\alpha_j^3+\sum_{j=t}^{\infty}\bigg(\sum_{i=j}^{\infty}\beta^i\bigg)\beta^{-j}\alpha_j^3$$
$$\leq\frac{\beta^t}{1-\beta}\sum_{j=2}^{t}\beta^{-j}\alpha_j^3+\frac{1}{1-\beta}\sum_{j=t}^{\infty}\alpha_j^3,\tag{81}$$

where the second inequality holds because $\sum_{j=2}^{i}a_j\leq\sum_{j=2}^{t}a_j+\sum_{j=t}^{i}a_j$ is valid for any $a_j\geq0$, and the last inequality follows from $0\leq\beta<1$. Thus, the quasi-addivity of $\rho(\cdot)$ and the non-decreasing of $[\rho(\cdot)]^{-1}$ indicate that

$$\mathrm{II}_t\overset{(79)}{=}\bigg[\rho'\bigg(3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2\bigg)\bigg]^{-1}$$
$$\overset{(81)}{\leq}\bigg[\rho'\bigg(3L^2R\bigg(\frac{\beta^t}{1-\beta}\sum_{j=2}^{t}\beta^{-j}\alpha_j^3+\frac{1}{1-\beta}\sum_{j=t}^{\infty}\alpha_j^3\bigg)\bigg)\bigg]^{-1}$$
$$\leq\ C_\rho\bigg[\rho'\bigg(\frac{3L^2R}{1-\beta}\,\beta^t\sum_{j=2}^{t}\beta^{-j}\alpha_j^3\bigg)\bigg]^{-1}+C_\rho\bigg[\rho'\bigg(\frac{3L^2R}{1-\beta}\sum_{j=t}^{\infty}\alpha_j^3\bigg)\bigg]^{-1}.\tag{82}$$

Upon substituting (80) and (82) in (79) gives

$$\sum_{t=t_3}^{\infty}\alpha_t\big[\rho'(u_t)\big]^{-1}\overset{(79)}{\leq}C_\rho\sum_{t=t_3}^{\infty}\alpha_t\big(\mathrm{I}_t+\mathrm{II}_t\big)\leq C_\rho\sum_{t=t_3}^{\infty}\alpha_t\bigg[\rho'\bigg(\frac{3nL^2D_1\alpha_1\beta^{t-1}}{2(1-\beta)}\bigg)\bigg]^{-1}$$
$$+C_\rho^2\sum_{t=t_3}^{\infty}\alpha_t\bigg[\rho'\bigg(\frac{3L^2R}{1-\beta}\,\beta^t\sum_{j=2}^{t}\beta^{-j}\alpha_j^3\bigg)\bigg]^{-1}+C_\rho^2\sum_{t=t_3}^{\infty}\alpha_t\bigg[\rho'\bigg(\frac{3L^2R}{1-\beta}\sum_{j=t}^{\infty}\alpha_j^3\bigg)\bigg]^{-1}<\infty,\tag{83}$$

where we apply the conditions that $\sum_{t=1}^{\infty}\alpha_t[\rho'(\beta^{t-1})]^{-1}<\infty$, $\sum_{t=2}^{\infty}\alpha_t[\rho'(\beta^t\sum_{j=2}^{t}\beta^{-j}\alpha_j^3)]^{-1}<\infty$ and $\sum_{t=2}^{\infty}\alpha_t[\rho'(\sum_{j=t}^{\infty}\alpha_j^3)]^{-1}<\infty$ in the last inequality. Upon recalling equation (78), we have

$$\sum_{t=t_3}^{\infty}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\| \overset{(78)}{\leq} 4\sqrt{10}C_\rho C_5+\frac{\sqrt{10}n}{2}\sum_{t=t_3}^{\infty}\alpha_t\big[\rho'(u_t)\big]^{-1} \overset{(83)}{<} \infty.$$

Thus, we obtain the finite length of $\{\tilde{\boldsymbol{x}}_t\}$, that is,

$$\sum_{t=1}^{\infty}\|\tilde{\boldsymbol{x}}_t-\tilde{\boldsymbol{x}}_{t-1}\|=\sum_{t=1}^{\infty}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|\leq\sum_{t=1}^{t_3}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|+\sum_{t=t_3}^{\infty}\|\boldsymbol{x}_0^{t+1}-\boldsymbol{x}_0^t\|<\infty, \tag{84}$$

where the first equality holds by the iterate of Algorithm 1.

Upon from (84), we can obtain that $\lim_{k\to\infty}\sum_{t=k}^{\infty}\|\tilde{\boldsymbol{x}}_t-\tilde{\boldsymbol{x}}_{t-1}\|=0$. Then for any $\epsilon>0$, there exists $t_4>0$ such that $\sum_{t=k}^{\infty}\|\tilde{\boldsymbol{x}}_t-\tilde{\boldsymbol{x}}_{t-1}\|\leq\epsilon$ holds for any $k\geq t_4$. Further, for any $q>p\geq t_4$, the application of triangle inequality gives

$$\|\tilde{\boldsymbol{x}}_q-\tilde{\boldsymbol{x}}_p\|=\left\|\sum_{t=p}^{q-1}\big(\tilde{\boldsymbol{x}}_{t+1}-\tilde{\boldsymbol{x}}_t\big)\right\|\leq\sum_{t=p}^{q-1}\|\tilde{\boldsymbol{x}}_{t+1}-\tilde{\boldsymbol{x}}_t\|\leq\sum_{t=p}^{\infty}\|\tilde{\boldsymbol{x}}_{t+1}-\tilde{\boldsymbol{x}}_t\|\leq\epsilon,$$

