Tighter Lower Bounds for Shuffling SGD: Random Permutations and Beyond

Jaeyoung Cha [1] Jaewook Lee [1] Chulhee Yun [1]

Abstract

We study convergence lower bounds of without-replacement stochastic gradient descent (SGD) for solving smooth (strongly-)convex finite-sum minimization problems. Unlike most existing results focusing on final iterate lower bounds in terms of the number of components n and the number of epochs K, we seek bounds for arbitrary weighted average iterates that are tight in all factors including the condition number κ. For SGD with *Random Reshuffling*, we present lower bounds that have tighter κ dependencies than existing bounds. Our results are the first to perfectly close the gap between lower and upper bounds for weighted average iterates in both strongly-convex and convex cases. We also prove weighted average iterate lower bounds for *arbitrary* permutation-based SGD, which apply to all variants that carefully choose the best permutation. Our bounds improve the existing bounds in factors of n and κ and thereby match the upper bounds shown for a recently proposed algorithm called GraB.

1. Introduction

One of the most common frameworks used in machine learning is the following finite-sum minimization problem,

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}). \tag{1}$$

Stochastic gradient descent (SGD), an algorithm first proposed by Robbins & Monro (1951), is highly capable of numerically solving finite-sum optimization problems. In the t-th iteration, SGD randomly samples a component index $i(t)$ and computes a gradient-based update equation of the form $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta_t \nabla f_{i(t)}(\boldsymbol{x}_{t-1})$, where η_t is a step size parameter, often set to a fixed constant.

Many prior studies on SGD have shown convergence results assuming *with-replacement* sampling of the component index $i(t)$ (Benaïm (1999); Bottou et al. (2018); Bubeck (2015) and many others), where we independently choose $i(t)$ from a uniform random distribution over the index set every time. This uniform sampling makes each step of SGD an unbiased noisy estimate of vanilla gradient descent (GD).

In real-world applications, however, it is much more common to use *without-replacement* SGD, where each epoch runs over the entire shuffled set of n components. Without-replacement SGD has gained popularity for both its simplicity and empirical observations of faster convergence rates (Bottou, 2009; Recht & Ré, 2013; Yun et al., 2021). However, theoretical analysis on without-replacement SGD remains quite elusive, especially because of the lack of independence between iterates. Nevertheless, recent works have managed to successfully deal with without-replacement SGD in theoretical aspects (Haochen & Sra, 2019; Nagaraj et al., 2019; Recht & Re, 2012).

A simple and popular method of without-replacement sampling is to randomly shuffle the n components independently on each epoch, often referred to as *Random Reshuffling* or SGD-RR. Some studies show upper bounds of convergence rates for a certain class of functions (Gürbüzbalaban et al., 2019; Ahn et al., 2020), while some others present lower bounds by analyzing a function contained in a certain class with a slow convergence rate (Safran & Shamir, 2020; Rajput et al., 2020). These preliminary results highlight that without-replacement SGD is in fact capable of converging provably faster than its with-replacement counterpart.

A recent line of work (Rajput et al., 2022; Lu et al., 2022b; Mohtashami et al., 2022) opens a new field of studies on *permutation-based SGD*, which covers all cases where the permutation of the n component functions is chosen according to a certain policy, instead of simple random reshuffling. The aim of this line of research is to design a policy that yields *faster* convergence compared to random permutations. Indeed, a recent result by Lu et al. (2022a) proposes GraB, a permutation-based SGD algorithm that uses the gradient information from previous epochs to manipulate the permutation of the current epoch, and shows that GraB provably

[1] Kim Jaechul Graduate School of AI, KAIST, Seoul, South Korea. Correspondence to: Chulhee Yun <chulhee.yun@kaist.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

converges faster than *Random Reshuffling*. This raises the following question:

$$\textit{Is GraB optimal, or can we find an even faster permutation-based SGD algorithm?} \tag{2}$$

1.1. Related Work

Before summarizing our contributions, we list up related prior results so as to better contextualize our results relative to them. In all convergence rates, we write $\mathcal{O}(\cdot)$ for upper bounds and $\Omega(\cdot)$ for lower bounds. The tilde notation $\tilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors. For simplicity, here we write convergence rates only with respect to the number of component functions n and the number of epochs K (i.e., number of passes through the entire components).

SGD with replacement is known to have a tight convergence rate of $\mathcal{O}\left(\frac{1}{T}\right)$ after T iterations, which translates to $\mathcal{O}\left(\frac{1}{nK}\right)$ in our notation. One of the first studies on SGD-RR by Gürbüzbalaban et al. (2019) shows an *upper bound* of $\tilde{\mathcal{O}}\left(\frac{1}{K^2}\right)$ for strongly convex objectives with smooth components, along with the assumption that n is a constant. Haochen & Sra (2019) show a convergence rate of $\tilde{\mathcal{O}}\left(\frac{1}{n^2K^2} + \frac{1}{K^3}\right)$ for functions with Lipschitz-continuous Hessians, which explicitly depends on both n and K. Rajput et al. (2020) further show that the upper bound for strongly convex quadratics is $\tilde{\mathcal{O}}\left(\frac{1}{n^2K^2} + \frac{1}{nK^3}\right)$. Follow-up studies prove upper bounds in broader settings, such as $\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$ for strongly convex (but not necessarily quadratic) functions (Nagaraj et al., 2019; Ahn et al., 2020; Mishchenko et al., 2020), or $\mathcal{O}\left(\frac{1}{n^{1/3}K^{2/3}}\right)$ under convex assumptions (Mishchenko et al., 2020). Some further generalize to other variants of SGD-RR, including Minibatch and Local SGD in federated learning (Yun et al., 2022), Nesterov's acceleration (Tran et al., 2022), or Stochastic Gradient Descent-Ascent used in minimax problems (Cho & Yun, 2023). Meanwhile, investigations on *lower bounds* have started from simple quadratic assumptions, where Safran & Shamir (2020) prove a lower bound of rate $\Omega\left(\frac{1}{n^2K^2} + \frac{1}{nK^3}\right)$. Lower bounds were then extended to smooth and strongly convex settings, as in Rajput et al. (2020) and Yun et al. (2022) which both derive a lower bound of $\Omega\left(\frac{1}{nK^2}\right)$.

