On the Last-Iterate Convergence of Shuffling Gradient Methods

Zijian Liu [1] **Zhengyuan Zhou** [1 2]

Abstract

Shuffling gradient methods are widely used in modern machine learning tasks and include three popular implementations: Random Reshuffle (RR), Shuffle Once (SO), and Incremental Gradient (IG). Compared to the empirical success, the theoretical guarantee of shuffling gradient methods was not well-understood for a long time. Until recently, the convergence rates had just been established for the average iterate for convex functions and the last iterate for strongly convex problems (using squared distance as the metric). However, when using the function value gap as the convergence criterion, existing theories cannot interpret the good performance of the last iterate in different settings (e.g., constrained optimization). To bridge this gap between practice and theory, we prove the first last-iterate convergence rates for shuffling gradient methods with respect to the objective value even without strong convexity. Our new results either (nearly) match the existing last-iterate lower bounds or are as fast as the previous best upper bounds for the average iterate.

1. Introduction

Solving the convex optimization problem in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq f(\mathbf{x}) + \psi(\mathbf{x}) \text{ where } f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$$

is arguably one of the most common tasks in machine learning. Many well-known problems such as the LAD regression (Pollard, 1991), SVMs (Cortes & Vapnik, 1995), the Lasso (Tibshirani, 1996), and any general problem from empirical risk minimization (ERM) (Shalev-Shwartz & Ben-David, 2014) fit the framework perfectly. A famous method for the problem with a large n (the standard case nowadays) is the stochastic gradient descent (SGD) algorithm, which in every step only computes one gradient differing from calculating n gradients in the classic gradient descent (GD) method and hence is more computationally affordable.

However, in practice, the shuffling gradient method is more widely implemented than SGD. Unlike SGD uniformly sampling the index in every step, shuffling gradient methods split the optimization procedure into K epochs each of which contains n updates where the index used in every step is determined by a permutation σ of $\{1, 2, \cdots, n\}$. Especially, three ways to generate σ are mainly used: (I) Random Reshuffle (RR) where σ is uniformly sampled without replacement in every epoch. (II) Shuffle Once (SO) also known as Single Shuffle where σ in every epoch is identical to a pre-generated permutation by uniform sampling without replacement. (III) Incremental Gradient (IG) where σ in every epoch is always the same as some deterministic order.

Though shuffling gradient methods are empirically popular and successful (Bottou, 2009; 2012; Bengio, 2012; Recht & Ré, 2013), the theoretical understanding has not been well-developed for a long time due to the biased gradient used in every step making the analysis notoriously harder than SGD. (Gürbüzbalaban et al., 2021) is the first to demonstrate that RR provably converges faster than SGD under certain conditions. Later on, different works give extensive studies for the theoretical convergence of shuffling gradient methods (see Section 2 for a detailed discussion).

Despite the existing substantial progress, the convergence of shuffling gradient methods is still not fully understood. Specifically, prior works heavily rely on the following problem form that significantly restricts the applicability:

The regularizer $\psi(\mathbf{x})$ is always omitted. Almost all previous studies narrow to the case of $\psi(\mathbf{x}) = 0$, which even does not include the common constrained optimization problems (i.e., taking $\psi(\mathbf{x}) = \mathrm{I}_{\mathbf{X}}(\mathbf{x})$ where $\mathrm{I}_{\mathbf{X}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathbf{X}$[1], otherwise, $+\infty$). The only exception we know is (Mishchenko et al., 2022), which allows the existence of a general $\psi(\mathbf{x})$. However, (Mishchenko et al., 2022) requires either $f_i(\mathbf{x})$ or $\psi(\mathbf{x})$ to be strongly convex, which can be easily violated in practice like the Lasso.

[1]Stern School of Business, New York University [2]Arena Technologies. Correspondence to: Zijian Liu <zl3067@stern.nyu.edu>, Zhengyuan Zhou <zhengyuanzhou24@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]This notation $\mathbf{X}$ always denotes a nonempty closed convex set in $\mathbb{R}^d$ throughout the paper.

Table 1. Summary of our new upper bounds and the existing lower bounds for L-smooth $f_i(\mathbf{x})$ for large K. If no lower bound was established before in the case, we instead state the previous best-known rate. Here, $\sigma^2_{\text{any}} \triangleq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_*)\|^2$, $\sigma^2_{\text{rand}} \triangleq \sigma^2_{\text{any}} + n\|\nabla f(\mathbf{x}_*)\|^2$ and $D \triangleq \|\mathbf{x}_* - \mathbf{x}_1\|$. All rates use the function value gap as the convergence criterion. In the column of "Type", "Any" means the rate holds for whatever permutation not limited to RR/SO/IG. "Random" refers to the uniformly sampled permutation but is not restricted to RR/SO (see Remark 4.5 for a detailed explanation). "Avg" and "Last" in the "Output" column stand for the average iterate and the last iterate, respectively. In the last column, "✓" means $\psi(\mathbf{x})$ can be taken arbitrarily and "✗" implies $\psi(\mathbf{x}) = 0$.

$F(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$ where $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ and $f_i(\mathbf{x})$ and $\psi(\mathbf{x})$ are convex					
Settings	References	Rate	Type	Output	$\psi(\mathbf{x})$
L-smooth $f_i(\mathbf{x})$	(Mishchenko et al., 2020)	$\mathcal{O}\left(\frac{L^{1/3}\sigma^{2/3}_{\text{any}}D^{4/3}}{K^{2/3}}\right)$	IG	Avg	✗
	Ours (Theorem 4.4)	$\widetilde{\mathcal{O}}\left(\frac{L^{1/3}\sigma^{2/3}_{\text{any}}D^{4/3}}{K^{2/3}}\right)$	Any	Last	✓
	Ours (Theorem 4.4)	$\widetilde{\mathcal{O}}\left(\frac{L^{1/3}\sigma^{2/3}_{\text{rand}}D^{4/3}}{n^{1/3}K^{2/3}}\right)$ [a]	Random	Last	✓
	(Cha et al., 2023)	$\Omega\left(\frac{L^{1/3}\sigma^{2/3}_{\text{any}}D^{4/3}}{n^{1/3}K^{2/3}}\right)$	RR	Last	✗
L-smooth $f_i(\mathbf{x})$, μ-strongly convex $f(\mathbf{x})$	**Ours** (Theorem 4.6)	$\widetilde{\mathcal{O}}\left(\frac{L\sigma^2_{\text{any}}}{\mu^2K^2}\right)$	Any	Last	✓
	(Safran & Shamir, 2020)	$\Omega\left(\frac{G^2}{\mu K^2}\right)$ [b]	IG	Last	✗
	Ours (Theorem 4.6)	$\widetilde{\mathcal{O}}\left(\frac{L\sigma^2_{\text{rand}}}{\mu^2nK^2}\right)$ [a]	Random	Last	✓
	(Safran & Shamir, 2021; Cha et al., 2023)	$\Omega\left(\frac{L\sigma^2_{\text{any}}}{\mu^2nK^2}\right)$	RR/SO	Last	✗

[a] Note that when $\psi(\mathbf{x}) = 0$, there is $\sigma^2_{\text{rand}} = \sigma^2_{\text{any}} + n\|\nabla f(\mathbf{x}_*)\|^2 = \sigma^2_{\text{any}}$ due to $\nabla f(\mathbf{x}_*) = \mathbf{0}$.

[b] This lower bound is established under $L = \mu$ and additionally requires $\|\nabla f_i(\mathbf{x})\| \le G$. Under the same condition, our above upper bound $\widetilde{\mathcal{O}}\left(\frac{L\sigma^2_{\text{any}}}{\mu^2K^2}\right)$ will be $\widetilde{\mathcal{O}}\left(\frac{G^2}{\mu K^2}\right)$ and hence almost matches this lower bound.

The components $f_i(\mathbf{x})$ are smooth. This assumption also reduces the generality of previous studies. Though, theoretically speaking, smoothness is known to be necessary for shuffling gradient methods to obtain better convergence rates than SGD (Nagaraj et al., 2019), lots of problems are actually non-smooth in reality, for example, the LAD regression and SVMs (with the hinge loss) mentioned before. However, the convergence of shuffling-based algorithms without smoothness is not yet fully established.

In addition to the above problems, there exists another critical gap between practice and theory, i.e., the empirical success of the last iterate versus the lack of theoretical interpretation. Returning the last iterate is cheaper and hence commonly used in practice, but proving the last-iterate rate is instead difficult. Even for GD and SGD, the optimal last-iterate bounds had only been established recently (see the discussion in Section 2). However, as for shuffling gradient methods (regardless of which method in RR/SO/IG is used), no existing bounds can be directly applied to the last-iterate objective value (i.e., $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)$ where $\mathbf{x}_*$ is the optimal solution) and still match the lower bounds at the same time. Even for the strongly convex problem, the previous rates are studied for the metric $\|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2$ and thus cannot be converted to a bound on $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)$[2] whenever $F(\mathbf{x})$ is non-smooth (the case like the Lasso) or $\nabla f(\mathbf{x}_*) \neq \mathbf{0}$ happens in constrained optimization (i.e., $\psi(\mathbf{x}) = \mathrm{I}_{\mathbf{X}}(\mathbf{x})$). Motivated by this, we would like to ask the following core question considered in this paper:

For smooth/non-smooth $f_i(\mathbf{x})$ and general $\psi(\mathbf{x})$, does the last iterate of shuffling gradient methods provably converge in terms of the function value gap? If so, how fast is it?

1.1. Our Contributions

We provide an affirmative answer to the question by proving the first last-iterate convergence rates for shuffling gradient methods. We particularly focus on two cases, smooth $f_i(\mathbf{x})$ and Lipschitz $f_i(\mathbf{x})$, but without special requirements on $\psi(\mathbf{x})$.

Our new last-iterate rates for smooth components are summarized in Table 1 (see Table 3 in Appendix A for a more detailed summary due to limited space). They almost match the existing lower bounds or the fastest upper bounds of the

[2] Conversely, $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) = \Omega(\|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2)$ always holds under strong convexity due to the first-order optimality. Thus, the function value gap is always a stronger convergence criterion than the squared distance from the optimum.

Table 2. Summary of our new upper bounds and the previous fastest rates for G-Lipschitz $f_i(\mathbf{x})$ for large K. The lower bound in this case has not been proved as far as we know. Here, $D \triangleq \|\mathbf{x}_* - \mathbf{x}_1\|$. All rates use the function value gap as the convergence criterion. In the column of "Type", "Any" means the rate holds for whatever permutation not limited to RR/SO/IG. "Avg" and "Last" in the "Output" column stand for the average iterate and the last iterate, respectively. In the last column, "✓" means $\psi(\mathbf{x})$ can be taken arbitrarily and "✗" implies $\psi(\mathbf{x}) = 0$.

$F(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$ where $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ and $f_i(\mathbf{x})$ and $\psi(\mathbf{x})$ are convex					
Settings	References	Rate	Type	Output	$\psi(\mathbf{x})$
G-Lipschitz $f_i(\mathbf{x})$	(Bertsekas, 2011)	$\mathcal{O}\left(\frac{GD}{\sqrt{K}}\right)$	IG	Avg	✓[a]
	Ours (Theorem 4.7)	$\mathcal{O}\left(\frac{GD}{\sqrt{K}}\right)$	Any	Last	✓
G-Lipschitz $f_i(\mathbf{x})$, $f(\mathbf{x}) - f(\mathbf{x}_*) \geq \mu \|\mathbf{x} - \mathbf{x}_*\|^2$ [b]	(Nedić & Bertsekas, 2001)	$\mathcal{O}\left(\frac{G^2}{\mu K}\right)$ [b]	IG	Last[b]	✓[b]
G-Lipschitz $f_i(\mathbf{x})$, μ-strongly convex $\psi(\mathbf{x})$	**Ours** (Theorem 4.10)	$\widetilde{\mathcal{O}}\left(\frac{G^2}{\mu K}\right)$	Any	Last	✓

[a]Only for $\psi(\mathbf{x}) = \varphi(\mathbf{x}) + \mathrm{I}_{\mathbf{X}}(\mathbf{x})$ where $\varphi(\mathbf{x})$ is Lipschitz on $\mathbf{X}$.

[b]The original rate $\mathcal{O}\left(\frac{G^2}{\mu^2 K}\right)$ is only for $\psi(\mathbf{x}) = \mathrm{I}_{\mathbf{X}}(\mathbf{x})$ and proved w.r.t. $\|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2$. We first remark that this setting implies that $\mathbf{X}$ has to be compact due to $\mu \|\mathbf{x} - \mathbf{x}_*\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}_*) \leq G \|\mathbf{x} - \mathbf{x}_*\|$. Next, the original rate cannot give the $\mathcal{O}(1/K)$ bound on the function value gap since $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) \geq \mu \|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2$. The result reported here is only for the convenience of comparison.

average iterate.

Our new last-iterate rates for Lipschitz components are summarized in Table 2. They are as fast as the previous best-known upper bounds.

Limitations in our work are also discussed in Section 6.

2. Related Work

When stating the convergence rate, we only provide the dependence on n and K (for large K) without problem-dependent parameters to save space unless the discussion will be beneficial from specifying them in some cases. In the last-iterate bound of SGD, nK serves as time T.

Convergence of shuffling gradient methods. We first review the prior works for general smooth components since smoothness is known to be necessary for obtaining better rates compared to SGD (Nagaraj et al., 2019). The existing results in the non-smooth case will be discussed afterwards. As for the special quadratic functions (note that they can be covered by smooth components), we do not include a detailed discussion here due to limited space. The reader could refer to (Gurbuzbalaban et al., 2019; Gürbüzbalaban et al., 2021; Haochen & Sra, 2019; Rajput et al., 2020; Safran & Shamir, 2020; Ahn et al., 2020; Safran & Shamir, 2021; Rajput et al., 2022) for recent progress.

For general L-smooth components that are not necessarily quadratic, without strong convexity, (Nagaraj et al., 2019) first demonstrates the rate $\mathcal{O}(1/\sqrt{nK})$ for the averaged output of RR when $\psi(\mathbf{x}) = \mathrm{I}_{\mathbf{X}}(\mathbf{x})$. However, it does not only extra require bounded gradients but also needs a compact $\mathbf{X}$. Later, in the unconstrained setting (i.e., $\psi(\mathbf{x}) = 0$) without the Lipschitz condition, (Mishchenko et al., 2020; Nguyen et al., 2021) improve the rate to $\mathcal{O}(1/n^{1/3}K^{2/3})$ for RR/SO and $\mathcal{O}(1/K^{2/3})$ for IG but still only analyze the average iterate. Their rate for RR/SO has the optimal order on both n, K and other problem-dependent parameters as illustrated by the lower bound in (Cha et al., 2023).

When $f(\mathbf{x})$ is assumed to be μ-strongly convex, (Gürbüzbalaban et al., 2021; Haochen & Sra, 2019) justify that RR provably converges under the unusual condition of Lipschitz Hessian matrices. If assuming bounded gradients (which implicitly means a bounded domain) instead of Lipschitz Hessian matrices, (Shamir, 2016) proves the rate $\widetilde{\mathcal{O}}(1/n)$ for RR/SO when $K = 1$. For RR, (Nagaraj et al., 2019) and (Ahn et al., 2020)[3] respectively obtain the rates $\widetilde{\mathcal{O}}\left(\frac{L^2}{\mu^3 nK^2}\right)$ for the tail average iterate and $\mathcal{O}\left(\frac{L^3}{\mu^4 nK^2}\right)$ for the last iterate. (Mishchenko et al., 2020; Nguyen et al., 2021)[3] further shave off the Lipschitz condition and still show the last-iterate rate $\widetilde{\mathcal{O}}\left(\frac{L^2}{\mu^3 nK^2}\right)$ for RR/SO. However, none of them matches the lower bound $\Omega\left(\frac{L}{\mu^2 nK^2}\right)$ for RR/SO (Rajput et al., 2020; Safran & Shamir, 2021; Cha et al., 2023). Until recently, (Cha et al., 2023) demonstrates that the tail average iterate of RR indeed guarantees the rate $\widetilde{\mathcal{O}}\left(\frac{L}{\mu^2 nK^2}\right)$.

As for IG, the first asymptotic last-iterate bound $\mathcal{O}(1/K^2)$ is shown in (Gurbuzbalaban et al., 2019)[3] while it extra assumes bounded gradients. (Ying et al., 2018)[3] provides

[3]The rates in (Gurbuzbalaban et al., 2019; Ying et al., 2018; Ahn et al., 2020; Mishchenko et al., 2020; Nguyen et al., 2021) are all obtained under $\psi(\mathbf{x}) = 0$ and use $\|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2$ as the metric (except (Nguyen et al., 2021)). Thus, they cannot measure the convergence of the objective value in general. Nevertheless, we here write them as the last-iterate convergence for convenience.

the last-iterate rate $\widetilde{\mathcal{O}}\left(\frac{L^3}{\mu^4 K^2}\right)$ without the Lipschitz condition. (Mishchenko et al., 2020; Nguyen et al., 2021)[3] then improve it to $\widetilde{\mathcal{O}}\left(\frac{L^2}{\mu^3 K^2}\right)$. Though it seems like their rate has already matched the lower bound $\Omega\left(\frac{1}{\mu K^2}\right)$ established under $L = \mu$ (Safran & Shamir, 2020), however, this is not true as indicated by our new last-iterate bound $\widetilde{\mathcal{O}}\left(\frac{L}{\mu^2 K^2}\right)$ holding for any kind of permutation.

We would also like to mention some other relevant results. The lower bound $\Omega(1/n^2K^2)$ in (Cha et al., 2023) considers the best possible permutation in the algorithm. A near-optimal rate in this scenario is achieved by GraB (Lu et al., 2022), in which the permutation used in every epoch is chosen manually instead of RR/SO/IG. In addition, (Mishchenko et al., 2022) is the only work we know that can apply to a general $\psi(\mathbf{x})$. However, their results are all presented for the squared distance from the optimum, which cannot be used in our settings as explained in Section 1.

For non-smooth components, fewer results are known. The first convergence rate $\mathcal{O}(1/\sqrt{K})$ established for the averaged output of IG dates back to (Nedic & Bertsekas, 2001) where the authors consider constrained optimization (i.e., $\psi(\mathbf{x}) = \mathrm{I}_{\mathbf{X}}(\mathbf{x})$). (Bertsekas, 2011) obtains the same $\mathcal{O}(1/\sqrt{K})$ bound for IG with a general $\psi(\mathbf{x}) = \varphi(\mathbf{x}) + \mathrm{I}_{\mathbf{X}}(\mathbf{x})$ but where $\varphi(\mathbf{x})$ needs to be Lipschitz on $\mathbf{X}$. The reader could refer to (Bertsekas et al., 2011) for a detailed survey. For the special case $K = 1$, (Shamir, 2016) presents a rate $\mathcal{O}(1/\sqrt{n})$ for the averaged output of RR and SO, whose proof is based on online optimization. If strong convexity holds, the only general result that we are aware of is (Nedić & Bertsekas, 2001) showing the rate $\mathcal{O}(1/K)$ for $\|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2$ output by IG, which cannot be converted to a bound on the objective value due to $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) \geq \Omega(\mu \|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2)$.

Apart from the basic shuffling gradient method considered in this work, (Cai et al., 2023) studies the primal-dual interpretation of shuffled SGD and obtains fine-grained dependence on problem-dependent parameters. (Tran et al., 2022) incorporates Nesterov's momentum to improve the convergence rate. Besides the convex problem, shuffling-based methods also provably converge in non-convex optimization, for example, see (Solodov, 1998; Li et al., 2019; Tran et al., 2021; Pauwels, 2021; Mohtashami et al., 2022; Li et al., 2023). Particularly, (Yu & Li, 2023) provides the last-iterate analysis in the non-convex regime. Further generalization of shuffling gradient methods has also been done in different works, e.g., distributed and federated learning (Meng et al., 2019; Yun et al., 2022; Malinovsky & Richtárik, 2022; Sadiev et al., 2023; Huang et al., 2023) and minimax optimization (Das et al., 2022; Cho & Yun, 2023).

Last-iterate convergence of SGD. (Zhang, 2004) obtains the first finite-time bound, however, which is limited to the linear prediction problem. (Rakhlin et al., 2011) proves an $\mathcal{O}(1/nK)$ rate for Lipschitz strongly convex functions but additionally requires smoothness. The extra smooth condition is then removed by (Shamir & Zhang, 2013), showing $\widetilde{\mathcal{O}}(1/\sqrt{nK})$ and $\widetilde{\mathcal{O}}(1/nK)$ rates for Lipschitz convex and strongly convex problems, respectively. The first optimal high-probability bounds for Lipschitz (strongly) convex functions are later established in (Harvey et al., 2019; Jain et al., 2021). Whereas, all works mentioned so far are subject to compact domains. (Orabona, 2020) is the first to remove this restriction and still gets the $\widetilde{\mathcal{O}}(1/\sqrt{nK})$ rate for Lipschitz convex optimization. Recently, a new proof by (Zamani & Glineur, 2023) gives the tight rate $\mathcal{O}(1/\sqrt{nK})$ for GD without compact domains either. (Liu & Zhou, 2024) then extends the idea in (Zamani & Glineur, 2023) to stochastic optimization and shows a unified proof working for various settings. Results in (Liu & Zhou, 2024) can apply to smooth optimization without the Lipschitz condition and give the $\widetilde{\mathcal{O}}(1/\sqrt{nK})$ rate for convex objectives (improving upon the previous best rate $\mathcal{O}(1/\sqrt[3]{nK})$ for smooth stochastic optimization (Moulines & Bach, 2011)) and the $\widetilde{\mathcal{O}}(1/nK)$ bound for strongly convex functions.

Independent work. A manuscript (Cai & Diakonikolas, 2024) appearing on arXiv after the submission deadline of ICML 2024 also studied the last-iterate convergence of shuffling methods based on the proof technique initialized in (Zamani & Glineur, 2023) and later developed by our previous work (Liu & Zhou, 2024). Compared to (Cai & Diakonikolas, 2024), there are several differences we would like to discuss. We first remark that the problem itself and the assumptions are different. Precisely, the objective in (Cai & Diakonikolas, 2024) does not contain the regularizer (i.e., $\psi(\mathbf{x}) = 0$ in Assumption 3.2). In addition, every component $f_i(\mathbf{x})$ is assumed to have the same smooth/Lipschitz parameter (i.e., $L_i \equiv L$ in Assumption 3.3 or $G_i \equiv G$ in Assumption 3.5). Moreover, (Cai & Diakonikolas, 2024) does not consider the strongly convex case (i.e., $\mu_f = 0$ in Assumption 3.4). Consequently, when each $f_i(\mathbf{x})$ is smooth, their last-iterate convergence results for RR/SO/IG of the plain shuffling gradient method (i.e., taking $\psi(\mathbf{x}) = 0$ in Algorithm 1) can be covered by our Theorem 4.4; when each $f_i(\mathbf{x})$ is Lipschitz, (Cai & Diakonikolas, 2024) does not provide a last-iterate convergence result for the plain shuffling gradient method. Hence, our Theorems 4.6, 4.7, 4.9, and 4.10 are still independently interesting.

It is noteworthy that (Cai & Diakonikolas, 2024) obtains the convergence rate for the weighted average iterate of the plain shuffling gradient method for smooth $f_i(\mathbf{x})$, which is not considered in our paper. They also develop the last-iterate convergence for the incremental proximal method, which is a different algorithm from our proximal shuffling gradient method (Algorithm 1).

3. Preliminaries

Notation. $\mathbb{N}$ denotes the set of natural numbers (excluding 0). $[K] \triangleq \{1, \cdots, K\}$ for any $K \in \mathbb{N}$. $a \vee b$ and $a \wedge b$ represent $\max\{a, b\}$ and $\min\{a, b\}$. $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are floor and ceiling functions. $\langle \cdot, \cdot \rangle$ stands for the Euclidean inner product on $\mathbb{R}^d$. $\|\cdot\| \triangleq \sqrt{\langle \cdot, \cdot \rangle}$ is the standard ℓ_2 norm. Given an extended real-valued convex function $g(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the domain of $g(\mathbf{x})$ is defined as $\mathrm{dom} g \triangleq \left\{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) < +\infty\right\}$. $\partial g(\mathbf{x})$ is the set of subgradients at $\mathbf{x}$. $\nabla g(\mathbf{x})$ denotes an element in $\partial g(\mathbf{x})$ if $\partial g(\mathbf{x}) \neq \varnothing$. The Bregman divergence induced by $g(\mathbf{x})$ is $B_g(\mathbf{x}, \mathbf{y}) \triangleq g(\mathbf{x}) - g(\mathbf{y}) - \langle \nabla g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ satisfying $\partial g(\mathbf{y}) \neq \varnothing$. $\mathrm{I}_{\mathbf{X}}(\mathbf{x})$ is the indicator function of the set $\mathbf{X}$, i.e., $\mathrm{I}_{\mathbf{X}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathbf{X}$, otherwise, $\mathrm{I}_{\mathbf{X}}(\mathbf{x}) = +\infty$.