which implies that $\{\tilde{\boldsymbol{x}}_t\}$ is a Cauchy sequence. Therefore, the iterative sequence $\{\tilde{\boldsymbol{x}}_t\}$ is convergent. Recalling the conclusion (b) of Lemma 3 that $\mathcal{C}\subseteq\{\boldsymbol{x}\in\mathbb{R}^d:\nabla F(\boldsymbol{x})=\boldsymbol{0}\}$, which indicates that for the sub-sequence $\{\tilde{\boldsymbol{x}}_{t_k}\}\subseteq\{\tilde{\boldsymbol{x}}_t\}$, if $\lim_{k\to\infty}\tilde{\boldsymbol{x}}_{t_k}=\boldsymbol{x}^*\in\mathcal{C}$, then $\nabla F(\boldsymbol{x}^*)=\boldsymbol{0}$. Thus, it follows from the convergence of $\tilde{\boldsymbol{x}}_t$ that

$$\lim_{T\to\infty}\tilde{\boldsymbol{x}}_{_T}=\lim_{k\to\infty}\tilde{\boldsymbol{x}}_{t_k}=\boldsymbol{x}^*\ \text{ with }\ \nabla F(\boldsymbol{x}^*)=\boldsymbol{0},$$

which means that $\{\tilde{\boldsymbol{x}}_{_T}\}$ converges to some stationary point of F. This completes the proof. ■

C.3 Proof of Corollary 3

Proof It follows from $\rho(x)=cx^{1-\theta}$ that

$$\rho'(x)=c(1-\theta)x^{-\theta}=\frac{c(1-\theta)}{x^\theta}\implies[\rho'(x)]^{-1}=\frac{x^\theta}{c(1-\theta)}.$$

Then, we perform the proof by checking the stepsize conditions of Theorem 5 in (15), and the proof is divided in the following four cases.

(a) Since $p\leq 1$, we then have $\sum_{t=1}^{\infty}\frac{1}{t^p}=\infty$, which meets the condition $\sum_{t=1}^{\infty}\alpha_t=\infty$. Then by the condition that $p>\frac{\theta+1}{3\theta+1}>\frac{1}{3}$, we have $3p>1$, thus the condition $\sum_{t=1}^{\infty}\alpha_t^3=\sum_{t=1}^{\infty}\frac{1}{t^{3p}}<\infty$ is valid. Moreover, it is easy to see that the condition $\alpha_t\leq\frac{1}{nL\sqrt{K}}$ holds for any $t\geq(nL\sqrt{K})^{1/p}$.

(b) It follows from $[\rho'(x)]^{-1}=\frac{x^\theta}{c(1-\theta)}$ that

$$\sum_{t=1}^{\infty}\alpha_t\left[\rho'\left(\beta^{t-1}\right)\right]^{-1}=\sum_{t=1}^{\infty}\alpha_t\frac{\beta^{\theta(t-1)}}{c(1-\theta)}\leq\frac{\alpha_1}{c(1-\theta)}\sum_{t=1}^{\infty}\beta^{\theta(t-1)}\leq\frac{\alpha_1}{c(1-\theta)}\frac{1}{1-\beta^\theta}<\infty,$$

where the first inequality follows from $\alpha_t\leq\alpha_1$, and the second inequality holds because $0\leq\beta<1$. Therefore, the condition $\sum_{t=1}^{\infty}\alpha_t[\rho'(\beta^{t-1})]^{-1}<\infty$ is satisfied.

(c) Since $\alpha_t = \frac{1}{t^p}$ and $[\rho'(x)]^{-1} = \frac{x^\theta}{c(1-\theta)}$, we then have

$$\left[\rho'\Big(\beta^t \sum_{j=2}^{t} \beta^{-j}\alpha_j^3\Big)\right]^{-1} = \left[\rho'\Big(\beta^t \sum_{j=2}^{t} \beta^{-j}\frac{1}{j^{3p}}\Big)\right]^{-1} = \frac{1}{c(1-\theta)}\,\beta^{\theta t}\Big(\sum_{j=2}^{t}\beta^{-j}\frac{1}{j^{3p}}\Big)^{\theta}$$
$$\leq \frac{1}{c(1-\theta)}\,\beta^{\theta t}\sum_{j=2}^{t}\Big(\beta^{-j}\frac{1}{j^{3p}}\Big)^{\theta} = \frac{1}{c(1-\theta)}\,\beta^{\theta t}\sum_{j=2}^{t}\beta^{-\theta j}\frac{1}{j^{\,3\theta p}}, \tag{85}$$

where the last inequality follows from $0<\theta<1$ and $q_j = \beta^{-j}\frac{1}{j^{3p}} > 0$, then $\big(\sum_{j=2}^{t} q_j\big)^\theta \leq \sum_{j=2}^{t} q_j^\theta$. Further, we use $\alpha_t \leq \alpha_j = \frac{1}{j^p}$ for any $t \geq j$ to obtain

$$\sum_{t=2}^{T}\alpha_t\left[\rho'\Big(\beta^t\sum_{j=2}^{t}\beta^{-j}\alpha_j^3\Big)\right]^{-1} \overset{(85)}{\leq} \sum_{t=2}^{T}\alpha_t\Big(\frac{1}{c(1-\theta)}\,\beta^{\theta t}\sum_{j=2}^{t}\beta^{-\theta j}\frac{1}{j^{\,3\theta p}}\Big)$$
$$= \frac{1}{c(1-\theta)}\sum_{t=2}^{T}\alpha_t\beta^{\theta t}\sum_{j=2}^{t}\beta^{-\theta j}\frac{1}{j^{\,3\theta p}} \;\leq\; \frac{1}{c(1-\theta)}\sum_{t=2}^{T}\beta^{\theta t}\sum_{j=2}^{t}\beta^{-\theta j}\frac{1}{j^p}\frac{1}{j^{\,3\theta p}}$$
$$= \frac{1}{c(1-\theta)}\sum_{t=2}^{T}\beta^{\theta t}\sum_{j=2}^{t}\beta^{-\theta j}\frac{1}{j^{(3\theta+1)p}} \;=\; \frac{1}{c(1-\theta)}\sum_{j=2}^{T}\Big(\sum_{t=j}^{T}\beta^{\theta t}\Big)\beta^{-\theta j}\frac{1}{j^{\,(3\theta+1)p}}$$
$$\leq \frac{1}{c(1-\theta)}\sum_{j=2}^{T}\Big(\sum_{t=j}^{\infty}\beta^{\theta t}\Big)\beta^{-\theta j}\frac{1}{j^{\,(3\theta+1)p}} \leq \frac{1}{c(1-\theta)(1-\beta^\theta)}\sum_{t=2}^{T}\frac{1}{t^{\,(3\theta+1)p}}.$$