Recent works provide evidence of designing algorithms that converge faster than SGD-RR. Concretely, Rajput et al. (2022) introduce a permutation-based SGD algorithm called FlipFlop and prove that it can outperform SGD-RR for quadratic objectives. The authors also propose a lower bound applicable to arbitrary permutation-based SGD, by proving that no algorithm can converge faster than $\Omega\left(\frac{1}{n^3K^2}\right)$ for some strongly convex objectives. Lu et al. (2022b) and Mohtashami et al. (2022) propose methods to find "good" permutations via a greedy strategy. Extending their previous work, Lu et al. (2022a) propose GraB and gain a convergence rate $\tilde{\mathcal{O}}\left(\frac{1}{n^2K^2}\right)$ for PŁ functions which is faster than $\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$ for SGD-RR (Ahn et al., 2020).

Most prior results (Rajput et al., 2020; 2022) mainly concern achieving tight convergence rates with respect to n and K, while recent studies delve deeper to unveil how other parameters can also affect the convergence properties. The *condition number* κ (defined in Section 2) is an example of such parameters, which is closely related to the problem's geometry. If we take κ into account[1] in the strongly convex case, the best known upper and lower bounds for SGD-RR are $\tilde{\mathcal{O}}\left(\frac{\kappa^3}{nK^2}\right)$ (Nagaraj et al., 2019; Ahn et al., 2020; Mishchenko et al., 2020) and $\Omega\left(\frac{\kappa}{nK^2}\right)$ (Rajput et al., 2020; Yun et al., 2022), which differ by a factor of κ^2, and those for permutation-based SGD are $\tilde{\mathcal{O}}\left(\frac{\kappa^3}{n^2K^2}\right)$ (Lu et al., 2022a) and $\Omega\left(\frac{1}{n^3K^2}\right)$ (Rajput et al., 2022), which differ by both n and some factors of κ—that is, the bounds are *not* completely tight for all factors yet.

While it is tempting to neglect the looseness in κ by treating factors in κ as "leading constants," characterizing the right dependence on κ becomes imperative for understanding the regimes in which without-replacement SGD is faster than the with-replacement version. For example, the aforementioned rate $\tilde{\mathcal{O}}\left(\frac{\kappa^3}{nK^2}\right)$ of SGD-RR improves upon the known tight rate $\mathcal{O}\left(\frac{\kappa}{nK}\right)$ of with-replacement SGD only if $K \gtrsim \kappa^2$. It turns out that this requirement of *large enough* K is in fact unavoidable in the strongly convex case (Safran & Shamir, 2021); by developing a lower bound, Safran & Shamir (2021) show that SGD-RR cannot converge faster than with-replacement SGD when $K \lesssim \kappa$. Characterizing the exact threshold (κ vs. κ^2) for faster convergence of SGD-RR requires a tighter analysis of the κ dependence of its convergence rate.

1.2. Summary of Our Contributions

Towards a complete understanding of SGD-RR and permutation-based SGD in general, we seek to close the existing gaps outlined above by developing tighter lower bounds with matching upper bounds. We present results under two different kinds of algorithm settings: Section 3 contains lower bounds obtained for SGD-RR, and Section 4 presents lower bounds that are applicable to *arbitrary permutation-based SGD algorithms*.

Our lower bounds are obtained for without-replacement SGD with constant step size, which is also the case in other existing results in the literature (Safran & Shamir, 2020; 2021; Rajput et al., 2020; 2022; Yun et al., 2022). While all

[1] For this section, we treat $\kappa = \Theta(1/\mu)$ for simplicity, following the convention of other existing results in the literature (Haochen & Sra, 2019; Nagaraj et al., 2019; Safran & Shamir, 2021).

Table 1. A comparison of existing convergence rates and our results for permutation-based SGD. Parameters L, μ, ν, and D are defined in Section 2. Algorithm outputs $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{x}}_{\text{tail}}$, and $\hat{\boldsymbol{x}}_{\text{avg}}$ are defined in Section 3. Function classes $\mathcal{F}$ and $\mathcal{F}_{\text{PŁ}}$ are defined in Sections 2 and 4, respectively. The herding bound H, which closely relates to the convergence rate of Algorithm 1, is defined in Section 4. The upper bound results are colored white and the lower bound results are colored gray. For a more detailed comparison with prior work, please refer to Table 2 in Appendix A.

Random Reshuffling				
Function Class	Output	References	Convergence Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	Mishchenko et al. (2020)	$\tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 nK^2}\right)$	$K \gtrsim \kappa$
		Ours, Theorem 3.1	$\Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$\kappa \geq c$, $K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}_{\text{tail}}$	Ours, Proposition 3.4	$\tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Ours, Theorem 3.3†	$\Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$\kappa \geq c$, $K \gtrsim \kappa$
$\mathcal{F}(L, 0, 0, \nu)$	$\hat{\boldsymbol{x}}_{\text{avg}}$	Mishchenko et al. (2020)	$\mathcal{O}\left(\frac{L^{1/3}\nu^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$	$K \gtrsim \frac{L^2D^2n}{\nu^2}$
	$\hat{\boldsymbol{x}}$	Ours, Corollary 3.5†	$\Omega\left(\frac{L^{1/3}\nu^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$	$K \gtrsim \max\{\frac{L^2D^2n}{\nu^2}, \frac{\nu}{\mu D n^{1/2}}\}$