We study the following optimization problem in this paper

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) \triangleq f(\mathbf{x}) + \psi(\mathbf{x}) \text{ where } f(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}).$$

Next, we list the assumptions used in the analysis.

Assumption 3.1. $\exists \mathbf{x}_* \in \mathbb{R}^d$ such that $F(\mathbf{x}_*) = \inf_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) \in \mathbb{R}$.

Assumption 3.2. Each $f_i(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is convex. $\psi(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper, closed and convex[4]. Moreover, $\exists \mu_\psi \geq 0$ such that $B_\psi(\mathbf{x}, \mathbf{y}) \geq \frac{\mu_\psi}{2} \|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ satisfying $\partial \psi(\mathbf{y}) \neq \varnothing$.

Assumptions 3.1 and 3.2 are standard in the optimization literature. Note that $\partial f_i(\mathbf{x}) \neq \varnothing$ for all $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$ since $f_i(\mathbf{x})$ is convex on $\mathbb{R}^d$, so we use $\nabla f_i(\mathbf{x})$ to denote an element in $\partial f_i(\mathbf{x})$ throughout the remaining paper. We allow $\psi(\mathbf{x})$ to take the value of $+\infty$ and hence include constrained optimization, e.g., setting $\psi(\mathbf{x}) = \varphi(\mathbf{x}) + \mathrm{I}_{\mathbf{X}}(\mathbf{x})$ where $\varphi(\mathbf{x})$ can be some other regularizer. Moreover, the parameter μ_ψ is possibly to be zero to better fit different tasks, e.g., $\psi(\mathbf{x}) = \|\mathbf{x}\|_1$ in the Lasso implies $\mu_\psi = 0$.

Assumption 3.3. $\exists L_i > 0$ such that $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L_i \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, i \in [n]$.

Assumption 3.4. $\exists \mu_f \geq 0$ such that $B_f(\mathbf{x}, \mathbf{y}) \geq \frac{\mu_f}{2} \|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Assumption 3.5. $\exists G_i > 0$ such that $\|\nabla f_i(\mathbf{x})\| \leq G_i$, $\forall \mathbf{x} \in \mathbb{R}^d, i \in [n]$.

Assumptions 3.3, 3.4 and 3.5 are commonly used in the related field. Note that Assumptions 3.4 (when $\mu_f > 0$) and 3.5 cannot hold together on $\mathbb{R}^d$ (e.g., see (Nguyen et al., 2018)). Hence, they will not be assumed to be true simultaneously. In addition, our analysis relies on the following well-known fact for smooth convex functions, whose proof is omitted and can refer to, for example, Theorem 2.1.5 in (Nesterov et al., 2018).

Lemma 3.6. *Given a convex and differentiable function $g(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ satisfying $\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ for some $L > 0$, then $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$\frac{\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\|^2}{2L} \leq B_g(\mathbf{x}, \mathbf{y}) \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 .$$

[4]Recall that this means the epigraph of $\psi(\mathbf{x})$, i.e., $\mathrm{epi}\psi \triangleq \left\{(\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R} : \psi(\mathbf{x}) \leq t\right\}$, is a nonempty closed convex set.

4. Proximal Shuffling Gradient Method

Algorithm 1 Proximal Shuffling Gradient Method

Input: initial point $\mathbf{x}_1 \in \mathrm{dom}\psi$, stepsize $\eta_k > 0$.
for $k = 1$ **to** K **do**
  Generate a permutation $\left\{\sigma_k^i : i \in [n]\right\}$ of $[n]$
  $\mathbf{x}_k^1 = \mathbf{x}_k$
  for $i = 1$ **to** n **do**
    $\mathbf{x}_k^{i+1} = \mathbf{x}_k^i - \eta_k \nabla f_{\sigma_k^i}(\mathbf{x}_k^i)$
  $\mathbf{x}_{k+1} = \mathrm{argmin}_{\mathbf{x} \in \mathbb{R}^d} n\psi(\mathbf{x}) + \frac{\|\mathbf{x} - \mathbf{x}_k^{n+1}\|^2}{2\eta_k}$
Return $\mathbf{x}_{K+1}$

The algorithmic framework of the proximal shuffling gradient method is shown in Algorithm 1 where the number of epochs K is assumed to satisfy $K \geq 2$ to avoid algebraic issues in the proof. Unlike (Mishchenko et al., 2022), we do not specify how the permutation σ_k is generated in the k-th epoch. Hence, Algorithm 1 actually includes several different algorithms depending on how σ_k is defined.

Example 4.1. Algorithm 1 recovers the ProxRR algorithm studied by (Mishchenko et al., 2022) if σ_k is uniformly sampled without replacement in every epoch.

Example 4.2. Algorithm 1 reduces to the ProxSO algorithm introduced by (Mishchenko et al., 2022) when $\sigma_k = \sigma_0$ for all epochs where σ_0 is a pre-generated permutation by uniform sampling without replacement.

Example 4.3. Algorithm 1 includes the proximal IG method developed in (Kibardin, 1979) as a subcase by considering $\sigma_k = \sigma_0$ for all epochs where σ_0 is a deterministic pre-generated permutation (e.g., $\sigma_0^i = i, \forall i \in [n]$).

It is worth noting that $\mathbf{x}_{k+1} \in \mathrm{dom}\psi$ holds for all $k \in [K]$ since we assume $\psi(\mathbf{x})$ is proper. Therefore, all $F(\mathbf{x}_k)$ take values on $\mathbb{R}$ and hence are well-defined.

4.1. Last-Iterate Convergence Rates

In the convergence rate, we denote the initial distance from the optimum by $D \triangleq \|\mathbf{x}_* - \mathbf{x}_1\|$ and the optimal function value by $F_* \triangleq F(\mathbf{x}_*)$. The full statements of all theorems presented in this subsection along with their proofs are deferred into Appendix C. The hidden log factor in the tilde $\widetilde{\mathcal{O}}(\cdot)$ will also be explicitly given in the full theorem.

4.1.1. SMOOTH FUNCTIONS

This subsection focuses on smooth components $f_i(\mathbf{x})$. To simplify expressions, we define the following symbols. $\bar{L} \triangleq \frac{1}{n}\sum_{i=1}^{n} L_i$ and $L^* \triangleq \max_{i\in[n]} L_i$ respectively stand for the averaged and the largest smooth parameter. In addition, similar to (Mishchenko et al., 2020; Nguyen et al., 2021), we use the quantity $\sigma_{\text{any}}^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_*)\|^2$ to measure the uncertainty induced by shuffling. Besides, we introduce another term $\sigma_{\text{rand}}^2 \triangleq \sigma_{\text{any}}^2 + n\|\nabla f(\mathbf{x}_*)\|^2$, where the appearance of the extra $n\|\nabla f(\mathbf{x}_*)\|^2$ is due to the proximal step. We remark that the term $n\|\nabla f(\mathbf{x}_*)\|^2$ also showed up when studying ProxRR and ProxSO before in (Mishchenko et al., 2022).

Theorem 4.4. *Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$) and 3.3:*

Regardless of how the permutation σ_k is generated in every epoch, taking the stepsize $\eta_k = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}} \wedge \frac{D^{2/3}}{n\sqrt[3]{\bar{L}\sigma_{\text{any}}^2 K(1+\log K)}}, \forall k \in [K]$, Algorithm 1 guarantees

$$F(\mathbf{x}_{K+1}) - F_* \le \widetilde{\mathcal{O}}\left(\frac{\sqrt{\bar{L}L^*}D^2}{K} + \frac{\bar{L}^{1/3}\sigma_{\text{any}}^{2/3}D^{4/3}}{K^{2/3}}\right).$$

Suppose the permutation σ_k is uniformly sampled without replacement, taking the stepsize $\eta_k = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}} \wedge \frac{D^{2/3}}{\sqrt[3]{n^2\bar{L}\sigma_{\text{rand}}^2 K(1+\log K)}}, \forall k \in [K]$, Algorithm 1 guarantees

$$\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F_*\right] \le \widetilde{\mathcal{O}}\left(\frac{\sqrt{\bar{L}L^*}D^2}{K} + \frac{\bar{L}^{1/3}\sigma_{\text{rand}}^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right).$$

Remark 4.5. When we say the permutation σ_k is uniformly sampled without replacement hereinafter, it is not necessary to think that σ_k is re-sampled in every epoch as in Example 4.1. Instead, it can also be sampled in advance like Example 4.2. Or even more generally, one can sample a permutation in advance and use it for an arbitrary number of epochs then re-sample a new permutation and repeat this procedure. Hence, our results are not only true for ProxRR and ProxSO but also hold for the general form of Algorithm 1.

As far as we know, Theorem 4.4 is the first to provide the last-iterate convergence rate of Algorithm 1 for smooth components but without strong convexity. Even in the simplest case of $\psi(\mathbf{x}) = 0$, the previous best-known result only works for the averaged output $\mathbf{x}_{K+1}^{\text{avg}} \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{k+1}$ (Mishchenko et al., 2020; Nguyen et al., 2021).

Next, we would like to discuss the convergence rate in more detail. For the first case (i.e., regardless of how the permutation σ_k is generated in every epoch), up to a logarithmic factor, our last-iterate result matches the fastest existing rate of $\mathbf{x}_{K+1}^{\text{avg}}$ output by the IG method when $\psi(\mathbf{x}) = 0$ (Mishchenko et al., 2020). But as one can see, our theorem is more general and can apply to various situations (e.g., $\psi(\mathbf{x}) = \|\mathbf{x}\|_1$) for whatever the permutation σ_k is used.

When the permutation σ_k is uniformly sampled without replacement (including RR and SO as special cases), let us first consider the unconstrained setting without any regularizer, i.e., $\psi(\mathbf{x}) = 0$. Note that now $\nabla f(\mathbf{x}_*) = \mathbf{0}$ and it implies $\sigma_{\text{rand}}^2 = \sigma_{\text{any}}^2 + n\|\nabla f(\mathbf{x}_*)\|^2 = \sigma_{\text{any}}^2$. Hence, our rate matches the last-iterate lower bound $\Omega\left(\frac{L^{1/3}\sigma_{\text{any}}^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$ established under the condition $L_i \equiv L$ and $\psi(\mathbf{x}) = 0$ for large K (which is only proved for RR) in (Cha et al., 2023) up to a logarithmic term. When a general $\psi(\mathbf{x})$ exists, the only difference is to replace σ_{any}^2 with the larger quantity σ_{rand}^2. A similar penalty also appeared in (Mishchenko et al., 2022) when studying ProxRR and ProxSO. Whether σ_{rand}^2 can be improved to σ_{any}^2 remains unclear to us.

Moreover, we would like to talk about the stepsize. Suppose $L_i \equiv L$ for simplicity, then our stepsizes for both cases are almost the same as the choices in (Mishchenko et al., 2020) and only different by $1 + \log K$ in the denominator. However, this distinction plays a key role in our analysis of proving the last-iterate bound. The reader could refer to Theorem C.1 and its proof for why we need it.

Theorem 4.6. *Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$), 3.3 and 3.4 (with $\mu_f > 0$), let $\bar{\kappa}_f \triangleq \frac{\bar{L}}{\mu_f}$ and $\kappa_f^* \triangleq \frac{L^*}{\mu_f}$:*

Regardless of how the permutation σ_k is generated in every epoch, taking the stepsize $\eta_k = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}} \wedge \frac{u}{n\mu_f K}$ where $u = 1 \vee \log\frac{\mu_f^3 D^2 K^2}{\bar{L}\sigma_{\text{any}}^2(1+\log K)}, \forall k \in [K]$, Algorithm 1 guarantees

$$F(\mathbf{x}_{K+1}) - F_* \le \widetilde{\mathcal{O}}\left(\frac{\sqrt{\bar{L}L^*}D^2}{Ke^{\frac{K}{\sqrt{\bar{\kappa}_f\kappa_f^*}}}} + \frac{\bar{L}\sigma_{\text{any}}^2}{\mu_f^2 K^2}\right).$$

Suppose the permutation σ_k is uniformly sampled without replacement, taking the stepsize $\eta_k = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}} \wedge \frac{u}{n\mu_f K}$ where $u = 1 \vee \log\frac{n\mu_f^3 D^2 K^2}{\bar{L}\sigma_{\text{rand}}^2(1+\log K)}, \forall k \in [K]$, Algorithm 1 guarantees

$$\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F_*\right] \le \widetilde{\mathcal{O}}\left(\frac{\sqrt{\bar{L}L^*}D^2}{Ke^{\frac{K}{\sqrt{\bar{\kappa}_f\kappa_f^*}}}} + \frac{\bar{L}\sigma_{\text{rand}}^2}{\mu_f^2 nK^2}\right).$$

Another result in this subsection is Theorem 4.6 showing the last-iterate convergence guarantee when strong convexity is additionally assumed. For simplicity, we only present the case of $\mu_\psi = 0$ and $\mu_f > 0$ here. We refer the reader to Theorem C.2 in the appendix for the full statement. To our best knowledge, this is also the first last-iterate convergence result w.r.t. the function value gap for shuffling gradient methods under strong convexity when a general $\psi(\mathbf{x})$ is allowed. In fact, even for the special case $\psi(\mathbf{x}) = \mathrm{I}_{\mathbf{X}}(\mathbf{x})$, i.e., constrained optimization, the previous bounds could fail as explained in Section 1.

As before, let us first take a look at the case of whatever the permutation σ_k is used. If we specify to the IG method, (Safran & Shamir, 2020) provided a last-iterate lower bound $\Omega\left(\frac{G^2}{\mu_f K^2}\right)$ for large K with $\psi(\mathbf{x}) = 0$, $L_i \equiv \mu_f$ and an additional requirement $\|\nabla f_i(\mathbf{x})\| \leq G$. Under the same condition, our rate degenerates to $\widetilde{\mathcal{O}}\left(\frac{G^2}{\mu_f K^2}\right)$ almost matching the lower bound in (Safran & Shamir, 2020) when K is large. However, for the arbitrary permutation case, the last-iterate lower bound under the condition $\psi(\mathbf{x}) = 0$ and $L_i \equiv L$ for large K is $\Omega\left(\frac{L\sigma^2_{\text{any}}}{\mu_f^2 n^2 K^2}\right)$ (Cha et al., 2023), which is better than our rate by a factor $\mathcal{O}\left(\frac{1}{n^2}\right)$. We remark that this is to be expected because, roughly speaking, our upper bound is equivalent to say (for $L_i \equiv L$ and large K)

$$\inf_{\text{stepsizes } \eta_k} \sup_{\substack{\text{functions } F(\mathbf{x}) \\ \text{permutations } \sigma_k}} F(\mathbf{x}_{K+1}) - F_* \leq \widetilde{\mathcal{O}}\left(\frac{L\sigma^2_{\text{any}}}{\mu_f^2 K^2}\right),$$

whereras the lower bound in (Cha et al., 2023) is proved for

$$\sup_{\text{functions } F(\mathbf{x})} \inf_{\substack{\text{stepsizes } \eta_k \\ \text{permutations } \sigma_k}} F(\mathbf{x}_{K+1}) - F_* \geq \Omega\left(\frac{L\sigma^2_{\text{any}}}{\mu_f^2 n^2 K^2}\right).$$

Hence, to achieve the lower bound in (Cha et al., 2023), one does not only need to specify stepsizes η_k but also has to choose the permutation σ_k according to the problem. Recently, (Lu et al., 2022) proposed GraB, in which the permutation for the current epoch is selected based on the information from previous epochs. GraB is provably faster than RR/SO/IG and achieves the rate $\widetilde{\mathcal{O}}\left(\frac{H^2L^2\sigma^2_{\text{any}}}{\mu_f^3 n^2 K^2}\right)$ differing from the lower bound in (Cha et al., 2023) by a factor of $\mathcal{O}\left(\frac{H^2 L}{\mu_f}\right)$ (the constant H here is named herding bound, see (Lu et al., 2022) for details).

Next, for the case of uniform sampling without replacement (including RR and SO as special cases), we again first check the rate when $\psi(\mathbf{x}) = 0$. Recall now $\sigma^2_{\text{rand}} = \sigma^2_{\text{any}}$ due to $\|\nabla f(\mathbf{x}_*)\| = 0$. Thus, our bound for large K is $\widetilde{\mathcal{O}}\left(\frac{\bar{L}\sigma^2_{\text{any}}}{\mu_f^2 n K^2}\right)$, which is nearly optimal compared with the lower bound $\Omega\left(\frac{L\sigma^2_{\text{any}}}{\mu_f^2 n K^2}\right)$ established for RR/SO under $L_i \equiv L$ (Safran & Shamir, 2021; Cha et al., 2023). As far as we are aware, this is the first last-iterate result not only being nearly tight in n and K but also optimal in the parameter $\Omega\left(\frac{L}{\mu_f^2}\right)$. In contrast, the previous best bound with the correct dependence only applies to the tail average iterate $\mathbf{x}^{\text{tail}}_{K+1} \triangleq \frac{1}{\lfloor \frac{K}{2} \rfloor + 1} \sum_{k=\lceil \frac{K}{2} \rceil}^{K} \mathbf{x}_{k+1}$ (Cha et al., 2023). Therefore, Theorem 4.6 fulfills the existing gap when $\psi(\mathbf{x}) = 0$ for large K. More importantly, our result can also be used for any general $\psi(\mathbf{x})$ with the only tradeoff of changing σ^2_{any} to σ^2_{rand} but not hurting any other term.

Lastly, we would like to talk about the stepsize used in Theorem 4.6. As one can check, when $L_i \equiv L$, it is similar to the stepsize used in (Cha et al., 2023) to obtain the rate for $\mathbf{x}^{\text{tail}}_{K+1}$. The only significant difference is still the extra term $1 + \log K$. However, as mentioned in the discussion about the stepsize used in Theorem 4.4 (i.e., the general convex case), this change is crucial in our analysis.

4.1.2. Lipschitz Functions

In this subsection, we consider the case of $f_i(\mathbf{x})$ all being Lipschitz. We employ the notation $\bar{G} \triangleq \frac{1}{n}\sum_{i=1}^{n} G_i$ to denote the averaged Lipschitz parameter.

Theorem 4.7. *Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$) and 3.5, regardless of how the permutation σ_k is generated in every epoch:*

Taking the stepsize $\eta_k = \frac{\eta}{\sqrt{k}}, \forall k \in [K]$ with $\eta = \frac{D}{n\bar{G}}$ or $\eta_k = \frac{\eta}{\sqrt{K}}, \forall k \in [K]$ with $\eta = \frac{D}{n\bar{G}\sqrt{1+\log K}}$, Algorithm 1 guarantees

$$F(\mathbf{x}_{K+1}) - F_* \leq \widetilde{\mathcal{O}}\left(\frac{\bar{G}D}{\sqrt{K}}\right).$$

Taking the stepsize $\eta_k = \eta\frac{K-k+1}{K^{3/2}}, \forall k \in [K]$ with $\eta = \frac{D}{n\bar{G}}$, Algorithm 1 guarantees

$$F(\mathbf{x}_{K+1}) - F_* \leq \mathcal{O}\left(\frac{\bar{G}D}{\sqrt{K}}\right).$$

Remark 4.8. The values of η in Theorem 4.7 are set to obtain the best dependence on parameters like $\bar{G}$ and D. In fact, Algorithm 1 converges for all three stepsizes with arbitrary $\eta > 0$ but suffers a worse dependence on parameters. We refer the reader to Theorem C.3 for the full statement.

We first focus on the general convex case. As shown in Theorem 4.7, regardless of how the permutation σ_k is generated, the last-iterate convergence rate is $\widetilde{\mathcal{O}}(\bar{G}D/\sqrt{K})$ for stepsizes $\eta_k = \frac{\eta}{\sqrt{k}}$ or $\eta_k = \frac{\eta}{\sqrt{K}}$ with a theoretically fine-tuned η. The extra log factor can be removed by taking the stepsize $\eta_k = \eta\frac{K-k+1}{K^{3/2}}$ that has a linear decay rate. This kind of stepsize was originally introduced by (Zamani & Glineur, 2023) to show the $\mathcal{O}(\bar{G}D/\sqrt{K})$ last-iterate rate can be achieved for the projected subgradient descent method, i.e., $n = 1$ and $\psi(\mathbf{x}) = \mathrm{I}_{\mathbf{X}}(\mathbf{x})$. Here, we prove it can also be applied to the proximal shuffling gradient method.

Theorem 4.7 provides the first concrete theoretical evidence that the last iterate of Algorithm 1 for any permutation is comparable with the averaged output by the proximal IG method, which was known to be able to achieve the rate $\mathcal{O}(GD/\sqrt{K})$ when $G_i \equiv G$ (Bertsekas, 2011).

Moreover, we compare the number of individual gradient evaluations required to make the function value gap of the last iterate be at most ϵ between Algorithm 1, GD and SGD.

- Algorithm 1: $\mathcal{O}\left(\frac{(\sum_{i=1}^n G_i)^2 D^2}{n\epsilon^2}\right)$ by Theorem 4.7.
- GD: $\mathcal{O}\left(\frac{(\sum_{i=1}^n G_i)^2 D^2}{n\epsilon^2}\right)$ by prior works in Section 2.
- SGD: $\mathcal{O}\left(\frac{(\sum_{i=1}^n G_i^2) D^2}{n\epsilon^2}\right)$ by prior works in Section 2.

This indicates that the sample complexity of Algorithm 1 matches GD's but is worse than SGD's by at most a factor of $\mathcal{O}(n)$ due to $\sum_{i=1}^n G_i^2 \le (\sum_{i=1}^n G_i)^2 \le n\sum_{i=1}^n G_i^2$.

Theorem 4.9. *Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$) and 3.5, regardless of how the permutation σ_k is generated in every epoch, taking the stepsize $\eta_k = \frac{r_k}{4\sqrt{3}n\bar{G}\sqrt{k}(1+\log k)}, \forall k \in \mathbb{N}$ where $r_k = r \vee \max_{\ell\in[k]} \|\mathbf{x}_\ell - \mathbf{x}_1\|$ for some $r > 0$, then Algorithm 1 guarantees for large enough K*

$$F(\mathbf{x}_{K+1}) - F_* \le \widetilde{\mathcal{O}}\left(\frac{\bar{G}(D \vee r)}{\sqrt{K}}\right).$$

In Theorem 4.7, one may have noticed that it needs the prior knowledge of $\bar{G}$ and D to obtain the optimal linear dependence on them. In practice, estimating $\bar{G}$ could be relatively easy. In contrast, the requirement of knowing D is however hard to satisfy. Fortunately, several methods have been proposed to overcome this difficulty. Here, we borrow the idea from (Ivgi et al., 2023) that is to introduce the term $r_k = r \vee \max_{\ell\in[k]} \|\mathbf{x}_\ell - \mathbf{x}_1\|$ in the stepsize as shown in Theorem 4.9. As a result, we can provide an asymptotic last-iterate convergence rate with a linear dependence on D without knowing it. However, we are not able to establish a finite-time rate that also has a linear dependence on D, which is left as a future research direction.

Theorem 4.10. *Under Assumptions 3.1, 3.2 (with $\mu_\psi > 0$) and 3.5, regardless of how the permutation σ_k is generated in every epoch, taking the stepsize $\eta_k = \frac{2}{n\mu_\psi k}, \forall k \in [K]$, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1}) - F_* \le \widetilde{\mathcal{O}}\left(\frac{\mu_\psi D^2}{K^2} + \frac{\bar{G}^2}{\mu_\psi K}\right).$$

Lastly, we consider the case of strongly convex $\psi(\mathbf{x})$. As stated in Theorem 4.10, Algorithm 1 guarantees the rate of $\widetilde{\mathcal{O}}(1/K)$ w.r.t. the function value gap for the stepsize $\eta_k = \frac{2}{n\mu_\psi k}$. In contrast, the same rate was previously only known to hold for $\|\mathbf{x}_{K+1} - \mathbf{x}^*\|^2$ in IG (Nedić & Bertsekas, 2001). Theorem C.5 in the appendix will generalize to $\eta_k = \frac{m}{n\mu_\psi k}$ for any $m \in \mathbb{N}$ and prove the rate $\widetilde{O}(m/K)$.

