Proceedings of Machine Learning Research vol 336:1–24, 2026 39th Annual Conference on Learning Theory

Random Reshuffling Dominates Stochastic Gradient Descent

Zijian Liu ZL3067@STERN.NYU.EDU
Stern School of Business, New York University

Editors: Steve Hanneke and Tor Lattimore

Abstract

Stochastic Gradient Descent (SGD) is one of the most classical optimization algorithms with favorable theoretical guarantees, yet the practical implementation of SGD differs subtly from its well-known form and is often referred to as Shuffling Stochastic Gradient Descent (Shuffling SGD). A particularly popular strategy in Shuffling SGD is Random Reshuffling (RR), which has achieved great empirical success across numerous experiments. Despite its strong performance, RR has long been considered a heuristic due to a lack of theoretical support. Over the last decade, people have finally established provable convergence rates for RR, thus justifying its observed superiority. However, for smooth convex optimization, two clouds over the convergence theory of RR remain to this day. More precisely, according to the current theory, Shuffling SGD under RR converges only when the stepsize is smaller than a threshold proportional to $1/n$, where n is the number of summands in the objective (or the number of data points). Consequently, the optimally tuned theoretical rate of Shuffling SGD under RR is strictly worse than that of SGD when the number of epochs is smaller than another threshold proportional to n. These two restrictions heavily limit the applicability of existing theories and leave a critical mismatch with practice. In this work, for the first time, we prove that RR dominates SGD in smooth convex optimization under any reasonable stepsize after any finite number of epochs, thereby addressing a longstanding open question.

Keywords: Convex Optimization, Stochastic Optimization, Random Reshuffling

1. Introduction

One of the fundamental tasks in machine learning is to optimize functions in a finite-sum form, i.e., $f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$. Among different optimization algorithms, Stochastic Gradient Descent (SGD), proposed in the seminal work of Robbins and Monro (1951), is arguably one of the most classical methods. Due to its easy implementation and computational efficiency, SGD is particularly popular when n is large, the standard case nowadays. More importantly, the convergence guarantees of SGD have been extensively studied, yielding provable rates in various settings (Polyak, 1987; Bottou et al., 2018; Lan, 2020), thereby providing a theoretical backbone for SGD.

However, compared with the standard form of SGD analyzed in theory, which uniformly samples a function to perform a gradient descent step at each iteration, the practical implementation differs subtly and is often referred to as Shuffling Stochastic Gradient Descent (Shuffling SGD). In Shuffling SGD, the optimization procedure is divided into K epochs, and within each epoch, the order in which functions are processed is determined by a permutation π of $\{1, \ldots, n\}$. A widely implemented strategy for generating π is Random Reshuffling (RR), which, in each epoch, independently and uniformly draws a new permutation from all possible ones.

Although Shuffling SGD under RR has achieved great empirical success across numerous experiments, it has long been considered a heuristic due to a lack of theoretical support. Over

0. In this work, we say that one optimization algorithm dominates another if the order of its convergence rate is no worse than that of the latter and is strictly better in certain regimes.

the last decade, beginning with the pioneering work of Gürbüzbalaban et al. (2021), people have finally established provable convergence rates for RR, thereby justifying its observed superiority over standard SGD.

In particular, for smooth convex optimization (i.e., each f_i is convex and L-smooth[1]), RR with a constant stepsize[2] η is known to converge in expectation at the rate $\frac{D^2}{\eta n K}+\eta^2 nL\sigma_\star^2$ (e.g., Mishchenko et al. (2020); Nguyen et al. (2021)), provided that the stepsize satisfies $\eta \lesssim \frac{1}{nL}$, where D denotes the distance between the initial point and the optimal solution, and $\sigma_\star^2$ is the gradient variance at the minimizer. In comparison, SGD under the same setting guarantees the in-expectation convergence rate $\frac{D^2}{\eta n K}+\eta\sigma_\star^2$ (Garrigos and Gower, 2023) but only requires $\eta \lesssim \frac{1}{L}$. Clearly, RR converges faster than SGD in the regime $\eta \lesssim \frac{1}{nL}$, which has been recognized as theoretical evidence demonstrating the strong performance of RR.

Despite the progress discussed above, some important issues remain unaddressed. The most critical longstanding open question is that people still do not understand what happens to RR when η falls into the regime $\eta \gtrsim \frac{1}{nL}$. This point is critical because, given that n is typically large in modern tasks, the threshold $\frac{1}{nL}$ can be extremely small or even vanish, whereas stepsizes used in practice are usually at a constant level, thereby leaving a significant gap between theory and practice. More crucially, even if one temporarily assumes that the existing rate for RR mentioned earlier could be extended to allow $\eta \lesssim \frac{1}{L}$ (though no such theory has been established), it would still fail to explain the advantage of RR over SGD, as the term $\eta^2 nL\sigma_\star^2$ for RR is worse than the term $\eta\sigma_\star^2$ for SGD when $\eta \gtrsim \frac{1}{nL}$. This hints that an analysis different from existing ones may be needed.

Another issue implied by the above discussion is that the optimally tuned rate for RR induced by the existing bound is only $\frac{LD^2}{K}+(\frac{L\sigma_\star^2 D^4}{nK^2})^{\frac{1}{3}}$, which is better than the best tuned rate $\frac{LD^2}{nK}+\frac{\sigma_\star D}{\sqrt{nK}}$ for SGD only when K is larger than a threshold proportional to n. Moreover, in the case of $\sigma_\star = 0$ (i.e., all f_i's share a common optimal solution), the rate of RR reduces to only $\frac{LD^2}{K}$, which is even worse than the $\frac{LD^2}{nK}$ rate of SGD by a factor of $\frac{1}{n}$.

The above restrictions on the stepsize η or the number of epochs K heavily limit the applicability of existing theories and cannot fully explain the favorable performance of Shuffling SGD under RR compared with standard SGD. Therefore, we are naturally led to the following question:

In smooth convex optimization, does RR dominate SGD without these two restrictions?

1.1. Our Contributions

This work provides an affirmative answer to the above question.

- Concretely, we show that Shuffling SGD under RR provably converges at a rate of $\frac{D^2}{\eta n K}+\min\{1,\eta nL\}\eta\sigma_\star^2$ for any stepsize satisfying $\eta \lesssim \frac{1}{L}$ (see Theorem 1 for the formal version with nonuniform smoothness parameters and dynamic stepsizes that depend on the epoch number). We highlight that this rate is not only the first provable result for RR that allows $\eta \gtrsim \frac{1}{nL}$, but also provably dominates the $\frac{D^2}{\eta n K}+\eta\sigma_\star^2$ bound of SGD under any stepsize $\eta \lesssim \frac{1}{L}$. It is noteworthy that this rate is not merely a simple extension of the previously best bound for RR, since the latter is slower than SGD when $\eta \gtrsim \frac{1}{nL}$, as discussed before.

1. For simplicity, we adopt a uniform smoothness parameter in the discussion, as in most of the existing literature.
2. We also use a constant stepsize in the discussion for convenience.

- Consequently, the optimally tuned rate for Shuffling SGD under RR is improved to $\frac{LD^2}{nK} + \min\{\frac{\sigma_\star D}{\sqrt{nK}}, (\frac{L\sigma_\star^2 D^4}{nK^2})^{\frac{1}{3}}\}$ (see Corollary 1 for the formal version with nonuniform smoothness parameters), which dominates the best tuned bound $\frac{LD^2}{nK} + \frac{\sigma_\star D}{\sqrt{nK}}$ of SGD for any finite K. Moreover, the rate reduces to $\frac{LD^2}{nK}$ when $\sigma_\star = 0$, improving upon the best known result $\frac{LD^2}{K}$ by a factor of $\frac{1}{n}$.

In summary, for the first time, we prove that RR dominates standard SGD in smooth convex optimization under any reasonable stepsize after any finite number of epochs, resolving a longstanding open question.

1.2. Related Work

We provide a brief overview of Shuffling SGD under RR and defer further details to Appendix A.

Over the past decades, the effectiveness of RR has been reported in many works (e.g., Bottou (2009, 2012); Bengio (2012)). However, the theoretical understanding of it has long lagged behind. The first breakthrough is by Gürbüzbalaban et al. (2021), which provides the first theoretical evidence that RR can beat SGD in smooth strongly convex optimization under certain additional assumptions. Since then, RR has been extensively studied (e.g., Nagaraj et al. (2019); Haochen and Sra (2019)). To date, the best known rate in smooth convex optimization is $\frac{D^2}{\eta nK} + \eta^2 nL\sigma_\star^2$ under the condition $\eta \lesssim \frac{1}{nL}$ (Mishchenko et al., 2020; Nguyen et al., 2021). The only existing lower bound in smooth convex optimization is $(\frac{L\sigma_\star^2 D^4}{nK^2})^{\frac{1}{3}}$, due to Cha et al. (2023), which holds for constant stepsizes $\eta \lesssim \frac{1}{nL}$ and a large number of epochs at least satisfying $K \gtrsim \frac{nL^2D^2}{\sigma_\star^2}$.

2. Preliminary

Notation. $\mathbb{N}$ denotes the set of natural numbers (excluding 0). Given $n \in \mathbb{N}$, we write $[n] \triangleq \{1, \ldots, n\}$. $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product, and $\|\cdot\| \triangleq \sqrt{\langle \cdot, \cdot \rangle}$ is the ℓ_2 norm. Given a real-valued differentiable function $h : \mathbb{R}^d \to \mathbb{R}$, $\nabla h(\mathbf{x})$ denotes the gradient at $\mathbf{x} \in \mathbb{R}^d$. The Bregman divergence induced by h is defined as $\mathrm{B}_h(\mathbf{x}, \mathbf{y}) \triangleq h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$, which is nonnegative if h is additionally convex.

Objective. We study the following finite-sum optimization problem in this work

$$\inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}),$$

where $n \in \mathbb{N}$ and each $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable.