Arbitrary Permutations				
Function Class	Output	References	Convergence Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	Lu et al. (2022a) (GraB)	$\tilde{\mathcal{O}}\left(\frac{H^2L^2\nu^2}{\mu^3 n^2K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Ours, Theorem 4.1	$\Omega\left(\frac{L\nu^2}{\mu^2 n^2K^2}\right)$	-
$\mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$	$\boldsymbol{x}_n^K$	Ours, Proposition 4.6 (GraB)	$\tilde{\mathcal{O}}\left(\frac{H^2L^2\nu^2}{\mu^3 n^2K^2}\right)$	$n \geq H$, $K \gtrsim \kappa(\tau+1)$
	$\hat{\boldsymbol{x}}$	Ours, Theorem 4.5	$\Omega\left(\frac{L^2\nu^2}{\mu^3 n^2K^2}\right)$	$\tau = \kappa \geq 8n$, $K \geq \max\{\frac{\kappa^2}{n}, \kappa^{\frac{3}{2}}n^{\frac{1}{2}}\}$

† Additionally assumes $\eta \leq \frac{1}{c_2 Ln}$

lower bounds proved in the aforementioned papers are only applicable to the *final iterate* of the algorithm, many of our results in this paper apply to arbitrary *weighted average* of end-of-epoch iterates, which can be used to show tightness of matching upper bounds that employ iterate averaging.

Our main contributions are as follows. Here we include $\kappa = \Theta\left(1/\mu\right)$ in the convergence rates to better describe the results. Please refer to Table 1 for a complete summary.

- Theorem 3.1 derives a lower bound of rate $\Omega\left(\frac{\kappa^2}{nK^2}\right)$ for the final iterate of SGD-RR in the strongly convex case, which matches the best-known corresponding upper bound $\tilde{\mathcal{O}}\left(\frac{\kappa^3}{nK^2}\right)$ up to a factor of κ.
- Theorem 3.3 extends the lower bound $\Omega\left(\frac{\kappa^2}{nK^2}\right)$ under strongly convex settings to arbitrary weighted average iterates of SGD-RR. Proposition 3.4 shows a matching upper bound $\tilde{\mathcal{O}}\left(\frac{\kappa^2}{nK^2}\right)$ for the tail average iterate, achieving tightness up to logarithmic factors.
- Corollary 3.5 shows a lower bound $\Omega\left(\frac{1}{n^{1/3}K^{2/3}}\right)$ for the average iterate of SGD-RR in the convex case, which matches the corresponding upper bound in Mishchenko et al. (2020).
- Theorem 4.1 provides a lower bound $\Omega\left(\frac{\kappa^2}{n^2K^2}\right)$ on arbitrary permutation-based SGD, which, to the best of our knowledge, is the first to match the best-known upper bound of GraB (Lu et al., 2022a) in terms of n and K.
- Theorem 4.5 relaxes the assumption of individual convexity and obtains a stronger lower bound $\Omega\left(\frac{\kappa^3}{n^2K^2}\right)$ in the scenario of arbitrary permutation-based SGD. This lower bound exactly matches the upper bound in all factors, including κ. Our results therefore answer the question in (2): *Yes,* GraB *is an optimal permutation-based SGD algorithm.*

2. Preliminaries

First we summarize some basic notations used throughout the paper. For a positive integer N, we use the notation $[N] := \{1, 2, \ldots, N\}$. For $\boldsymbol{v} \in \mathbb{R}^d$, we denote its L_2 and L_∞ norm as $\|\boldsymbol{v}\|$ and $\|\boldsymbol{v}\|_\infty$, respectively. We denote the number of component functions as n and the number of epochs as K, where n and K are both positive integers.

Some of our results require large K and we will use $K \gtrsim x$ to express such an assumption. We use $K \gtrsim x$ to denote the condition $K \geq Cx \log\left(\text{poly}(n, K, \mu, L, ...)\right)$ when C is a numerical constant.

2.1. Function Class

The following definitions help us to formally define the class of problems to which our objective function belongs.

Definition 2.1 (Smoothness)**.** A differentiable function f is *L-smooth* if

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Definition 2.2 (Strong convexity)**.** A differentiable function f is *μ-strongly convex* if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. If the inequality holds for $\mu = 0$, then we say that f is *convex*.

Definition 2.3 (PŁ condition)**.** A differentiable function f satisfies the *μ-Polyak-Łojasiewicz (PŁ) condition* if

$$\frac{1}{2} \|\nabla f(\boldsymbol{x})\|^2 \geq \mu(f(\boldsymbol{x}) - f^*), \; \forall \boldsymbol{x} \in \mathbb{R}^d,$$

where $f^* := \inf_{\boldsymbol{x}} f(\boldsymbol{x}) > -\infty$ is the global minimum value of f.

Additionally, we define the *condition number* as $\kappa := L/\mu$, where L is the smoothness constant and μ is either the strong convexity constant or the PŁ constant.

We also make a common assumption regarding the finite-sum setup in (1), which is that the gradients of the objective function and its components are not too far from each other.

Assumption 2.4 (Bounded gradient errors)**.** There exists $\tau \geq 0$ and $\nu \geq 0$ such that for all $i = 1, \dots, n$,

$$\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\| \leq \tau \|\nabla F(\boldsymbol{x})\| + \nu, \; \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Now we define the function class $\mathcal{F}$ as follows.