By the fact that $p > \frac{\theta+1}{3\theta+1} > \frac{1}{3\theta+1}$, we have $(3\theta+1)p > 1$ and $\sum_{t=2}^{\infty}\frac{1}{t^{\,(3\theta+1)p}} < \infty$. Thus, we can conclude from the above equation that $\sum_{t=2}^{\infty}\alpha_t[\rho'(\beta^t\sum_{j=2}^{t}\beta^{-j}\alpha_j^3)]^{-1} < \infty$ is satisfied.

(d) For any $t \geq 2$, we use the integral test inequality to get

$$\sum_{j=t}^{\infty}\alpha_j^3 = \sum_{j=t}^{\infty}\frac{1}{j^{3p}} \leq \int_{t-1}^{\infty}\frac{1}{x^{3p}}\,dx = \frac{-x^{\,1-3p}}{3p-1}\Big|_{t-1}^{\infty} \leq \frac{(t-1)^{\,1-3p}}{3p-1},$$

where the last inequality follows from $p > \frac{\theta+1}{3\theta+1} > \frac{1}{3}$, thus $1-3p<0$. Upon by the non-decreasing of $[\rho'(\cdot)]^{-1}$, we can obtain

$$\left[\rho'\Big(\sum_{j=t}^{\infty}\alpha_j^3\Big)\right]^{-1} \leq \left[\rho'\Big(\frac{(t-1)^{\,1-3p}}{3p-1}\Big)\right]^{-1} = \frac{1}{c(1-\theta)}\Big(\frac{(t-1)^{\,1-3p}}{3p-1}\Big)^{\theta} = \frac{(t-1)^{\,(1-3p)\theta}}{c(1-\theta)(3p-1)^\theta}, \tag{86}$$

where the first equality follows from $[\rho'(x)]^{-1} = \frac{x^\theta}{c(1-\theta)}$. Thus, it follows that

$$\sum_{t=2}^{\infty}\alpha_t\left[\rho'\Big(\sum_{j=t}^{\infty}\alpha_j^3\Big)\right]^{-1} \overset{(86)}{\leq} \sum_{t=2}^{\infty}\frac{1}{t^p}\frac{(t-1)^{\,(1-3p)\theta}}{c(1-\theta)(3p-1)^\theta} \leq \frac{1}{c(1-\theta)(3p-1)^\theta}\sum_{t=2}^{\infty}\frac{(t-1)^{\,(1-3p)\theta}}{(t-1)^p}$$
$$= \frac{1}{c(1-\theta)(3p-1)^\theta}\sum_{t=2}^{\infty}\frac{1}{(t-1)^{\,(3p-1)\theta+p}} < \infty,$$

where the second inequality is due to $\frac{1}{t^p} \leq \frac{1}{(t-1)^p}$ for any $t \geq 2$, and the last inequality holds by $p > \frac{\theta+1}{3\theta+1}$, then $(3p-1)\theta + p > 1$. Hence, the condition $\sum_{t=2}^{\infty}\alpha_t[\rho'(\sum_{j=t}^{\infty}\alpha_j^3)]^{-1}$ is satisfied.

Finally, combining the above cases (a), (b), (c) and (d), we know that the conditions of Theorem 5 in (15) hold. Therefore, we can apply the conclusion of Theorem 5 to get that $\{\tilde{\boldsymbol{x}}_T\}$ has finite length and converges to some stationary point $\boldsymbol{x}^*$ of F. This completes the proof. ■

C.4 Proof of Proposition 4

Proof It follows from $e_{t+1} \leq (1-\lambda_t)e_t + \gamma_t$ and $\lambda_t \leq 1$ that

$$\begin{aligned} e_{T+1} &\leq \prod_{t=1}^{T}(1-\lambda_t)e_1 + \sum_{t=1}^{T}\Big(\prod_{i=t+1}^{T}(1-\lambda_i)\Big)\gamma_t \\ &\leq \prod_{t=1}^{T}\exp(-\lambda_t)e_1 + \sum_{t=1}^{T}\Big(\prod_{i=t+1}^{T}\exp(-\lambda_i)\Big)\gamma_t \\ &= \exp\Big(-\sum_{t=1}^{T}\lambda_t\Big)e_1 + \sum_{t=1}^{T}\exp\Big(-\sum_{i=t+1}^{T}\lambda_i\Big)\gamma_t, \end{aligned}$$

where the second inequality follows from $1-x \leq \exp(-x)$. This completes the proof. ■