By a similar comparison (after Theorem 4.7), the sample complexity of Algorithm 1 in the case $\mu_\psi > 0$ is as good as GD's but still worse than SGD's by at most an $\mathcal{O}(n)$ term.

5. Proof Idea

In this section, we outline some key steps in the analysis of Theorem 4.4 and highlight our novel techniques.

From a high-level view, our proof is inspired by the recent progress on the last-iterate convergence of GD (Zamani & Glineur, 2023). (Zamani & Glineur, 2023) designs an auxiliary sequence $\mathbf{z}_k$ for $k \in [K]$, each of which is a convex combination of $\mathbf{x}_*, \mathbf{x}_1, \cdots, \mathbf{x}_k$ (say $\mathbf{z}_k = w_{k,0}\mathbf{x}_* + \sum_{\ell=1}^k w_{k,\ell}\mathbf{x}_\ell$ where $w_{k,\ell} \ge 0$ and $\sum_{\ell=0}^k w_{k,\ell} = 1$), and then bounds $F(\mathbf{x}_{k+1}) - F(\mathbf{z}_k)$ instead of $F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*)$. By using $F(\mathbf{z}_k) \le w_{k,0}F(\mathbf{x}_*) + \sum_{\ell=1}^k w_{k,\ell}\mathbf{x}_\ell$, one can finally prove $\sum_{k=1}^K p_k(F(\mathbf{x}_{k+1}) - F(\mathbf{z}_k)) \ge \Omega(F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*))$ for properly picked $w_{k,\ell}$ where p_k is another carefully chosen sequence and obtain the rate for the last iterate.

However, to make the above argument work, it indeed requires $F(\mathbf{x}_{k+1}) - F(\mathbf{z}_k) \le$ good terms, where "good terms" means some quantities we can finally bound (e.g., a telescoping sum). But in shuffling gradient methods, a key difference appears: there is only (see Lemmas D.1 and D.2)

$$F(\mathbf{x}_{k+1}) - F(\mathbf{z}_k) \le \text{good terms} + \mathcal{O}(B_f(\mathbf{z}_k, \mathbf{x}_*) + R_k),$$

where R_k is a "bad" residual due to shuffling. Fortunately, R_k can always be bounded (Lemma E.1) and thus can be included in "good terms". In contrast, bounding another new extra term $B_f(\mathbf{z}_k, \mathbf{x}_*)$ is more tricky, which is significantly distinct from the previous works. Though the existence of $B_f(\mathbf{z}_k, \mathbf{x}_*)$, we still sum from $k = 1$ to K to obtain

$$F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) \le \text{good terms} + \mathcal{O}\left(\sum_{k=1}^K B_f(\mathbf{z}_k, \mathbf{x}_*)\right). \tag{1}$$

The first key observation in the analysis is that $B_f(\mathbf{z}_k, \mathbf{x}_*) = \mathcal{O}(\sum_{\ell=1}^k B_f(\mathbf{x}_\ell, \mathbf{x}_*))$ because $\mathbf{z}_k$ is a convex combination of $\mathbf{x}_*, \mathbf{x}_1, \cdots, \mathbf{x}_k$ and $B_f(\cdot, \cdot)$ is convex in the first argument. The second important notice is that $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) \ge B_f(\mathbf{x}_{K+1}, \mathbf{x}_*)$ always holds for whatever $\psi(\mathbf{x})$ is (see the proof of Theorem C.1 for details). Therefore, we have

$$B_f(\mathbf{x}_{K+1}, \mathbf{x}_*) \le \text{good terms} + \mathcal{O}\left(\sum_{k=1}^K B_f(\mathbf{x}_k, \mathbf{x}_*)\right).$$

But this is still not enough to bound $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)$. To deal with this issue, departing from the previous works that only prove one inequality for time K, we instead apply the above procedure for every $k \in [K]$ to get

$$B_f(\mathbf{x}_{k+1}, \mathbf{x}_*) \leq \text{good terms} + \mathcal{O}\left(\sum_{\ell=1}^{k} B_f(\mathbf{x}_\ell, \mathbf{x}_*)\right).$$

We remark that this new way requires us to define the auxiliary sequence $\{\mathbf{z}_\ell, \forall \ell \in [k]\}$ carefully. Specifically, for each $k \in [K]$, we let the auxiliary sequence $\{\mathbf{z}_\ell, \forall \ell \in [k]\}$ depend on the current time k. Hence, we indeed construct a total of K different auxiliary sequences. Next, we develop a new algebraic inequality (Lemma E.2) to recursively bound all $B_f(\mathbf{x}_k, \mathbf{x}_*)$. Equipped with the bound on $B_f(\mathbf{x}_k, \mathbf{x}_*)$, we finally invoke (1) again to get the desired result. The reader could refer to the appendix for detailed proof.

6. Limitation

Here we list some limitations in our work and look forward to them being addressed in the future. For smooth optimization, the extra factor $1 + \log K$ in the stepsizes seems necessary in our analysis. Whether, and if so how, it can be removed is an important problem. In addition, our results are limited to constant stepsizes depending on the number of epochs K. Finding time-varying stepsizes that still achieve the optimal dependence on both n, K and other problem-dependent parameters is another interesting task. For non-smooth optimization, i.e., Lipschitz components, our theorems are stated for any kind of permutation and can only match the sample complexity of GD but be worse than SGD's. Hence, it is worth investigating whether shuffling gradient methods can benefit from random permutations and be as good as SGD under certain types of shuffling, for example, RR and SO.

Acknowledgements

This work is supported by the NSF grant CCF-2106508. Zhengyuan Zhou also gratefully acknowledges the 2024 NYU CGEB faculty grant.

Impact Statement

This paper presents a theory work showing the last-iterate convergence of shuffling gradient methods. There are no ethical impacts and expected societal implications that we feel must be specifically highlighted here.

References

Ahn, K., Yun, C., and Sra, S. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33:17526–17535, 2020.

Bengio, Y. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade: Second Edition*, pp. 437–478. Springer, 2012.

Bertsekas, D. P. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

Bertsekas, D. P. et al. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

Bottou, L. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pp. 2624–2633. Citeseer, 2009.

Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pp. 421–436. Springer, 2012.

Cai, X. and Diakonikolas, J. Last iterate convergence of incremental methods and applications in continual learning. *arXiv preprint arXiv:2403.06873*, 2024.

Cai, X., Lin, C. Y., and Diakonikolas, J. Empirical risk minimization with shuffled sgd: A primal-dual perspective and improved bounds. *arXiv preprint arXiv:2306.12498*, 2023.

Cha, J., Lee, J., and Yun, C. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3855–3912. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cha23a.html.

Cho, H. and Yun, C. SGDA with shuffling: faster convergence for nonconvex-pł minimax optimization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6xXtM8bFFJ.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Das, A., Schölkopf, B., and Muehlebach, M. Sampling without replacement leads to faster rates in finite-sum minimax optimization. *Advances in Neural Information Processing Systems*, 35:6749–6762, 2022.

Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P. A. Convergence rate of incremental gradient and incremental newton methods. *SIAM Journal on Optimization*, 29(4): 2542–2565, 2019.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. A. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186:49–84, 2021.

Haochen, J. and Sra, S. Random shuffling beats SGD after finite epochs. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2624–2633. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/haochen19a.html.

Harvey, N. J., Liaw, C., Plan, Y., and Randhawa, S. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

Huang, K., Li, X., Milzarek, A., Pu, S., and Qiu, J. Distributed random reshuffling over networks. *IEEE Transactions on Signal Processing*, 71:1143–1158, 2023.

Ivgi, M., Hinder, O., and Carmon, Y. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14465–14499. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ivgi23a.html.

Jain, P., Nagaraj, D. M., and Netrapalli, P. Making the last iterate of sgd information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130, 2021. doi: 10.1137/19M128908X. URL https://doi.org/10.1137/19M128908X.

Kibardin, V. Decomposition into functions in the minimization problem. *Automation and Remote Control*, 1979, 01 1979.

Li, X., Zhu, Z., So, A. M.-C., and Lee, J. D. Incremental methods for weakly convex optimization. *arXiv preprint arXiv:1907.11687*, 2019.

Li, X., Milzarek, A., and Qiu, J. Convergence of random reshuffling under the kurdyka–łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023.

Liu, Z. and Zhou, Z. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xxaEhwC1I4.

Lu, Y., Guo, W., and De Sa, C. M. Grab: Finding provably better data permutations than random reshuffling. *Advances in Neural Information Processing Systems*, 35: 8969–8981, 2022.

Malinovsky, G. and Richtárik, P. Federated random reshuffling with compression and variance reduction. 2022.

Meng, Q., Chen, W., Wang, Y., Ma, Z.-M., and Liu, T.-Y. Convergence analysis of distributed stochastic gradient descent with shuffling. *Neurocomputing*, 337:46–57, 2019.

Mishchenko, K., Khaled, A., and Richtárik, P. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33: 17309–17320, 2020.

Mishchenko, K., Khaled, A., and Richtárik, P. Proximal and federated random reshuffling. In *International Conference on Machine Learning*, pp. 15718–15749. PMLR, 2022.

Mohtashami, A., Stich, S., and Jaggi, M. Characterizing & finding good data orderings for fast convergence of sequential gradient methods. *arXiv preprint arXiv:2202.01838*, 2022.

Moulines, E. and Bach, F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Nagaraj, D., Jain, P., and Netrapalli, P. SGD without replacement: Sharper rates for general smooth convex functions. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4703–4711. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/nagaraj19a.html.

Nedić, A. and Bertsekas, D. Convergence rate of incremental subgradient algorithms. *Stochastic optimization: algorithms and applications*, pp. 223–264, 2001.

Nedic, A. and Bertsekas, D. P. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Nguyen, L., Nguyen, P. H., Dijk, M., Richtárik, P., Scheinberg, K., and Takác, M. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pp. 3750–3758. PMLR, 2018.

Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and Van Dijk, M. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):9397–9440, 2021.

Orabona, F. Last iterate of sgd converges (even in unbounded domains). 2020.

Pauwels, E. Incremental without replacement sampling in nonconvex optimization. *Journal of Optimization Theory and Applications*, 190(1):274–299, 2021.

Pollard, D. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

Rajput, S., Gupta, A., and Papailiopoulos, D. Closing the convergence gap of SGD without replacement. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7964–7973. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rajput20a.html.

Rajput, S., Lee, K., and Papailiopoulos, D. Permutation-based SGD: Is random optimal? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YiBa9HKTyXE.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Recht, B. and Ré, C. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

Sadiev, A., Malinovsky, G., Gorbunov, E., Sokolov, I., Khaled, A., Burlachenko, K. P., and Richtárik, P. Federated optimization algorithms with random reshuffling and gradient compression. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023. URL https://openreview.net/forum?id=izNFOlSK5H.

Safran, I. and Shamir, O. How good is sgd with random shuffling? In *Conference on Learning Theory*, pp. 3250–3284. PMLR, 2020.

Safran, I. and Shamir, O. Random shuffling beats sgd only after many epochs on ill-conditioned problems. *Advances in Neural Information Processing Systems*, 34:15151–15161, 2021.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shamir, O. Without-replacement sampling for stochastic gradient methods. *Advances in neural information processing systems*, 29, 2016.

Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pp. 71–79. PMLR, 2013.

Solodov, M. V. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11:23–35, 1998.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Tran, T. H., Nguyen, L. M., and Tran-Dinh, Q. Smg: A shuffling gradient-based method with momentum. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10379–10389. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tran21b.html.

Tran, T. H., Scheinberg, K., and Nguyen, L. M. Nesterov accelerated shuffling gradient method for convex optimization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21703–21732. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tran22a.html.

Ying, B., Yuan, K., Vlaski, S., and Sayed, A. H. Stochastic learning under random reshuffling with constant stepsizes. *IEEE Transactions on Signal Processing*, 67(2):474–489, 2018.

Yu, H. and Li, X. High probability guarantees for random reshuffling, 2023.

Yun, C., Rajput, S., and Sra, S. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=LdlwbBP2mlq.

Zamani, M. and Glineur, F. Exact convergence rate of the last iterate in subgradient methods. *arXiv preprint arXiv:2307.11134*, 2023.

Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116, 2004.

A. Detailed Comparison

We provide a more detailed comparison for smooth components in Table 3, where the case of strongly convex $\psi(\mathbf{x})$ and the previous fastest upper bounds for all cases are included.

Table 3. Summary of our new upper bounds and the previous best upper and lower bounds for L-smooth $f_i(\mathbf{x})$ for large K. Here, $\sigma_{\mathrm{any}}^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_*)\|^2$, $\sigma_{\mathrm{rand}}^2 \triangleq \sigma_{\mathrm{any}}^2 + n\|\nabla f(\mathbf{x}_*)\|^2$ and $D \triangleq \|\mathbf{x}_* - \mathbf{x}_1\|$. All rates use the function value gap as the convergence criterion. In the column of "Type", "Any" means the rate holds for whatever permutation not limited to RR/SO/IG. "Random" refers to the uniformly sampled permutation but is not restricted to RR/SO (see Remark 4.5 for a detailed explanation). "Avg", "Last" and "Tail" in the "Output" column stand for $\mathbf{x}_{K+1}^{\mathrm{avg}} \triangleq \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{k+1}$, $\mathbf{x}_{K+1}$ and $\mathbf{x}_{K+1}^{\mathrm{tail}} \triangleq \frac{1}{\lfloor \frac{K}{2} \rfloor + 1}\sum_{k=\lceil \frac{K}{2} \rceil}^{K}\mathbf{x}_{k+1}$, respectively. In the last column, "✓" means $\psi(\mathbf{x})$ can be taken arbitrarily and "✗" implies $\psi(\mathbf{x}) = 0$.

$F(\mathbf{x}) = f(\mathbf{x}) + \psi(\mathbf{x})$ where $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ and $f_i(\mathbf{x})$ and $\psi(\mathbf{x})$ are convex					
Settings	References	Rate	Type	Output	$\psi(\mathbf{x})$
L-smooth $f_i(\mathbf{x})$	(Mishchenko et al., 2020)	$\mathcal{O}\left(\frac{L^{1/3}\sigma_{\mathrm{any}}^{2/3}D^{4/3}}{K^{2/3}}\right)$	IG	Avg	✗
	Ours (Theorem 4.4)	$\widetilde{\mathcal{O}}\left(\frac{L^{1/3}\sigma_{\mathrm{any}}^{2/3}D^{4/3}}{K^{2/3}}\right)$	Any	Last	✓
	(Mishchenko et al., 2020; Nguyen et al., 2021)	$\mathcal{O}\left(\frac{L^{1/3}\sigma_{\mathrm{any}}^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$	RR/SO	Avg	✗
	Ours (Theorem 4.4)	$\widetilde{\mathcal{O}}\left(\frac{L^{1/3}\sigma_{\mathrm{rand}}^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$ [a]	Random	Last	✓
	(Cha et al., 2023)	$\Omega\left(\frac{L^{1/3}\sigma_{\mathrm{any}}^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$	RR	Last	✗
L-smooth $f_i(\mathbf{x})$, μ-strongly convex $f(\mathbf{x})$	(Mishchenko et al., 2020; Nguyen et al., 2021)	$\widetilde{\mathcal{O}}\left(\frac{L^2\sigma_{\mathrm{any}}^2}{\mu^3K^2}\right)$	IG	Last	✗
	Ours (Theorem 4.6)	$\widetilde{\mathcal{O}}\left(\frac{L\sigma_{\mathrm{any}}^2}{\mu^2K^2}\right)$	Any	Last	✓
	(Safran & Shamir, 2020)	$\Omega\left(\frac{G^2}{\mu K^2}\right)$ [b]	IG	Last	✗
	(Cha et al., 2023)	$\widetilde{\mathcal{O}}\left(\frac{L\sigma_{\mathrm{any}}^2}{\mu^2nK^2}\right)$	RR	Tail	✗
	Ours (Theorem 4.6)	$\widetilde{\mathcal{O}}\left(\frac{L\sigma_{\mathrm{rand}}^2}{\mu^2nK^2}\right)$ [a]	Random	Last	✓
	(Safran & Shamir, 2021; Cha et al., 2023)	$\Omega\left(\frac{L\sigma_{\mathrm{any}}^2}{\mu^2nK^2}\right)$	RR/SO	Last	✗
L-smooth $f_i(\mathbf{x})$, μ-strongly convex $\psi(\mathbf{x})$	**Ours** (Theorem C.2)	$\widetilde{\mathcal{O}}\left(\frac{L\sigma_{\mathrm{any}}^2}{\mu^2K^2}\right)$	Any	Last	✓
	(Mishchenko et al., 2022)	$\widetilde{\mathcal{O}}\left(\frac{L^2\sigma_{\mathrm{rand}}^2}{\mu^3nK^2}\right)$ [c]	RR/SO	Last[c]	✓
	Ours (Theorem C.2)	$\widetilde{\mathcal{O}}\left(\frac{L\sigma_{\mathrm{rand}}^2}{\mu^2nK^2}\right)$	Random	Last	✓

[a] Note that when $\psi(\mathbf{x}) = 0$, there is $\sigma_{\mathrm{rand}}^2 = \sigma_{\mathrm{any}}^2 + n\|\nabla f(\mathbf{x}_*)\|^2 = \sigma_{\mathrm{any}}^2$ due to $\nabla f(\mathbf{x}_*) = \mathbf{0}$.

[b] This lower bound is established under $L = \mu$ and additionally requires $\|\nabla f_i(\mathbf{x})\| \le G$. Under the same condition, our above upper bound $\widetilde{\mathcal{O}}\left(\frac{L\sigma_{\mathrm{any}}^2}{\mu^2K^2}\right)$ will be $\widetilde{\mathcal{O}}\left(\frac{G^2}{\mu K^2}\right)$ and hence almost matches this lower bound.

[c] The original rate $\widetilde{\mathcal{O}}\left(\frac{L\sigma_{\mathrm{rand}}^2}{\mu^3nK^2}\right)$ is only proved for $\|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2$ and hence cannot measure the function value gap when a general $\psi(\mathbf{x})$ exists as explained in Section 1. However, for the convenience of comparison, we still consider the traditional conversion $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) = \mathcal{O}(L\|\mathbf{x}_{K+1} - \mathbf{x}_*\|^2)$ here, though it does not hold in general.

B. Important Notations

We summarize important notations used in the analysis as follows for the sake of readability.

- Objective function: $F(x) \triangleq f(\mathbf{x}) + \psi(\mathbf{x})$ where $f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$.
- Number of epochs: $K \ge 2$.

- Initial distance: $D \triangleq \|\mathbf{x}_* - \mathbf{x}_1\|$.
- Optimal function value: $F_* \triangleq F(\mathbf{x}_*)$.
- The averaged smooth parameter: $\bar{L} \triangleq \frac{1}{n}\sum_{i=1}^n L_i$.
- The largest smooth parameter: $L^* \triangleq \max_{i\in[n]} L_i$.
- Two quantities measuring the uncertainty: $\sigma_{\text{any}}^2 \triangleq \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_*)\|^2$ and $\sigma_{\text{rand}}^2 \triangleq \sigma_{\text{any}}^2 + n\|\nabla f(\mathbf{x}_*)\|^2$.
- The averaged Lipschitz parameter: $\bar{G} \triangleq \frac{1}{n}\sum_{i=1}^n G_i$.
- The Bregman divergence induced by $f_{\sigma_k^i}(\mathbf{x})$: $B_{\sigma_k^i}(\cdot,\cdot) \triangleq B_{f_{\sigma_k^i}}(\cdot,\cdot)$.

C. Full Theorems and Proofs

In this section, we provide full statements of all theorems with proofs. Lemmas used in the proof can be found in Sections D and E.

C.1. Smooth Functions

All results presented in this subsection are for smooth components $f_i(\mathbf{x})$.

First, we consider the general convex case, i.e., $\mu_f = \mu_\psi = 0$. The key tools in the proof are Lemmas D.4 and E.2.

Theorem C.1. *(Full version of Theorem 4.4) Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$) and 3.3, taking the stepsize $\eta_k = \eta \le \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}}, \forall k \in [K]$:*

- *Regardless of how the permutation σ_k is generated in every epoch, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\frac{D^2}{n\eta K} + (n\eta)^2\bar{L}\sigma_{\text{any}}^2(1+\log K)\right).$$

  Setting $\eta = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}} \wedge \frac{D^{2/3}}{n\sqrt[3]{\bar{L}\sigma_{\text{any}}^2 K(1+\log K)}}$ to get

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\frac{D^2\sqrt{\bar{L}L^*(1+\log K)}}{K} + \frac{D^{4/3}\sqrt[3]{\bar{L}\sigma_{\text{any}}^2(1+\log K)}}{K^{2/3}}\right).$$

- *Suppose the permutation σ_k is uniformly sampled without replacement, Algorithm 1 guarantees*

$$\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F_*\right] \le \mathcal{O}\left(\frac{D^2}{n\eta K} + n\eta^2\bar{L}\sigma_{\text{rand}}^2(1+\log K)\right).$$

  Setting $\eta = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}} \wedge \frac{D^{2/3}}{\sqrt[3]{n^2\bar{L}\sigma_{\text{rand}}^2 K(1+\log K)}}$ to get

$$\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F_*\right] \le \mathcal{O}\left(\frac{D^2\sqrt{\bar{L}L^*(1+\log K)}}{K} + \frac{D^{4/3}\sqrt[3]{\bar{L}\sigma_{\text{rand}}^2(1+\log K)}}{n^{1/3}K^{2/3}}\right).$$

Proof. Note that $\eta_k = \eta \le \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}} \le \frac{1}{2n\sqrt{\bar{L}L^*}}$ satisfies the requirement of Lemma D.4. Hence, for any $k \in [K]$

$$\begin{aligned}F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) &\le \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\sum_{\ell=1}^k \eta_\ell} + \sum_{\ell=1}^k \frac{4\eta_\ell^3 R_\ell}{\sum_{s=\ell}^k \eta_s} + \sum_{\ell=2}^k \frac{8n^2\bar{L}^2\eta_{\ell-1}\left(\sum_{s=\ell}^k \eta_s^3\right)}{\left(\sum_{s=\ell}^k \eta_s\right)\left(\sum_{s=\ell-1}^k \eta_s\right)} B_f(\mathbf{x}_\ell, \mathbf{x}_*)\\ &= \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta k} + \sum_{\ell=1}^k \frac{4\eta^2 R_\ell}{k-\ell+1} + 8(n\eta\bar{L})^2\sum_{\ell=2}^k \frac{B_f(\mathbf{x}_\ell,\mathbf{x}_*)}{k-\ell+2},\end{aligned} \tag{2}$$

where $R_\ell = \sum_{i=2}^{n} \frac{L_{\sigma_\ell^i}}{n} \left\| \sum_{j=1}^{i-1} \nabla f_{\sigma_\ell^j}(\mathbf{x}_*) \right\|^2$. The definition of $\mathbf{x}_* \in \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$ implies $\exists \nabla\psi(\mathbf{x}_*) \in \partial\psi(\mathbf{x}_*)$ such that $\nabla f(\mathbf{x}_*) + \nabla\psi(\mathbf{x}_*) = \mathbf{0}$. Thus, for any $k \in [K]$

$$\begin{aligned} F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) &= F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) - \langle \nabla f(\mathbf{x}_*) + \nabla\psi(\mathbf{x}_*), \mathbf{x}_{k+1} - \mathbf{x}_* \rangle \\ &= B_f(\mathbf{x}_{k+1}, \mathbf{x}_*) + B_\psi(\mathbf{x}_{k+1}, \mathbf{x}_*) \geq B_f(\mathbf{x}_{k+1}, \mathbf{x}_*), \end{aligned}$$

which implies

$$B_f(\mathbf{x}_{k+1}, \mathbf{x}_*) \leq \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta k} + \sum_{\ell=1}^{k} \frac{4\eta^2 R_\ell}{k-\ell+1} + 8(n\eta\bar{L})^2 \sum_{\ell=2}^{k} \frac{B_f(\mathbf{x}_\ell, \mathbf{x}_*)}{k-\ell+2}, \forall k \in [K]. \tag{3}$$