Definition 2.5 (Function Class)**.** We define the function class $\mathcal{F}(L, \mu, \tau, \nu)$ of objective functions F as:

$$\begin{aligned} \mathcal{F}(L, \mu, \tau, \nu) := \{F : \; & f_i \text{ are } L\textit{-smooth} \text{ and } \textit{convex}, \\ & F \text{ is } \mu\textit{-strongly convex}, \\ & F \text{ and } f_i \text{ satisfy Assumption 2.4}\}. \end{aligned}$$

Note that Definition 2.5 takes into account not only the properties of F but that of the components f_i as well. Also, as seen in Definition 2.2, $\mathcal{F}(L, 0, \tau, \nu)$ corresponds to the case where F is convex.

Remark. One may concern that Assumption 2.4 is too "strong" compared to common assumptions used for upper bounds, e.g., the bounded variance assumption:

$$\mathbb{E}[\|\nabla f_i(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|^2] \leq \tau' \|\nabla F(\boldsymbol{x})\|^2 + \nu'.$$

However, we would like to emphasize that posing stronger assumptions does *not* lead to weaker results in the case of lower bounds. This is because for two function classes with $\mathcal{F} \subset \mathcal{F}'$, a lower-bound-achieving function $f \in \mathcal{F}$ must also satisfy $f \in \mathcal{F}'$, i.e., f also establishes the same lower bound for $\mathcal{F}'$. For our case, if the components satisfy Assumption 2.4, then the function will also satisfy the bounded variance assumption for constants $\tau' = 2\tau^2$ and $\nu' = 2\nu^2$.

Algorithm 1 Offline GraB (Lu et al., 2022a)

```
Input: Initial point $\boldsymbol{x}_0 \in \mathbb{R}^d$, Learning rate $\eta > 0$, Number of epochs $K$, Nonnegative weights $\{\alpha_k\}_{k=1}^{K+1}$, Initial order $\sigma_1$
Initialize $\boldsymbol{x}_0^1 = \boldsymbol{x}_0$
for $k = 1, \dots, K$ do
  for $i = 1, \dots, n$ do
    Compute gradient: $\nabla f_{\sigma_k(i)}\left(\boldsymbol{x}_{i-1}^k\right)$.
    Store the gradient: $\boldsymbol{z}_i \leftarrow \nabla f_{\sigma_k(i)}\left(\boldsymbol{x}_{i-1}^k\right)$.
    Optimizer step: $\boldsymbol{x}_i^k = \boldsymbol{x}_{i-1}^k - \eta \boldsymbol{z}_i$
  end for
  $\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_n^k$
  Compute gradient mean: $\boldsymbol{z} \leftarrow \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i$
  Generate new order: $\sigma_{k+1} \leftarrow$ Herding $\left(\{\boldsymbol{z}_i - \boldsymbol{z}\}_{i=1}^n\right)$
end for
Output: $\hat{\boldsymbol{x}} = \sum_{k=1}^{K+1} \alpha_k \boldsymbol{x}_0^k / \sum_{k=1}^{K+1} \alpha_k$
```

2.2. Algorithms

We denote the i-th iterate of the k-th epoch of permutation-based SGD by $\boldsymbol{x}_i^k$, where $i = 0, \dots, n$ and $k = 1, \dots, K$. We denote the distance between the initial point $\boldsymbol{x}_0^1$ and the optimal point $\boldsymbol{x}^*$ as $D := \|\boldsymbol{x}_0^1 - \boldsymbol{x}^*\|$. We also follow the conventional notation $\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_n^k$, which indicates that the final result of an epoch becomes the initial point of its subsequent epoch. At the beginning of the k-th epoch, we choose a permutation $\sigma_k : [n] \to [n]$. The algorithm then accesses the component functions in the order of $f_{\sigma_k(1)}, \dots, f_{\sigma_k(n)}$. That is, we use the following update equation:

$$\boldsymbol{x}_i^k = \boldsymbol{x}_{i-1}^k - \eta \nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)$$

for $i = 1, \dots, n$, where $\eta > 0$ is a constant step size.

We particularly focus on two different types of permutation-based SGD. Section 3 states theoretical results based on SGD-RR, which assumes that the components are randomly shuffled independently in each epoch.

In Section 4, we study the case when permutations can be carefully chosen to gain faster convergence. We provide lower bounds that are applicable to *any* kind of permutation-based SGD. To show our lower bound is tight, it suffices to show that a *specific* permutation-based SGD algorithm provides a matching upper bound. To this end, we use *offline herding* SGD (Lu et al., 2022a), where the components are manually ordered to "balance" the gradients.

Specifically, Lu et al. (2022b) prove that as the gap between the partial sums of consecutive stochastic gradients and the full gradient diminishes faster, the optimizer converges faster as well. In their subsequent work (Lu et al., 2022a), they first propose *offline herding SGD*, a permutation-based SGD algorithm that manages this gap via the herding algorithm but requires intensive memory consumption, and devise *online herding SGD* (or GraB) that successfully overcomes the memory challenges. They prove that both algorithms guarantee the same convergence rate $\tilde{\mathcal{O}}\left(\frac{1}{n^2K^2}\right)$. In our setting, since we are not interested in the usability of algorithms, we will focus on *offline herding SGD* (or Offline GraB) just for simplicity. Algorithm 1 provides a pseudocode of Offline GraB. For the description of Herding subroutine in Algorithm 1, see Assumption 4.2 and Section 4.3.

3. Random Reshuffling

Here we show lower bounds of SGD-RR on the last iterate and arbitrary weighted averaged iterates for strongly convex objectives and then extend results to convex functions. We stress that the lower bounds on weighted average iterates tightly match the upper bounds both for the strongly convex and convex case.

3.1. Lower Bound for the Final Iterate

Theorem 3.1 provides a lower bound for the final iterate of SGD-RR for arbitrary step sizes $\eta > 0$ in the μ-strongly convex case.