Remark 1 *To ease notation, hereinafter we use* $\mathrm{B} \triangleq \mathrm{B}_f$ *and* $\mathrm{B}_i \triangleq \mathrm{B}_{f_i}$ *to denote the Bregman divergences induced by* f *and* f_i*, respectively.*

Assumptions. Our analysis relies on the following three assumptions.

Assumption 1 (Minimizer) $\exists \mathbf{x}_\star \in \mathbb{R}^d$ *such that* $f_\star \triangleq f(\mathbf{x}_\star) = \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \in \mathbb{R}$.

Assumption 2 (Convexity) *Each* f_i *is convex.*

Assumption 3 (Smoothness) *Each f_i is L_i-smooth, i.e., $\exists L_i > 0$ such that $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L_i \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

All three of the above assumptions are standard and commonly adopted in the literature (Polyak, 1987; Nesterov et al., 2018; Bottou et al., 2018; Lan, 2020). Notably, we do not impose any assumptions on the difference between the individual gradient ∇f_i and the full gradient ∇f, such as the popular finite variance condition.

Next, we introduce three more notations to simplify the expressions in the subsequent sections. $\sigma_\star^2$ denotes the variance of the gradient at the minimizer $\mathbf{x}_\star$, and $\bar{L}$ (resp. $\hat{L}$) represents the average (resp. maximum) smoothness parameter, i.e.,

$$\sigma_\star^2 \triangleq \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}_\star)\|^2, \quad \bar{L} \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^{n} L_i, \quad \hat{L} \triangleq \max_{i\in[n]} L_i.$$

We note that the quantity $\sigma_\star^2$ is widely used in prior works on shuffling gradient methods (e.g., Ying et al. (2019); Mishchenko et al. (2020); Nguyen et al. (2021)) and remains invariant even when f has multiple minimizers (see Lemma 4.17 of Garrigos and Gower (2023)). In particular, $\sigma_\star^2 = 0$ corresponds to the case in which all f_i share a common optimal solution.

To finish this section, we state a classical result in convex optimization, known as the co-coercivity property of smooth convex functions, which serves as a key tool in our analysis. As for its proof, see, for example, Theorem 2.15 of Nesterov et al. (2018).

Lemma 1 (Co-coercivity) *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function that is also L-smooth, then we have, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$\|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\|^2 \leq 2L\mathrm{B}_h(\mathbf{x}, \mathbf{y}) \quad \textit{and} \quad \|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\|^2 \leq L \langle \nabla h(\mathbf{x}) - \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle.$$

3. Shuffling Stochastic Gradient Descent

Algorithm 1 Shuffling Stochastic Gradient Descent (Shuffling SGD)

Input: initial point $\mathbf{x}_1^1 \in \mathbb{R}^d$, stepsize $\eta_k > 0$
for $k = 1$ **to** K **do**
  Generate a permutation π_k of $[n]$
  for $i = 1$ **to** n **do**
    $\mathbf{x}_k^{i+1} = \mathbf{x}_k^i - \eta_k \nabla f_{\pi_k^i}(\mathbf{x}_k^i)$
  end for
  $\mathbf{x}_{k+1}^1 = \mathbf{x}_k^{n+1}$
end for

The method studied in this work, Shuffling Stochastic Gradient Descent (Shuffling SGD), is given in Algorithm 1. Compared with the standard SGD algorithm, which uniformly samples a function to process at each step, Shuffling SGD determines the order in which functions are passed in each epoch based on a permutation. In particular, three strategies for generating permutations are popular in practice, as illustrated in the following examples, where S_n denotes the symmetric group of $[n]$.

Example 1 (Random Reshuffling (RR)) *Each π_k is drawn independently and uniformly from S_n.*

Example 2 (Single Shuffling (SS)) *Each $\pi_k = \pi$, a permutation drawn uniformly from S_n.*

Example 3 (Incremental Gradient (IG)) *Each $\pi_k = \pi$, a deterministic permutation from S_n.*

3.1. New Rate for RR

We are now ready to provide the main result, Theorem 1, a new convergence rate for RR.

Theorem 1 *Under Assumptions 1, 2, and 3, suppose RR is employed with $\eta_k \le \frac{1}{6\hat{L}}, \forall k \in [K]$, let*

$$\bar{\mathbf{x}}_K \triangleq \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\eta_k}{nH_K}\mathbf{x}_k^i \quad \text{where} \quad H_K \triangleq \sum_{k=1}^{K}\eta_k, \tag{1}$$

then Shuffling SGD (Algorithm 1) guarantees that

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K) - f_\star\right] \le \frac{6\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|^2}{n\sum_{k=1}^{K}\eta_k} + 51\min\left\{\frac{\sum_{k=1}^{K}\eta_k^2}{\sum_{k=1}^{K}\eta_k}, \frac{\sum_{k=1}^{K}\eta_k^3 n\bar{L}}{\sum_{k=1}^{K}\eta_k}\right\}\sigma_\star^2.$$

Remark 2 *We make no effort to optimize the constants in the bounds obtained in this work.*

Proof The proof is deferred to Subsection 4.4. ∎

To the best of our knowledge, in smooth convex optimization, Theorem 1 shows the first convergence rate for RR under any reasonable stepsize, i.e., $\eta_k \lesssim 1/\hat{L}$, thereby improving existing results in different aspects. Previously, the best known convergence rate for RR is $\frac{D^2}{n\sum_{k=1}^{K}\eta_k} + \frac{\sum_{k=1}^{K}\eta_k^3 n\bar{L}\sigma_\star^2}{\sum_{k=1}^{K}\eta_k}$ (where $D \triangleq \left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|$) under the condition $\eta_k \lesssim 1/(n\sqrt{\bar{L}\hat{L}})$ (Liu and Zhou, 2024; Cai et al., 2024). However, this requirement on η_k is highly unrealistic, as n is typically pretty large in modern machine learning tasks, meaning that the stepsize has to be extremely small or even vanishing. This contradicts the constant-level stepsizes commonly used in practice and leads to a gap between theory and practice.

In comparison, Theorem 1 not only allows the stepsize to lie in the constant regime $\eta_k \lesssim 1/\hat{L}$ but also shows a fundamental improvement rather than a mere extension of the previously best known rate, since the latter becomes slower than SGD once $\eta_k \gtrsim 1/(n\bar{L})$ (the rate of SGD is stated below), while our Theorem 1 never does.

In addition, compared with the $\frac{D^2}{n\sum_{k=1}^{K}\eta_k} + \frac{\sum_{k=1}^{K}\eta_k^2\sigma_\star^2}{\sum_{k=1}^{K}\eta_k}$ rate of standard SGD (Garrigos and Gower, 2023), our Theorem 1 is never worse for any reasonable stepsize (i.e., $\eta_k \lesssim 1/\hat{L}$), and is strictly better when $\eta_k \lesssim 1/(n\bar{L})$. This feature has an important implication: if both methods employ their own optimally tuned stepsizes, Shuffling SGD under RR provably achieves a better upper bound that dominates the $\frac{\hat{L}D^2}{nK} + \frac{\sigma_\star D}{\sqrt{nK}}$ rate of SGD after any finite number of epochs, as evidenced by Corollary 1 below.

Corollary 1 *Under the same setting as in Theorem 1, with the optimally tuned constant stepsize $\eta_k = \eta_\star, \forall k \in [K]$, where $\eta_\star \le \frac{1}{6\hat{L}}$, Shuffling SGD (Algorithm 1) guarantees that*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K) - f_\star\right] \lesssim \frac{\hat{L}\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|^2}{nK} + \min\left\{\frac{\sigma_\star\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|}{\sqrt{nK}}, \left(\frac{\bar{L}\sigma_\star^2\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|^4}{nK^2}\right)^{\frac{1}{3}}\right\}.$$

Proof The proof is deferred to Subsection 4.4. ■

In contrast, the known optimally tuned rate for RR is only $\frac{\sqrt{\bar{L}\hat{L}}D^2}{K}+(\frac{\bar{L}\sigma_\star^2D^4}{nK^2})^{\frac{1}{3}}$. To better understand the differences among these optimally tuned rates, let $\mathsf{Order}_{\mathsf{SGD}}(K)$ and $\mathsf{Order}_{\mathsf{RR}}^{\mathsf{Old}}(K)$ denote the dominant terms in the existing optimally tuned rates for SGD and Shuffling SGD under RR, respectively, i.e.,

$$\mathsf{Order}_{\mathsf{SGD}}(K)\triangleq\max\left\{\frac{\hat{L}D^2}{nK},\frac{\sigma_\star D}{\sqrt{nK}}\right\}\text{ and }\mathsf{Order}_{\mathsf{RR}}^{\mathsf{Old}}(K)\triangleq\max\left\{\frac{\sqrt{\bar{L}\hat{L}}D^2}{K},\left(\frac{\bar{L}\sigma_\star^2D^4}{nK^2}\right)^{\frac{1}{3}}\right\}.$$

Similarly, $\mathsf{Order}_{\mathsf{RR}}^{\mathsf{New}}(K)$ denotes the dominant term in the rate obtained in Corollary 1, i.e.,

$$\mathsf{Order}_{\mathsf{RR}}^{\mathsf{New}}(K)\triangleq\max\left\{\frac{\hat{L}D^2}{nK},\min\left\{\frac{\sigma_\star D}{\sqrt{nK}},\left(\frac{\bar{L}\sigma_\star^2D^4}{nK^2}\right)^{\frac{1}{3}}\right\}\right\}.$$