C.5 Proof of Theorem 6

Proof From the stepsize choice of $\alpha_t = \frac{10c^2}{nt}$, we can obtain that α_t is non-increasing, $\sum_{t=1}^{\infty}\alpha_t^3 = \sum_{t=1}^{\infty}\frac{10^3c^6}{n^3t^3} < \infty$, $\sum_{t=1}^{\infty}\alpha_t = \sum_{t=1}^{\infty}\frac{10c^2}{nt} = \infty$, and $\alpha_t \leq \frac{1}{nL\sqrt{K}}$ is satisfied for any $t \geq 10c^2L\sqrt{K}$. In the rest of the proof, we assume that $t \geq t_5 = \max\{t_3, 10c^2L\sqrt{K}\}$, where t_3 is defined in the proof of Theorem 5, that is, t_3 is a constant such that $\mathrm{dist}(\boldsymbol{x}_0^t, \mathcal{C}) < \epsilon$ and $|F(\boldsymbol{x}_0^t) - \bar{F}| < \eta$ holds for any $t \geq t_3$. Thus, we begin with (76) in the proof of Theorem 5, that is,

$$\begin{aligned} \frac{1}{\sqrt{10n\alpha_t}}\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| + \frac{\sqrt{n\alpha_t}}{\sqrt{2}}\|\nabla F(\boldsymbol{x}_0^t)\| &\overset{(76)}{\leq} \frac{2C_\rho\big(\delta_t - \delta_{t+1}\big)}{\sqrt{n\alpha_t}} + \frac{\sqrt{n\alpha_t}}{2}\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{\sqrt{n\alpha_t}}{2}\big[\rho'(u_t)\big]^{-1} \\ &= \frac{2\big(\delta_t - \delta_{t+1}\big)}{\sqrt{n\alpha_t}} + \frac{\sqrt{n\alpha_t}}{2}\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{\sqrt{n\alpha_t}}{2}\big[\rho'(u_t)\big]^{-1}, \end{aligned}$$

where $\delta_t = \rho(F(\boldsymbol{x}_0^t) - \bar{F} + u_t)$ and the sequence u_t is defined in (68), that is, $u_t = \frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i + 3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2$ with two finite values $D_1 = \frac{1}{L^2}(\sigma_1\|\nabla F(\boldsymbol{x}_0^1)\|^2 + \sigma_2)$ and $R = n^3(\beta G + \sigma_1(1-\beta)G + \sigma_2(1-\beta))$, and the last equality holds by equation (13), that is, if $\rho(x) = cx^{1-\theta}$, then $C_\rho = 1$. Now, rearranging the above equation, we have

$$\begin{aligned} \frac{1}{\sqrt{10n\alpha_t}}\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| &\leq \frac{2\big(\delta_t - \delta_{t+1}\big)}{\sqrt{n\alpha_t}} - \frac{(\sqrt{2}-1)\sqrt{n\alpha_t}}{2}\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{\sqrt{n\alpha_t}}{2}\big[\rho'(u_t)\big]^{-1} \\ &\leq \frac{2\big(\delta_t - \delta_{t+1}\big)}{\sqrt{n\alpha_t}} - \frac{\sqrt{n\alpha_t}}{5}\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{\sqrt{n\alpha_t}}{2}\big[\rho'(u_t)\big]^{-1}, \end{aligned}$$

where the last inequality follows from $\sqrt{2}-1 \geq \frac{2}{5}$. Thus, it follows that

$$\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| \leq 2\sqrt{10}\big(\delta_t - \delta_{t+1}\big) - \frac{\sqrt{10}n\alpha_t}{5}\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{\sqrt{10}n\alpha_t}{2}\big[\rho'(u_t)\big]^{-1}.$$

Upon summing up the above equation, we have

$$\sum_{k=t}^{T}\|\boldsymbol{x}_0^{k+1}-\boldsymbol{x}_0^k\| \le 2\sqrt{10}\sum_{k=t}^{T}\big(\delta_k-\delta_{k+1}\big)-\frac{\sqrt{10}n}{5}\sum_{k=t}^{T}\alpha_k\|\nabla F(\boldsymbol{x}_0^k)\|+\frac{\sqrt{10}n}{2}\sum_{k=t}^{T}\alpha_k\big[\rho'(u_k)\big]^{-1}$$
$$=2\sqrt{10}\big(\delta_t-\delta_{T+1}\big)-\frac{\sqrt{10}n}{5}\sum_{k=t}^{T}\alpha_k\|\nabla F(\boldsymbol{x}_0^k)\|+\frac{\sqrt{10}n}{2}\sum_{k=t}^{T}\alpha_k\big[\rho'(u_k)\big]^{-1}.$$

Taking $T\to\infty$ and using the fact that $\lim_{T\to\infty}\delta_{T+1}=0$ obtained in (77), we arrive at

$$\sum_{k=t}^{\infty}\|\boldsymbol{x}_0^{k+1}-\boldsymbol{x}_0^k\| \le 2\sqrt{10}\delta_t-\frac{\sqrt{10}n}{5}\sum_{k=t}^{\infty}\alpha_k\|\nabla F(\boldsymbol{x}_0^k)\|+\frac{\sqrt{10}n}{2}\sum_{k=t}^{\infty}\alpha_k\big[\rho'(u_k)\big]^{-1}$$
$$=2\sqrt{10}\rho(F(\boldsymbol{x}_0^t)-\bar{F}+u_t)-\frac{\sqrt{10}n}{5}\sum_{k=t}^{\infty}\alpha_k\|\nabla F(\boldsymbol{x}_0^k)\|+\frac{\sqrt{10}n}{2}\sum_{k=t}^{\infty}\alpha_k\big[\rho'(u_k)\big]^{-1}, \tag{87}$$

where the last equality follows from the definition of δ_t. Upon since $\rho(x)=c\sqrt{x}$, we then obtain that $\rho(x)$ is increasing. By the fact that $F(\boldsymbol{x}_0^t)-\bar{F}+u_t\le|F(\boldsymbol{x}_0^t)-\bar{F}|+u_t$, we have