Theorem 3.1. *For any $n \geq 2$ and $\kappa \geq c_1$, there exists a 3-dimensional function $F \in \mathcal{F}(L, \mu, 0, \nu)$ and an initialization point $\boldsymbol{x}_0$ such that for any constant step size η, the final iterate $\boldsymbol{x}_n^K$ of SGD-RR satisfies*

$$\mathbb{E}\left[F(\boldsymbol{x}_n^K) - F^*\right] = \begin{cases} \Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right), & \text{if } K \geq c_2\kappa, \\ \Omega\left(\frac{\nu^2}{\mu nK}\right), & \text{if } K < c_2\kappa, \end{cases}$$

for some universal constants c_1, c_2.

We take an approach similar to Yun et al. (2022), which is to construct F by aggregating three functions, each showing a lower bound for a different step size regime. The proof of Theorem 3.1 is deferred to Appendix B.

We can compare Theorem 3.1 with results of Yun et al. (2022) for $M = B = 1$, which establishes a lower bound of $\Omega(\frac{\nu^2}{\mu nK^2})$ for the large epoch regime $K \gtrsim \kappa$ and $\Omega(\frac{\nu^2}{\mu nK})$ for the small epoch regime $K \lesssim \kappa$. We can observe that the lower bound in Theorem 3.1 for the large epoch regime is tightened by a factor of κ. In fact, the bound can be compactly written as:

$$\mathbb{E}\left[F(\boldsymbol{x}_n^K) - F^*\right] = \Omega\left(\frac{\nu^2}{\mu nK} \cdot \min\left\{1, \frac{\kappa}{K}\right\}\right),$$

which can be interpreted as a *continuous* change from $\Omega\left(\frac{\nu^2}{\mu nK}\right)$ to $\Omega\left(\frac{\kappa\nu^2}{\mu nK^2}\right)$ as K gradually increases past the threshold $K \geq c_2\kappa$.

We can also compare our results with Safran & Shamir (2021), which provide a lower bound of rate $\Omega\left(\frac{\nu^2}{\mu nK} \cdot \min\left\{1, \frac{\kappa}{K}\left(\frac{1}{n} + \frac{\kappa}{K}\right)\right\}\right)$ under a stronger assumption that the objective and components are all *quadratic*. The lower bound for the small $K \lesssim \kappa$ regime is identical to ours since for this case our lower bound also relies on quadratic functions. However, if K grows past $\Omega(\kappa)$, then we can observe that the lower bound in Theorem 3.1 derived from non-quadratic functions is tighter by a factor of $\left(\frac{1}{n} + \frac{\kappa}{K}\right)$.

An upper bound for SGD-RR in the μ-strongly convex case under the step-size condition $\eta = \mathcal{O}(\frac{1}{Ln})$ is introduced in Theorem 2 of Mishchenko et al. (2020).

Proposition 3.2 (Corollary of Mishchenko et al. (2020), Theorem 2)**.** *Suppose that F and $f_1, \dots, f_n$ are all L-smooth, $f_1, \dots, f_n$ are convex, and F is μ-strongly convex. Also, let us define*

$$\sigma_*^2 := \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{x}^*)\|^2. \tag{3}$$

Then, for SGD-RR with constant step size

$$\eta = \min\left\{\frac{2}{Ln}, \frac{1}{\mu nK}\log\left(\frac{\mu^3 nD^2K^2}{L\sigma_*^2}\right)\right\},$$

the final iterate $\boldsymbol{x}_n^K$ satisfies

$$\mathbb{E}\left[F(\boldsymbol{x}_n^K) - F^*\right] = \tilde{\mathcal{O}}\left(LD^2 e^{-\frac{K}{\kappa}} + \frac{L^2\sigma_*^2}{\mu^3 nK^2}\right).$$

Note that the above proposition uses σ_*^2 which only relies on the gradients at the optimal point $\boldsymbol{x}^*$, while our lower bounds involve ν^2 which bounds the gradients for all $\boldsymbol{x}$. However, we can easily observe that Assumption 2.4 with $\tau = 0$ and $\boldsymbol{x} = \boldsymbol{x}^*$ implies that $\|\nabla f_i(\boldsymbol{x}^*)\|^2 \leq \nu^2$ for all i, and hence $\sigma_*^2 \leq \nu^2$. Therefore it is safe to compare this upper bound with our lower bounds by simply substituting the σ_*^2 terms with ν^2. Note that the same applies to Proposition 3.6.

Now, assuming $K \gtrsim \kappa$ so that the learning rate becomes

$$\eta = \frac{1}{\mu nK}\log\left(\frac{\mu^3 nD^2K^2}{L\nu^2}\right) \leq \frac{2}{Ln},$$

then we have $\mathbb{E}\left[F(\boldsymbol{x}_n^K) - F^*\right] = \tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 nK^2}\right)$, and for this case we can observe that lower bound shown in Theorem 3.1 matches the upper bound in Proposition 3.2 up to a factor of $\kappa = \frac{L}{\mu}$ and some polylogarithmic factors.

3.2. Lower Bound for Weighted Average Iterates

For small step sizes $\eta = \mathcal{O}\left(\frac{1}{Ln}\right)$, we can extend Theorem 3.1 to *arbitrary weighted average (end-of-epoch) iterates*. That is, Theorem 3.3 provides a lower bound which is applicable to all linear combinations of the following form,

$$\hat{\boldsymbol{x}} = \frac{\sum_{k=1}^{K+1} \alpha_k \boldsymbol{x}_0^k}{\sum_{k=1}^{K+1} \alpha_k}, \tag{4}$$

for nonnegative weights $\alpha_k \geq 0$ for all $k = 1, \dots, K+1$.