In the nondegenerate case $\sigma_\star\neq 0$ (which implies that $n\geq 2$), the following comparisons hold:

$$\begin{cases}\mathsf{Order}_{\mathsf{RR}}^{\mathsf{New}}(K)=\mathsf{Order}_{\mathsf{SGD}}(K)\overset{(a)}{\leq}\mathsf{Order}_{\mathsf{RR}}^{\mathsf{Old}}(K), & K\leq\frac{n\bar{L}^2D^2}{\sigma_\star^2},\\ \mathsf{Order}_{\mathsf{RR}}^{\mathsf{New}}(K)<\mathsf{Order}_{\mathsf{SGD}}(K)\overset{(b)}{\leq}\mathsf{Order}_{\mathsf{RR}}^{\mathsf{Old}}(K), & \frac{n\bar{L}^2D^2}{\sigma_\star^2}<K\leq\frac{n\bar{L}\hat{L}D^2}{\sigma_\star^2},\\ \mathsf{Order}_{\mathsf{RR}}^{\mathsf{New}}(K)\overset{(c)}{\leq}\mathsf{Order}_{\mathsf{RR}}^{\mathsf{Old}}(K)<\mathsf{Order}_{\mathsf{SGD}}(K), & \frac{n\bar{L}\hat{L}D^2}{\sigma_\star^2}<K\leq\frac{n\bar{L}^{1/2}\hat{L}^{3/2}D^2}{\sigma_\star^2},\\ \mathsf{Order}_{\mathsf{RR}}^{\mathsf{New}}(K)=\mathsf{Order}_{\mathsf{RR}}^{\mathsf{Old}}(K)<\mathsf{Order}_{\mathsf{SGD}}(K), & K>\frac{n\bar{L}^{1/2}\hat{L}^{3/2}D^2}{\sigma_\star^2},\end{cases}$$

where (a) becomes an equality if and only if $K=\frac{n\bar{L}^2D^2}{\sigma_\star^2}$ and $\bar{L}=\hat{L}$, (b) becomes an equality if and only if $K=\frac{n\bar{L}\hat{L}D^2}{\sigma_\star^2}$, and (c) becomes an equality if and only if $K=\frac{n\bar{L}^{1/2}\hat{L}^{3/2}D^2}{\sigma_\star^2}$. As one can see, unlike the existing optimally tuned bound for RR in the literature, which can be slower than standard SGD when $K\leq\frac{n\bar{L}\hat{L}D^2}{\sigma_\star^2}$, the new result in Corollary 1 is never worse than the optimally tuned rate of standard SGD and strictly beats it once $K>\frac{n\bar{L}^2D^2}{\sigma_\star^2}$, thereby improving the threshold from $\frac{n\bar{L}\hat{L}D^2}{\sigma_\star^2}$ to $\frac{n\bar{L}^2D^2}{\sigma_\star^2}$.

Moreover, in the special case where all f_i's share a common minimizer, or equivalently when $\sigma_\star=0$, Corollary 1 achieves the same $\frac{\hat{L}D^2}{nK}$ rate as SGD. However, the prior bound for RR can only be reduced to a slower rate in the order of $\frac{\sqrt{\bar{L}\hat{L}}D^2}{K}$ due to the stepsize restriction $\eta_k\lesssim 1/(n\sqrt{\bar{L}\hat{L}})$, as discussed before. Formally, once $n\geq 2$, we have

$$\mathsf{Order}_{\mathsf{RR}}^{\mathsf{New}}(K)=\mathsf{Order}_{\mathsf{SGD}}(K)<\mathsf{Order}_{\mathsf{RR}}^{\mathsf{Old}}(K),\forall K\in\mathbb{N}.$$

In summary, Theorem 1 and Corollary 1 together imply that RR dominates SGD in smooth convex optimization under any reasonable stepsize after any finite number of epochs.

4. Theoretical Analysis

In this section, we lay the groundwork for proving Theorem 1 and complete its proof at the end. The section is organized into four parts. First, we provide the two most important lemmas in Subsection

4.1. Next, in Subsection 4.2, we establish an upper bound in Theorem 2, which indicates that Shuffling SGD under RR never converges more slowly than SGD under any reasonable stepsize. Then, Theorem 3 in Subsection 4.3 presents an alternative convergence rate for Shuffling SGD under RR, which demonstrates that Algorithm 1 under RR provably converges faster than SGD when the stepsize is sufficiently small. Finally, in Subsection 4.4, we conclude Theorem 1 and then use it to prove Corollary 1.

4.1. Two Core Lemmas

This subsection contains two core lemmas, both of which are critical to our analysis.

Before presenting the lemmas, we introduce two notions. Given a permutation π of $[n]$ and two indices $i, j \in [n]$ satisfying $j \leq i$, we define the following new permutation

$$\pi(i,j) \triangleq \left(\pi^1, \ldots, \pi^{j-1}, \pi^i, \pi^{j+1}, \ldots, \pi^{i-1}, \pi^j, \pi^{i+1}, \ldots, \pi^n\right). \tag{2}$$

In words, $\pi(i,j)$ is the permutation generated by exchanging the elements π_i and π_j in π. Equipped with the notion of $\pi(i,j)$, we introduce the following virtual sequence, for any given $k \in [K]$,

$$\mathbf{x}_k^{l+1}(i,j) \triangleq \mathbf{x}_k^l(i,j) - \eta_k \nabla f_{\pi_k^l(i,j)}(\mathbf{x}_k^l(i,j)), \forall l \in [n], \quad \text{where } \mathbf{x}_k^1(i,j) \triangleq \mathbf{x}_k^1. \tag{3}$$

This means that the sequence $\mathbf{x}_k^l(i,j), \forall l \in [n+1]$ denotes the trajectory of the k-th epoch starting from $\mathbf{x}_k^1$, produced by Algorithm 1, but under the permutation $\pi_k(i,j)$. This virtual iterate can be viewed as a coupled sequence of the real output and, to the best of our knowledge, was first introduced by Sherman et al. (2021). It plays a fundamental role in our proof, as will become clear.

With these two concepts in hand, we proceed to state the two core lemmas. The first is Lemma 2, which is based on Lemma 2 of Sherman et al. (2021). For completeness, we reproduce the proof of Lemma 2 in Appendix C.

Lemma 2 *Given an arbitrary finite-sum function $\ell(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbf{x})$, suppose RR is employed, then for any $k \in [K]$ and $i \in [n]$, Shuffling SGD (Algorithm 1) guarantees that*

$$\mathbb{E}\left[\ell(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right] = \frac{1}{n}\sum_{j<i} \mathbb{E}\left[\ell_{\pi_k^i}(\mathbf{x}_k^i(i,j)) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right],$$

where $\mathbf{x}_k(i,j)$ is defined in (3).

In the analysis of shuffling gradient methods, a well-known major challenge is to properly bound the term $\mathbb{E}\left[f(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_k^i)\right]$, unlike in SGD, which no longer equals 0 due to the nature of shuffling-based algorithms. Lemma 2 provides a possible approach by relating the term we want to control (in a slightly more general form, applicable to any finite-sum function ℓ) to another quantity involving the virtual sequence introduced earlier in (3).

For the convenience of the discussion, temporarily assume $\ell_i = f_i$ in Lemma 2. Then, under the smoothness assumption, one would expect the difference between $\ell_{\pi_k^i}(\mathbf{x}_k^i(i,j))$ and $\ell_{\pi_k^i}(\mathbf{x}_k^i)$ to be small whenever $\mathbf{x}_k^i(i,j)$ and $\mathbf{x}_k^i$ are close. This observation naturally leads us to the other core Lemma 3 stated below.

Lemma 3 *Under Assumptions 2 and 3, suppose $\eta_k \leq \frac{2}{\hat{L}}, \forall k \in [K]$, then for any $k \in [K]$, $i \in [n]$, and $j \in [i-1]$, Shuffling SGD (Algorithm 1) guarantees that*

$$\left\|\mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\| \leq \eta_k \left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|,$$

where $\mathbf{x}_k(i,j)$ is defined in (3).

Lemma 3 quantifies how close the virtual iterate and the true trajectory can be, under the widely required condition of $\eta_k \leq 2/\hat{L}$ in smooth optimization. Although the inequality does not directly offer a bound on the distance between $\mathbf{x}_k^i(i,j)$ and $\mathbf{x}_k^i$ that depends only on deterministic terms (e.g., the stepsize η_k), it is sufficient for our proof when combined with a careful analysis.

The proof of Lemma 3 is given in Appendix C and relies on the co-coercivity property (i.e., Lemma 1), which is closely related to the nonexpansiveness of the update rule in gradient methods (Polyak, 1987; Nesterov et al., 2018).

4.2. Bound I: Never Worse than SGD under Reasonable Stepsize

In this subsection, we give the first bound for Shuffling SGD under RR, stated in Theorem 2 below.

Theorem 2 *Under Assumptions 1, 2, and 3, suppose RR is employed with $\eta_k \leq \frac{1}{6\hat{L}}, \forall k \in [K]$, then Shuffling SGD (Algorithm 1) guarantees that*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K) - f_\star\right] \leq \frac{\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|^2}{2n\sum_{k=1}^K \eta_k} + \frac{51\sum_{k=1}^K \eta_k^2 \sigma_\star^2}{\sum_{k=1}^K \eta_k},$$

where $\bar{\mathbf{x}}_K$ is defined in (1).

Discussion on Theorem 2. To the best of our knowledge, Theorem 2 offers the first theoretical evidence that Shuffling SGD under RR shares surprising similarities with SGD, as reflected in the two aspects elaborated below.

First, Theorem 2 states that, similar to SGD, Shuffling SGD under RR does converge under any reasonable stepsize (i.e., $\eta_k \lesssim 1/\hat{L}$). In contrast, as far as we know, all prior works that provide provable rates of multi-epoch RR for smooth convex optimization require the stepsize η_k to be smaller than a threshold proportional to $1/n$, with only one exception (Nagaraj et al., 2019), which, however, assumes each f_i to be additionally Lipschitz, thereby limiting the applicability of their theory and even excluding common quadratic optimization problems over $\mathbb{R}^d$.