Besides, by taking $k = K$ in (2), we obtain

$$F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) \leq \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta K} + \sum_{\ell=1}^{K} \frac{4\eta^2 R_\ell}{K-\ell+1} + 8(n\eta\bar{L})^2 \sum_{\ell=2}^{K} \frac{B_f(\mathbf{x}_\ell, \mathbf{x}_*)}{K-\ell+2}. \tag{4}$$

First, by Lemma E.1, for any permutation σ_ℓ used in the ℓ-th epoch we have

$$R_\ell \leq n^2 \bar{L} \sigma_{\text{any}}^2, \forall \ell \in [K]. \tag{5}$$

Hence, for any $k \in [K]$

$$\begin{aligned} B_f(\mathbf{x}_{k+1}, \mathbf{x}_*) &\overset{(3),(5)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta k} + \sum_{\ell=1}^{k} \frac{4\eta^2 n^2 \bar{L} \sigma_{\text{any}}^2}{k-\ell+1} + 8(n\eta\bar{L})^2 \sum_{\ell=2}^{k} \frac{B_f(\mathbf{x}_\ell, \mathbf{x}_*)}{k-\ell+2} \\ &\leq \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta k} + 4(n\eta)^2 \bar{L} \sigma_{\text{any}}^2 (1 + \log k) + 8(n\eta\bar{L})^2 \sum_{\ell=2}^{k} \frac{B_f(\mathbf{x}_\ell, \mathbf{x}_*)}{k-\ell+2}. \end{aligned}$$

In addition,

$$\begin{aligned} F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\overset{(4),(5)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta K} + \sum_{\ell=1}^{K} \frac{4\eta^2 n^2 \bar{L} \sigma_{\text{any}}^2}{K-\ell+1} + 8(n\eta\bar{L})^2 \sum_{\ell=2}^{K} \frac{B_f(\mathbf{x}_\ell, \mathbf{x}_*)}{K-\ell+2} \\ &\leq \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta K} + 4(n\eta)^2 \bar{L} \sigma_{\text{any}}^2 (1 + \log K) + 8(n\eta\bar{L})^2 \sum_{\ell=2}^{K} \frac{B_f(\mathbf{x}_\ell, \mathbf{x}_*)}{K-\ell+2}. \end{aligned}$$

We apply Lemma E.2 with $d_{k+1} = \begin{cases} B_f(\mathbf{x}_{k+1}, \mathbf{x}_*) & k \in [K-1] \\ F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) & k = K \end{cases}$, $a = \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta}$, $b = 4(n\eta)^2 \bar{L} \sigma_{\text{any}}^2$, $c = 8(n\eta\bar{L})^2$ to obtain

$$B_f(\mathbf{x}_{k+1}, \mathbf{x}_*) \leq \left(\frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta k} + 4(n\eta)^2 \bar{L} \sigma_{\text{any}}^2 (1 + \log k) \right) \sum_{i=0}^{k-1} \left(16(n\eta\bar{L})^2 (1 + \log k) \right)^i, \forall k \in [K-1],$$

and

$$F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) \leq \left(\frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta K} + 4(n\eta)^2 \bar{L} \sigma_{\text{any}}^2 (1 + \log K) \right) \sum_{i=0}^{K-1} \left(16(n\eta\bar{L})^2 (1 + \log K) \right)^i.$$

By $\eta_k = \eta \leq \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}}$, we have for any $k \in [K]$,

$$\sum_{i=0}^{k-1} \left(16(n\eta\bar{L})^2 (1 + \log k) \right)^i \leq \sum_{i=0}^{k-1} \left(\frac{\bar{L}(1 + \log k)}{2L^*(1 + \log K)} \right)^i \leq \sum_{i=0}^{\infty} \frac{1}{2^i} = 2.$$

So there is

$$B_f(\mathbf{x}_{k+1},\mathbf{x}_*)\le\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{n\eta k}+8(n\eta)^2\bar{L}\sigma_{\text{any}}^2(1+\log k),\forall k\in[K-1]\,,$$

and

$$\begin{aligned}F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)&\le\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{n\eta K}+8(n\eta)^2\bar{L}\sigma_{\text{any}}^2(1+\log K)\\&=\mathcal{O}\left(\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{n\eta K}+(n\eta)^2\bar{L}\sigma_{\text{any}}^2(1+\log K)\right).\end{aligned}$$

Finally, taking $\eta=\frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}}\wedge\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^{2/3}}{n\sqrt[3]{\bar{L}\sigma_{\text{any}}^2K(1+\log K)}}$ to obtain

$$F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\le\mathcal{O}\left(\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2\sqrt{\bar{L}L^*(1+\log K)}}{K}+\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^{4/3}\sqrt[3]{\bar{L}\sigma_{\text{any}}^2(1+\log K)}}{K^{2/3}}\right).$$

For the case of randomly generated permutations, by Lemma E.1, we have

$$\mathbb{E}\left[R_\ell\right]\le\frac{2}{3}n\bar{L}\sigma_{\text{rand}}^2,\forall\ell\in[K]\,.\tag{6}$$

Hence, for any $k\in[K]$,

$$\begin{aligned}\mathbb{E}\left[B_f(\mathbf{x}_{k+1},\mathbf{x}_*)\right]&\overset{(3),(6)}{\le}\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{2n\eta k}+\sum_{\ell=1}^{k}\frac{\frac{8}{3}\eta^2n\bar{L}\sigma_{\text{rand}}^2}{k-\ell+1}+8(n\eta\bar{L})^2\sum_{\ell=2}^{k}\frac{\mathbb{E}\left[B_f(\mathbf{x}_\ell,\mathbf{x}_*)\right]}{k-\ell+2}\\&\le\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{2n\eta k}+\frac{8}{3}n\eta^2\bar{L}\sigma_{\text{rand}}^2(1+\log k)+8(n\eta\bar{L})^2\sum_{\ell=2}^{k}\frac{\mathbb{E}\left[B_f(\mathbf{x}_\ell,\mathbf{x}_*)\right]}{k-\ell+2}.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\right]&\overset{(4),(6)}{\le}\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{2n\eta K}+\sum_{\ell=1}^{K}\frac{\frac{8}{3}\eta^2n\bar{L}\sigma_{\text{rand}}^2}{K-\ell+1}+8(n\eta\bar{L})^2\sum_{\ell=2}^{K}\frac{\mathbb{E}\left[B_f(\mathbf{x}_\ell,\mathbf{x}_*)\right]}{K-\ell+2}\\&\le\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{2n\eta K}+\frac{8}{3}n\eta^2\bar{L}\sigma_{\text{rand}}^2(1+\log K)+8(n\eta\bar{L})^2\sum_{\ell=2}^{K}\frac{B_f(\mathbf{x}_\ell,\mathbf{x}_*)}{K-\ell+2}.\end{aligned}$$

We apply Lemma E.2 with $d_{k+1}=\begin{cases}\mathbb{E}\left[B_f(\mathbf{x}_{k+1},\mathbf{x}_*)\right]&k\in[K-1]\\\mathbb{E}\left[F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\right]&k=K\end{cases}$, $a=\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{2n\eta}$, $b=\frac{8}{3}n\eta^2\bar{L}\sigma_{\text{rand}}^2$, $c=8(n\eta\bar{L})^2$ and then follow the similar steps used before to obtain

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\right]&\le\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{n\eta K}+\frac{16}{3}n\eta^2\bar{L}\sigma_{\text{rand}}^2(1+\log K)\\&=\mathcal{O}\left(\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{n\eta K}+n\eta^2\bar{L}\sigma_{\text{rand}}^2(1+\log K)\right).\end{aligned}$$

Finally, taking $\eta=\frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}}\wedge\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^{2/3}}{\sqrt[3]{n^2\bar{L}\sigma_{\text{rand}}^2K(1+\log K)}}$ to obtain

$$F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\le\mathcal{O}\left(\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2\sqrt{\bar{L}L^*(1+\log K)}}{K}+\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^{4/3}\sqrt[3]{\bar{L}\sigma_{\text{rand}}^2(1+\log K)}}{n^{1/3}K^{2/3}}\right).$$

□

Next, we provide the full theorem when at least one of $f(\mathbf{x})$ and $\psi(\mathbf{x})$ is strongly convex. The key step is to bound $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)$ by $\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2$ using Theorem C.1 and bound $\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2$ by $\|\mathbf{x}_* - \mathbf{x}_1\|^2$ using Lemma D.5.

Theorem C.2. *(Full version of Theorem 4.6) Under Assumptions 3.1, 3.2, 3.3 and 3.4 with $\mu_F \triangleq \mu_f + 2\mu_\psi > 0$, taking the stepsize $\eta_k = \eta \le \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)} \vee \mu_\psi}, \forall k \in [K]$:*

- *Regardless of how the permutation σ_k is generated in every epoch, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\frac{D^2}{n\eta K} e^{-n\eta\mu_F K} + \frac{n\eta\bar{L}\sigma_{\text{any}}^2}{\mu_F K} + (n\eta)^2\bar{L}\sigma_{\text{any}}^2(1+\log K)\right).$$

  Setting $\eta = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)} \vee \mu_\psi} \wedge \frac{u}{n\mu_F K}$ where $u = 1 \vee \log \frac{\mu_F^3 D^2 K^2}{\bar{L}\sigma_{\text{any}}^2(1+\log K)}$ to get

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\frac{D^2\left(\sqrt{\bar{L}L^*(1+\log K)} \vee \mu_\psi\right)}{K e^{\frac{\mu_F K}{\sqrt{\bar{L}L^*(1+\log K)} \vee \mu_\psi}}} + \frac{(1+u^2)\bar{L}\sigma_{\text{any}}^2(1+\log K)}{\mu_F^2 K^2}\right).$$

- *Suppose the permutation σ_k is uniformly sampled without replacement, Algorithm 1 guarantees*

$$\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F_*\right] \le \mathcal{O}\left(\frac{D^2}{n\eta K} e^{-n\eta\mu_F K} + \frac{\eta\bar{L}\sigma_{\text{rand}}^2}{\mu_F K} + n\eta^2\bar{L}\sigma_{\text{rand}}^2(1+\log K)\right).$$

  Setting $\eta = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)} \vee \mu_\psi} \wedge \frac{u}{n\mu_F K}$ where $u = 1 \vee \log \frac{n\mu_F^3 D^2 K^2}{\bar{L}\sigma_{\text{rand}}^2(1+\log K)}$ to get

$$\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F_*\right] \le \mathcal{O}\left(\frac{D^2\left(\sqrt{\bar{L}L^*(1+\log K)} \vee \mu_\psi\right)}{K e^{\frac{\mu_F K}{\sqrt{\bar{L}L^*(1+\log K)} \vee \mu_\psi}}} + \frac{(1+u^2)\bar{L}\sigma_{\text{rand}}^2(1+\log K)}{\mu_F^2 n K^2}\right).$$

Proof. Suppose we run Algorithm 1 by K epochs with the stepsize $\eta_k = \eta, \forall k \in [K]$. Then one can recognize $\mathbf{x}_{K+1}$ as obtained by running Algorithm 1 with the initial point $\mathbf{x}_{\lceil \frac{K}{2} \rceil+1}$ by $K - \lceil \frac{K}{2} \rceil = \lfloor \frac{K}{2} \rfloor$ epochs. Note that $\eta \le \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)} \vee \mu_\psi} \le \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log \lfloor \frac{K}{2} \rfloor)}}$, then we can apply Theorem C.1 to obtain that for any permutation σ_k

$$\begin{aligned}
F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\le \mathcal{O}\left(\frac{\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2}{n\eta\lfloor \frac{K}{2} \rfloor} + (n\eta)^2\bar{L}\sigma_{\text{any}}^2\left(1+\log\left\lfloor\frac{K}{2}\right\rfloor\right)\right) \\
&= \mathcal{O}\left(\frac{\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2}{n\eta K} + (n\eta)^2\bar{L}\sigma_{\text{any}}^2(1+\log K)\right).
\end{aligned} \tag{7}$$

If the permutation σ_k is uniformly sampled without replacement, there is

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)\right] &\le \mathcal{O}\left(\frac{\mathbb{E}\left[\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2\right]}{n\eta\lfloor \frac{K}{2} \rfloor} + n\eta^2\bar{L}\sigma_{\text{rand}}^2\left(1+\log\left\lfloor\frac{K}{2}\right\rfloor\right)\right) \\
&= \mathcal{O}\left(\frac{\mathbb{E}\left[\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2\right]}{n\eta K} + n\eta^2\bar{L}\sigma_{\text{rand}}^2(1+\log K)\right).
\end{aligned} \tag{8}$$

Now we use Lemma D.5 for $k=\lceil\frac{K}{2}\rceil$ to obtain

$$\begin{aligned}
\left\|\mathbf{x}_*-\mathbf{x}_{\lceil\frac{K}{2}\rceil+1}\right\|^2 &\leq \frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{\prod_{s=1}^{\lceil\frac{K}{2}\rceil}(1+n\eta_s(\mu_f+2\mu_\psi))}+\sum_{\ell=1}^{\lceil\frac{K}{2}\rceil}\frac{8n\eta_\ell^3R_\ell}{\prod_{s=\ell}^{\lceil\frac{K}{2}\rceil}(1+n\eta_s(\mu_f+2\mu_\psi))}\\
&=\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{(1+n\eta\mu_F)^{\lceil\frac{K}{2}\rceil}}+8n\eta^3\sum_{\ell=1}^{\lceil\frac{K}{2}\rceil}\frac{R_\ell}{(1+n\eta\mu_F)^{\lceil\frac{K}{2}\rceil-\ell+1}}\\
&\overset{(a)}{\leq}\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-\frac{n\eta\mu_F\lceil\frac{K}{2}\rceil}{1+n\eta\mu_F}}+8n\eta^3\sum_{\ell=1}^{\lceil\frac{K}{2}\rceil}\frac{R_\ell}{(1+n\eta\mu_F)^{\lceil\frac{K}{2}\rceil-\ell+1}}\\
&\overset{(b)}{\leq}\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-\frac{5n\eta\mu_FK}{32}}+8n\eta^3\sum_{\ell=1}^{\lceil\frac{K}{2}\rceil}\frac{R_\ell}{(1+n\eta\mu_F)^{\lceil\frac{K}{2}\rceil-\ell+1}},
\end{aligned} \tag{9}$$

where (a) is by $\frac{1}{1+x}\leq\exp\left(-\frac{x}{1+x}\right)$ and (b) is due to $\lceil\frac{K}{2}\rceil\geq\frac{K}{2}$ and $n\eta\mu_F\leq\frac{\mu_f+2\mu_\psi}{4\sqrt{2\bar{L}L^*(1+\log K)}\vee\mu_\psi}\leq\frac{1}{4\sqrt{2}}+2\leq\frac{11}{5}$.

By Lemma E.1, there is $R_\ell\leq n^2\bar{L}\sigma^2_{\text{any}}$ for any permutation σ_ℓ, which implies

$$\begin{aligned}
\left\|\mathbf{x}_*-\mathbf{x}_{\lceil\frac{K}{2}\rceil+1}\right\|^2 &\overset{(9)}{\leq}\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-\frac{5n\eta\mu_FK}{32}}+8(n\eta)^3\bar{L}\sigma^2_{\text{any}}\sum_{\ell=1}^{\lceil\frac{K}{2}\rceil}\frac{1}{(1+n\eta\mu_F)^{\lceil\frac{K}{2}\rceil-\ell+1}}\\
&\leq\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-\frac{5n\eta\mu_FK}{32}}+\frac{8(n\eta)^2\bar{L}\sigma^2_{\text{any}}}{\mu_F}\\
&=\mathcal{O}\left(\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-n\eta\mu_FK}+\frac{(n\eta)^2\bar{L}\sigma^2_{\text{any}}}{\mu_F}\right).
\end{aligned} \tag{10}$$

Additionally, if σ_ℓ is uniformly sampled without replacement, then $\mathbb{E}\left[R_\ell\right]\leq\frac{2}{3}n\bar{L}\sigma^2_{\text{rand}}$. Therefore

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_*-\mathbf{x}_{\lceil\frac{K}{2}\rceil+1}\right\|^2\right] &\overset{(9)}{\leq}\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-\frac{5n\eta\mu_FK}{32}}+\frac{16}{3}n^2\eta^3\bar{L}\sigma^2_{\text{rand}}\sum_{\ell=1}^{\lceil\frac{K}{2}\rceil}\frac{1}{(1+n\eta\mu_F)^{\lceil\frac{K}{2}\rceil-\ell+1}}\\
&\leq\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-\frac{5n\eta\mu_FK}{32}}+\frac{16n\eta^2\bar{L}\sigma^2_{\text{rand}}}{3\mu_F}\\
&=\mathcal{O}\left(\|\mathbf{x}_*-\mathbf{x}_1\|^2e^{-n\eta\mu_FK}+\frac{n\eta^2\bar{L}\sigma^2_{\text{rand}}}{\mu_F}\right).
\end{aligned} \tag{11}$$

Combining (7) and (10), we obtain for any permutation

$$F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\leq\mathcal{O}\left(\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{n\eta K}e^{-n\eta\mu_FK}+\frac{n\eta\bar{L}\sigma^2_{\text{any}}}{\mu_FK}+(n\eta)^2\bar{L}\sigma^2_{\text{any}}(1+\log K)\right).$$

Taking $\eta=\frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}\vee\mu_\psi}\wedge\frac{u}{n\mu_FK}$ where $u=1\vee\log\frac{\mu_F^3\|\mathbf{x}_*-\mathbf{x}_1\|^2K^2}{\bar{L}\sigma^2_{\text{any}}(1+\log K)}$ to obtain

$$F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\leq\mathcal{O}\left(\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2\left(\sqrt{\bar{L}L^*(1+\log K)}\vee\mu_\psi\right)}{Ke^{\frac{\mu_FK}{\sqrt{\bar{L}L^*(1+\log K)}\vee\mu_\psi}}}+\frac{(1+u^2)\bar{L}\sigma^2_{\text{any}}(1+\log K)}{\mu_F^2K^2}\right).$$

Combining (8) and (11), we obtain for uniform sampling without replacement

$$\mathbb{E}\left[F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\right]\leq\mathcal{O}\left(\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{n\eta K}e^{-n\eta\mu_FK}+\frac{\eta\bar{L}\sigma^2_{\text{rand}}}{\mu_FK}+n\eta^2\bar{L}\sigma^2_{\text{rand}}(1+\log K)\right).$$

Taking $\eta = \frac{1}{4n\sqrt{2\bar{L}L^*(1+\log K)}\vee\mu_\psi} \wedge \frac{u}{n\mu_F K}$ where $u = 1 \vee \log\frac{n\mu_F^3\|\mathbf{x}_*-\mathbf{x}_1\|^2K^2}{\bar{L}\sigma^2_{\text{rand}}(1+\log K)}$ to get

$$\mathbb{E}\left[F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)\right] \le \mathcal{O}\left(\frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2\left(\sqrt{\bar{L}L^*(1+\log K)}\vee\mu_\psi\right)}{Ke^{\frac{\mu_F K}{\sqrt{\bar{L}L^*(1+\log K)}\vee\mu_\psi}}} + \frac{(1+u^2)\bar{L}\sigma^2_{\text{rand}}(1+\log K)}{\mu_F^2 nK^2}\right).$$

□

C.2. Lipschitz Functions

We focus on Lipschitz components $f_i(\mathbf{x})$ in this subsection.

First, we analyze the case of $\mu_\psi = 0$ in Theorem C.3 with three different stepsizes. It is worth noting that the last stepsize schedule, which is inspired by (Zamani & Glineur, 2023), can remove the extra $\mathcal{O}(\log K)$ term.

Theorem C.3. *(Full version of Theorem 4.7) Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$) and 3.5, regardless of how the permutation σ_k is generated in every epoch:*

- *Taking the stepsize $\eta_k = \frac{\eta}{\sqrt{k}}, \forall k \in [K]$, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\left(\frac{D^2}{n\eta} + n\eta\bar{G}^2(1+\log K)\right)\frac{1}{\sqrt{K}}\right).$$

  Setting $\eta = \frac{D^2}{n\bar{G}}$ to get the best dependence on parameters.

- *Taking the stepsize $\eta_k = \frac{\eta}{\sqrt{K}}, \forall k \in [K]$, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\left(\frac{D^2}{n\eta} + n\eta\bar{G}^2(1+\log K)\right)\frac{1}{\sqrt{K}}\right).$$

  Setting $\eta = \frac{D}{n\bar{G}\sqrt{1+\log K}}$ to get the best dependence on parameters.

- *Taking the stepsize $\eta_k = \eta\frac{K-k+1}{K^{3/2}}, \forall k \in [K]$, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\left(\frac{D^2}{n\eta} + n\eta\bar{G}^2\right)\frac{1}{\sqrt{K}}\right).$$

  Setting $\eta = \frac{D}{n\bar{G}}$ to get the best dependence on parameters.

Proof. We invoke Lemma D.6 to get

$$\begin{aligned}F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\le \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2 - \|\mathbf{x}_* - \mathbf{x}_{K+1}\|^2}{2n\sum_{k=1}^K \gamma_k} + 3\bar{G}^2 n\sum_{k=1}^K \frac{\gamma_k\eta_k}{\sum_{\ell=k}^K \gamma_\ell} \\ &\le \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\sum_{k=1}^K \gamma_k} + 3\bar{G}^2 n\sum_{k=1}^K \frac{\gamma_k\eta_k}{\sum_{\ell=k}^K \gamma_\ell},\end{aligned}$$

where $\gamma_k = \eta_k\prod_{\ell=2}^k(1+n\eta_{\ell-1}\mu_\psi), \forall k \in [K]$. Note that $\mu_\psi = 0 \Rightarrow \gamma_k = \eta_k, \forall k \in [K]$, therefore

$$F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) \le \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\sum_{k=1}^K \eta_k} + 3\bar{G}^2 n\sum_{k=1}^K \frac{\eta_k^2}{\sum_{\ell=k}^K \eta_\ell}. \quad (12)$$

If $\eta_k = \frac{\eta}{\sqrt{k}}, \forall k \in [K]$, then

$$\begin{aligned}
F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\overset{(12)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta \sum_{k=1}^{K} 1/\sqrt{k}} + 3\bar{G}^2 n\eta \sum_{k=1}^{K} \frac{1}{k\sum_{\ell=k}^{K} 1/\sqrt{\ell}} \\
&\overset{(a)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{4n\eta(\sqrt{K+1}-1)} + 3\bar{G}^2 n\eta \sum_{k=1}^{K} \frac{1}{2k(\sqrt{K+1}-\sqrt{k})} \\
&\overset{(b)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{4n\eta(\sqrt{K+1}-1)} + \frac{3\bar{G}^2 n\eta}{\sqrt{K+1}} \sum_{k=1}^{K} \frac{1}{k} + \frac{1}{K+1-k} \\
&= \mathcal{O}\left(\left(\frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{n\eta} + n\eta\bar{G}^2(1+\log K)\right)\frac{1}{\sqrt{K}}\right),
\end{aligned}$$

where (a) is by $\sum_{\ell=k}^{K} 1/\sqrt{\ell} \geq \int_{k}^{K+1} 1/\sqrt{\ell}\mathrm{d}\ell = 2\left(\sqrt{K+1}-\sqrt{k}\right)$ and (b) is due to

$$\frac{1}{2k(\sqrt{K+1}-\sqrt{k})} = \frac{\sqrt{K+1}+\sqrt{k}}{2k(K+1-k)} \leq \frac{\sqrt{K+1}}{k(K+1-k)} = \frac{1}{\sqrt{K+1}}\left(\frac{1}{k} + \frac{1}{K+1-k}\right).$$

Taking $\eta = \frac{\|\mathbf{x}_* - \mathbf{x}_1\|}{n\bar{G}}$ to obtain

$$F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) = \mathcal{O}\left(\frac{\bar{G}\|\mathbf{x}_* - \mathbf{x}_1\|(1+\log K)}{\sqrt{K}}\right).$$

If $\eta_k = \frac{\eta}{\sqrt{K}}, \forall k \in [K]$, then

$$\begin{aligned}
F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\overset{(12)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta \sum_{k=1}^{K} 1/\sqrt{K}} + \frac{3\bar{G}^2 n\eta}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{K-k+1} \\
&= \mathcal{O}\left(\left(\frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{n\eta} + n\eta\bar{G}^2(1+\log K)\right)\frac{1}{\sqrt{K}}\right).
\end{aligned}$$

Taking $\eta = \frac{\|\mathbf{x}_* - \mathbf{x}_1\|}{n\bar{G}\sqrt{1+\log K}}$ to obtain

$$F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) = \mathcal{O}\left(\frac{\bar{G}\|\mathbf{x}_* - \mathbf{x}_1\|\sqrt{1+\log K}}{\sqrt{K}}\right).$$

If $\eta_k = \eta\frac{K-k+1}{K^{3/2}}, \forall k \in [K]$, then

$$\begin{aligned}
F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\overset{(12)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{2n\eta \sum_{k=1}^{K} (K-k+1)/K^{3/2}} + \frac{3\bar{G}^2 n\eta}{K^{3/2}} \sum_{k=1}^{K} \frac{(K-k+1)^2}{\sum_{\ell=k}^{K} K-\ell+1} \\
&= \frac{\sqrt{K}\|\mathbf{x}_* - \mathbf{x}_1\|^2}{n\eta(K+1)} + \frac{6\bar{G}^2 n\eta}{K^{3/2}} \sum_{k=1}^{K} \frac{K-k+1}{K-k+2} \\
&= \mathcal{O}\left(\left(\frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{n\eta} + \bar{G}^2 n\eta\right)\frac{1}{\sqrt{K}}\right),
\end{aligned}$$

Taking $\eta = \frac{\|\mathbf{x}_* - \mathbf{x}_1\|}{n\bar{G}}$ to obtain

$$F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) = \mathcal{O}\left(\frac{\bar{G}\|\mathbf{x}_* - \mathbf{x}_1\|}{\sqrt{K}}\right).$$

□

Next, Theorem C.4 shows that, under the particularly carefully designed stepsize, it is possible to obtain an asymptotic rate with an optimal linear dependence on D without knowing it. The key is to use $r_k = r \vee \max_{\ell\in[k]} \|\mathbf{x}_\ell - \mathbf{x}_1\|$ to approximate D, which was orginally introduced by (Ivgi et al., 2023).