Theorem 3.3. *For any $n \geq 2$ and $\kappa \geq c_1$, there exists a 2-dimensional function $F \in \mathcal{F}(L, \mu, 0, \nu)$ and an initialization point $\boldsymbol{x}_0$ such that for any constant step size $\eta \leq \frac{1}{c_2 Ln}$, any weighted average iterate $\hat{\boldsymbol{x}}$ of SGD-RR of the form as in (4) satisfies*

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}) - F^*\right] = \begin{cases} \Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right), & \text{if } K \geq c_2\kappa, \\ \Omega\left(\frac{\nu^2}{\mu}\right), & \text{if } K < c_2\kappa, \end{cases}$$

for the same universal constants c_1, c_2 as in Theorem 3.1.

The full proof of Theorem 3.3 can be found in Appendix C. Note that, for weighted average iterates, we restrict ourselves to small step sizes $\eta = \mathcal{O}(\frac{1}{Ln})$; while this could look restrictive, such a choice of step size is commonly used in existing upper bounds, and we will see shortly that our lower bound exactly matches an upper bound when $K \gtrsim \kappa$ (Proposition 3.4). The tightness also extends to general convex cases, as seen in Section 3.3.

One might wonder why the lower bound becomes a *constant* for small $K \lesssim \kappa$. This is because in the $\eta = \mathcal{O}(\frac{1}{Ln})$ regime, $K < c_2\kappa$ implies $\eta < \frac{1}{c_2 Ln} \leq \frac{1}{\mu nK}$, i.e., the step size is too small for SGD to reach the optimum in K steps. For instance, K epochs of SGD on $F(x) = f_i(x) = \frac{\mu}{2}x^2$ initialized at $x = x_0$ reaches the point $(1 - \eta\mu)^{nK}x_0 > (1 - \frac{1}{nK})^{nK}x_0 \geq \frac{x_0}{4}$. Hence the iterate cannot get past $\frac{x_0}{4}$, rendering it impossible to reach the optimal point $x^* = 0$.

The difficulty of extending the $\eta = \Omega(\frac{1}{Ln})$ regime in Theorem 3.1 to arbitrary weighted average iterates originates from our proof strategy: for small enough η, we can show for our worst-case examples that all $\boldsymbol{x}_0^k$'s (in expectation) stay on the positive side bounded away from zero, thereby proving that any weighted average also stays sufficiently far from zero. However, for larger η, the iterates may oscillate between positive and negative regions, making it possible for an average iterate to converge faster than individual $\boldsymbol{x}_0^k$'s.

Note that our definition in (4) *includes* the final iterate, as the choice $\alpha_k = 0$ for $1 \leq k \leq K$ and $\alpha_{K+1} = 1$ yields $\hat{\boldsymbol{x}} = \boldsymbol{x}_0^{K+1} = \boldsymbol{x}_n^K$. Different forms of algorithm outputs other than the final iterate also frequently appear in prior works, especially regarding upper bounds for SGD-RR. For instance, we may choose $\alpha_k = 1$ for all $2 \leq k \leq K+1$ and $\alpha_1 = 0$ to represent the *average iterate* $\hat{\boldsymbol{x}}_{\text{avg}} := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_n^k$ (Mishchenko et al., 2020). We may also set $\alpha_k = 1$ for $\lceil \frac{K}{2} \rceil + 1 \leq k \leq K+1$ and $\alpha_k = 0$ otherwise to recover the *tail average iterate* (Nagaraj et al., 2019), defined as $\hat{\boldsymbol{x}}_{\text{tail}} := \frac{1}{K - \lceil \frac{K}{2} \rceil + 1}\sum_{k=\lceil \frac{K}{2} \rceil}^{K} \boldsymbol{x}_n^k$.

We further show that the lower bound in Theorem 3.3 tightly matches the upper bound suggested in Proposition 3.4.

Proposition 3.4. *Suppose that $F \in \mathcal{F}(L, \mu, 0, \nu)$, and that we choose η as*

$$\eta = \min\left\{\frac{1}{\sqrt{2}Ln}, \frac{9}{\mu nK}\max\left\{1, \log\left(\frac{\mu^3 nD^2K^2}{L\nu^2}\right)\right\}\right\}.$$

Then, for SGD-RR with constant step size η and $K \geq 5$, the tail average iterate $\hat{\boldsymbol{x}}_{\text{tail}}$ satisfies:

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] = \tilde{\mathcal{O}}\left(\frac{LD^2}{K}e^{-\frac{1}{9\sqrt{2}}\frac{K}{\kappa}} + \frac{L\nu^2}{\mu^2 nK^2}\right).$$

See Appendix D for a full proof of Proposition 3.4.

Assuming $K \gtrsim \kappa$ so that the learning rate becomes

$$\eta = \frac{9}{\mu nK}\max\left\{1, \log\left(\frac{\mu^3 nD^2K^2}{L\nu^2}\right)\right\} \leq \frac{1}{\sqrt{2}Ln},$$

then we have $\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] = \tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$ (see Cases (c), (d) in the proof). Then we can observe that the lower bound shown in Theorem 3.3 exactly matches the upper bound, ignoring polylogarithmic terms.

By introducing the tail average $\hat{\boldsymbol{x}}_{\text{tail}}$, we can obtain a rate of $\tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$ which is tighter than the rate $\tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 nK^2}\right)$ for the final iterate $\boldsymbol{x}_n^K$ by a factor of κ. Whether we can achieve the same, stronger upper bound for the final iterate $\boldsymbol{x}_n^K$ or not is still an open problem.

3.3. Extension to Convex Objectives

One important implication of Theorem 3.3 is that we can carefully choose a small value of μ to derive a lower bound that exactly matches the upper bound for *convex* objectives. Corollary 3.5 extends Theorem 3.3 to the convex case.