Second, we highlight that Theorem 2 gives the same convergence upper bound (up to constant factors) as SGD (Garrigos and Gower, 2023) in smooth convex optimization, while allowing a stepsize that depends on the epoch number. This result thus fills a gap in the literature.

Putting these together, Theorem 2 indicates that, for smooth convex optimization, RR under any reasonable stepsize never converges more slowly than SGD.

Analysis. In the following, we present the analysis for Theorem 2 and finally prove it. The core idea underlying the proof is, as one might expect, to analyze Shuffling SGD in a manner analogous to SGD. In other words, we aim to quantify the progress made by Algorithm 1 at each iteration. Although this perspective is natural, it has been less explored in prior studies. The main reason is that, as discussed earlier, the permutation in Shuffling SGD causes the most important property of

SGD, unbiasedness, to no longer hold. To overcome this barrier, we develop a novel analysis that avoids any additional assumptions, such as Lipschitz continuity considered in Nagaraj et al. (2019).

We start with the following Lemma 4, a standard step in characterizing the per-iterate progress of Shuffling SGD (or SGD). The proof of Lemma 4 follows directly from expanding both sides. To make the work self-contained, we include it in Appendix D.

Lemma 4 *Under Assumption 1, for any $k \in [K]$ and $i \in [n]$, Shuffling SGD (Algorithm 1) guarantees that*

$$f_{\pi_k^i}(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_\star) = \frac{\left\|\mathbf{x}_k^i - \mathbf{x}_\star\right\|^2 - \left\|\mathbf{x}_k^{i+1} - \mathbf{x}_\star\right\|^2}{2\eta_k} + \frac{\eta_k}{2}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 - \mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i).$$

By the co-coercivity property (i.e., Lemma 1), the term $\frac{\eta_k}{2}\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\|^2 - \mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i)$ can be easily upper bounded by $\eta_k\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\|^2$ (ignoring constant factors) once $\eta_k \lesssim 1/\hat{L}$, which further yields a desired residual term $\eta_k\sigma_\star^2$ after taking expectations. Therefore, the only difficulty is to relate $f_{\pi_k^i}(\mathbf{x}_k^i)$ to $f(\mathbf{x}_k^i)$, which is, again, the main challenge in the analysis of Shuffling SGD.

To address the issue mentioned, the prior work of Nagaraj et al. (2019) applies an argument based on Wasserstein distance, which additionally requires the Lipschitz continuity of each f_i. In comparison, we tackle this problem by establishing the following new inequality in Lemma 5.

Lemma 5 *Under Assumptions 2 and 3, suppose RR is employed with $\eta_k \le \frac{1}{6\hat{L}}, \forall k \in [K]$, then for any $k \in [K]$ and $i \in [n]$, Shuffling SGD (Algorithm 1) guarantees that*

$$\mathbb{E}\left[f(\mathbf{x}_k^i)\right] \le \mathbb{E}\left[f_{\pi_k^i}(\mathbf{x}_k^i)\right] + \eta_k\mathbb{E}\left[\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2\right] + \frac{4\eta_k}{3n}\sum_{j<i}\mathbb{E}\left[\left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\right].$$

Remark 3 *Lemma 5 is stronger than the existing bound of Nagaraj et al. (2019) derived via the Wasserstein distance, since imposing the additional condition $\|\nabla f_i(\mathbf{x})\| \le G$, as in Nagaraj et al. (2019), recovers their Lemma 4.*

Lemma 5 provides a novel inequality that measures the difference between $\mathbb{E}\left[f(\mathbf{x}_k^i)\right]$ and $\mathbb{E}[f_{\pi_k^i}(\mathbf{x}_k^i)]$ by the second moment of the stochastic gradients up to time i. Note that the second term on the R.H.S. can be absorbed by the R.H.S. of the inequality in Lemma 4. For the remaining term, the coefficient η_k/n is key to the final proof, which ensures that the accumulated error in one epoch is controlled by $\mathbb{E}[\eta_k\sum_{i=1}^n\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\|^2]$.

The proof of Lemma 5 builds on the two core results, Lemmas 2 and 3, presented before. To save space, we defer it to Appendix D.

Final proof. With Lemmas 4 and 5 stated above, we are ready to prove Theorem 2.

Proof of Theorem 2 We sum the inequality in Lemma 4 from $i = 1$ to n and use $\mathbf{x}_{k+1}^1 = \mathbf{x}_k^{n+1}$ to obtain

$$\sum_{i=1}^n f_{\pi_k^i}(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_\star) = \frac{\left\|\mathbf{x}_k^1 - \mathbf{x}_\star\right\|^2 - \left\|\mathbf{x}_{k+1}^1 - \mathbf{x}_\star\right\|^2}{2\eta_k} + \frac{\eta_k}{2}\sum_{i=1}^n\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 - \sum_{i=1}^n \mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i).$$

Take expectations on both sides and note that $\mathbb{E}\left[f_{\pi_k^i}(\mathbf{x}_\star)\right] = f_\star$ to yield

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left[f_{\pi_k^i}(\mathbf{x}_k^i) - f_\star\right] =& \frac{\mathbb{E}\left[\left\|\mathbf{x}_k^1 - \mathbf{x}_\star\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{x}_{k+1}^1 - \mathbf{x}_\star\right\|^2\right]}{2\eta_k} \\ &+ \mathbb{E}\left[\sum_{i=1}^{n}\frac{\eta_k}{2}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 - \mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i)\right].\end{aligned} \tag{4}$$

Next, we invoke Lemma 5 and sum it up from $i = 1$ to n to have

$$\sum_{i=1}^{n}\mathbb{E}\left[f(\mathbf{x}_k^i)\right] \le \sum_{i=1}^{n}\mathbb{E}\left[f_{\pi_k^i}(\mathbf{x}_k^i)\right] + \mathbb{E}\left[\frac{7\eta_k}{3}\sum_{i=1}^{n}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2\right]. \tag{5}$$

Combine (4) and (5) to obtain

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left[f(\mathbf{x}_k^i) - f_\star\right] \le& \frac{\mathbb{E}\left[\left\|\mathbf{x}_k^1 - \mathbf{x}_\star\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{x}_{k+1}^1 - \mathbf{x}_\star\right\|^2\right]}{2\eta_k} \\ &+ \mathbb{E}\left[\sum_{i=1}^{n}\frac{17\eta_k}{6}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 - \mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i)\right].\end{aligned} \tag{6}$$

One more step, we observe that

$$\begin{aligned}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 &\le \frac{18}{17}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i) - \nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 18\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 \\ &\overset{(a)}{\le} \frac{36}{17}L_{\pi_k^i}\mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i) + 18\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 \\ \Rightarrow \mathbb{E}\left[\frac{17\eta_k}{6}\sum_{i=1}^{n}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2\right] &\overset{(b)}{\le} \mathbb{E}\left[\sum_{i=1}^{n}\mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i)\right] + 51\eta_k n\sigma_\star^2,\end{aligned} \tag{7}$$

where (a) is by Lemma 1 and (b) is due to $\eta_k \le \frac{1}{6\hat{L}}$ and $\mathbb{E}\left[\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2\right] = \sigma_\star^2, \forall k \in [K], i \in [n]$.

Finally, we plug (7) back into (6), multiply both sides by η_k, sum over $k = 1$ to K, divide both sides by $n\sum_{k=1}^{K}\eta_k$, apply the convexity of f, and use the definition of $\bar{\mathbf{x}}_K$ in (1) to conclude. ■

4.3. Bound II: Always Better than SGD under Small Stepsize

This subsection presents the other rate of Shuffling SGD under RR, as shown in Theorem 3 below.

Theorem 3 *Under Assumptions 1, 2, and 3, suppose RR is employed with $\eta_k \le \frac{1}{2\hat{L}}, \forall k \in [K]$, then Shuffling SGD (Algorithm 1) guarantees that*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K) - f_\star\right] \le \frac{6\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|^2}{n\sum_{k=1}^{K}\eta_k} + \frac{8\sum_{k=1}^{K}\eta_k^3 n\bar{L}\sigma_\star^2}{\sum_{k=1}^{K}\eta_k},$$

where $\bar{\mathbf{x}}_K$ is defined in (1).

Discussion on Theorem 3. Readers familiar with the literature on shuffling gradient methods may readily figure out that the rate given in Theorem 3 perfectly matches the known bound for Shuffling SGD under RR in smooth convex optimization (e.g., Mishchenko et al. (2020); Nguyen et al. (2021)). However, we emphasize a key difference here, that is, the stepsize in our Theorem 3 is allowed to satisfy $\eta_k \lesssim 1/\hat{L}$, in contrast to all existing results that require η_k to be at most inversely proportional to n.

More importantly, in the setting of nonuniform L_i considered in this work, the largest threshold on the stepsize in the literature that guarantees a rate similar to Theorem 3 is in the order of $1/(n\sqrt{\bar{L}\hat{L}})$ (Liu and Zhou, 2024; Cai et al., 2024). But as indicated by our Theorem 3, the superiority of RR over SGD already exists once $\eta_k \lesssim 1/(n\bar{L})$. Especially, this improvement can be significant when a dominant smoothness parameter exists, leading to $\hat{L} \approx n\bar{L}$.

Therefore, Theorem 3 is the first result to extend the known bound in smooth convex optimization to any reasonable stepsize while preserving the favorable property of Algorithm 1, i.e., Shuffling SGD under RR provably converges faster than SGD when the stepsize is sufficiently small.

Analysis. The roadmap for establishing Theorem 3 differs wildly from that used before to prove Theorem 2. This time, our proof strategy is to check how close Shuffling SGD can be to Gradient Descent. More concretely, we will view each epoch of Algorithm 1 (containing n iterations) as a single step and analyze the progress made by it at once. This kind of approach has appeared in different previous works (e.g., Mishchenko et al. (2020); Nguyen et al. (2021)) and always yields a convergence rate in the order of $1/(n^{1/3}K^{2/3})$, faster than SGD when K is large.