$$\rho(F(\boldsymbol{x}_0^t)-\bar{F}+u_t)\le\rho(|F(\boldsymbol{x}_0^t)-\bar{F}|+u_t)=\frac{c^2}{2}\frac{1}{\rho'(|F(\boldsymbol{x}_0^t)-\bar{F}|+u_t)}$$
$$\le\frac{c^2}{2}\Big(\frac{1}{\rho'(|F(\boldsymbol{x}_0^t)-\bar{F}|)}+\frac{1}{\rho'(u_t)}\Big)=\frac{c^2}{2}\Big(\frac{2\sqrt{|F(\boldsymbol{x}_0^t)-\bar{F}|}}{c}+\frac{2\sqrt{u_t}}{c}\Big)$$
$$\le\frac{c^2}{2}\Big(\|\nabla F(\boldsymbol{x}_0^t)\|+\frac{2\sqrt{u_t}}{c}\Big), \tag{88}$$

where the first equality is due to $\rho(x)=\frac{c^2}{2\rho'(x)}$, the second inequality holds by $u_t\ge0$ and equation (13), that is, $\frac{1}{\rho'(x+y)}\le\frac{1}{\rho'(x)}+\frac{1}{\rho'(y)}$, the second equality follows from $\rho'(x)=\frac{c}{2\sqrt{x}}$, and the last inequality follows from the Łojasiewicz inequality with $\theta=\frac{1}{2}$, that is, $\|\nabla F(\boldsymbol{x}_0^t)\|\ge\frac{2}{c}\sqrt{|F(\boldsymbol{x}_0^t)-\bar{F}|}$.

Next, substituting (88) in (87), we arrive at

$$\sum_{k=t}^{\infty}\|\boldsymbol{x}_0^{k+1}-\boldsymbol{x}_0^k\| \le 2\sqrt{10}\Big(\frac{c^2}{2}\Big(\|\nabla F(\boldsymbol{x}_0^t)\|+\frac{2\sqrt{u_t}}{c}\Big)\Big)$$
$$-\frac{\sqrt{10}n}{5}\sum_{k=t}^{\infty}\alpha_k\|\nabla F(\boldsymbol{x}_0^k)\|+\frac{\sqrt{10}n}{2}\sum_{k=t}^{\infty}\alpha_k\big[\rho'(u_k)\big]^{-1}$$
$$=\sqrt{10}c^2\Big(\|\nabla F(\boldsymbol{x}_0^t)\|+\frac{2\sqrt{u_t}}{c}\Big)-\frac{\sqrt{10}n}{5}\sum_{k=t}^{\infty}\alpha_k\|\nabla F(\boldsymbol{x}_0^k)\|+\frac{\sqrt{10}n}{c}\sum_{k=t}^{\infty}\alpha_k\sqrt{u_k}.$$

Upon adding $\frac{2\sqrt{10}n}{5c}\sum_{k=t}^{\infty}\alpha_k\sqrt{u_k}$ to both sides of the above equation and rearranging it to get

$$\sum_{k=t}^{\infty}\|\boldsymbol{x}_0^{k+1}-\boldsymbol{x}_0^k\|+\frac{\sqrt{10}n}{5}\sum_{k=t}^{\infty}\alpha_k\Big(\|\nabla F(\boldsymbol{x}_0^k)\|+\frac{2\sqrt{u_k}}{c}\Big)$$
$$\le\sqrt{10}c^2\Big(\|\nabla F(\boldsymbol{x}_0^t)\|+\frac{2\sqrt{u_t}}{c}\Big)+\frac{\sqrt{10}n}{c}\sum_{k=t}^{\infty}\alpha_k\sqrt{u_k}+\frac{2\sqrt{10}n}{5c}\sum_{k=t}^{\infty}\alpha_k\sqrt{u_k}$$
$$=\sqrt{10}c^2\Big(\|\nabla F(\boldsymbol{x}_0^t)\|+\frac{2\sqrt{u_t}}{c}\Big)+\frac{7\sqrt{10}n}{5c}\sum_{k=t}^{\infty}\alpha_k\sqrt{u_k}. \tag{89}$$

Further, let's set

$$\Upsilon_t = \sum_{k=t}^{\infty} \|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}_0^k\| + \frac{\sqrt{10}n}{5} \sum_{k=t}^{\infty} \alpha_k \bigg(\|\nabla F(\boldsymbol{x}_0^k)\| + \frac{2\sqrt{u_k}}{c} \bigg). \tag{90}$$

Since α_t is non-increasing and $\sum_{t=1}^{\infty} \alpha_t^3 < \infty$, then we can obtain that $\|\nabla F(\boldsymbol{x}_0^t)\| \leq \sqrt{G}$ is bounded from Corollary 1, u_t is bounded from $\lim_{t\to\infty} u_t = 0$ in (69), and $\sum_{k=t}^{\infty} \alpha_k \sqrt{u}_k$ is bounded from (83) with $\rho(x) = c\sqrt{x}$. Thus, there exists a constant $\Upsilon > 0$ such that

$$\Upsilon_t \overset{(89)}{\leq} \sqrt{10}c^2 \bigg(\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{2\sqrt{u_t}}{c} \bigg) + \frac{7\sqrt{10}n}{5c} \sum_{k=t}^{\infty} \alpha_k \sqrt{u_k} \leq \Upsilon. \tag{91}$$

In addition, we have $\Upsilon_{t+1} \leq \Upsilon_t$ and

$$\Upsilon_t - \Upsilon_{t+1} = \|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| + \frac{\sqrt{10}n\alpha_t}{5} \bigg(\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{2\sqrt{u_t}}{c} \bigg).$$