Corollary 3.5. *For any $n \geq 2$, there exists a 2-dimensional function $F \in \mathcal{F}(L, 0, 0, \nu)$ such that if*

$$K \geq c_3 \max\left\{\frac{L^2D^2n}{\nu^2}, \frac{\nu}{\mu Dn^{1/2}}\right\}, \tag{5}$$

then for any constant step size $\eta \leq \frac{1}{c_2 Ln}$, any weighted average iterate $\hat{\boldsymbol{x}}$ of SGD-RR of the form as in (4) satisfies

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}) - F^*\right] = \Omega\left(\frac{L^{1/3}\nu^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right),$$

for some universal constants c_2 and c_3.

We defer the proof of Corollary 3.5 to Appendix C.3.

A matching upper bound for SGD-RR for the convex case under the step-size condition $\eta = \mathcal{O}(\frac{1}{Ln})$ is introduced in Theorem 3 of Mishchenko et al. (2020).

Proposition 3.6 (Mishchenko et al. (2020), Theorem 3)**.** *Suppose that F and $f_1, \dots, f_n$ are all L-smooth and $f_1, \dots, f_n$ are convex. Also, suppose that we define σ_*^2 as in (3). Then, for* SGD-RR *with constant step size*

$$\eta = \min\left\{ \frac{1}{\sqrt{2}Ln}, \left(\frac{D^2}{L\sigma_*^2 n^2 K} \right)^{1/3} \right\},$$

the ***average iterate*** $\hat{\boldsymbol{x}}_{\text{avg}} := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_n^k$ *satisfies*

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{avg}}) - F^*\right] = \mathcal{O}\left(\frac{LD^2}{K} + \frac{L^{1/3}\sigma_*^{2/3}D^{4/3}}{n^{1/3}K^{2/3}} \right).$$

With the same logic as in Proposition 3.2, we can compare the above results with our lower bounds by substituting σ_*^2 with ν^2.

In Proposition 3.6, if we have a large number of epochs with $K \geq \frac{2\sqrt{2}L^2D^2n}{\nu^2}$, then $\eta = \left(\frac{D^2}{L\nu^2n^2K}\right)^{1/3} \leq \frac{1}{\sqrt{2}Ln}$ yields

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{avg}}) - F^*\right] = \mathcal{O}\left(\frac{L^{1/3}\nu^{2/3}D^{4/3}}{n^{1/3}K^{2/3}} \right).$$

For a large K regime of $K = \Omega\left(\frac{L^2D^2n}{\nu^2} + \frac{\nu}{\mu D n^{1/2}}\right)$, we may choose $\alpha_k = 1$ for all $k = 2, \dots, K+1$ and $\alpha_1 = 0$ so that $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{\text{avg}}$, and then observe that the lower bound in Corollary 3.5 exactly matches the upper bound in Proposition 3.6. Note that the lower bound of K in (5) reduces to $\Omega\left(\frac{L^2D^2n}{\nu^2}\right)$ when $n = \Omega\left(\frac{\nu^2}{\mu^{2/3}L^{4/3}D^2}\right)$, which then matches the epoch requirement that arises in the upper bound.

4. Arbitrary Permutation-based SGD

So far, we have considered the case where permutations are randomly shuffled for each epoch. In this section, we study the scenario when permutations can be chosen *manually* rather than randomly. We provide lower bounds that are applicable to any arbitrary permutation-based SGD. Our lower bounds match the previously established upper bound in terms of n and K, and can further match with respect to κ when the objective is ill-conditioned.

4.1. Lower Bound with Component Convexity

Theorem 4.1 establishes a lower bound on arbitrary weighted average (end-of-epoch) iterates applicable to any permutation-based SGD.

Theorem 4.1. *For any $n \geq 2$ and $\kappa \geq 4$, there exists a 4-dimensional function $F \in \mathcal{F}(L, \mu, 0, \nu)$ and an initialization point $\boldsymbol{x}_0$ such that for any permutation-based SGD with any constant step size η, any weighted average iterate $\hat{\boldsymbol{x}}$ of the form as in Equation (4) satisfies*

$$F(\hat{\boldsymbol{x}}) - F^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2} \right).$$

The main technical difficulty in proving Theorem 4.1 is that we must construct an objective that demonstrates a "slow" convergence rate for every permutation over K epochs. To achieve this, we design an objective that pushes $\boldsymbol{x}_n^k$ toward a constant direction, regardless of the permutation. The constructed objective belongs to the class $\mathcal{F}(L, \mu, 0, \nu)$ and satisfies component convexity. Here we note that our proof technique does *not* require any assumptions about large epochs. Furthermore, in contrast to the SGD-RR case (Theorem 3.3 and Corollary 3.5), this lower bound covers the entire range of step sizes. The full proof of Theorem 4.1 is written in Appendix E.

As mentioned in Section 3.1, applying $\alpha_k = 0$ for $1 \leq k \leq K$ and $\alpha_{K+1} = 1$ yields the lower bound for the last iterate. Our result significantly improves the previous lower bound and also matches the known upper bound of permutation-based SGD which will be discussed later in this section.

Comparison with the Previous Work. To the best of our knowledge, the best-known lower bound that holds for any arbitrary permutation-based SGD is proved by Rajput et al. (2022) prior to our work. Specifically, the authors show that there exists a $(2n+1)$-dimensional function $F \in \mathcal{F}\left(2L, \frac{n-1}{n}L, 1, \nu\right)$ such that for any permutation-based SGD with any constant step size,

$$F(\boldsymbol{x}_n^K) - F^* = \Omega\left(\frac{\nu^2}{Ln^3K^2} \right). \tag{6}$$

Thus, Theorem 4.1 improves the lower bound rate by a factor of n and sharpens the dependency on κ.

Before we state the matching upper bound, we define an additional assumption and a function class.