However, all works that follow the idea described above share the same issue, that is, they require the stepsize to be in the order of $1/n$, which is, however, not enough for our purpose. In the following, we develop a new analysis to bypass this critical obstacle.

We now formally begin the analysis by introducing another virtual sequence, defined as follows, for any $k \in [K]$,

$$\mathbf{y}_k^{i+1} \triangleq \mathbf{y}_k^i - \eta_k \nabla f_{\pi_k^i}(\mathbf{x}_\star), \forall i \in [n], \quad \mathbf{y}_{k+1}^1 \triangleq \mathbf{y}_k^{n+1}, \quad \text{where } \mathbf{y}_1^1 \triangleq \mathbf{x}_\star. \tag{8}$$

Under the above definition, and noting that $\nabla f(\mathbf{x}_\star) = \mathbf{0}$, one can find

$$\mathbf{y}_k^{n+1} = \mathbf{y}_k^1 - \eta_k \sum_{i=1}^{n} \nabla f_{\pi_k^i}(\mathbf{x}_\star) = \mathbf{y}_k^1 - \eta_k n \nabla f(\mathbf{x}_\star) = \mathbf{y}_k^1, \forall k \in [K].$$

Combine the above line and $\mathbf{y}_1^1 = \mathbf{x}_\star$ as defined in (8) to have

$$\mathbf{y}_k^{n+1} = \mathbf{y}_k^1 = \mathbf{x}_\star, \forall k \in [K]. \tag{9}$$

The above virtual sequence is inspired by the work of Mishchenko et al. (2020), which, as far as we know, was the first to propose a similar term under the constant stepsize. Here, we slightly extend their idea to accommodate the case where the stepsize can depend on the current epoch number.

Remark 4 *We note that Mishchenko et al. (2020) introduced the virtual sequence to handle the case of individual strong convexity, i.e., each f_i is required to be strongly convex. However, in our setting, only individual convexity is assumed. This means that their proof cannot be applied. As such, our analysis substantially departs from the existing approach.*

Equipped with the new virtual sequence introduced above, we first present the following Lemma 6.

Lemma 6 *Under Assumptions 1, 2, and 3, suppose $\eta_k \leq \frac{1}{2\hat{L}}, \forall k \in [K]$, then for any $k \in [K]$,* Shuffling SGD *(Algorithm 1) guarantees that*

$$\eta_k \sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) \leq \left\|\mathbf{x}_k^1 - \mathbf{x}_\star\right\|^2 - \left\|\mathbf{x}_{k+1}^1 - \mathbf{x}_\star\right\|^2 + 2\eta_k \sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star),$$

where $\mathbf{y}_k^i$ is defined in (8).

As discussed earlier, we intentionally treat each epoch of Shuffling SGD as a single step. Hence, compared with Lemma 4 used to prove the first bound, Lemma 6 is in a different flavor, which shows the progress made by Algorithm 1 over an entire epoch.

Based on the form of Lemma 6, two tasks naturally arise. The first is to lower bound the term $\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)$ on the L.H.S. by $\sum_{i=1}^{n} \mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star) = \sum_{i=1}^{n} f(\mathbf{x}_k^i) - f_\star$. The second is to upper bound the residual term $\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star)$ on the R.H.S.

We address the first task by establishing the following Lemma 7, a novel inequality upper bounding each $\mathbb{E}\left[\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star)\right]$ in terms of $\mathbb{E}\left[\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right]$, $\frac{1}{n}\sum_{j<i} \mathbb{E}\left[\mathrm{B}_{\pi_k^j}(\mathbf{x}_k^j, \mathbf{x}_\star)\right]$, and $\sigma_\star^2$.

Lemma 7 *Under Assumptions 1, 2, and 3, suppose* RR *is employed with $\eta_k \leq \frac{1}{2\hat{L}}, \forall k \in [K]$, then for any $k \in [K]$ and $i \in [n]$,* Shuffling SGD *(Algorithm 1) guarantees that*

$$\mathbb{E}\left[\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] \leq 2\mathbb{E}\left[\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] + \frac{4}{n}\sum_{j<i} \mathbb{E}\left[\mathrm{B}_{\pi_k^j}(\mathbf{x}_k^j, \mathbf{x}_\star)\right] + 12\eta_k^2(i-1)\bar{L}\sigma_\star^2.$$

To the best of our knowledge, we are the first to obtain an inequality in such a form. Intuitively, Lemma 7 says that each $\mathbb{E}\left[\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star)\right]$ differs from its stochastic counterpart $\mathbb{E}\left[\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right]$ by at most a multiplicative constant, together with an average of the preceding terms $\mathbb{E}\left[\mathrm{B}_{\pi_k^j}(\mathbf{x}_k^j, \mathbf{x}_\star)\right]$ satisfying $j < i$ (not an exact average due to the coefficient $4/n$), plus an additional term involving $\sigma_\star^2$. Summing the inequality in Lemma 7 from $i = 1$ to n yields a meaningful lower bound on $\mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right]$ based on $\mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star)\right]$ and $\sigma_\star^2$.

The proof of Lemma 7 is rather technical, so we skip the discussion here. For details, we kindly refer the interested reader to Appendix E.

Lastly, we need to bound the residual term $\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star)$ in Lemma 6. The inequality obtained is given in Lemma 8 below.

Lemma 8 *Under Assumptions 1 and 3, suppose* RR *is employed, then for any $k \in [K]$,* Shuffling SGD *(Algorithm 1) guarantees that*

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star)\right] \leq \frac{\eta_k^2 n^2 \bar{L}\sigma_\star^2}{6},$$

where $\mathbf{y}_k^i$ is defined in (8).

Lemma 8 can be derived in a relatively easy way, as it can be deduced from existing works. In particular, thanks to Liu and Zhou (2024), we obtain a bound that depends only on the average smoothness parameter $\bar{L}$ rather than the maximum smoothness parameter $\hat{L}$.

Final proof. Armed with Lemmas 6, 7, and 8 above, we are finally able to prove Theorem 3.
Proof of Theorem 3 First, we invoke Lemma 6 to have

$$\eta_k \sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) \le \left\|\mathbf{x}_k^1 - \mathbf{x}_\star\right\|^2 - \left\|\mathbf{x}_{k+1}^1 - \mathbf{x}_\star\right\|^2 + 2\eta_k \sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star),$$

where $\mathbf{y}_k^i$ is defined in (8). Take expectations on both sides and apply Lemma 8 to yield

$$\eta_k \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] \le \mathbb{E}\left[\left\|\mathbf{x}_k^1 - \mathbf{x}_\star\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{x}_{k+1}^1 - \mathbf{x}_\star\right\|^2\right] + \frac{\eta_k^3 n^2 \bar{L} \sigma_\star^2}{3}. \tag{10}$$

Next, we multiply both sides of the inequality in Lemma 7 by η_k and sum it up from $i = 1$ to n to obtain

$$\eta_k \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] \le 6\eta_k \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] + 6\eta_k^3 n^2 \bar{L} \sigma_\star^2. \tag{11}$$

Combine (10) and (11) to have

$$\begin{aligned} \eta_k \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] &\le 6\mathbb{E}\left[\left\|\mathbf{x}_k^1 - \mathbf{x}_\star\right\|^2\right] - 6\mathbb{E}\left[\left\|\mathbf{x}_{k+1}^1 - \mathbf{x}_\star\right\|^2\right] + 8\eta_k^3 n^2 \bar{L} \sigma_\star^2 \\ \Rightarrow \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{n} \eta_k \mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] &\le 6\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|^2 + \sum_{k=1}^{K} 8\eta_k^3 n^2 \bar{L} \sigma_\star^2. \end{aligned}$$

Finally, we observe that $\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star) = f(\mathbf{x}_k^i) - f_\star$, divide both sides by $n\sum_{k=1}^{K}\eta_k$, apply the convexity of f, and use the definition of $\bar{\mathbf{x}}_K$ in (1) to conclude. ■

4.4. Proofs of Theorem 1 and Corollary 1

With the previous preparation, Theorem 1 follows immediately.
Proof of Theorem 1 Combine Theorems 2 and 3 to conclude. ■

We next derive Corollary 1 directly from Theorem 1.
Proof of Corollary 1 With a constant stepsize $\eta_k = \eta \le \frac{1}{6\hat{L}}, \forall k \in [K]$, Theorem 1 reduces to

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_K) - f_\star\right] \le \frac{6\left\|\mathbf{x}_1^1 - \mathbf{x}_\star\right\|^2}{\eta n K} + 51\min\left\{\eta, \eta^2 n \bar{L}\right\}\sigma_\star^2.$$

Optimizing the R.H.S. of the above inequality over $0 < \eta \le \frac{1}{6\hat{L}}$ yields the desired result. ■

5. Conclusion and Future Work

In this work, we prove that Shuffling SGD under RR dominates SGD in smooth convex optimization under any reasonable stepsize after any finite number of epochs. Our main Theorem 1 follows from combining two novel convergence results, whose analysis may each be of independent interest.

Our work suggests several new directions for future research. From an upper-bound perspective, our current proof is split into two distinct parts. It is therefore worthwhile to investigate whether Theorem 1 can be obtained via a unified analysis. From a lower-bound perspective, the only existing hardness result for RR in smooth convex optimization by Cha et al. (2023) is established for the stepsize η_k at most inversely proportional to n and for the number of epochs K at least proportional to n. As such, providing a complete characterization of the lower bound for RR under any reasonable stepsize and any number of epochs remains an important task for future work.

Acknowledgments

The author thanks the anonymous reviewers for their valuable feedback.

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17526–17535. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/cb8acb1dc9821bf74e6ca9068032d623-Paper.pdf.

Yoshua Bengio. *Practical Recommendations for Gradient-Based Training of Deep Architectures*, pages 437–478. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_26. URL https://doi.org/10.1007/978-3-642-35289-8_26.