Thus, it follows that

$$\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{2\sqrt{u_t}}{c} = \frac{5}{\sqrt{10}n\alpha_t} \big(\Upsilon_t - \Upsilon_{t+1} \big) - \frac{5}{\sqrt{10}n\alpha_t} \|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| \leq \frac{5}{\sqrt{10}n\alpha_t} \big(\Upsilon_t - \Upsilon_{t+1} \big). \tag{92}$$

Upon substituting the definition of Υ_t from (90) and equation (92) into (89), we arrive at

$$\begin{aligned} \Upsilon_t &\overset{(89)}{\leq} \sqrt{10}c^2 \left(\|\nabla F(\boldsymbol{x}_0^t)\| + \frac{2\sqrt{u_t}}{c} \right) + \frac{7\sqrt{10}n}{5c} \sum_{k=t}^{\infty} \alpha_k \sqrt{u_k} \\ &\overset{(92)}{\leq} \frac{5c^2}{n} \frac{\big(\Upsilon_t - \Upsilon_{t+1} \big)}{\alpha_t} + \frac{7\sqrt{10}n}{5c} \sum_{k=t}^{\infty} \alpha_k \sqrt{u_k}. \end{aligned} \tag{93}$$

Upon multiplying both sides of (93) by $\frac{n\alpha_t}{5c^2}$ and rearranging it, we can get

$$\Upsilon_{t+1} \leq \bigg(1 - \frac{n\alpha_t}{5c^2} \bigg) \Upsilon_t + \frac{7\sqrt{10}n^2}{25c^3} \alpha_t \sum_{k=t}^{\infty} \alpha_k \sqrt{u_k} = \bigg(1 - \frac{2}{t} \bigg) \Upsilon_t + \underbrace{28\sqrt{10}c \, \frac{1}{t} \sum_{k=t}^{\infty} \frac{\sqrt{u_k}}{k}}_{\gamma_t}, \tag{94}$$

where the last equality is due to $\alpha_t = \frac{10c^2}{nt}$. By setting $\lambda_t = \frac{2}{t}$ and $\gamma_t = 28\sqrt{10}c \, \frac{1}{t} \sum_{k=t}^{\infty} \frac{\sqrt{u_k}}{k}$, the equation (94) becomes $\Upsilon_{t+1} \leq (1 - \lambda_t)\Upsilon_t + \gamma_t$, thus we use the conclusion of Proposition 4 to get

$$\Upsilon_{T+1} \leq \exp\bigg(- \sum_{k=t_5}^{T} \lambda_k \bigg) \Upsilon_{t_5} + \sum_{k=t_5}^{T} \exp\bigg(- \sum_{i=k+1}^{T} \lambda_i \bigg) \gamma_k,$$

where $t_5 = \max\{t_3, 10c^2 L\sqrt{K}\}$ is a constant. Without loss of generality, assuming that $t_5 \geq 2$ is an integer, thus $\lambda_t \leq 1$ holds. Then it follows from the integral test inequality that

$$\sum_{k=t_5}^{T} \lambda_k = \sum_{k=t_5}^{T} \frac{2}{k} \geq \int_{t_5}^{T+1} \frac{2}{x} \, dx = 2 \ln \bigg(\frac{T+1}{t_5} \bigg), \quad \sum_{i=k+1}^{T} \lambda_i = \sum_{i=k+1}^{T} \frac{2}{i} \geq \int_{k+1}^{T+1} \frac{2}{x} \, dx = 2 \ln \bigg(\frac{T+1}{k+1} \bigg).$$

Thus, it follows that

$$\begin{aligned}\Upsilon_{T+1} &\leq \exp\left(-2\ln\left(\frac{T+1}{t_5}\right)\right)\Upsilon_{t_5} + \sum_{k=t_5}^{T}\exp\left(-2\ln\left(\frac{T+1}{k+1}\right)\right)\gamma_k \\ &= \frac{t_5^2\,\Upsilon_{t_5}}{(T+1)^2} + \frac{1}{(T+1)^2}\sum_{k=t_5}^{T}(k+1)^2\gamma_k. \end{aligned} \tag{95}$$

Further, we estimate the second term in (95) as follows. From the definitions of γ_k in (94) and u_t in (68), we can obtain

$$\begin{aligned}&\sum_{k=t_5}^{T}(k+1)^2\gamma_k \overset{(94)}{=} 28\sqrt{10}c\sum_{k=t_5}^{T}\frac{(k+1)^2}{k}\sum_{t=k}^{\infty}\frac{\sqrt{u_t}}{t} \\ &\overset{(68)}{=} 28\sqrt{10}c\sum_{k=t_5}^{T}\frac{(k+1)^2}{k}\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i + 3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2} \\ &\leq 28\sqrt{10}c\sum_{k=t_5}^{T}\frac{(k+1)^2}{k}\sum_{t=k}^{\infty}\frac{1}{t}\left(\sqrt{\frac{3nL^2D_1}{2}\sum_{i=t}^{\infty}\beta^{i-1}\alpha_i} + \sqrt{3L^2R\sum_{i=t}^{\infty}\alpha_i\sum_{j=2}^{i}\beta^{i-j}\alpha_j^2}\right) \\ &= \underbrace{28\sqrt{10}c\sqrt{\frac{3nL^2D_1}{2}\frac{10c^2}{n}}}_{C_7}\underbrace{\sum_{k=t_5}^{T}\frac{(k+1)^2}{k}\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{\beta^{i-1}}{i}}}_{\mathrm{I}_t} \\ &+ \underbrace{28\sqrt{10}c\sqrt{3L^2R\frac{10^3c^6}{n^3}}}_{C_8}\underbrace{\sum_{k=t_5}^{T}\frac{(k+1)^2}{k}\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}}}_{\mathrm{II}_t}, \end{aligned} \tag{96}$$