Theorem C.4. *(Full version of Theorem 4.9) Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$) and 3.5, regardless of how the permutation σ_k is generated in every epoch, taking the stepsize $\eta_k = r_k\widetilde{\eta}_k, \forall k \in \mathbb{N}$ where $r_k = r \vee \max_{\ell\in[k]} \|\mathbf{x}_\ell - \mathbf{x}_1\|$ for some $r > 0$ and $\widetilde{\eta}_k = \frac{c}{n\bar{G}\sqrt{6(1+\delta^{-1})k(1+\log k)^{1+\delta}}}, \forall k \in \mathbb{N}$ for some $\delta > 0, 0 < c < 1$, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1}) - F_* \le \left(\frac{r_{K+1}}{c\sum_{k=1}^K r_k/K} + \frac{3c}{1-c}\right)\bar{G}\,(D \vee r)\sqrt{\frac{6(1+\delta^{-1})(1+\log K)^{1+\delta}}{K}}.$$

Moreover, for large enough K

$$F(\mathbf{x}_{K+1}) - F_* \le \mathcal{O}\left(\frac{3c^2 - c + 1}{c(1-c)}\bar{G}\,(D \vee r)\sqrt{\frac{(1+\delta^{-1})(1+\log K)^{1+\delta}}{K}}\right).$$

Proof. We first invoke Lemma D.6 to obtain

$$\begin{aligned}F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\le \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2 - \|\mathbf{x}_* - \mathbf{x}_{K+1}\|^2}{2n\sum_{k=1}^K \gamma_k} + 3\bar{G}^2 n \sum_{k=1}^K \frac{\gamma_k \eta_k}{\sum_{\ell=k}^K \gamma_\ell}\\ &= \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2 - \|\mathbf{x}_* - \mathbf{x}_{K+1}\|^2}{2n\sum_{k=1}^K \eta_k} + 3\bar{G}^2 n \sum_{k=1}^K \frac{\eta_k^2}{\sum_{\ell=k}^K \eta_\ell},\end{aligned}$$

where the last equation holds due to $\gamma_k = \eta_k \prod_{\ell=2}^k (1 + n\eta_{\ell-1}\mu_\psi) = \eta_k$ when $\mu_\psi = 0$. Now if $\|\mathbf{x}_* - \mathbf{x}_1\| \ge \|\mathbf{x}_* - \mathbf{x}_{K+1}\|$, we have

$$\begin{aligned}\|\mathbf{x}_* - \mathbf{x}_1\|^2 - \|\mathbf{x}_* - \mathbf{x}_{K+1}\|^2 &= (\|\mathbf{x}_* - \mathbf{x}_1\| - \|\mathbf{x}_* - \mathbf{x}_{K+1}\|)(\|\mathbf{x}_* - \mathbf{x}_1\| + \|\mathbf{x}_* - \mathbf{x}_{K+1}\|)\\ &\le 2\|\mathbf{x}_1 - \mathbf{x}_{K+1}\|\|\mathbf{x}_* - \mathbf{x}_1\| \le 2r_{K+1}\|\mathbf{x}_* - \mathbf{x}_1\|,\end{aligned}$$

where we use $r_{K+1} = r \vee \max_{k\in[K+1]} \|\mathbf{x}_1 - \mathbf{x}_k\| \ge \|\mathbf{x}_1 - \mathbf{x}_{K+1}\|$ in the last step. If $\|\mathbf{x}_* - \mathbf{x}_1\| < \|\mathbf{x}_* - \mathbf{x}_{K+1}\|$, there is still $\|\mathbf{x}_* - \mathbf{x}_1\|^2 - \|\mathbf{x}_* - \mathbf{x}_{K+1}\|^2 \le 0 \le 2r_{K+1}\|\mathbf{x}_* - \mathbf{x}_1\|$. Hence, we always have

$$\|\mathbf{x}_* - \mathbf{x}_1\|^2 - \|\mathbf{x}_* - \mathbf{x}_{K+1}\|^2 \le 2r_{K+1}\|\mathbf{x}_* - \mathbf{x}_1\|.$$

Now there is

$$\begin{aligned}F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\le \frac{r_{K+1}\|\mathbf{x}_* - \mathbf{x}_1\|}{n\sum_{k=1}^K \eta_k} + 3\bar{G}^2 n \sum_{k=1}^K \frac{\eta_k^2}{\sum_{\ell=k}^K \eta_\ell}\\ &= \frac{r_{K+1}\|\mathbf{x}_* - \mathbf{x}_1\|}{n\sum_{k=1}^K r_k\widetilde{\eta}_k} + 3\bar{G}^2 n \sum_{k=1}^K \frac{r_k^2\widetilde{\eta}_k^2}{\sum_{\ell=k}^K r_\ell\widetilde{\eta}_\ell}\\ &\overset{(a)}{\le} \frac{r_{K+1}\|\mathbf{x}_* - \mathbf{x}_1\|}{n\sum_{k=1}^K r_k\widetilde{\eta}_k} + 3r_K\bar{G}^2 n \sum_{k=1}^K \frac{\widetilde{\eta}_k^2}{\sum_{\ell=k}^K \widetilde{\eta}_\ell}\\ &\overset{(b)}{\le} \frac{r_{K+1}}{\sum_{k=1}^K r_k}\cdot\frac{\|\mathbf{x}_* - \mathbf{x}_1\|}{n\widetilde{\eta}_K} + 3r_K\bar{G}^2 n \sum_{k=1}^K \frac{\widetilde{\eta}_k^2}{\sum_{\ell=k}^K \widetilde{\eta}_\ell},\end{aligned} \tag{13}$$

where (a) is due to $\frac{r_k^2\widetilde{\eta}_k^2}{\sum_{\ell=k}^K r_\ell\widetilde{\eta}_\ell} \le \frac{r_k\widetilde{\eta}_k^2}{\sum_{\ell=k}^K \widetilde{\eta}_\ell} \le \frac{r_K\widetilde{\eta}_k^2}{\sum_{\ell=k}^K \widetilde{\eta}_\ell}, \forall k \in [K]$ because r_k is non-decreasing and (b) is by $\widetilde{\eta}_K \le \widetilde{\eta}_k, \forall k \in [K]$ because $\widetilde{\eta}_k$ is non-increasing.

From $\widetilde{\eta}_k = \frac{c}{n\bar{G}\sqrt{6(1+\delta^{-1})k(1+\log k)^{1+\delta}}}$, we have

$$
\begin{aligned}
\sum_{k=1}^{K} \frac{\widetilde{\eta}_k^2}{\sum_{\ell=k}^{K} \widetilde{\eta}_\ell} &= \frac{c}{n\bar{G}\sqrt{6(1+\delta^{-1})}} \sum_{k=1}^{K} \frac{1}{k(1+\log k)^{1+\delta} \sum_{\ell=k}^{K} 1/\sqrt{\ell(1+\log \ell)^{1+\delta}}} \\
&\leq \frac{c}{n\bar{G}} \sqrt{\frac{(1+\log K)^{1+\delta}}{6(1+\delta^{-1})}} \sum_{k=1}^{K} \frac{1}{k(1+\log k)^{1+\delta} \sum_{\ell=k}^{K} 1/\sqrt{\ell}} \\
&\overset{(c)}{\leq} \frac{c}{n\bar{G}} \sqrt{\frac{(1+\log K)^{1+\delta}}{6(1+\delta^{-1})}} \sum_{k=1}^{K} \frac{1}{2k(1+\log k)^{1+\delta}(\sqrt{K+1}-\sqrt{k})} \\
&\overset{(d)}{\leq} \frac{c}{n\bar{G}} \sqrt{\frac{(1+\log K)^{1+\delta}}{6(1+\delta^{-1})(K+1)}} \sum_{k=1}^{K} \frac{1}{(1+\log k)^{1+\delta}} \left(\frac{1}{k} + \frac{1}{K+1-k}\right) \\
&\overset{(e)}{\leq} \frac{c}{n\bar{G}} \sqrt{\frac{(1+\log K)^{1+\delta}}{6(1+\delta^{-1})(K+1)}} \sum_{k=1}^{K} \frac{2}{k(1+\log k)^{1+\delta}} \\
&\overset{(f)}{\leq} \frac{c}{n\bar{G}} \sqrt{\frac{2(1+\delta^{-1})(1+\log K)^{1+\delta}}{3K}},
\end{aligned} \tag{14}
$$

where (c) is by $\sum_{\ell=k}^{K} 1/\sqrt{\ell} \geq \int_{k}^{K+1} 1/\sqrt{\ell}\mathrm{d}\ell = 2\left(\sqrt{K+1}-\sqrt{k}\right)$ and (d) is due to

$$
\frac{1}{2k(\sqrt{K+1}-\sqrt{k})} = \frac{\sqrt{K+1}+\sqrt{k}}{2k(K+1-k)} \leq \frac{\sqrt{K+1}}{k(K+1-k)} = \frac{1}{\sqrt{K+1}}\left(\frac{1}{k} + \frac{1}{K+1-k}\right).
$$

(e) is by noticing that both $\frac{1}{k}$ and $\frac{1}{(1+\log k)^{1+\delta}}$ are decreasing sequences, then by rearrangement inequality

$$
\sum_{k=1}^{K} \frac{1}{(1+\log k)^{1+\delta}} \times \frac{1}{K+1-k} \leq \sum_{k=1}^{K} \frac{1}{(1+\log k)^{1+\delta}} \times \frac{1}{k}.
$$

(f) is because

$$
\begin{aligned}
\sum_{k=1}^{K} \frac{1}{k(1+\log k)^{1+\delta}} \leq \sum_{k=1}^{\infty} \frac{1}{k(1+\log k)^{1+\delta}} &\leq 1 + \int_{1}^{\infty} \frac{1}{k(1+\log k)^{1+\delta}}\mathrm{d}k \\
&= 1 + \left(-\delta^{-1}(1+\log k)^{-\delta}\right)|_1^\infty = 1 + \delta^{-1}.
\end{aligned} \tag{15}
$$

Next let us check the conditions in Lemma D.7. For Condition 1, we notice that $\widetilde{\eta}_k$ is positive defined on $\mathbb{N}$ from its definition. For Condition 2, we compute

$$
\sum_{k=1}^{\infty} 6\bar{G}^2 n^2 \widetilde{\eta}_k^2 = \frac{c^2}{1+\delta^{-1}} \sum_{k=1}^{\infty} \frac{1}{k(1+\log k)^{1+\delta}} \overset{(15)}{\leq} c^2 < 1.
$$

Then we can invoke Lemma D.7 to get

$$
\|\mathbf{x}_k - \mathbf{x}_1\| \leq \frac{2}{1-c}\|\mathbf{x}_* - \mathbf{x}_1\| + \frac{c}{1-c} r, \forall k \in \mathbb{N},
$$

which implies

$$
\begin{aligned}
r_K = r \vee \max_{k\in[K]} \|\mathbf{x}_k - \mathbf{x}_1\| &\leq r + \frac{2}{1-c}\|\mathbf{x}_* - \mathbf{x}_1\| + \frac{c}{1-c} r \\
&= \frac{2\|\mathbf{x}_* - \mathbf{x}_1\| + r}{1-c} \leq \frac{3}{1-c}\left(\|\mathbf{x}_* - \mathbf{x}_1\| \vee r\right).
\end{aligned} \tag{16}
$$

Plugging (14) and (16) into (13) to get

$$\begin{aligned}
F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\le&\frac{r_{K+1}}{\sum_{k=1}^{K}r_k}\cdot\frac{\|\mathbf{x}_*-\mathbf{x}_1\|}{n\widetilde{\eta}_K}+\frac{3c}{1-c}\bar{G}\left(\|\mathbf{x}_*-\mathbf{x}_1\|\vee r\right)\sqrt{\frac{6(1+\delta^{-1})(1+\log K)^{1+\delta}}{K}}\\
=&\frac{r_{K+1}}{\sum_{k=1}^{K}r_k}\cdot\frac{\|\mathbf{x}_*-\mathbf{x}_1\|\,\bar{G}\sqrt{6(1+\delta^{-1})K(1+\log K)^{1+\delta}}}{c}\\
&+\frac{3c}{1-c}\bar{G}\left(\|\mathbf{x}_*-\mathbf{x}_1\|\vee r\right)\sqrt{\frac{6(1+\delta^{-1})(1+\log K)^{1+\delta}}{K}}\\
\le&\left(\frac{r_{K+1}}{c\sum_{k=1}^{K}r_k/K}+\frac{3c}{1-c}\right)\bar{G}\left(\|\mathbf{x}_*-\mathbf{x}_1\|\vee r\right)\sqrt{\frac{6(1+\delta^{-1})(1+\log K)^{1+\delta}}{K}},
\end{aligned}\tag{17}$$

where we use the definition of $\widetilde{\eta}_k=\frac{c}{n\bar{G}\sqrt{6(1+\delta^{-1})k(1+\log k)^{1+\delta}}}$ in the second line.

Finally, note that r_k is non-decreasing with a uniform upper bound $\frac{3}{1-c}\left(\|\mathbf{x}_*-\mathbf{x}_1\|\vee r\right)$ by (16). Hence, there exists $\widetilde{r}=\lim_{k\to\infty}r_k<\infty$, which implies

$$\lim_{K\to\infty}\frac{r_{K+1}}{\sum_{k=1}^{K}r_k/K}=\frac{\widetilde{r}}{\widetilde{r}}=1.$$

Thus, for large enough K, $\frac{r_{K+1}}{\sum_{k=1}^{K}r_k/K}=\mathcal{O}(1)$. Combining (17), the following bound holds asymptotically

$$F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\le\mathcal{O}\left(\frac{3c^2-c+1}{c(1-c)}\bar{G}\left(\|\mathbf{x}_*-\mathbf{x}_1\|\vee r\right)\sqrt{\frac{(1+\delta^{-1})(1+\log K)^{1+\delta}}{K}}\right).$$

□

Finally, we turn the attention to the case of strongly convex $\psi(\mathbf{x})$. Instead of $\eta_k=\frac{2}{n\mu_\psi k}$ presented in Theorem 4.10, we show that $\eta_k=\frac{m}{n\mu_\psi k}$ for any $m\in\mathbb{N}$ guarantees the last-iterate convergence rate $O(m\log K/K)$.

Theorem C.5. *(Full version of Theorem 4.10) Under Assumptions 3.1, 3.2 (with $\mu_\psi>0$) and 3.5, regardless of how the permutation σ_k is generated in every epoch, taking the stepsize $\eta_k=\frac{m}{n\mu_\psi k},\forall k\in[K]$ where $m\in\mathbb{N}$ can be chosen arbitrarily, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1})-F_*\le\mathcal{O}\left(\frac{\mu_\psi D^2}{\binom{K+m}{m}}+\frac{\bar{G}^2m\log K}{\mu_\psi K}\right).$$

Proof. We invoke Lemma D.6 to get

$$\begin{aligned}
F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)&\le\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2-\|\mathbf{x}_*-\mathbf{x}_{K+1}\|^2}{2n\sum_{k=1}^{K}\gamma_k}+3\bar{G}^2n\sum_{k=1}^{K}\frac{\gamma_k\eta_k}{\sum_{\ell=k}^{K}\gamma_\ell}\\
&\le\frac{\|\mathbf{x}_*-\mathbf{x}_1\|^2}{2n\sum_{k=1}^{K}\gamma_k}+3\bar{G}^2n\sum_{k=1}^{K}\frac{\gamma_k\eta_k}{\sum_{\ell=k}^{K}\gamma_\ell},
\end{aligned}$$

where $\gamma_k=\eta_k\prod_{\ell=2}^{k}(1+n\eta_{\ell-1}\mu_\psi),\forall k\in[K]$. When $\eta_k=\frac{m}{n\mu_\psi k},\forall k\in[K]$, we have

$$\gamma_k=\frac{m}{n\mu_\psi k}\prod_{\ell=2}^{k}\frac{\ell-1+m}{\ell-1}=\frac{1}{n\mu_\psi}\cdot\frac{(k+m-1)!}{(m-1)!k!}=\frac{\binom{k+m-1}{m-1}}{n\mu_\psi}.$$

Thus

$$
\begin{aligned}
F(\mathbf{x}_{K+1})-F(\mathbf{x}_*) &\le \frac{\mu_\psi\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{2\sum_{k=1}^K\binom{k+m-1}{m-1}}+\frac{3\bar{G}^2m}{\mu_\psi}\sum_{k=1}^K\frac{\binom{k+m-1}{m-1}}{k\sum_{\ell=k}^K\binom{\ell+m-1}{m-1}}\\
&\overset{(a)}{=}\frac{\mu_\psi\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{2\left(\binom{K+m}{m}-1\right)}+\frac{3\bar{G}^2m}{\mu_\psi}\sum_{k=1}^K\frac{\binom{k+m-1}{m-1}}{k\sum_{\ell=k}^K\binom{\ell+m-1}{m-1}}\\
&\overset{(b)}{\le}\frac{\mu_\psi\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{2\left(\binom{K+m}{m}-1\right)}+\frac{3\bar{G}^2m}{\mu_\psi}\sum_{k=1}^K\frac{1}{k(K-k+1)}\\
&=\frac{\mu_\psi\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{2\left(\binom{K+m}{m}-1\right)}+\frac{3\bar{G}^2m}{\mu_\psi(K+1)}\sum_{k=1}^K\frac{1}{k}+\frac{1}{K-k+1}\\
&=\mathcal{O}\left(\frac{\mu_\psi\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{\binom{K+m}{m}}+\frac{\bar{G}^2m\log K}{\mu_\psi K}\right).
\end{aligned}
$$

where (a) is due to

$$
\sum_{k=1}^K\binom{k+m-1}{m-1}=\sum_{k=1}^K\binom{k+m}{m}-\binom{k+m-1}{m}=\binom{K+m}{m}-1,
$$

and (b) is by $\binom{\ell+m-1}{m-1}\ge\binom{k+m-1}{m-1},\forall\ell\ge k$. □

D. Theoretical Analysis

In this section, we provide our analysis and several lemmas.

D.1. Core Lemmas

This subsection contains three core lemmas that play fundamental roles in the analysis.

Inspired by (Mishchenko et al., 2020), we first introduce the following descent inequality, which provides the progress on the objective value in one epoch. We emphasize that Lemma D.1 holds for any point $\mathbf{z}\in\mathbb{R}^d$ rather than only $\mathbf{x}_*$. This important fact ensures that we can use the key idea proposed by (Zamani & Glineur, 2023) and later developed by (Liu & Zhou, 2024). However, some barriers would appear if we apply the technique used in (Zamani & Glineur, 2023) directly. Hence, our analysis departs from the previous paper and needs new tools, e.g., Lemma E.2. The reader could refer to Lemmas D.4 and D.6 for how our analysis proceeds.

Lemma D.1. *Under Assumption 3.2, for any $k\in[K]$, permutation σ_k and $\mathbf{z}\in\mathbb{R}^d$, Algorithm 1 guarantees*

$$
\begin{aligned}
F(\mathbf{x}_{k+1})-F(\mathbf{z})\le&\frac{\left\|\mathbf{z}-\mathbf{x}_k\right\|^2}{2n\eta_k}-\left(\frac{1}{\eta_k}+n\mu_\psi\right)\frac{\left\|\mathbf{z}-\mathbf{x}_{k+1}\right\|^2}{2n}-\frac{\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|^2}{2n\eta_k}\\
&+\frac{1}{n}\sum_{i=1}^n B_{\sigma_k^i}(\mathbf{x}_{k+1},\mathbf{x}_k^i)-B_{\sigma_k^i}(\mathbf{z},\mathbf{x}_k^i).
\end{aligned}
$$

Proof. It is enough to only consider the case $\mathbf{z}\in\mathrm{dom}\psi$. Let $\mathbf{g}_k\triangleq\sum_{i=1}^n\nabla f_{\sigma_k^i}(\mathbf{x}_k^i)$, then $\mathbf{x}_k^{n+1}=\mathbf{x}_k^1-\eta_k\mathbf{g}_k=\mathbf{x}_k-\eta_k\mathbf{g}_k$. Observe that

$$
\mathbf{x}_{k+1}=\mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d}n\psi(\mathbf{x})+\frac{\left\|\mathbf{x}-\mathbf{x}_k^{n+1}\right\|^2}{2\eta_k}=\mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d}n\psi(\mathbf{x})+\langle\mathbf{g}_k,\mathbf{x}-\mathbf{x}_k\rangle+\frac{\left\|\mathbf{x}-\mathbf{x}_k\right\|^2}{2\eta_k}.
$$

By the first-order optimality condition, there exists $\nabla\psi(\mathbf{x}_{k+1})\in\partial\psi(\mathbf{x}_{k+1})$ such that

$$
\mathbf{0}=n\nabla\psi(\mathbf{x}_{k+1})+\mathbf{g}_k+\frac{\mathbf{x}_{k+1}-\mathbf{x}_k}{\eta_k}.
$$

Therefore for any $\mathbf{z} \in \mathrm{dom}\psi$

$$\begin{aligned}
\langle \mathbf{g}_k, \mathbf{x}_{k+1} - \mathbf{z} \rangle &= n\langle \nabla\psi(\mathbf{x}_{k+1}), \mathbf{z} - \mathbf{x}_{k+1} \rangle + \frac{\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{x}_{k+1} - \mathbf{z} \rangle}{\eta_k} \\
&\overset{(a)}{\leq} n\left(\psi(\mathbf{z}) - \psi(\mathbf{x}_{k+1}) - \frac{\mu_\psi}{2}\left\|\mathbf{z} - \mathbf{x}_{k+1}\right\|^2\right) + \frac{\langle \mathbf{x}_k - \mathbf{x}_{k+1}, \mathbf{x}_{k+1} - \mathbf{z} \rangle}{\eta_k} \\
&= n\left(\psi(\mathbf{z}) - \psi(\mathbf{x}_{k+1}) - \frac{\mu_\psi}{2}\left\|\mathbf{z} - \mathbf{x}_{k+1}\right\|^2\right) + \frac{\left\|\mathbf{z} - \mathbf{x}_k\right\|^2 - \left\|\mathbf{z} - \mathbf{x}_{k+1}\right\|^2 - \left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2}{2\eta_k} \\
&= n\left(\psi(\mathbf{z}) - \psi(\mathbf{x}_{k+1})\right) + \frac{\left\|\mathbf{z} - \mathbf{x}_k\right\|^2}{2\eta_k} - \left(\frac{1}{\eta_k} + n\mu_\psi\right)\frac{\left\|\mathbf{z} - \mathbf{x}_{k+1}\right\|^2}{2} - \frac{\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2}{2\eta_k}, \qquad (18)
\end{aligned}$$

where (a) is by the strong convexity of ψ. Note that (recall $B_{\sigma_k^i} \triangleq B_{f_{\sigma_k^i}}$)

$$\langle \nabla f_{\sigma_k^i}(\mathbf{x}_k^i), \mathbf{x}_{k+1} - \mathbf{z} \rangle = f_{\sigma_k^i}(\mathbf{x}_{k+1}) - f_{\sigma_k^i}(\mathbf{z}) - B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) + B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i),$$

which implies

$$\begin{aligned}
\langle \mathbf{g}_k, \mathbf{x}_{k+1} - \mathbf{z} \rangle &= \sum_{i=1}^n \langle \nabla f_{\sigma_k^i}(\mathbf{x}_k^i), \mathbf{x}_{k+1} - \mathbf{z} \rangle = \sum_{i=1}^n f_{\sigma_k^i}(\mathbf{x}_{k+1}) - f_{\sigma_k^i}(\mathbf{z}) - B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) + B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i) \\
&= n\left(f(\mathbf{x}_{k+1}) - f(\mathbf{z})\right) - \sum_{i=1}^n B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i). \qquad (19)
\end{aligned}$$

Plugging (19) into (18) and rearranging the terms to obtain

$$\begin{aligned}
F(\mathbf{x}_{k+1}) - F(\mathbf{z}) \leq &\frac{\left\|\mathbf{z} - \mathbf{x}_k\right\|^2}{2n\eta_k} - \left(\frac{1}{\eta_k} + n\mu_\psi\right)\frac{\left\|\mathbf{z} - \mathbf{x}_{k+1}\right\|^2}{2n} - \frac{\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2}{2n\eta_k} \\
&+ \frac{1}{n}\sum_{i=1}^n B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i).
\end{aligned}$$

□

Note that Lemma D.1 includes a term $\frac{1}{n}\sum_{i=1}^n B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i)$, which can be bounded easily when $n = 1$. However, for a general n, we need to make extra efforts to find a proper bound on it.