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

Léon Bottou. *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_25. URL https://doi.org/10.1007/978-3-642-35289-8_25.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods as a model of forgetting. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 102613–102647, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fea9f93f4cec99f65a8b4d575fc353a8-Paper-Conference.pdf.

Xufeng Cai, Cheuk Yin Lin, and Jelena Diakonikolas. Tighter convergence bounds for shuffled sgd via primal-dual perspective. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 72475–72524. Curran Associates, Inc., 2024. doi: 10.52202/079017-2310. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/84d395725a9b40cb4a49d84478ac24c7-Paper-Conference.pdf.

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3855–3912. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cha23a.html.

Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

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

Jeff Haochen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2624–2633. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/haochen19a.html.

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

Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 19605–19617. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7bc4f74e35bcfe8cfe43b0a860786d6a-Paper-Conference.pdf.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32471–32508. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24cg.html.

Zijian Liu and Zhengyuan Zhou. Improved last-iterate convergence of shuffling gradient methods for nonsmooth convex optimization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings* *of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 40152–40193. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25ct.html.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtarik. Random reshuffling: Simple analysis with vast improvements. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17309–17320. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8cc6e90ccbff44c9cee23611711cdc4-Paper.pdf.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4703–4711. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/nagaraj19a.html.

Angelia Nedic and Dimitri P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001. doi: 10.1137/S1052623499362111. URL https://doi.org/10.1137/S1052623499362111.

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

Lam M. Nguyen, Quoc Tran-Dinh, Dzung T. Phan, Phuong Ha Nguyen, and Marten van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021. URL http://jmlr.org/papers/v22/20-1238.html.

Boris T. Polyak. *Introduction to optimization*. New York, Optimization Software, 1987.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7964–7973. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rajput20a.html.

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

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3250–3284. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/safran20a.html.

Itay Safran and Ohad Shamir. Random shuffling beats sgd only after many epochs on ill-conditioned problems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15151–15161. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/803ef56843860e4a48fc4cdb3065e8ce-Paper.pdf.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

Uri Sherman, Tomer Koren, and Yishay Mansour. Optimal rates for random order online optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2097–2108. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/107030ca685076c0ed5e054e2c3ed940-Paper.pdf.

Bicheng Ying, Kun Yuan, Stefan Vlaski, and Ali H. Sayed. Stochastic learning under random reshuffling with constant step-sizes. *IEEE Transactions on Signal Processing*, 67(2):474–489, 2019. doi: 10.1109/TSP.2018.2878551.

Appendix A. Additional Related Work

This section provides additional discussion of the related work. We mainly focus on smooth convex optimization under $\textsf{RR}$. As for nonsmooth convex optimization under $\textsf{RR}/\textsf{SS}/\textsf{IG}$, the reader could refer to Kibardin (1979); Nedic and Bertsekas (2001); Koren et al. (2022); Liu and Zhou (2025). See also Shamir (2016) for $\textsf{RR}$ under structured problems.

The first breakthrough on $\textsf{RR}$ is by Gürbüzbalaban et al. (2021) for smooth strongly convex optimization, which, however, requires each f_i to be quadratic or to have a Lipschitz Hessian. Since then, extensive studies have emerged. Among them, a series of works continues to study the convergence behavior of $\textsf{RR}$ in smooth strongly convex optimization for quadratic objectives or under other additional conditions (Ying et al., 2019; Haochen and Sra, 2019; Safran and Shamir, 2020; Rajput et al., 2020; Ahn et al., 2020; Safran and Shamir, 2021; Rajput et al., 2022).

To the best of our knowledge, the first work that drops the strong convexity assumption is Nagaraj et al. (2019), which provides a convergence rate of $\frac{D^2}{\eta nK} + \eta G^2$ under the requirement $\eta \leq \frac{2}{L}$ for L-smooth G-Lipschitz convex f_i, where we remind the reader that D denotes the distance between the initial point and the optimal solution, η represents the stepsize, and K is the number of epochs. However, this rate cannot reflect any advantage of $\textsf{RR}$ over standard $\textsf{SGD}$. Subsequently, two works (Mishchenko et al., 2020; Nguyen et al., 2021) further remove the extra Lipschitz assumption and establish the bound $\frac{D^2}{\eta nK} + \eta^2 nL\sigma_\star^2$ under the condition $\eta \lesssim \frac{1}{nL}$, where $\sigma_\star^2$ is the gradient variance at the optimal solution. This rate is faster than standard $\textsf{SGD}$ under its required regime and remains the best bound so far. In fact, it is unimprovable for small $\eta \lesssim \frac{1}{nL}$ and large $K \gtrsim \frac{nL^2D^2}{\sigma_\star^2}$ due to the lower bound of $(\frac{L\sigma_\star^2 D^4}{nK^2})^{\frac{1}{3}}$ by Cha et al. (2023). Recently, Liu and Zhou (2024) and Cai and Diakonikolas (2025) extend the above rate from the average iterate to the last iterate (up to additional polylogarithmic factors).

Appendix B. Summary of Notation

For readability, we recall and summarize the notation used in the paper.

- $f = \frac{1}{n}\sum_{i=1}^{n} f_i$ is the objective, where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex.
- $\mathbf{x}_\star$ denotes the minimizer of f. $f_\star = f(\mathbf{x}_\star)$ is the optimal function value. $\sigma_\star^2 = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}_\star)\|^2$ is the variance of the gradient at the optimal solution.
- $L_i > 0$ is the smoothness parameter of f_i. $\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$ is the average smoothness parameter. $\hat{L} = \max_{i\in[n]} L_i$ is the maximum smoothness parameter.
- $\mathrm{B}_h(\mathbf{x},\mathbf{y}) = h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle$ denotes the Bregman divergence induced by a real-valued differentiable function h (not necessarily convex). We write B (resp. B_i) as shorthand for B_f (resp. B_{f_i}).

Appendix C. Missing Proofs of Core Lemmas

This section contains the missing proofs of the two core lemmas presented in Subsection 4.1.

Before providing the proofs, we recall two notions introduced in (2) and (3), respectively. Given a permutation π of $[n]$ and two indices $i, j \in [n]$ satisfying $j \leq i$, $\pi(i,j)$ is the permutation generated by exchanging the elements π_i and π_j in π, i.e.,

$$\pi(i,j) = \left(\pi^1, \dots, \pi^{j-1}, \pi^i, \pi^{j+1}, \dots, \pi^{i-1}, \pi^j, \pi^{i+1}, \dots, \pi^n\right). \tag{12}$$

For any given $k \in [K]$, $\mathbf{x}_k^l(i,j), \forall l \in [n+1]$ denotes the trajectory of the k-th epoch starting from $\mathbf{x}_k^1$, produced by Shuffling SGD, but under the permutation $\pi_k(i,j)$, i.e.,

$$\mathbf{x}_k^{l+1}(i,j) = \mathbf{x}_k^l(i,j) - \eta_k \nabla f_{\pi_k^l(i,j)}(\mathbf{x}_k^l(i,j)), \forall l \in [n], \quad \text{where } \mathbf{x}_k^1(i,j) = \mathbf{x}_k^1. \tag{13}$$

C.1. Proof of Lemma 2

Proof Given $k \in [K]$ and $i \in [n]$, we first have the decomposition

$$\ell(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i) = \frac{1}{n}\left(\sum_{j<i} \ell_{\pi_k^j}(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i) + \sum_{j \geq i} \ell_{\pi_k^j}(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right).$$

Note that $\mathbb{E}\left[\ell_{\pi_k^j}(\mathbf{x}_k^i)\right] = \mathbb{E}\left[\ell_{\pi_k^i}(\mathbf{x}_k^i)\right]$ holds for any $j \in \{i, \dots, n\}$ under RR, since $\ell_{\pi_k^j}(\mathbf{x}_k^i)$ and $\ell_{\pi_k^i}(\mathbf{x}_k^i)$ are equal in distribution conditioning on π_k^1 to π_k^{i-1} and π_1 to π_{k-1}. Therefore, we obtain

$$\mathbb{E}\left[\ell(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right] = \frac{1}{n}\sum_{j<i} \mathbb{E}\left[\ell_{\pi_k^j}(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right]. \tag{14}$$

For any fixed $j \in [i]$, it is known that π_k equals $\pi_k(i,j)$ in distribution (e.g., Lemma C.3 of Liu and Zhou (2025)), which implies that $(\pi_k, \mathbf{x}_k)$ also equals $(\pi_k(i,j), \mathbf{x}_k(i,j))$ in distribution, since the trajectory of the k-th epoch generated by Shuffling SGD is deterministically determined by the permutation and stepsize. This implies that

$$\mathbb{E}\left[\ell_{\pi_k^j}(\mathbf{x}_k^i)\right] = \mathbb{E}\left[\ell_{\pi_k^j(i,j)}(\mathbf{x}_k^i(i,j))\right] = \mathbb{E}\left[\ell_{\pi_k^i}(\mathbf{x}_k^i(i,j))\right],$$

where the last step is due to $\pi_k^j(i,j) = \pi_k^i$ from its definition (2). Hence, we finally obtain

$$\mathbb{E}\left[\ell(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right] = \frac{1}{n}\sum_{j<i} \mathbb{E}\left[\ell_{\pi_k^i}(\mathbf{x}_k^i(i,j)) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right].$$

■

C.2. Proof of Lemma 3

Proof By the definition of $\pi(i,j)$ in (2), we know

$$\pi_k^l(i,j) = \pi_k^l, \forall l \notin \{i,j\} \Rightarrow \nabla f_{\pi_k^l(i,j)} = \nabla f_{\pi_k^l}, \forall l \notin \{i,j\}. \tag{15}$$