where the first inequality holds by $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$, the last equality follows from $\alpha_t = \frac{10c^2}{nt}$, and we introduce the constants C_7 and C_8 in the last equality to simply the proof. Now, we estimate I_t and II_t in (96), respectively. For I_t, by the fact that $(k+1)^2 \leq (k+k)^2 = 4k^2$, we have

$$\begin{aligned}\mathrm{I}_t &= \sum_{k=t_5}^{T}\frac{(k+1)^2}{k}\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{\beta^{i-1}}{i}} \leq 4\sum_{k=t_5}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{\beta^{i-1}}{i}} \\ &\leq 4\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{\beta^{i-1}}{i}} \leq \frac{2\sqrt{2}}{\sqrt{1-\beta}}\,\frac{1}{(1-\sqrt{\beta})^2}, \end{aligned} \tag{97}$$

where the second inequality holds by the condition that $t_5 \geq 2$, and the last inequality follows from Lemma 8(a). Upon for II_t, we have

$$\begin{aligned}\mathrm{II}_t &= \sum_{k=t_5}^{T}\frac{(k+1)^2}{k}\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}} \leq 4\sum_{k=t_5}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}} \\ &\leq 4\sum_{k=2}^{T}k\sum_{t=k}^{\infty}\frac{1}{t}\sqrt{\sum_{i=t}^{\infty}\frac{1}{i}\sum_{j=2}^{i}\frac{\beta^{i-j}}{j^2}} \leq \frac{4}{\sqrt{1-\beta}}\frac{4\sqrt{T}}{1-\sqrt{\beta}} + \frac{4\sqrt{2}(\ln T + T)}{\sqrt{1-\beta}} + 4C_6, \end{aligned} \tag{98}$$

where $C_6 = \frac{1}{\sqrt{1-\beta}}\frac{4}{(1-\sqrt{\beta})^2} + \frac{\sqrt{2}}{\sqrt{1-\beta}}$ is a finite value, and the last inequality follows from Lemma 8(b). Upon substituting (96), (97) and (98) in (95), we arrive at

$$\begin{aligned}
\Upsilon_{T+1} &\overset{(95)}{\leq} \frac{t_5^2\,\Upsilon_{t_5}}{(T+1)^2} + \frac{1}{(T+1)^2}\sum_{k=t_5}^{T}(k+1)^2\gamma_k \overset{(91),(96)}{\leq} \frac{t_5^2\,\Upsilon}{(T+1)^2} + \frac{1}{(T+1)^2}\left(C_7\,\mathrm{I}_t + C_8\,\mathrm{II}_t\right)\\
&\leq \frac{t_5^2\,\Upsilon}{(T+1)^2} + \frac{C_7}{(T+1)^2}\frac{2\sqrt{2}}{\sqrt{1-\beta}}\frac{1}{(1-\sqrt{\beta})^2} + \frac{C_8}{(T+1)^2}\Big(\frac{4}{\sqrt{1-\beta}}\frac{4\sqrt{T}}{1-\sqrt{\beta}} + \frac{4\sqrt{2}(\ln T + T)}{\sqrt{1-\beta}} + 4C_6\Big)\\
&\leq \frac{t_5^2\,\Upsilon}{(T+1)^2} + \frac{C_7}{(T+1)^2}\frac{2\sqrt{2}}{\sqrt{1-\beta}(1-\sqrt{\beta})^2} + \frac{C_8 T}{(T+1)^2}\Big(\frac{4}{\sqrt{1-\beta}}\frac{4}{1-\sqrt{\beta}} + \frac{8\sqrt{2}}{\sqrt{1-\beta}} + 4C_6\Big), \qquad (99)
\end{aligned}$$

where the last inequality holds because $\sqrt{T} \leq T$ and $\ln T \leq T$. Upon recalling the definition of Υ_t in (90), we know that $\Upsilon_t = \sum_{k=t}^{\infty}\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}_0^k\| + \frac{\sqrt{10}n}{5}\sum_{k=t}^{\infty}\alpha_k(\|\nabla F(\boldsymbol{x}_0^k)\| + \frac{2\sqrt{u_k}}{c})$. Since Corollary 3 shows that when $\theta = \frac{1}{2}$ and $p = 1$, $\boldsymbol{x}_0^{T+1} \to \boldsymbol{x}^*\,(T \to \infty)$ for some stationary point $\boldsymbol{x}^*$ of F. We then use the triangle inequality to get

$$\|\boldsymbol{x}_0^{T+1} - \boldsymbol{x}^*\| \leq \sum_{t=T+1}^{\infty}\|\boldsymbol{x}_0^{t+1} - \boldsymbol{x}_0^t\| \leq \Upsilon_{T+1}.$$

Thus, we finally obtain

$$\begin{aligned}
\|\tilde{\boldsymbol{x}}_T - \boldsymbol{x}^*\| = \|\boldsymbol{x}_0^{T+1} - \boldsymbol{x}^*\| &\overset{(99)}{\leq} \frac{t_5^2\,\Upsilon}{(T+1)^2} + \frac{C_7}{(T+1)^2}\frac{2\sqrt{2}}{\sqrt{1-\beta}(1-\sqrt{\beta})^2}\\
&+ \frac{C_8 T}{(T+1)^2}\Big(\frac{4}{\sqrt{1-\beta}}\frac{4}{1-\sqrt{\beta}} + \frac{8\sqrt{2}}{\sqrt{1-\beta}} + 4C_6\Big) = \mathcal{O}\Big(\frac{1}{T^2}\Big) + \mathcal{O}\Big(\frac{1}{T}\Big).
\end{aligned}$$

This completes the proof. ■

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: Optimal rates without component convexity and large epoch requirements. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 17526–17535, 2020.

Anas Barakat and Pascal Bianchi. Convergence rates of a momentum algorithm with bounded adaptive step size for nonconvex optimization. In *Asian Conference on Machine Learning (ACML)*, volume 129, pages 225–240, 2020.

Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Symposium on Statistical Learning and Data Science (SLDS)*, volume 8, pages 2624–2633, 2009.

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

Hanseul Cho and Chulhee Yun. SGDA with shuffling: Faster convergence for nonconvex-PŁ minimax optimization. In *International Conference on Learning Representations (ICLR)*, 2023.

Aniket Das, Bernhard Schölkopf, and Michael Muehlebach. Sampling without replacement leads to faster rates in finite-sum minimax optimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 35, pages 6749–6762, 2022.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29, 2016.

Jiahong Guo, Xiao Wang, and Xiantao Xiao. Dynamical convergence analysis for nonconvex linearized proximal ADMM algorithms. *arXiv preprint arXiv:2309.07008*, 2023a.

Jiahong Guo, Xiao Wang, and Xiantao Xiao. Preconditioned primal-dual gradient methods for nonconvex composite and finite-sum optimization. *arXiv preprint arXiv:2309.13416*, 2023b.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo A Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021.

Kun Huang, Xiao Li, Andre Milzarek, Shi Pu, and Junwen Qiu. Distributed random reshuffling over networks. *IEEE Transactions on Signal Processing*, 71:1143–1158, 2023.

Cédric Josz and Lexiao Lai. Global stability of first-order methods for coercive tame functions. *Mathematical Programming*, 207(1):551–576, 2023.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 795–811, 2016.

Hiroyuki Kasai. SGDLibrary: A MATLAB library for stochastic optimization algorithms. *Journal of Machine Learning Research*, 18(215):1–5, 2018.

Anastasia Koloskova, Nikita Doikov, Sebastian U Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In *International Conference on Machine Learning (ICML)*, 2024.

Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4394–4400, 2020.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in Neural Information Processing Systems (NIPS)*, volume 28, 2015.

Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka-Łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 983–992, 2019.

Zhouchen Lin, Huan Li, and Cong Fang. *Accelerated Optimization for Machine Learning*. Springer Singapore, 2020.

Jinlan Liu, Jun Kong, Dongpo Xu, Miao Qi, and Yinghua Lu. Convergence analysis of AdaBound with relaxed bound functions for non-convex optimization. *Neural Networks*, 145:300–307, 2022.

Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory (COLT)*, volume 178, pages 2963–2983, 2022.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 18261–18271, 2020.

Grigory Malinovsky, Alibek Sailanbayev, and Peter Richtárik. Random reshuffling with variance reduction: New analysis and better rates. In *Uncertainty in Artificial Intelligence (UAI)*, volume 216, pages 1347–1357, 2023.

Danilo P. Mandic and Jonathon Chambers. *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*. Wiley, USA, 2001.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtarik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 17309–17320, 2020.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Proximal and federated random reshuffling. In *International Conference on Machine Learning (ICML)*, volume 162, pages 15718–15749, 2022.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning (ICML)*, volume 97, pages 4703–4711, 2019.

Lam M. Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takáč, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019.

Lam M. Nguyen, Quoc Tran-Dinh, Dzung T. Phan, Phuong Ha Nguyen, and Marten van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.

Nhan H. Pham, Lam M. Nguyen, Dzung T. Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

Zhen Qin, Zhishuai Liu, and Pan Xu. Convergence of sign-based random reshuffling algorithms for nonconvex optimization. *arXiv preprint arXiv:2310.15976*, 2023.

Ali Ramezani-Kebrya, Kimon Antonakopoulos, Volkan Cevher, Ashish Khisti, and Ben Liang. On the generalization of stochastic gradient descent with momentum. *Journal of Machine Learning Research*, 25(22):1–56, 2024.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Conference on Learning Theory (COLT)*, volume 125, pages 3250–3284, 2020.

Tao Sun, Linbo Qiao, and Dongsheng Li. Nonergodic complexity of proximal inertial gradient descents. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4613–4626, 2021.

Trang H Tran, Lam M Nguyen, and Quoc Tran-Dinh. SMG: A shuffling gradient-based method with momentum. In *International Conference on Machine Learning (ICML)*, volume 139, pages 10379–10389, 2021.

Trang H Tran, Katya Scheinberg, and Lam M Nguyen. Nesterov accelerated shuffling gradient method for convex optimization. In *International Conference on Machine Learning (ICML)*, volume 162, pages 21703–21732, 2022.

Trang H Tran, Quoc Tran-Dinh, and Lam M Nguyen. Shuffling momentum gradient algorithm for convex optimization. *Vietnam Journal of Mathematics*, pages 1–29, 2024.

Quoc Tran-Dinh and Marten van Dijk. Gradient descent-type methods: Background and simple unified convergence analysis. In *Federated Learning*, pages 3–28. Academic Press, 2024.

Thijs Vogels, Lie He, Anastasiia Koloskova, Sai Praneeth Karimireddy, Tao Lin, Sebastian U Stich, and Martin Jaggi. RelaySum for decentralized deep learning on heterogeneous data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 34, pages 28004–28015, 2021.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Conference on Learning Theory (COLT)*, volume 195, pages 161–190, 2023.

Xiaoyu Wang and Yaxiang Yuan. On the convergence of stochastic gradient descent with bandwidth-based step size. *Journal of Machine Learning Research*, 24(48):1–49, 2023.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019.

Yingzhen Yang and Ping Li. Projective proximal gradient descent for nonconvex nonsmooth optimization: Fast convergence without Kurdyka-Łojasiewicz (KL) property. In *International Conference on Learning Representations (ICLR)*, 2023.

Jinshan Zeng and Wotao Yin. On nonconvex decentralized gradient descent. *IEEE Transactions on Signal Processing*, 66(11):2834–2848, 2018.