In Lemma D.2, we first provide the bound on the extra term for smooth functions. There are several points we want to discuss here. First, we need the stepsize to satisfy $\eta_k \leq \frac{1}{2n\sqrt{\bar{L}L^*}}$. Similar requirements also existed in previous works (e.g., (Mishchenko et al., 2020; Cha et al., 2023)). Besides the normal term $\bar{L}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2$, there are two unusual quantities appearing in the R.H.S. of the inequality. One is the term $B_f(\mathbf{z}, \mathbf{x}_*)$, which leads to the main difficulty in the analysis (see Lemma D.4 for details). The other is the residual R_k. Fortunately, it can bounded as shown in Lemma E.1 in Section E.

Lemma D.2. *Under Assumptions 3.1, 3.2 and 3.3, suppose $\eta_k \leq \frac{1}{2n\sqrt{\bar{L}L^*}}, \forall k \in [K]$, then for any $k \in [K]$, permutation σ_k and $\mathbf{z} \in \mathbb{R}^d$, Algorithm 1 guarantees*

$$\frac{1}{n}\sum_{i=1}^n B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i) \leq \bar{L}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 + 8\eta_k^2 n^2 \bar{L}^2 B_f(\mathbf{z}, \mathbf{x}_*) + 4\eta_k^2 R_k,$$

where $R_k \triangleq \sum_{i=2}^n \frac{L_{\sigma_k^i}}{n}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2, \forall k \in [K]$.

Proof. By Lemma 3.6, there are

$$\begin{aligned}
B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) &\leq \frac{L_{\sigma_k^i}}{2}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k^i\right\|^2 \leq L_{\sigma_k^i}\left(\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 + \left\|\mathbf{x}_k^i - \mathbf{x}_k\right\|^2\right), \\
B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i) &\geq \frac{\left\|\nabla f_{\sigma_k^i}(\mathbf{x}_k^i) - \nabla f_{\sigma_k^i}(\mathbf{z})\right\|^2}{2L_{\sigma_k^i}}.
\end{aligned}$$

Therefore

$$\frac{1}{n}\sum_{i=1}^{n}B_{\sigma_k^i}(\mathbf{x}_{k+1},\mathbf{x}_k^i)-B_{\sigma_k^i}(\mathbf{z},\mathbf{x}_k^i)\leq\frac{1}{n}\sum_{i=1}^{n}L_{\sigma_k^i}\left(\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2+\left\|\mathbf{x}_k^i-\mathbf{x}_k\right\|^2\right)-\frac{\left\|\nabla f_{\sigma_k^i}(\mathbf{x}_k^i)-\nabla f_{\sigma_k^i}(\mathbf{z})\right\|^2}{2L_{\sigma_k^i}}$$
$$=\bar{L}\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|^2+\frac{1}{n}\sum_{i=1}^{n}L_{\sigma_k^i}\left\|\mathbf{x}_k^i-\mathbf{x}_k\right\|^2-\frac{\left\|\nabla f_{\sigma_k^i}(\mathbf{x}_k^i)-\nabla f_{\sigma_k^i}(\mathbf{z})\right\|^2}{2L_{\sigma_k^i}}. \quad (20)$$

Note that $\mathbf{x}_k=\mathbf{x}_k^1$, hence

$$\sum_{i=1}^{n}L_{\sigma_k^i}\left\|\mathbf{x}_k^i-\mathbf{x}_k\right\|^2=\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\mathbf{x}_k^i-\mathbf{x}_k^1\right\|^2=\eta_k^2\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)\right\|^2$$
$$=\eta_k^2\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)-\nabla f_{\sigma_k^j}(\mathbf{z})+\nabla f_{\sigma_k^j}(\mathbf{z})-\nabla f_{\sigma_k^j}(\mathbf{x}_*)+\nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2$$
$$\leq\eta_k^2\sum_{i=2}^{n}L_{\sigma_k^i}\left(2\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)-\nabla f_{\sigma_k^j}(\mathbf{z})\right\|^2+4\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{z})-\nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2+4\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2\right)$$
$$=2\eta_k^2\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)-\nabla f_{\sigma_k^j}(\mathbf{z})\right\|^2+4\eta_k^2\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{z})-\nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2$$
$$+4\eta_k^2\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2. \quad (21)$$

We bound

$$\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)-\nabla f_{\sigma_k^j}(\mathbf{z})\right\|^2$$
$$\leq\sum_{i=2}^{n}\sum_{j=1}^{i-1}L_{\sigma_k^i}(i-1)\left\|\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)-\nabla f_{\sigma_k^j}(\mathbf{z})\right\|^2=\sum_{j=1}^{n-1}\left(\sum_{i=j+1}^{n}L_{\sigma_k^i}(i-1)\right)\left\|\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)-\nabla f_{\sigma_k^j}(\mathbf{z})\right\|^2$$
$$\leq\sum_{j=1}^{n-1}n^2\bar{L}\left\|\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)-\nabla f_{\sigma_k^j}(\mathbf{z})\right\|^2\leq\sum_{i=1}^{n}n^2\bar{L}\left\|\nabla f_{\sigma_k^i}(\mathbf{x}_k^i)-\nabla f_{\sigma_k^i}(\mathbf{z})\right\|^2, \quad (22)$$

and

$$\sum_{i=2}^{n}L_{\sigma_k^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma_k^j}(\mathbf{z})-\nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2$$
$$\overset{(a)}{\leq}\sum_{i=2}^{n}L_{\sigma_k^i}\times 2\left(\sum_{j=1}^{i-1}L_{\sigma_k^j}\right)\left(\sum_{\ell=1}^{i-1}B_{\sigma_k^\ell}(\mathbf{z},\mathbf{x}_*)\right)=\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{\ell=1}^{i-1}2L_{\sigma_k^i}L_{\sigma_k^j}B_{\sigma_k^\ell}(\mathbf{z},\mathbf{x}_*)$$
$$=\sum_{j=1}^{n-1}\sum_{\ell=1}^{n-1}\left(\sum_{i=j\vee\ell+1}^{n}2L_{\sigma_k^i}\right)L_{\sigma_k^j}B_{\sigma_k^\ell}(\mathbf{z},\mathbf{x}_*)\leq 2n\bar{L}\sum_{j=1}^{n-1}\sum_{\ell=1}^{n-1}L_{\sigma_k^j}B_{\sigma_k^\ell}(\mathbf{z},\mathbf{x}_*)$$
$$\leq 2n^2\bar{L}^2\sum_{\ell=1}^{n-1}B_{\sigma_k^\ell}(\mathbf{z},\mathbf{x}_*)\overset{(b)}{\leq}2n^2\bar{L}^2\sum_{\ell=1}^{n}B_{\sigma_k^\ell}(\mathbf{z},\mathbf{x}_*)=2n^3\bar{L}^2B_f(\mathbf{z},\mathbf{x}_*), \quad (23)$$

where (a) is due to Lemma 3.6 and (b) is by $B_{\sigma_k^\ell} \geq 0$.

Plugging (22) and (23) into (21) to get

$$\sum_{i=1}^{n} L_{\sigma_k^i} \left\|\mathbf{x}_k^i - \mathbf{x}_k\right\|^2 \leq \sum_{i=1}^{n} 2\eta_k^2 n^2 \bar{L} \left\|\nabla f_{\sigma_k^i}(\mathbf{x}_k^i) - \nabla f_{\sigma_k^i}(\mathbf{z})\right\|^2 + 8\eta_k^2 n^3 \bar{L}^2 B_f(\mathbf{z}, \mathbf{x}_*) + 4\eta_k^2 \sum_{i=2}^{n} L_{\sigma_k^i} \left\|\sum_{j=1}^{i-1} \nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2.$$

Combing (20), we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i) \leq & \bar{L} \left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 + 8\eta_k^2 n^2 \bar{L}^2 B_f(\mathbf{z}, \mathbf{x}_*) + 4\eta_k^2 \sum_{i=2}^{n} \frac{L_{\sigma_k^i}}{n} \left\|\sum_{j=1}^{i-1} \nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2 \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(2\eta_k^2 n^2 \bar{L} - \frac{1}{2L_{\sigma_k^i}}\right)\left\|\nabla f_{\sigma_k^i}(\mathbf{x}_k^i) - \nabla f_{\sigma_k^i}(\mathbf{z})\right\|^2 \\
\leq & \bar{L} \left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 + 8\eta_k^2 n^2 \bar{L}^2 B_f(\mathbf{z}, \mathbf{x}_*) + 4\eta_k^2 \sum_{i=2}^{n} \frac{L_{\sigma_k^i}}{n} \left\|\sum_{j=1}^{i-1} \nabla f_{\sigma_k^j}(\mathbf{x}_*)\right\|^2,
\end{aligned}$$

where the last inequality is due to $2\eta_k^2 n^2 \bar{L} \leq \frac{1}{2L^*} \leq \frac{1}{2L_{\sigma_k^i}}, \forall i \in [n]$ from the condition $\eta_k \leq \frac{1}{2n\sqrt{\bar{L}L^*}}, \forall k \in [K]$. □

Next, let us consider the Lipschitz case, i.e., $\|\nabla f_i(\mathbf{x})\| \leq G_i, \forall \mathbf{x} \in \mathbb{R}^d, i \in [n]$. Unlike Lemma D.2, the following lemma shows that the extra term $\frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i)$ can be always bounded by $\bar{G}, n, \eta_k$ and $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ now regardless of what $\mathbf{z}$ is.

Lemma D.3. *Under Assumptions 3.2 and 3.5, for any $k \in [K]$, permutation σ_k and $\mathbf{z} \in \mathbb{R}^d$, Algorithm 1 guarantees*

$$\frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i) \leq 2\bar{G}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| + \bar{G}^2 n\eta_k.$$

Proof. By noticing $B_{\sigma_k^i} \geq 0$ because of the covnexity of $f_{\sigma_k^i}(\mathbf{x})$ from Assumption 3.2, we have

$$\frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{z}, \mathbf{x}_k^i) \leq \frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i). \tag{24}$$

Under Assumption 3.5

$$\begin{aligned}
B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) &= f_{\sigma_k^i}(\mathbf{x}_{k+1}) - f_{\sigma_k^i}(\mathbf{x}_k^i) - \langle \nabla f_{\sigma_k^i}(\mathbf{x}_k^i), \mathbf{x}_{k+1} - \mathbf{x}_k^i\rangle \\
&\overset{(a)}{\leq} \langle \nabla f_{\sigma_k^i}(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_k^i\rangle - \langle \nabla f_{\sigma_k^i}(\mathbf{x}_k^i), \mathbf{x}_{k+1} - \mathbf{x}_k^i\rangle \\
&\leq \left(\left\|\nabla f_{\sigma_k^i}(\mathbf{x}_{k+1})\right\| + \left\|\nabla f_{\sigma_k^i}(\mathbf{x}_k^i)\right\|\right)\left\|\mathbf{x}_{k+1} - \mathbf{x}_k^i\right\| \\
&\leq 2G_{\sigma_k^i}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k^i\right\| \leq 2G_{\sigma_k^i}\left(\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| + \left\|\mathbf{x}_k^i - \mathbf{x}_k\right\|\right),
\end{aligned}$$

where (a) is by the convexity of $f_{\sigma_k^i}(\mathbf{x})$ from Assumption 3.2. Therefore

$$\frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) \leq \frac{1}{n}\sum_{i=1}^{n} 2G_{\sigma_k^i}\left(\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| + \left\|\mathbf{x}_k^i - \mathbf{x}_k\right\|\right) = 2\bar{G}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\| + \frac{2}{n}\sum_{i=1}^{n} G_{\sigma_k^i}\left\|\mathbf{x}_k^i - \mathbf{x}_k\right\|. \tag{25}$$

Note that $\mathbf{x}_k = \mathbf{x}_k^1$, hence

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n} G_{\sigma_k^i}\left\|\mathbf{x}_k^i - \mathbf{x}_k\right\| &= \frac{2}{n}\sum_{i=2}^{n} G_{\sigma_k^i}\left\|\mathbf{x}_k^i - \mathbf{x}_k^1\right\| \leq \frac{2}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1} G_{\sigma_k^i}\left\|\mathbf{x}_k^{j+1} - \mathbf{x}_k^j\right\| = \frac{2\eta_k}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1} G_{\sigma_k^i}\left\|\nabla f_{\sigma_k^j}(\mathbf{x}_k^j)\right\| \\
&\leq \frac{2\eta_k}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1} G_{\sigma_k^i} G_{\sigma_k^j} = \frac{\eta_k}{n}\left[\left(\sum_{i=1}^{n} G_{\sigma_k^i}\right)^2 - \sum_{i=1}^{n} G_{\sigma_k^i}^2\right] \leq \bar{G}^2(n-1)\eta_k \leq \bar{G}^2 n\eta_k.
\end{aligned} \tag{26}$$

Combining (24), (25) and (26), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}B_{\sigma_k^i}(\mathbf{x}_{k+1},\mathbf{x}_k^i)-B_{\sigma_k^i}(\mathbf{z},\mathbf{x}_k^i)\leq 2\bar{G}\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|+\bar{G}^2 n\eta_k.$$

□

D.2. Smooth Functions

In this subsection, we deal with smooth functions and will prove two important results.

First, we present the most important result for the smooth case, Lemma D.4. The main difficulty is to deal with the extra term $B_f(\mathbf{z},\mathbf{x}_*)$ after using Lemmas D.1 and D.2. Suppose we follow the same way used in the previous works, i.e., setting $\mathbf{z}=\mathbf{z}_k$ for a carefully designed sequence $\{\mathbf{z}_k,\forall k\in[K]\}$. We can only bound $F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)$ by $\mathcal{O}(\sum_{k=1}^{K}B_f(\mathbf{z}_k,\mathbf{x}_*))$.

A key observation is that $B_f(\mathbf{z}_k,\mathbf{x}_*)=\mathcal{O}(\sum_{\ell=1}^{k}B_f(\mathbf{x}_\ell,\mathbf{x}_*))$ because $\mathbf{z}_k$ will be taken as a convex combination of $\mathbf{x}_*,\mathbf{x}_1,\cdots,\mathbf{x}_k$. Thus, we can bound $F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)$ by $\mathcal{O}(\sum_{k=1}^{K}B_f(\mathbf{x}_k,\mathbf{x}_*))$. However, this is not enough since we still need the bound on $B_f(\mathbf{x}_k,\mathbf{x}_*)$ for every $k\in[K]$. Temporarily assume $\psi(\mathbf{x})=0$, we can find there is $B_f(\mathbf{x}_k,\mathbf{x}_*)=F(\mathbf{x}_k)-F(\mathbf{x}_*)$. So if $F(\mathbf{x}_k)-F(\mathbf{x}_*)$ are small enough for every $k\in[K]$, we can hope that $F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)$ is also small. This thought inspires us to bound $F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)$ for every $k\in[K]$ instead of only bounding $F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)$. Hence, departing from the existing works that only bound the function value gap once for time K, we prove the following anytime inequality, which can finally help us prove the last-iterate rate (see the proof of Theorem C.1). In addition, we would like to emphasize that the sequence $\{\mathbf{z}_\ell,\forall \ell\in[k]\}$ now is defined differently for every $k\in[K]$ as mentioned in Section 5.

Lemma D.4. *Under Assumptions 3.1, 3.2 and 3.3, suppose* $\eta_k\leq\frac{1}{2n\sqrt{\bar{L}L^*}},\forall k\in[K]$, *then for any permutation* $\sigma_k,\forall k\in[K]$, *Algorithm 1 guarantees*

$$F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)\leq\frac{\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{2n\sum_{\ell=1}^{k}\eta_\ell}+\sum_{\ell=1}^{k}\frac{4\eta_\ell^3R_\ell}{\sum_{s=\ell}^{k}\eta_s}+\sum_{\ell=2}^{k}\frac{8n^2\bar{L}^2\eta_{\ell-1}\left(\sum_{s=\ell}^{k}\eta_s^3\right)}{\left(\sum_{s=\ell}^{k}\eta_s\right)\left(\sum_{s=\ell-1}^{k}\eta_s\right)}B_f(\mathbf{x}_\ell,\mathbf{x}_*),\forall k\in[K],$$

where R_ℓ *is defined in Lemma D.2.*

Proof. Fix $k\in[K]$, we define the non-decreasing sequence

$$v_s\triangleq\frac{\eta_k}{\sum_{\ell=s}^{k}\eta_\ell},\forall s\in[k]\,, \tag{27}$$

$$v_0\triangleq v_1, \tag{28}$$

and the auxiliary points $\mathbf{z}_0\triangleq\mathbf{x}_*$ and

$$\mathbf{z}_s\triangleq\left(1-\frac{v_{s-1}}{v_s}\right)\mathbf{x}_s+\frac{v_{s-1}}{v_s}\mathbf{z}_{s-1},\forall s\in[k]\,. \tag{29}$$

Equivalently, we can write $\mathbf{z}_s$ as

$$\mathbf{z}_s\triangleq\frac{v_0}{v_s}\mathbf{x}_*+\sum_{\ell=1}^{s}\frac{v_\ell-v_{\ell-1}}{v_s}\mathbf{x}_\ell,\forall s\in\{0\}\cup[k]\,. \tag{30}$$

Note that $\mathbf{z}_s\in\mathrm{dom}\psi,\forall s\in\{0\}\cup[k]$ since it is a convex combination of $\mathbf{x}_*,\mathbf{x}_1,\cdots\mathbf{x}_s$ due to $v_s\geq v_{s-1}$ and all of which fall into $\mathrm{dom}\psi$.

We invoke Lemma D.1 with $k=s$ (this k is for k in Lemma D.1 and is not our current fixed k) and $\mathbf{z}=\mathbf{z}_s$ to obtain

$$\begin{aligned}
F(\mathbf{x}_{s+1})-F(\mathbf{z}_s) &\le \frac{\|\mathbf{z}_s-\mathbf{x}_s\|^2}{2n\eta_s}-\left(\frac{1}{\eta_s}+n\mu_\psi\right)\frac{\|\mathbf{z}_s-\mathbf{x}_{s+1}\|^2}{2n}-\frac{\|\mathbf{x}_{s+1}-\mathbf{x}_s\|^2}{2n\eta_s}+\frac{1}{n}\sum_{i=1}^{n}B_{\sigma_s^i}(\mathbf{x}_{s+1},\mathbf{x}_s^i)-B_{\sigma_s^i}(\mathbf{z}_s,\mathbf{x}_s^i)\\
&\overset{(a)}{\le} \frac{\|\mathbf{z}_s-\mathbf{x}_s\|^2}{2n\eta_s}-\frac{\|\mathbf{z}_s-\mathbf{x}_{s+1}\|^2}{2n\eta_s}-\frac{\|\mathbf{x}_{s+1}-\mathbf{x}_s\|^2}{2n\eta_s}+\frac{1}{n}\sum_{i=1}^{n}B_{\sigma_s^i}(\mathbf{x}_{s+1},\mathbf{x}_s^i)-B_{\sigma_s^i}(\mathbf{z}_s,\mathbf{x}_s^i)\\
&\overset{(b)}{\le} \frac{\|\mathbf{z}_s-\mathbf{x}_s\|^2}{2n\eta_s}-\frac{\|\mathbf{z}_s-\mathbf{x}_{s+1}\|^2}{2n\eta_s}-\frac{\|\mathbf{x}_{s+1}-\mathbf{x}_s\|^2}{2n\eta_s}+\bar{L}\|\mathbf{x}_{s+1}-\mathbf{x}_s\|^2+8\eta_s^2n^2\bar{L}^2B_f(\mathbf{z}_s,\mathbf{x}_*)+4\eta_s^2R_s\\
&\overset{(c)}{\le} \frac{\|\mathbf{z}_s-\mathbf{x}_s\|^2}{2n\eta_s}-\frac{\|\mathbf{z}_s-\mathbf{x}_{s+1}\|^2}{2n\eta_s}+8\eta_s^2n^2\bar{L}^2B_f(\mathbf{z}_s,\mathbf{x}_*)+4\eta_s^2R_s,
\end{aligned}$$

where (a) is due to $\mu_\psi\ge 0$, (b) is by Lemma D.2 and (c) holds because of $\bar{L}\le\frac{1}{2n\eta_s}$ from the requirment of $\eta_k\le\frac{1}{2n\sqrt{\bar{L}L^*}},\forall k\in[K]$. Note that

$$\|\mathbf{z}_s-\mathbf{x}_s\|^2\overset{(29)}{=}\frac{v_{s-1}^2}{v_s^2}\|\mathbf{z}_{s-1}-\mathbf{x}_s\|^2\le\frac{v_{s-1}}{v_s}\|\mathbf{z}_{s-1}-\mathbf{x}_s\|^2,$$

where the last inequality is due to $v_{s-1}\le v_s$. Hence

$$\begin{aligned}
F(\mathbf{x}_{s+1})-F(\mathbf{z}_s)&\le\frac{\frac{v_{s-1}}{v_s}\|\mathbf{z}_{s-1}-\mathbf{x}_s\|^2}{2n\eta_s}-\frac{\|\mathbf{z}_s-\mathbf{x}_{s+1}\|^2}{2n\eta_s}+8\eta_s^2n^2\bar{L}^2B_f(\mathbf{z}_s,\mathbf{x}_*)+4\eta_s^2R_s\\
\Rightarrow \eta_sv_s\left(F(\mathbf{x}_{s+1})-F(\mathbf{z}_s)\right)&\le\frac{v_{s-1}\|\mathbf{z}_{s-1}-\mathbf{x}_s\|^2}{2n}-\frac{v_s\|\mathbf{z}_s-\mathbf{x}_{s+1}\|^2}{2n}+\underbrace{8\eta_s^3v_sn^2\bar{L}^2B_f(\mathbf{z}_s,\mathbf{x}_*)+4\eta_s^3v_sR_s}_{\triangleq Q_s},
\end{aligned}\tag{31}$$

Summing (31) from $s=1$ to k to obtain

$$\sum_{s=1}^{k}\eta_sv_s\left(F(\mathbf{x}_{s+1})-F(\mathbf{z}_s)\right)\le\frac{v_0\|\mathbf{z}_0-\mathbf{x}_1\|^2}{2n}-\frac{v_k\|\mathbf{z}_k-\mathbf{x}_{k+1}\|^2}{2n}+\sum_{s=1}^{k}Q_s\le\frac{v_0\|\mathbf{x}_*-\mathbf{x}_1\|^2}{2n}+\sum_{s=1}^{k}Q_s,\tag{32}$$

where the last line is by $\mathbf{z}_0=\mathbf{x}_*$ and $\|\mathbf{z}_k-\mathbf{x}_{k+1}\|^2\ge 0$.