Therefore, given $l \notin \{i, j\}$, by the definition of $\mathbf{x}_k(i,j)$ (see (3)) and the update rule of Shuffling SGD, we have

$$\begin{aligned}
\left\|\mathbf{x}_k^{l+1}(i,j) - \mathbf{x}_k^{l+1}\right\|^2 &= \left\|\mathbf{x}_k^l(i,j) - \mathbf{x}_k^l - \eta_k\left(\nabla f_{\pi_k^l(i,j)}(\mathbf{x}_k^l(i,j)) - \nabla f_{\pi_k^l}(\mathbf{x}_k^l)\right)\right\|^2 \\
&\overset{(15)}{=} \left\|\mathbf{x}_k^l(i,j) - \mathbf{x}_k^l - \eta_k\left(\nabla f_{\pi_k^l}(\mathbf{x}_k^l(i,j)) - \nabla f_{\pi_k^l}(\mathbf{x}_k^l)\right)\right\|^2 \\
&= \left\|\mathbf{x}_k^l(i,j) - \mathbf{x}_k^l\right\|^2 + \eta_k^2\left\|\nabla f_{\pi_k^l}(\mathbf{x}_k^l(i,j)) - \nabla f_{\pi_k^l}(\mathbf{x}_k^l)\right\|^2 \\
&\quad - 2\eta_k\left\langle \nabla f_{\pi_k^l}(\mathbf{x}_k^l(i,j)) - \nabla f_{\pi_k^l}(\mathbf{x}_k^l), \mathbf{x}_k^l(i,j) - \mathbf{x}_k^l\right\rangle \\
&\overset{(a)}{\leq} \left\|\mathbf{x}_k^l(i,j) - \mathbf{x}_k^l\right\|^2 + \left(\eta_k^2 - \frac{2\eta_k}{L_{\pi_k^l}}\right)\left\|\nabla f_{\pi_k^l}(\mathbf{x}_k^l(i,j)) - \nabla f_{\pi_k^l}(\mathbf{x}_k^l)\right\|^2 \\
&\overset{(b)}{\leq} \left\|\mathbf{x}_k^l(i,j) - \mathbf{x}_k^l\right\|^2, \qquad (16)
\end{aligned}$$

where (a) is due to Lemma 1 and (b) holds by $\eta_k \leq \frac{2}{\hat{L}} \Rightarrow \eta_k^2 - \frac{2\eta_k}{L_{\pi_k^l}} \leq 0$.

Apply (16) from $l = j+1$ to $l = i-1$ to obtain

$$\begin{aligned}
&\left\|\mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\|^2 \leq \left\|\mathbf{x}_k^{j+1}(i,j) - \mathbf{x}_k^{j+1}\right\|^2 \\
=& \left\|\mathbf{x}_k^j(i,j) - \mathbf{x}_k^j - \eta_k\left(\nabla f_{\pi_k^j(i,j)}(\mathbf{x}_k^j(i,j)) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right)\right\|^2 \\
\overset{(c)}{=}& \eta_k^2\left\|\nabla f_{\pi_k^j(i,j)}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2 \overset{(2)}{=} \eta_k^2\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2,
\end{aligned}$$

where (c) holds by $\mathbf{x}_k^j(i,j) = \mathbf{x}_k^j$, since $\mathbf{x}_k^l(i,j) = \mathbf{x}_k^l, \forall l \in [j]$ implied by (16) and $\mathbf{x}_k^1(i,j) \overset{(3)}{=} \mathbf{x}_k^1$ together. ■

Appendix D. Missing Proofs of Lemmas for Bound I

In this section, we provide the missing proofs of the lemmas presented in Subsection 4.2, which were used to prove the first convergence rate in Theorem 2.

D.1. Proof of Lemma 4

Proof Given $k \in [K]$ and $i \in [n]$, by the definition of $\mathrm{B}_{\pi_k^i}$,

$$\begin{aligned}
f_{\pi_k^i}(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_\star) &= \left\langle \nabla f_{\pi_k^i}(\mathbf{x}_k^i), \mathbf{x}_k^i - \mathbf{x}_\star\right\rangle - \mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i) \\
&= \frac{\left\|\mathbf{x}_k^i - \mathbf{x}_\star\right\|^2 - \left\|\mathbf{x}_k^{i+1} - \mathbf{x}_\star\right\|^2}{2\eta_k} + \frac{\eta_k}{2}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 - \mathrm{B}_{\pi_k^i}(\mathbf{x}_\star, \mathbf{x}_k^i),
\end{aligned}$$

where the second step holds by the update rule of Algorithm 1. ■

D.2. Proof of Lemma 5

Proof We apply Lemma 2 with $\ell = f$ and $\ell_i = f_i$ to have

$$\mathbb{E}\left[f(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_k^i)\right] = \frac{1}{n}\sum_{j<i}\mathbb{E}\left[f_{\pi_k^i}(\mathbf{x}_k^i(i,j)) - f_{\pi_k^i}(\mathbf{x}_k^i)\right], \tag{17}$$

where $\mathbf{x}_k(i,j)$ is defined in (3). Next, by the $L_{\pi_k^i}$-smoothness of $f_{\pi_k^i}$ (Assumption 3), we know

$$\begin{aligned}
&f_{\pi_k^i}(\mathbf{x}_k^i(i,j)) - f_{\pi_k^i}(\mathbf{x}_k^i)\\
&\le \left\langle \nabla f_{\pi_k^i}(\mathbf{x}_k^i), \mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\rangle + \frac{L_{\pi_k^i}}{2}\left\|\mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\|^2\\
&\overset{(a)}{\le} \eta_k \left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\| \left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\| + \frac{\eta_k^2 L_{\pi_k^i}}{2}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\\
&\overset{(b)}{\le} \eta_k \left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 + \frac{\eta_k + 2\eta_k^2 L_{\pi_k^i}}{4}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\\
&\le \eta_k \left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 + \frac{\eta_k + 2\eta_k^2 L_{\pi_k^i}}{2}\left(\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j)\right\|^2 + \left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\right)\\
&\overset{(c)}{\le} \eta_k \left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 + \frac{2}{3}\eta_k\left(\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j)\right\|^2 + \left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\right),
\end{aligned} \tag{18}$$

where (a) is by Cauchy-Schwarz inequality and Lemma 3, (b) is due to AM-GM inequality, and (c) holds by $\eta_k \le \frac{1}{6\hat{L}}$. Finally, we plug (18) back into (17) to obtain

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_k^i)\right] &\le \frac{\eta_k}{n}\sum_{j<i}\mathbb{E}\left[\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 + \frac{2}{3}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j)\right\|^2 + \frac{2}{3}\left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\right]\\
&\overset{(d)}{=} \frac{\eta_k}{n}\sum_{j<i}\mathbb{E}\left[\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2 + \frac{4}{3}\left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\right]\\
&\le \eta_k\mathbb{E}\left[\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i)\right\|^2\right] + \frac{4\eta_k}{3n}\sum_{j<i}\mathbb{E}\left[\left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\right],
\end{aligned}$$

where (d) is due to $\mathbb{E}\left[\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j)\right\|^2\right] = \mathbb{E}\left[\left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2\right]$ when $j < i$. ■

Appendix E. Missing Proofs of Lemmas for Bound II

In this section, we provide the missing proofs of the lemmas presented in Subsection 4.3, which were used to prove the second convergence rate in Theorem 3.

Before presenting the proofs, we recall the notion introduced in (8). For any $k \in [K]$, the virtual sequence $\mathbf{y}_k^i, \forall i \in [n+1]$ follows the equations,

$$\mathbf{y}_k^{i+1} = \mathbf{y}_k^i - \eta_k \nabla f_{\pi_k^i}(\mathbf{x}_\star), \forall i \in [n], \quad \mathbf{y}_{k+1}^1 = \mathbf{y}_k^{n+1}, \quad \text{where } \mathbf{y}_1^1 = \mathbf{x}_\star. \tag{19}$$

As shown in (9), the virtual sequence satisfies that

$$\mathbf{y}_k^{n+1} = \mathbf{y}_k^1 = \mathbf{x}_\star, \forall k \in [K]. \tag{20}$$

E.1. Proof of Lemma 6

Proof Given $k \in [K]$ and $i \in [n]$, by the update rule of Algorithm 1 and the definition of $\mathbf{y}_k^{i+1}$ in (8), we have

$$\begin{aligned}
&\left\|\mathbf{x}_k^{i+1} - \mathbf{y}_k^{i+1}\right\|^2 = \left\|\mathbf{x}_k^i - \mathbf{y}_k^i - \eta_k\left(\nabla f_{\pi_k^i}(\mathbf{x}_k^i) - \nabla f_{\pi_k^i}(\mathbf{x}_\star)\right)\right\|^2 \\
&= \left\|\mathbf{x}_k^i - \mathbf{y}_k^i\right\|^2 + \eta_k^2\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i) - \nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 2\eta_k\left\langle \nabla f_{\pi_k^i}(\mathbf{x}_\star) - \nabla f_{\pi_k^i}(\mathbf{x}_k^i), \mathbf{x}_k^i - \mathbf{y}_k^i\right\rangle \\
&= \left\|\mathbf{x}_k^i - \mathbf{y}_k^i\right\|^2 + \eta_k^2\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i) - \nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 2\eta_k\left(\mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star) - \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_k^i) - \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right) \\
&\overset{(a)}{\leq} \left\|\mathbf{x}_k^i - \mathbf{y}_k^i\right\|^2 + \eta_k^2\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i) - \nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 2\eta_k\left(\mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star) - \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right) \\
&\overset{(b)}{\leq} \left\|\mathbf{x}_k^i - \mathbf{y}_k^i\right\|^2 + 2\left(\eta_k^2 L_{\pi_k^i} - \eta_k\right)\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) + 2\eta_k \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star) \\
&\overset{(c)}{\leq} \left\|\mathbf{x}_k^i - \mathbf{y}_k^i\right\|^2 - \eta_k \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) + 2\eta_k \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star),
\end{aligned}$$

where we use $\mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_k^i) \geq 0$ in (a), apply Lemma 1 to $f_{\pi_k^i}$ in (b), and notice that $\eta_k \leq \frac{1}{2\hat{L}} \Rightarrow \eta_k^2 L_{\pi_k^i} \leq \frac{\eta_k}{2}$ in (c). Finally, we sum the above inequality from $i = 1$ to n, use $\mathbf{x}_{k+1}^1 = \mathbf{x}_k^{n+1}$ and $\mathbf{y}_k^{n+1} = \mathbf{y}_k^1 = \mathbf{x}_\star$ (see (9)), and rearrange terms to complete the proof. ■