By the convexity of $F(\mathbf{x})$ and (30), we can bound

$$F(\mathbf{z}_s)\le\frac{v_0}{v_s}F(\mathbf{x}_*)+\sum_{\ell=1}^{s}\frac{v_\ell-v_{\ell-1}}{v_s}F(\mathbf{x}_\ell)=F(\mathbf{x}_*)+\sum_{\ell=1}^{s}\frac{v_\ell-v_{\ell-1}}{v_s}\left(F(\mathbf{x}_\ell)-F(\mathbf{x}_*)\right),$$

which implies

$$\begin{aligned}
\sum_{s=1}^{k}\eta_sv_s\left(F(\mathbf{x}_{s+1})-F(\mathbf{z}_s)\right)\ge&\sum_{s=1}^{k}\eta_s\left(v_s\left(F(\mathbf{x}_{s+1})-F(\mathbf{x}_*)\right)-\sum_{\ell=1}^{s}(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_\ell)-F(\mathbf{x}_*)\right)\right)\\
=&\sum_{s=1}^{k}\eta_sv_s\left(F(\mathbf{x}_{s+1})-F(\mathbf{x}_*)\right)-\sum_{s=1}^{k}\sum_{\ell=1}^{s}\eta_s(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_\ell)-F(\mathbf{x}_*)\right)\\
=&\sum_{s=1}^{k}\eta_sv_s\left(F(\mathbf{x}_{s+1})-F(\mathbf{x}_*)\right)-\sum_{s=1}^{k}\left(\sum_{\ell=s}^{k}\eta_\ell\right)(v_s-v_{s-1})\left(F(\mathbf{x}_s)-F(\mathbf{x}_*)\right)\\
=&\eta_kv_k\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)\right)-\left(\sum_{\ell=1}^{k}\eta_\ell\right)(v_1-v_0)\left(F(\mathbf{x}_1)-F(\mathbf{x}_*)\right)\\
&+\sum_{s=2}^{k}\left[\eta_{s-1}v_{s-1}-\left(\sum_{\ell=s}^{k}\eta_\ell\right)(v_s-v_{s-1})\right]\left(F(\mathbf{x}_s)-F(\mathbf{x}_*)\right).
\end{aligned}$$

Note that $v_1 \overset{(28)}{=} v_0$ and for $2 \leq s \leq k$

$$\eta_{s-1}v_{s-1} - \left(\sum_{\ell=s}^{k}\eta_\ell\right)(v_s - v_{s-1}) = \left(\sum_{\ell=s-1}^{k}\eta_\ell\right)v_{s-1} - \left(\sum_{\ell=s}^{k}\eta_\ell\right)v_s \overset{(27)}{=} \eta_k - \eta_k = 0.$$

Hence, we know

$$\sum_{s=1}^{k}\eta_s v_s\left(F(\mathbf{x}_{s+1}) - F(\mathbf{z}_s)\right) \geq \eta_k v_k\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*)\right). \tag{33}$$

Plugging (33) into (32), we obtain

$$\begin{aligned}
\eta_k v_k\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*)\right) &\leq \frac{v_0\left\|\mathbf{x}_* - \mathbf{x}_1\right\|^2}{2n} + \sum_{s=1}^{k}Q_s \\
\Rightarrow F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) &\leq \frac{v_0}{\eta_k v_k}\cdot\frac{\left\|\mathbf{x}_* - \mathbf{x}_1\right\|^2}{2n} + \frac{1}{\eta_k v_k}\sum_{s=1}^{k}Q_s \overset{(27),(28)}{=} \frac{\left\|\mathbf{x} - \mathbf{x}_1\right\|^2}{2n\sum_{\ell=1}^{k}\eta_\ell} + \frac{1}{\eta_k}\sum_{s=1}^{k}Q_s \\
&= \frac{\left\|\mathbf{x}_* - \mathbf{x}_1\right\|^2}{2n\sum_{\ell=1}^{k}\eta_\ell} + \underbrace{\frac{1}{\eta_k}\sum_{s=1}^{k}4\eta_s^3 v_s R_s}_{(i)} + \underbrace{\frac{1}{\eta_k}\sum_{s=1}^{k}8\eta_s^3 v_s n^2\bar{L}^2 B_f(\mathbf{z}_s, \mathbf{x}_*)}_{(ii)}.
\end{aligned} \tag{34}$$

For term (i), we have

$$(i) \overset{(27)}{=} \sum_{s=1}^{k}\frac{4\eta_s^3 R_s}{\sum_{\ell=s}^{k}\eta_\ell} = \sum_{\ell=1}^{k}\frac{4\eta_\ell^3 R_\ell}{\sum_{s=\ell}^{k}\eta_s}. \tag{35}$$

For term (ii), by the convexity of the first argument of $B_f(\cdot,\cdot)$ (which is implied by the convexity of $f(\mathbf{x})$) and (30), we first bound

$$B_f(\mathbf{z}_s, \mathbf{x}_*) \leq \frac{v_0}{v_s}B_f(\mathbf{x}_*, \mathbf{x}_*) + \sum_{\ell=1}^{s}\frac{v_\ell - v_{\ell-1}}{v_s}B_f(\mathbf{x}_\ell, \mathbf{x}_*) = \sum_{\ell=1}^{s}\frac{v_\ell - v_{\ell-1}}{v_s}B_f(\mathbf{x}_\ell, \mathbf{x}_*),$$

which implies

$$\begin{aligned}
(ii) &\leq \frac{1}{\eta_k}\sum_{s=1}^{k}\sum_{\ell=1}^{s}8\eta_s^3 n^2\bar{L}^2(v_\ell - v_{\ell-1})B_f(\mathbf{x}_\ell, \mathbf{x}_*) \\
&\overset{(28)}{=} \frac{1}{\eta_k}\sum_{s=2}^{k}\sum_{\ell=2}^{s}8\eta_s^3 n^2\bar{L}^2(v_\ell - v_{\ell-1})B_f(\mathbf{x}_\ell, \mathbf{x}_*) \\
&= \frac{1}{\eta_k}\sum_{\ell=2}^{k}8n^2\bar{L}^2\left(\sum_{s=\ell}^{k}\eta_s^3\right)(v_\ell - v_{\ell-1})B_f(\mathbf{x}_\ell, \mathbf{x}_*) \\
&\overset{(27)}{=} \frac{1}{\eta_k}\sum_{\ell=2}^{k}8n^2\bar{L}^2\left(\sum_{s=\ell}^{k}\eta_s^3\right)\left(\frac{\eta_k}{\sum_{s=\ell}^{k}\eta_s} - \frac{\eta_k}{\sum_{s=\ell-1}^{k}\eta_s}\right)B_f(\mathbf{x}_\ell, \mathbf{x}_*) \\
&= \sum_{\ell=2}^{k}\frac{8n^2\bar{L}^2\eta_{\ell-1}\left(\sum_{s=\ell}^{k}\eta_s^3\right)}{\left(\sum_{s=\ell}^{k}\eta_s\right)\left(\sum_{s=\ell-1}^{k}\eta_s\right)}B_f(\mathbf{x}_\ell, \mathbf{x}_*).
\end{aligned} \tag{36}$$

Plugging (35) and (36) into (34), we finally obtain

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) \leq \frac{\left\|\mathbf{x}_* - \mathbf{x}_1\right\|^2}{2n\sum_{\ell=1}^{k}\eta_\ell} + \sum_{\ell=1}^{k}\frac{4\eta_\ell^3 R_\ell}{\sum_{s=\ell}^{k}\eta_s} + \sum_{\ell=2}^{k}\frac{8n^2\bar{L}^2\eta_{\ell-1}\left(\sum_{s=\ell}^{k}\eta_s^3\right)}{\left(\sum_{s=\ell}^{k}\eta_s\right)\left(\sum_{s=\ell-1}^{k}\eta_s\right)}B_f(\mathbf{x}_\ell, \mathbf{x}_*).$$

□

The second result, Lemma D.5, is particularly useful for the strongly convex case, i.e., $\mu_\psi > 0$ or $\mu_f > 0$. Though the distance from the optimum is not the desired bound we want, it can finally help us to bound the function value. The key idea here is using $\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2$ to bound $F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)$ and applying Lemma D.5 to bound $\left\|\mathbf{x}_* - \mathbf{x}_{\lceil \frac{K}{2} \rceil+1}\right\|^2$. A similar argument showed up in (Cha et al., 2023) but was used to bound $F(\mathbf{x}_{K+1}^{\text{tail}}) - F(\mathbf{x}_*)$ (recall $\mathbf{x}_{K+1}^{\text{tail}} \triangleq \frac{1}{\lceil \frac{K}{2} \rceil+1} \sum_{k=\lceil \frac{K}{2} \rceil}^{K} \mathbf{x}_{k+1}$). In contrast, our work directly bounds the last iterate instead of the tail average iterate.

Lemma D.5. *Under Assumptions 3.1, 3.2, 3.3 and 3.4, suppose $\eta_k \le \frac{1}{2n\sqrt{\bar{L}L^*}}, \forall k \in [K]$, then for any permutation $\sigma_k, \forall k \in [K]$, Algorithm 1 guarantees*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \le \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{\prod_{s=1}^{k}(1 + n\eta_s(\mu_f + 2\mu_\psi))} + \sum_{\ell=1}^{k} \frac{8n\eta_\ell^3 R_\ell}{\prod_{s=\ell}^{k}(1 + n\eta_s(\mu_f + 2\mu_\psi))}, \forall k \in [K],$$

where R_ℓ is defined in Lemma D.2.

Proof. We invoke Lemma D.1 with $\mathbf{z} = \mathbf{x}_*$ to obtain

$$\begin{aligned}
F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) \le& \frac{\|\mathbf{x}_* - \mathbf{x}_k\|^2}{2n\eta_k} - \left(\frac{1}{\eta_k} + n\mu_\psi\right) \frac{\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2}{2n} - \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2}{2n\eta_k} \\
&+ \frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{x}_*, \mathbf{x}_k^i) \\
\overset{(a)}{\le}& \frac{\|\mathbf{x}_* - \mathbf{x}_k\|^2}{2n\eta_k} - \left(\frac{1}{\eta_k} + n\mu_\psi\right) \frac{\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2}{2n} - \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2}{2n\eta_k} \\
&+ \bar{L}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + 8\eta_k^2 n^2 \bar{L}^2 B_f(\mathbf{x}_*, \mathbf{x}_*) + 4\eta_k^2 R_k \\
\overset{(b)}{\le}& \frac{\|\mathbf{x}_* - \mathbf{x}_k\|^2}{2n\eta_k} - \left(\frac{1}{\eta_k} + n\mu_\psi\right) \frac{\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2}{2n} + 4\eta_k^2 R_k,
\end{aligned}$$

where (a) is by Lemma D.2 and (b) holds due to $\bar{L} \le \frac{1}{2n\eta_k}$ from the requirment of $\eta_k \le \frac{1}{2n\sqrt{\bar{L}L^*}}, \forall k \in [K]$ and $B_f(\mathbf{x}_*, \mathbf{x}_*) = 0$. Note that $\mathbf{x}_* \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$ implies $\exists \nabla\psi(\mathbf{x}_*) \in \partial\psi(\mathbf{x}_*)$ such that $\nabla f(\mathbf{x}_*) + \nabla\psi(\mathbf{x}_*) = \mathbf{0}$, hence

$$\begin{aligned}
F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) &= F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) - \langle \nabla f(\mathbf{x}_*) + \nabla\psi(\mathbf{x}_*), \mathbf{x}_{k+1} - \mathbf{x}_* \rangle \\
&= B_f(\mathbf{x}_{k+1}, \mathbf{x}_*) + B_\psi(\mathbf{x}_{k+1}, \mathbf{x}_*) \ge \frac{\mu_f + \mu_\psi}{2}\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2,
\end{aligned}$$

where the last inequality is due to the strong convexity of f and ψ. So we know

$$\begin{aligned}
&\frac{\mu_f + \mu_\psi}{2}\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 \le \frac{\|\mathbf{x}_* - \mathbf{x}_k\|^2}{2n\eta_k} - \left(\frac{1}{\eta_k} + n\mu_\psi\right) \frac{\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2}{2n} + 4\eta_k^2 R_k \\
\Rightarrow& (1 + n\eta_k(\mu_f + 2\mu_\psi))\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 \le \|\mathbf{x}_* - \mathbf{x}_k\|^2 + 8n\eta_k^3 R_k \\
\Rightarrow& \|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 \le \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{\prod_{s=1}^{k}(1 + n\eta_s(\mu_f + 2\mu_\psi))} + \frac{\sum_{\ell=1}^{k} 8n\eta_\ell^3 R_\ell \prod_{s=1}^{\ell-1}(1 + n\eta_s(\mu_f + 2\mu_\psi))}{\prod_{s=1}^{k}(1 + n\eta_s(\mu_f + 2\mu_\psi))} \\
&= \frac{\|\mathbf{x}_* - \mathbf{x}_1\|^2}{\prod_{s=1}^{k}(1 + n\eta_s(\mu_f + 2\mu_\psi))} + \sum_{\ell=1}^{k} \frac{8n\eta_\ell^3 R_\ell}{\prod_{s=\ell}^{k}(1 + n\eta_s(\mu_f + 2\mu_\psi))}.
\end{aligned}$$

□

D.3. Lipschitz Functions

We focus on the Lipschitz case and prove two useful results in this subsection.

We first introduce Lemma D.6, which is a consequence of combining Lemmas D.1 and D.3. However, the proof of Lemma D.6 is different from both the prior works (Zamani & Glineur, 2023; Liu & Zhou, 2024) and Lemma D.4. For any fixed

$k\in[K]$ here, instead of setting $\mathbf{z}=\mathbf{z}_k$ directly, we will invoke Lemma D.1 $k+1$ times with $\mathbf{z}=\mathbf{x}_\ell$ for $\ell\in[k]$ and $\mathbf{z}=\mathbf{x}_*$ and then sum them up with some weights. The key reason is that we need a refined bound for proving Theorem C.4. To be more precise, one can see there is an extra negative term $-\left\|\mathbf{x}_*-\mathbf{x}_{K+1}\right\|^2$ showing up in Lemma D.6. It plays an important role in the proof of Theorem C.4. However, if we totally follow (Zamani & Glineur, 2023; Liu & Zhou, 2024), the term $-\left\|\mathbf{x}_*-\mathbf{x}_{K+1}\right\|^2$ will be replaced by a larger quantity, which cannot help us prove Theorem C.4.

Lemma D.6. *Under Assumptions 3.1, 3.2 and 3.5, let $\gamma_k\triangleq\eta_k\prod_{\ell=2}^{k}(1+n\eta_{\ell-1}\mu_\psi),\forall k\in[K]$. For any permutation $\sigma_k,\forall k\in[K]$, Algorithm 1 guarantees*

$$F(\mathbf{x}_{K+1})-F(\mathbf{x}_*)\le\frac{\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2-\left\|\mathbf{x}_*-\mathbf{x}_{K+1}\right\|^2}{2n\sum_{k=1}^{K}\gamma_k}+3\bar{G}^2n\sum_{k=1}^{K}\frac{\gamma_k\eta_k}{\sum_{\ell=k}^{K}\gamma_\ell}.$$

Proof. We define the non-decreasing sequence

$$v_k\triangleq\frac{\gamma_K}{\sum_{\ell=k}^{K}\gamma_\ell},\forall k\in[K]\,, \tag{37}$$

$$v_0\triangleq v_1. \tag{38}$$

Fix $k\in[K]$, we invoke Lemma D.1 to obtain for any $\mathbf{z}\in\mathbb{R}^d$

$$\begin{aligned}
F(\mathbf{x}_{k+1})-F(\mathbf{z})\le&\frac{\left\|\mathbf{z}-\mathbf{x}_k\right\|^2}{2n\eta_k}-\left(\frac{1}{\eta_k}+n\mu_\psi\right)\frac{\left\|\mathbf{z}-\mathbf{x}_{k+1}\right\|^2}{2n}-\frac{\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|^2}{2n\eta_k}\\
&+\frac{1}{n}\sum_{i=1}^{n}B_{\sigma_k^i}(\mathbf{x}_{k+1},\mathbf{x}_k^i)-B_{\sigma_k^i}(\mathbf{z},\mathbf{x}_k^i)\\
\overset{(a)}{\le}&\frac{\left\|\mathbf{z}-\mathbf{x}_k\right\|^2}{2n\eta_k}-\left(\frac{1}{\eta_k}+n\mu_\psi\right)\frac{\left\|\mathbf{z}-\mathbf{x}_{k+1}\right\|^2}{2n}-\frac{\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|^2}{2n\eta_k}\\
&+2\bar{G}\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|+\bar{G}^2n\eta_k\\
\overset{(b)}{\le}&\frac{\left\|\mathbf{z}-\mathbf{x}_k\right\|^2}{2n\eta_k}-\left(\frac{1}{\eta_k}+n\mu_\psi\right)\frac{\left\|\mathbf{z}-\mathbf{x}_{k+1}\right\|^2}{2n}+3\bar{G}^2n\eta_k\\
\Rightarrow\gamma_k\left(F(\mathbf{x}_{k+1})-F(\mathbf{z})\right)\le&\frac{\gamma_k\eta_k^{-1}\left\|\mathbf{z}-\mathbf{x}_k\right\|^2}{2n}-\gamma_k\left(\frac{1}{\eta_k}+n\mu_\psi\right)\frac{\left\|\mathbf{z}-\mathbf{x}_{k+1}\right\|^2}{2n}+3\bar{G}^2n\gamma_k\eta_k\\
\overset{(c)}{=}&\frac{\gamma_k\eta_k^{-1}\left\|\mathbf{z}-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}\left\|\mathbf{z}-\mathbf{x}_{k+1}\right\|^2}{2n}+3\bar{G}^2n\gamma_k\eta_k,
\end{aligned} \tag{39}$$

where (a) is due to Lemma D.3, (b) is by AM-GM inequality and (c) is because of

$$\gamma_k\left(\frac{1}{\eta_k}+n\mu_\psi\right)=\left(\frac{1}{\eta_k}+n\mu_\psi\right)\times\eta_k\prod_{\ell=2}^{k}(1+n\eta_{\ell-1}\mu_\psi)=\prod_{\ell=2}^{k+1}(1+n\eta_{\ell-1}\mu_\psi)=\gamma_{k+1}\eta_{k+1}^{-1}.$$

Taking $\mathbf{z}=\mathbf{x}_\ell$ for $\ell\in[k]$ in (39) and multiplying both sides by $v_\ell-v_{\ell-1}$ (it is non-negative due to $v_\ell\ge v_{\ell-1}$) to obtain

$$\gamma_k(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_\ell)\right)\le(v_\ell-v_{\ell-1})\left(\frac{\gamma_k\eta_k^{-1}\left\|\mathbf{x}_\ell-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}\left\|\mathbf{x}_\ell-\mathbf{x}_{k+1}\right\|^2}{2n}+3\bar{G}^2n\gamma_k\eta_k\right).$$

Summing it up from $\ell=1$ to k

$$\begin{aligned}
&\sum_{\ell=1}^{k}\gamma_k(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_\ell)\right)\\
\le&\frac{\gamma_k\eta_k^{-1}\sum_{\ell=1}^{k-1}(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}\sum_{\ell=1}^{k}(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_{k+1}\right\|^2}{2n}+3\bar{G}^2n\gamma_k\eta_k(v_k-v_0).
\end{aligned} \tag{40}$$

Next, taking $\mathbf{z}=\mathbf{x}_*$ in (39) and multiplying both sides by v_0 to obtain

$$\gamma_k v_0\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)\right)\leq\frac{\gamma_k\eta_k^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_{k+1}\right\|^2}{2n}+3\bar{G}^2n\gamma_k\eta_k v_0. \tag{41}$$

Adding (41) to (40) to get

$$\begin{aligned}
&\sum_{\ell=1}^{k}\gamma_k(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_\ell)\right)+\gamma_k v_0\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)\right)\\
\leq&\frac{\gamma_k\eta_k^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_{k+1}\right\|^2}{2n}+3\bar{G}^2n\gamma_k\eta_k v_k\\
&+\frac{\gamma_k\eta_k^{-1}\sum_{\ell=1}^{k-1}(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}\sum_{\ell=1}^{k}(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_{k+1}\right\|^2}{2n}.
\end{aligned}\tag{42}$$

Summing up (42) from $k=1$ to K to get

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{\ell=1}^{k}\gamma_k(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_\ell)\right)+\sum_{k=1}^{K}\gamma_k v_0\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)\right)\\
\leq&\sum_{k=1}^{K}\frac{\gamma_k\eta_k^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_{k+1}\right\|^2}{2n}+\sum_{k=1}^{K}3\bar{G}^2n\gamma_k\eta_k v_k\\
&+\sum_{k=1}^{K}\frac{\gamma_k\eta_k^{-1}\sum_{\ell=1}^{k-1}(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_k\right\|^2}{2n}-\frac{\gamma_{k+1}\eta_{k+1}^{-1}\sum_{\ell=1}^{k}(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_{k+1}\right\|^2}{2n}\\
=&\frac{\gamma_1\eta_1^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{2n}-\frac{\gamma_{K+1}\eta_{K+1}^{-1}v_0\left\|\mathbf{x}_*-\mathbf{x}_{K+1}\right\|^2}{2n}+\sum_{k=1}^{K}3\bar{G}^2n\gamma_k\eta_k v_k\\
&-\frac{\gamma_{K+1}\eta_{K+1}^{-1}\sum_{\ell=1}^{K}(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_{K+1}\right\|^2}{2n}\\
\leq&\frac{v_0\left\|\mathbf{x}_*-\mathbf{x}_1\right\|^2}{2n}-\frac{v_0\left\|\mathbf{x}_*-\mathbf{x}_{K+1}\right\|^2}{2n}+\sum_{k=1}^{K}3\bar{G}^2n\gamma_k\eta_k v_k,
\end{aligned}\tag{43}$$

where the last line is by $\gamma_1\eta_1^{-1}=1$, $\gamma_{K+1}\eta_{K+1}^{-1}=\prod_{\ell=2}^{K+1}(1+n\eta_{\ell-1}\mu_\psi)\geq1$ and $(v_\ell-v_{\ell-1})\left\|\mathbf{x}_\ell-\mathbf{x}_{K+1}\right\|^2\geq0$. For the L.H.S. of (43), we observe that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{\ell=1}^{k}\gamma_k(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_\ell)\right)\\
=&\sum_{k=1}^{K}\sum_{\ell=1}^{k}\gamma_k(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)\right)-\sum_{k=1}^{K}\sum_{\ell=1}^{k}\gamma_k(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_\ell)-F(\mathbf{x}_*)\right)\\
=&\sum_{k=1}^{K}\gamma_k\left(v_k-v_0\right)\left(F(\mathbf{x}_{k+1})-F(\mathbf{x}_*)\right)-\sum_{\ell=1}^{K}\left(\sum_{k=\ell}^{K}\gamma_k\right)(v_\ell-v_{\ell-1})\left(F(\mathbf{x}_\ell)-F(\mathbf{x}_*)\right).
\end{aligned}$$

Hence

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{\ell=1}^{k}\gamma_k(v_\ell - v_{\ell-1})\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x}_\ell)\right) + \sum_{k=1}^{K}\gamma_k v_0\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*)\right)\\
=&\sum_{k=1}^{K}\gamma_k v_k\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*)\right) - \sum_{\ell=1}^{K}\left(\sum_{k=\ell}^{K}\gamma_k\right)(v_\ell - v_{\ell-1})\left(F(\mathbf{x}_\ell) - F(\mathbf{x}_*)\right)\\
=&\gamma_K v_K\left(F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)\right) - \left(\sum_{k=1}^{K}\gamma_k\right)(v_1 - v_0)\left(F(\mathbf{x}_1) - F(\mathbf{x}_*)\right)\\
&+\sum_{k=2}^{K}\left[\gamma_{k-1}v_{k-1} - \left(\sum_{\ell=k}^{K}\gamma_\ell\right)(v_k - v_{k-1})\right]\left(F(\mathbf{x}_k) - F(\mathbf{x}_*)\right).
\end{aligned}
$$

Note that $v_1 \overset{(38)}{=} v_0$ and for $2 \le k \le K$

$$
\gamma_{k-1}v_{k-1} - \left(\sum_{\ell=k}^{K}\gamma_\ell\right)(v_k - v_{k-1}) = \left(\sum_{\ell=k-1}^{K}\gamma_\ell\right)v_{k-1} - \left(\sum_{\ell=k}^{K}\gamma_\ell\right)v_k \overset{(37)}{=} \gamma_K - \gamma_K = 0.
$$

Thus, we know

$$
\sum_{k=1}^{K}\sum_{\ell=1}^{k}\gamma_k(v_\ell - v_{\ell-1})\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x}_\ell)\right) + \sum_{k=1}^{K}\gamma_k v_0\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*)\right) = \gamma_K v_K\left(F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)\right). \tag{44}
$$

Plugging (44) into (43), we finally obtain

$$
\begin{aligned}
\gamma_K v_K\left(F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*)\right) &\le \frac{v_0\left\|\mathbf{x}_* - \mathbf{x}_1\right\|^2}{2n} - \frac{v_0\left\|\mathbf{x}_* - \mathbf{x}_{K+1}\right\|^2}{2n} + \sum_{k=1}^{K}3\bar{G}^2 n\gamma_k\eta_k v_k\\
\Rightarrow F(\mathbf{x}_{K+1}) - F(\mathbf{x}_*) &\le \frac{v_0}{\gamma_K v_K}\cdot\frac{\left\|\mathbf{x}_* - \mathbf{x}_1\right\|^2 - \left\|\mathbf{x}_* - \mathbf{x}_{K+1}\right\|^2}{2n} + 3\bar{G}^2 n\sum_{k=1}^{K}\frac{\gamma_k\eta_k v_k}{\gamma_K v_K}\\
&\overset{(37),(38)}{=} \frac{\left\|\mathbf{x}_* - \mathbf{x}_1\right\|^2 - \left\|\mathbf{x}_* - \mathbf{x}_{K+1}\right\|^2}{2n\sum_{k=1}^{K}\gamma_k} + 3\bar{G}^2 n\sum_{k=1}^{K}\frac{\gamma_k\eta_k}{\sum_{\ell=k}^{K}\gamma_\ell}.
\end{aligned}
$$

□

Next, for a special class of stepsizes (including the stepsize used in Theorem C.4 as a special case), the following Lemma D.7 gives a uniform bound on the movement of $\mathbf{x}_k$ from the initial point $\mathbf{x}_1$. A simple but useful fact implied by this result is that r_k is also uniformly upper bounded. This important corollary will finally lead us to an asymptotic rate having a linear dependence on D (see the proof of Theorem C.4 for details).

Lemma D.7. *Under Assumptions 3.1, 3.2 (with $\mu_\psi = 0$) and 3.5, suppose the following two conditions hold:*

1. $\eta_k = r_k\widetilde{\eta}_k, \forall k \in \mathbb{N}$ *where* $r_k = r \vee \max_{\ell\in[k]}\left\|\mathbf{x}_\ell - \mathbf{x}_1\right\|$ *for some* $r > 0$ *and* $\widetilde{\eta}_k$ *is a positive sequence defined on* $\mathbb{N}$.
2. $\sum_{k=1}^{\infty}6\bar{G}^2 n^2\widetilde{\eta}_k^2 \le c^2 < 1$ *for some constant* $c > 0$.

Then for any permutation $\sigma_k, \forall k \in \mathbb{N}$, Algorithm 1 guarantees

$$
\left\|\mathbf{x}_k - \mathbf{x}_1\right\| \le \frac{2}{1-c}\left\|\mathbf{x}_* - \mathbf{x}_1\right\| + \frac{c}{1-c}r, \forall k \in \mathbb{N}.
$$

Proof. We invoke Lemma D.1 with $\mathbf{z} = \mathbf{x}_*$ and $\mu_\psi = 0$ to obtain

$$
\begin{aligned}
F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) &\leq \frac{\|\mathbf{x}_* - \mathbf{x}_k\|^2}{2n\eta_k} - \frac{\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2}{2n\eta_k} - \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2}{2n\eta_k} + \frac{1}{n}\sum_{i=1}^{n} B_{\sigma_k^i}(\mathbf{x}_{k+1}, \mathbf{x}_k^i) - B_{\sigma_k^i}(\mathbf{x}_*, \mathbf{x}_k^i) \\
&\overset{(a)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_k\|^2}{2n\eta_k} - \frac{\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2}{2n\eta_k} - \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2}{2n\eta_k} + 2\bar{G}\|\mathbf{x}_{k+1} - \mathbf{x}_k\| + \bar{G}^2 n\eta_k \\
&\overset{(b)}{\leq} \frac{\|\mathbf{x}_* - \mathbf{x}_k\|^2}{2n\eta_k} - \frac{\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2}{2n\eta_k} + 3\bar{G}^2 n\eta_k,
\end{aligned}
$$

where the (a) is due to Lemma D.3 and (b) is by AM-GM inequality. Note that $F(\mathbf{x}_{k+1}) - F(\mathbf{x}_*) \geq 0$, hence

$$
\begin{aligned}
&\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 \leq \|\mathbf{x}_* - \mathbf{x}_k\|^2 + 6\bar{G}^2 n^2 \eta_k^2 \\
\Rightarrow\ &\|\mathbf{x}_* - \mathbf{x}_{k+1}\|^2 \leq \|\mathbf{x}_* - \mathbf{x}_1\|^2 + \sum_{\ell=1}^{k} 6\bar{G}^2 n^2 \eta_\ell^2 \\
\Rightarrow\ &\|\mathbf{x}_{k+1} - \mathbf{x}_1\| \leq \|\mathbf{x}_* - \mathbf{x}_1\| + \|\mathbf{x}_* - \mathbf{x}_{k+1}\| \leq \|\mathbf{x}_* - \mathbf{x}_1\| + \sqrt{\|\mathbf{x}_* - \mathbf{x}_1\|^2 + \sum_{\ell=1}^{k} 6\bar{G}^2 n^2 \eta_\ell^2} \\
&\leq 2\|\mathbf{x}_* - \mathbf{x}_1\| + \sqrt{\sum_{\ell=1}^{k} 6\bar{G}^2 n^2 \eta_\ell^2} \overset{(c)}{\leq} 2\|\mathbf{x}_* - \mathbf{x}_1\| + r_k\sqrt{\sum_{\ell=1}^{k} 6\bar{G}^2 n^2 \widetilde{\eta}_\ell^2}, \\
&\overset{(d)}{\leq} 2\|\mathbf{x}_* - \mathbf{x}_1\| + c r_k \leq 2\|\mathbf{x}_* - \mathbf{x}_1\| + cr + c\max_{\ell \in [k]} \|\mathbf{x}_\ell - \mathbf{x}_1\|,
\end{aligned} \tag{45}
$$

where (c) is by $\eta_\ell^2 = r_\ell^2 \widetilde{\eta}_\ell^2 \leq r_k^2 \widetilde{\eta}_\ell^2, \forall \ell \in [k]$ and (d) is from the requirement $\sum_{k=1}^{\infty} 6\bar{G}^2 n^2 \widetilde{\eta}_k^2 \leq c^2$.

Now we use induction to prove (45) implies the following fact when $c < 1$,

$$
\|\mathbf{x}_k - \mathbf{x}_1\| \leq \frac{2}{1-c}\|\mathbf{x}_* - \mathbf{x}_1\| + \frac{c}{1-c} r, \forall k \in \mathbb{N}. \tag{46}
$$

First if $k = 1$, (46) reduces to $0 \leq \frac{2}{1-c}\|\mathbf{x}_* - \mathbf{x}_1\| + \frac{c}{1-c} r$, which is true automatically. Suppose (46) holds for 1 to k. Then for $k+1$, by (45)

$$
\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_1\| &\leq 2\|\mathbf{x}_* - \mathbf{x}_1\| + cr + c\max_{\ell \in [k]} \|\mathbf{x}_\ell - \mathbf{x}_1\| \\
&\leq 2\|\mathbf{x}_* - \mathbf{x}_1\| + cr + c\left(\frac{2}{1-c}\|\mathbf{x}_* - \mathbf{x}_1\| + \frac{c}{1-c} r\right) \\
&= \frac{2}{1-c}\|\mathbf{x}_* - \mathbf{x}_1\| + \frac{c}{1-c} r.
\end{aligned}
$$

Therefore, the induction is completed. □

E. Technical Lemmas

In this section, we introduce two helpful lemmas used in the analysis.

First, we provide an upper bound for the residual term $R_k \triangleq \sum_{i=2}^{n} \frac{L_{\sigma_k^i}}{n} \left\| \sum_{j=1}^{i-1} \nabla f_{\sigma_k^j}(\mathbf{x}_*) \right\|^2, \forall k \in [K]$ defined in Lemma D.2. Specifically, we will prove two bounds for this term: one in (47) holds for any permutation, and the other in (48) holds for the permutation uniformly sampled without replacement. It is worth noting that if $L_i = L$ for any $i \in [n]$, then one can apply Lemma 1 in (Mishchenko et al., 2020) to get the desired bound. However, L_i can be different in our setting, which requires a more careful analysis.

Lemma E.1. *Under Assumptions 3.1 and 3.3, for any permutation σ of $[n]$, there is*

$$
\sum_{i=2}^{n} \frac{L_{\sigma^i}}{n} \left\| \sum_{j=1}^{i-1} \nabla f_{\sigma^j}(\mathbf{x}_*) \right\|^2 \leq n^2 \bar{L} \sigma_{\text{any}}^2. \tag{47}
$$

If the permutation σ is uniformly sampled without replacement, there is

$$\mathbb{E}\left[\sum_{i=2}^{n}\frac{L_{\sigma^i}}{n}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]\le\frac{2}{3}n\bar{L}\sigma_{\mathrm{rand}}^2,\tag{48}$$

where the expectation is taken over the permutation σ.

Proof. First note that

$$\begin{aligned}\sum_{i=2}^{n}\frac{L_{\sigma^i}}{n}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2&\le\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{L_{\sigma^i}}{n}(i-1)\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2=\sum_{j=1}^{n-1}\sum_{i=j+1}^{n}\frac{L_{\sigma^i}}{n}(i-1)\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\\&\le n\bar{L}\sum_{j=1}^{n-1}\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\le n\bar{L}\sum_{j=1}^{n}\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2=n^2\bar{L}\sigma_{\mathrm{any}}^2.\end{aligned}$$

Next, suppose σ is uniformly sampled without replacement.

When $n=2$

$$\begin{aligned}\mathbb{E}\left[\sum_{i=2}^{n}\frac{L_{\sigma^i}}{n}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]&=\mathbb{E}\left[\frac{L_{\sigma^2}}{2}\left\|\nabla f_{\sigma^1}(\mathbf{x}_*)\right\|^2\right]=\frac{L_1\left\|\nabla f_2(\mathbf{x}_*)\right\|^2+L_2\left\|\nabla f_1(\mathbf{x}_*)\right\|^2}{4}\\&=\underbrace{\frac{L_1+L_2}{2}\cdot\frac{\left\|\nabla f_1(\mathbf{x}_*)\right\|^2+\left\|\nabla f_2(\mathbf{x}_*)\right\|^2}{2}}_{=\bar{L}\sigma_{\mathrm{any}}^2}-\frac{L_1\left\|\nabla f_1(\mathbf{x}_*)\right\|^2+L_2\left\|\nabla f_2(\mathbf{x}_*)\right\|^2}{4}\\&\le\bar{L}\sigma_{\mathrm{any}}^2\le\frac{2}{3}n\bar{L}\sigma_{\mathrm{rand}}^2,\end{aligned}$$

where the last inequality holds due to $\sigma_{\mathrm{any}}^2\le\sigma_{\mathrm{any}}^2+2\left\|\nabla f(\mathbf{x}_*)\right\|^2=\sigma_{\mathrm{rand}}^2$ and $\frac{2}{3}n=\frac{4}{3}>1$.

When $n\ge 3$

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=2}^{n}\frac{L_{\sigma^i}}{n}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]=\frac{1}{n}\sum_{i=2}^{n}\mathbb{E}\left[L_{\sigma^i}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]\\=&\frac{1}{n}\sum_{i=2}^{n}\mathbb{E}\left[\sum_{j=1}^{i-1}L_{\sigma^i}\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2+\sum_{1\le p\ne q\le i-1}L_{\sigma^i}\left\langle\nabla f_{\sigma^p}(\mathbf{x}_*),\nabla f_{\sigma^q}(\mathbf{x}_*)\right\rangle\right]\\=&\frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}\left[L_{\sigma^i}\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]+\frac{1}{n}\sum_{i=2}^{n}\sum_{1\le p\ne q\le i-1}\mathbb{E}\left[L_{\sigma^i}\left\langle\nabla f_{\sigma^p}(\mathbf{x}_*),\nabla f_{\sigma^q}(\mathbf{x}_*)\right\rangle\right].\end{aligned}\tag{49}$$

For $j<i$, we have

$$\begin{aligned}\mathbb{E}\left[L_{\sigma^i}\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]&=\mathbb{E}\left[\mathbb{E}\left[L_{\sigma^i}\mid\sigma^j\right]\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]=\mathbb{E}\left[\frac{n\bar{L}-L_{\sigma^j}}{n-1}\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]\\&=\frac{n}{n-1}\bar{L}\sigma_{\mathrm{any}}^2-\frac{\sum_{\ell=1}^{n}L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{n(n-1)}\\\Rightarrow\frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbb{E}\left[L_{\sigma^i}\left\|\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right]&=\frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\frac{n}{n-1}\bar{L}\sigma_{\mathrm{any}}^2-\frac{\sum_{\ell=1}^{n}L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{n(n-1)}\right)\\&=\frac{n\bar{L}\sigma_{\mathrm{any}}^2}{2}-\frac{\sum_{\ell=1}^{n}L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{2n}.\end{aligned}\tag{50}$$

For $1 \le p \ne q \le i-1$, we have

$$\begin{aligned}
&\mathbb{E}\left[L_{\sigma^i}\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[L_{\sigma^i} \mid \sigma^p, \sigma^q\right]\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] \\
=&\mathbb{E}\left[\frac{n\bar{L} - L_{\sigma^p} - L_{\sigma^q}}{n-2}\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] \\
=&\frac{n\bar{L}}{n-2}\mathbb{E}\left[\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] - \frac{1}{n-2}\mathbb{E}\left[(L_{\sigma^p} + L_{\sigma^q})\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right]. \qquad (51)
\end{aligned}$$

Note that

$$\mathbb{E}\left[\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] = \mathbb{E}\left[\langle\frac{n\nabla f(\mathbf{x}_*) - \nabla f_{\sigma^q}(\mathbf{x}_*)}{n-1}, \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] = \frac{n\left\|\nabla f(\mathbf{x}_*)\right\|^2 - \sigma_{\text{any}}^2}{n-1}, \qquad (52)$$

and

$$\begin{aligned}
&\mathbb{E}\left[(L_{\sigma^p} + L_{\sigma^q})\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] \\
=&2\mathbb{E}\left[L_{\sigma^p}\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] \\
=&2\mathbb{E}\left[L_{\sigma^p}\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \frac{n\nabla f(\mathbf{x}_*) - \nabla f_{\sigma^p}(\mathbf{x}_*)}{n-1}\rangle\right] \\
=&\frac{2\langle\sum_{\ell=1}^n L_\ell\nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{n-1} - \frac{2\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{(n-1)n}. \qquad (53)
\end{aligned}$$

Plugging (52) and (53) into (51) to obtain

$$\begin{aligned}
&\mathbb{E}\left[L_{\sigma^i}\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] \\
=&\frac{n\bar{L}}{n-2}\cdot\frac{n\left\|\nabla f(\mathbf{x}_*)\right\|^2 - \sigma_{\text{any}}^2}{n-1} - \frac{1}{n-2}\left(\frac{2\langle\sum_{\ell=1}^n L_\ell\nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{n-1} - \frac{2\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{(n-1)n}\right) \\
=&\frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2 - n\bar{L}\sigma_{\text{any}}^2 - 2\langle\sum_{\ell=1}^n L_\ell\nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{(n-2)(n-1)} + \frac{2\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{(n-2)(n-1)n},
\end{aligned}$$

which implies

$$\begin{aligned}
&\frac{1}{n}\sum_{i=2}^n\sum_{1\le p\ne q\le i-1}\mathbb{E}\left[L_{\sigma^i}\langle\nabla f_{\sigma^p}(\mathbf{x}_*), \nabla f_{\sigma^q}(\mathbf{x}_*)\rangle\right] \\
=&\frac{1}{n}\sum_{i=2}^n\sum_{1\le p\ne q\le i-1}\frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2 - n\bar{L}\sigma_{\text{any}}^2 - 2\langle\sum_{\ell=1}^n L_\ell\nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{(n-2)(n-1)} + \frac{2\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{(n-2)(n-1)n} \\
=&\frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2 - n\bar{L}\sigma_{\text{any}}^2 - 2\langle\sum_{\ell=1}^n L_\ell\nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{3} + \frac{2\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{3n}. \qquad (54)
\end{aligned}$$

Finally, plugging (50) and (54) into (49), we know

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=2}^n\frac{L_{\sigma^i}}{n}\left\|\sum_{j=1}^{i-1}\nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right] \\
=&\frac{n\bar{L}\sigma_{\text{any}}^2}{2} - \frac{\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{2n} + \frac{2\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{3n} \\
&+\frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2 - n\bar{L}\sigma_{\text{any}}^2 - 2\langle\sum_{\ell=1}^n L_\ell\nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{3} \\
=&\frac{n\bar{L}\sigma_{\text{any}}^2}{6} + \frac{\sum_{\ell=1}^n L_\ell\left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{6n} + \frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2}{3} - \frac{2\langle\sum_{\ell=1}^n L_\ell\nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{3}.
\end{aligned}$$

Note that

$$\begin{aligned}\frac{\sum_{\ell=1}^{n} L_\ell \left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{6n} &\leq \frac{n\bar{L}\sigma_{\text{any}}^2}{6},\\ -\frac{2\langle\sum_{\ell=1}^{n} L_\ell \nabla f_\ell(\mathbf{x}_*), \nabla f(\mathbf{x}_*)\rangle}{3} &\leq \frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2}{3} + \frac{\left\|\sum_{\ell=1}^{n} L_\ell \nabla f_\ell(\mathbf{x}_*)\right\|^2}{3n^2\bar{L}}\\ &\leq \frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2}{3} + \frac{\sum_{\ell=1}^{n} L_\ell \left\|\nabla f_\ell(\mathbf{x}_*)\right\|^2}{3n}\\ &\leq \frac{n^2\bar{L}\left\|\nabla f(\mathbf{x}_*)\right\|^2}{3} + \frac{n\bar{L}\sigma_{\text{any}}^2}{3}.\end{aligned}$$

Hence, there is

$$\mathbb{E}\left[\sum_{i=2}^{n} \frac{L_{\sigma^i}}{n} \left\|\sum_{j=1}^{i-1} \nabla f_{\sigma^j}(\mathbf{x}_*)\right\|^2\right] \leq \frac{2}{3}n\bar{L}\left(\sigma_{\text{any}}^2 + n\left\|\nabla f(\mathbf{x}_*)\right\|^2\right) = \frac{2}{3}n\bar{L}\sigma_{\text{rand}}^2.$$

□

Next, we introduce the following algebraic inequality, which is a useful tool when proving Theorem C.1.

Lemma E.2. *Given a sequence $d_2, \cdots, d_K, d_{K+1}$, suppose there exist positive constants a, b, c satisfying*

$$d_{k+1} \leq \frac{a}{k} + b(1+\log k) + c\sum_{\ell=2}^{k} \frac{d_\ell}{k-\ell+2}, \forall k \in [K], \tag{55}$$

then the following inequality holds

$$d_{k+1} \leq \left(\frac{a}{k} + b(1+\log k)\right)\sum_{i=0}^{k-1} \left(2c(1+\log k)\right)^i, \forall k \in [K]. \tag{56}$$

Proof. We use induction to prove (56) holds for every $k \in [K]$. First, for $k = 1$, we need to show

$$d_2 \leq a + b,$$

which is true by taking $k = 1$ in (55). Suppose (56) holds for 1 to $k-1$ (where $2 \leq k \leq K$), i.e.,

$$d_\ell \leq \left(\frac{a}{\ell-1} + b(1+\log(\ell-1))\right)\sum_{i=0}^{\ell-2} \left(2c(1+\log(\ell-1))\right)^i, \forall 2 \leq \ell \leq k.$$

which implies

$$d_\ell \leq \left(\frac{a}{\ell-1} + b(1+\log k)\right)\sum_{i=0}^{\ell-2} \left(2c(1+\log k)\right)^i, \forall 2 \leq \ell \leq k. \tag{57}$$

Now for d_{k+1}, from (55)

$$\begin{aligned}d_{k+1} \leq& \frac{a}{k} + b(1+\log k) + c\sum_{\ell=2}^{k} \frac{d_\ell}{k-\ell+2}\\ \overset{(57)}{\leq}& \frac{a}{k} + ac\sum_{\ell=2}^{k}\sum_{i=0}^{\ell-2} \frac{\left(2c(1+\log k)\right)^i}{(k-\ell+2)(\ell-1)}\\ &+ b(1+\log k)\left(1 + c\sum_{\ell=2}^{k}\sum_{i=0}^{\ell-2} \frac{\left(2c(1+\log k)\right)^i}{k-\ell+2}\right).\end{aligned} \tag{58}$$

Note that

$$\begin{aligned}
c\sum_{\ell=2}^{k}\sum_{i=0}^{\ell-2}\frac{\left(2c(1+\log k)\right)^{i}}{(k-\ell+2)(\ell-1)} &= c\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i}\left(\sum_{\ell=2+i}^{k}\frac{1}{(k-\ell+2)(\ell-1)}\right)\\
&= \frac{c}{k+1}\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i}\left(\sum_{\ell=2+i}^{k}\frac{1}{k-\ell+2}+\frac{1}{\ell-1}\right)\\
&\leq \frac{c}{k+1}\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i}\sum_{\ell=1}^{k}\frac{2}{\ell}\leq\frac{\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i+1}}{k+1}\\
&= \frac{\sum_{i=1}^{k-1}\left(2c(1+\log k)\right)^{i}}{k+1}\leq\frac{\sum_{i=1}^{k-1}\left(2c(1+\log k)\right)^{i}}{k}, \qquad (59)
\end{aligned}$$

and

$$\begin{aligned}
c\sum_{\ell=2}^{k}\sum_{i=0}^{\ell-2}\frac{\left(2c(1+\log k)\right)^{i}}{k-\ell+2} &= c\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i}\sum_{\ell=2+i}^{k}\frac{1}{k-\ell+2}\leq c\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i}\sum_{\ell=1}^{k}\frac{1}{\ell}\\
&\leq c(1+\log k)\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i}\leq\sum_{i=0}^{k-2}\left(2c(1+\log k)\right)^{i+1}\\
&= \sum_{i=1}^{k-1}\left(2c(1+\log k)\right)^{i}. \qquad (60)
\end{aligned}$$

Combining (58), (59) and (60), we obtain the following inequality and finish the induction

$$\begin{aligned}
d_{k+1} &\leq \frac{a}{k}+\frac{a}{k}\sum_{i=1}^{k-1}\left(2c(1+\log k)\right)^{i}+b(1+\log k)\left(1+\sum_{i=1}^{k-1}\left(2c(1+\log k)\right)^{i}\right)\\
&= \left(\frac{a}{k}+b(1+\log k)\right)\sum_{i=0}^{k-1}\left(2c(1+\log k)\right)^{i}.
\end{aligned}$$

□