E.2. Proof of Lemma 7

Proof Expanding the definitions of B and $\mathrm{B}_{\pi_k^i}$, we know

$$\begin{aligned}
\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star) - \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) =& f(\mathbf{x}_k^i) - f(\mathbf{x}_\star) - \left\langle \nabla f(\mathbf{x}_\star), \mathbf{x}_k^i - \mathbf{x}_\star\right\rangle \\
&- \left(f_{\pi_k^i}(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_\star) - \left\langle \nabla f_{\pi_k^i}(\mathbf{x}_\star), \mathbf{x}_k^i - \mathbf{x}_\star\right\rangle\right) \\
=& f(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_k^i) - f(\mathbf{x}_\star) + f_{\pi_k^i}(\mathbf{x}_\star) \\
&- \left\langle \nabla f(\mathbf{x}_\star) - \nabla f_{\pi_k^i}(\mathbf{x}_\star), \mathbf{x}_k^i - \mathbf{x}_\star\right\rangle.
\end{aligned}$$

Since $\mathbb{E}\left[f_{\pi_k^i}(\mathbf{x}_\star)\right] = f(\mathbf{x}_\star)$ and $\mathbb{E}\left[\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right] = \nabla f(\mathbf{x}_\star)$ under RR, after taking expectations on both sides, we obtain

$$\mathbb{E}\left[\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star) - \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] = \mathbb{E}\left[f(\mathbf{x}_k^i) - f_{\pi_k^i}(\mathbf{x}_k^i) - \left\langle \nabla f(\mathbf{x}_\star) - \nabla f_{\pi_k^i}(\mathbf{x}_\star), \mathbf{x}_k^i\right\rangle\right]. \tag{21}$$

Now, we denote by $\ell_i(\mathbf{x}) \triangleq f_i(\mathbf{x}) - \langle \nabla f_i(\mathbf{x}_\star), \mathbf{x}\rangle, \forall i \in [n]$ and $\ell(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbf{x}) = f(\mathbf{x}) - \langle \nabla f(\mathbf{x}_\star), \mathbf{x}\rangle$. Then, (21) implies that,

$$\mathbb{E}\left[\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star) - \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] = \mathbb{E}\left[\ell(\mathbf{x}_k^i) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right] \overset{(a)}{=} \frac{1}{n}\sum_{j<i}\mathbb{E}\left[\ell_{\pi_k^i}(\mathbf{x}_k^i(i,j)) - \ell_{\pi_k^i}(\mathbf{x}_k^i)\right], \tag{22}$$

where (a) holds by Lemma 2.

Note that $\ell_{\pi_k^i}$ is $L_{\pi_k^i}$-smooth by its definition and Assumption 3, we therefore have, for any $j \in [i-1]$,

$$\begin{aligned}
\ell_{\pi_k^i}(\mathbf{x}_k^i(i,j)) - \ell_{\pi_k^i}(\mathbf{x}_k^i) &\le \left\langle \nabla \ell_{\pi_k^i}(\mathbf{x}_k^i), \mathbf{x}_k^i(i,j) - \mathbf{x}_k^i \right\rangle + \frac{L_{\pi_k^i}}{2}\left\|\mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\|^2 \\
&= \left\langle \nabla f_{\pi_k^i}(\mathbf{x}_k^i) - \nabla f_{\pi_k^i}(\mathbf{x}_\star), \mathbf{x}_k^i(i,j) - \mathbf{x}_k^i \right\rangle + \frac{L_{\pi_k^i}}{2}\left\|\mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\|^2 \\
&\overset{(b)}{\le} \frac{\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^i) - \nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2}{2L_{\pi_k^i}} + L_{\pi_k^i}\left\|\mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\|^2 \\
&\overset{(c)}{\le} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) + L_{\pi_k^i}\left\|\mathbf{x}_k^i(i,j) - \mathbf{x}_k^i\right\|^2 \\
&\overset{(d)}{\le} \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) + \eta_k^2 L_{\pi_k^i}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2, \qquad (23)
\end{aligned}$$

where (b) is by Cauchy-Schwarz inequality and AM-GM inequality, (c) is due to Lemma 1, and (d) holds by Lemma 3. Furthermore, we can bound

$$\begin{aligned}
&\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_k^j)\right\|^2 \\
\le& 4\left\|\nabla f_{\pi_k^i}(\mathbf{x}_k^j) - \nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 4\left\|\nabla f_{\pi_k^j}(\mathbf{x}_k^j) - \nabla f_{\pi_k^j}(\mathbf{x}_\star)\right\|^2 + 4\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 4\left\|\nabla f_{\pi_k^j}(\mathbf{x}_\star)\right\|^2 \\
\le& 8L_{\pi_k^i}\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^j, \mathbf{x}_\star) + 8L_{\pi_k^j}\mathrm{B}_{\pi_k^j}(\mathbf{x}_k^j, \mathbf{x}_\star) + 4\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 4\left\|\nabla f_{\pi_k^j}(\mathbf{x}_\star)\right\|^2,
\end{aligned}$$

where the last step is by Lemma 1 again. Plug the above inequality back into (23) and use $\eta_k \le \frac{1}{2\hat{L}}$ to have

$$\begin{aligned}
\ell_{\pi_k^i}(\mathbf{x}_k^i(i,j)) - \ell_{\pi_k^i}(\mathbf{x}_k^i) \le& \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) + 2\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^j, \mathbf{x}_\star) + 2\mathrm{B}_{\pi_k^j}(\mathbf{x}_k^j, \mathbf{x}_\star) \\
&+ 4\eta_k^2 L_{\pi_k^i}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + 4\eta_k^2 L_{\pi_k^i}\left\|\nabla f_{\pi_k^j}(\mathbf{x}_\star)\right\|^2. \qquad (24)
\end{aligned}$$

Combine (22) and (24) to obtain

$$\begin{aligned}
\mathbb{E}\left[\mathrm{B}(\mathbf{x}_k^i, \mathbf{x}_\star) - \mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star)\right] \le& \frac{1}{n}\sum_{j<i}\mathbb{E}\left[\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^i, \mathbf{x}_\star) + 2\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^j, \mathbf{x}_\star) + 2\mathrm{B}_{\pi_k^j}(\mathbf{x}_k^j, \mathbf{x}_\star)\right] \\
&+ \frac{4\eta_k^2}{n}\sum_{j<i}\mathbb{E}\left[L_{\pi_k^i}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2 + L_{\pi_k^i}\left\|\nabla f_{\pi_k^j}(\mathbf{x}_\star)\right\|^2\right].
\end{aligned}$$

When $j < i$, we observe that $\mathbb{E}\left[\mathrm{B}_{\pi_k^i}(\mathbf{x}_k^j, \mathbf{x}_\star)\right] = \mathbb{E}\left[\mathrm{B}_{\pi_k^j}(\mathbf{x}_k^j, \mathbf{x}_\star)\right]$ and the following two inequalities hold

$$\begin{aligned}
\mathbb{E}\left[L_{\pi_k^i}\left\|\nabla f_{\pi_k^i}(\mathbf{x}_\star)\right\|^2\right] &= \frac{\sum_{l=1}^n L_l\left\|\nabla f_l(\mathbf{x}_\star)\right\|^2}{n} \le \hat{L}\sigma_\star^2 \le n\bar{L}\sigma_\star^2, \\
\mathbb{E}\left[L_{\pi_k^i}\left\|\nabla f_{\pi_k^j}(\mathbf{x}_\star)\right\|^2\right] &= \frac{\sum_{l=1}^n \frac{n\bar{L}-L_l}{n-1}\left\|\nabla f_l(\mathbf{x}_\star)\right\|^2}{n} \le 2\bar{L}\sigma_\star^2.
\end{aligned}$$

Put everything together, rearrange terms, and use $1 \le n$ to conclude the desired inequality. ∎

E.3. Proof of Lemma 8

Proof By smoothness (i.e., Assumption 3), we have

$$\mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star) \le \frac{L_{\pi_k^i}}{2} \left\| \mathbf{y}_k^i - \mathbf{x}_\star \right\|^2 \stackrel{(8)}{=} \frac{L_{\pi_k^i}}{2} \left\| \mathbf{y}_k^i - \mathbf{y}_k^1 \right\|^2 \stackrel{(8)}{=} \frac{\eta_k^2 L_{\pi_k^i}}{2} \left\| \sum_{j=1}^{i-1} \nabla f_{\pi_k^j}(\mathbf{x}_\star) \right\|^2 .$$

Therefore, we can bound

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathrm{B}_{\pi_k^i}(\mathbf{y}_k^i, \mathbf{x}_\star) \right] \le \frac{\eta_k^2}{2} \mathbb{E}\left[\sum_{i=1}^{n} L_{\pi_k^i} \left\| \sum_{j=1}^{i-1} \nabla f_{\pi_k^j}(\mathbf{x}_\star) \right\|^2 \right] \le \frac{\eta_k^2 n^2 \bar{L} \sigma_\star^2}{6},$$

where the last step is due to Lemma E.1 of Liu and Zhou (2024) (the constant here is slightly better, since $\nabla f(\mathbf{x}_\star) = \mathbf{0}$ in our setting leads to a provable improvement). ■