Assumption 4.2 (Herding bound)**.** There exists an algorithm that has the following property: Given $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n \in \mathbb{R}^d$ satisfying $\|\boldsymbol{z}_i\| \leq 1$ for $\forall i \in [n]$ and $\sum_{i=1}^{n} \boldsymbol{z}_i = 0$, the algorithm outputs a permutation $\sigma : [n] \to [n]$ such that $\max_{k \in \{1,\dots,n\}} \left\| \sum_{i=1}^{k} \boldsymbol{z}_{\sigma(i)} \right\| \leq H$.

We call an algorithm considered in Assumption 4.2 as the *Herding algorithm*, used as a subroutine in Algorithm 1.

Definition 4.3 (Function class)**.** We define the function class $\mathcal{F}_{P\L}$ as follows.

$$\begin{aligned}\mathcal{F}_{P\L}(L, \mu, \tau, \nu) := \{F :\ & f_i \text{ are } L\text{-smooth},\\ & F \text{ satisfies } \mu\text{-P\L condition},\\ & F \text{ and } f_i \text{ satisfy Assumption 2.4}\}.\end{aligned}$$

Note that $\mathcal{F}_{\text{PŁ}}$ is a relaxation of $\mathcal{F}$ in Definition 2.5. Compared to $\mathcal{F}$, we relax μ-strong convexity to μ-PŁ, and we also do not assume convexity of component functions f_i.

We now state the following proposition, provided in Theorem 1 of Lu et al. (2022a), which gives the convergence rate of Algorithm 1 for objectives belonging to $\mathcal{F}_{\text{PŁ}}(L, \mu, 0, \nu)$.

Proposition 4.4 (Lu et al. (2022a), Theorem 1)**.** *Suppose that $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, 0, \nu)$. Under Assumption 4.2, with constant step size η as*

$$\eta = \frac{2}{\mu n K} W_0 \left(\frac{\left(F(\boldsymbol{x}_0^1) - F^* + \nu^2/L\right) \mu^3 n^2 K^2}{192 H^2 L^2 \nu^2} \right),$$

where W_0 denotes the Lambert W function, Algorithm 1 converges at the rate

$$F(\boldsymbol{x}_n^K) - F^* = \tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2} \right)$$

for $K \gtrsim \kappa$.

Proposition 4.4 is a slightly different version compared to the original paper (Lu et al., 2022a); the differences are discussed in Section 4.3. We emphasize that Theorem 4.1 provides a lower bound $\Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right)$ for arbitrary permutation-based SGD and Proposition 4.4 shows that there exists an algorithm that converges at the rate of $\tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$. Note that the function class considered in the lower bound is a subset of the function class handled in the upper bound. Thus, Theorem 4.1 matches the upper bound up to a factor of κ, if we ignore the term H and some polylogarithmic terms. Therefore, we can conclude that Algorithm 1 is optimal in terms of the convergence rate with respect to n and K. We defer the discussion of herding constant H to Section 4.3.

4.2. Lower Bound without Component Convexity

Section 4.1 leads us to wonder if it is possible to tighten this κ gap. Our next theorem drops the assumption of component convexity in the lower bound and shows that we can close the gap and perfectly match the upper bound, if the problem is sufficiently ill-conditioned and the number of epochs is large enough.

Theorem 4.5. *For any $n \geq 104$, L and μ satisfying $\kappa \geq 8n$, and $K \geq \max\left\{\frac{\kappa^2}{n}, \kappa^{3/2} n^{1/2}\right\}$, there exists a 4-dimensional function $F \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$ and an initialization point $\boldsymbol{x}_0$ such that for any permutation-based SGD with any constant step size η, any weighted average iterate $\hat{\boldsymbol{x}}$ of the form as in Equation (4) satisfies*

$$F(\hat{\boldsymbol{x}}) - F^* = \Omega\left(\frac{L^2 \nu^2}{\mu^3 n^2 K^2} \right).$$

The proof is in Appendix F. Theorem 4.5 provides a sharper lower bound than the previous result with respect to κ. In our construction, some of the components f_i are nonconvex but the constructed objective F is actually strongly convex; however, for simplicity of exposition, we stated F as a member of a larger class $\mathcal{F}_{\text{PŁ}}$. Here we discuss the effect of nonconvex components on the convergence rate.

Nonconvex components. Some of our component functions constructed in Theorem 4.5 are concave in particular directions, and this is the key to obtaining an additional κ factor. Rajput et al. (2022) also observe that the presence of nonconvex components can slow down convergence. They prove that for a 1-dimensional objective $F(x) = \frac{1}{n}\sum_{i=1}^n \frac{a_i}{2}x^2 - b_i x$, where all a_i's are nonnegative, there exists a permutation that leads to exponential convergence, but also that this no longer holds if a_i's are allowed to be negative. It is an open problem whether the convergence rate of Algorithm 1 could be improved to match the lower bound in Theorem 4.1 with respect to κ if we additionally assume component convexity.

Theorem 4.5 provides a sharper lower bound compared to Theorem 4.1 with respect to κ. One should be aware, however, that the function classes considered in the upper bound (Proposition 4.4) and the construction in Theorem 4.5 mismatch. Therefore, Proposition 4.4 does *not* guarantee the $\mathcal{O}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2}\right)$ convergence rate for the function constructed in Theorem 4.5. However, we argue that this issue can be addressed by extending Proposition 4.4 to a wider function class, which is done in Proposition 4.6.

Proposition 4.6 (Extended version of Lu et al. (2022a), Theorem 1)**.** *Suppose that $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ and $n \geq H$. Under Assumption 4.2, with constant step size η as*

$$\eta = \frac{2}{\mu n K} W_0 \left(\frac{\left(F(\boldsymbol{x}_0^1) - F^* + \nu^2/L\right) \mu^3 n^2 K^2}{192 H^2 L^2 \nu^2} \right),$$

where W_0 denotes the Lambert W function, Algorithm 1 converges at the rate

$$F(\boldsymbol{x}_n^K) - F^* = \tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2} \right)$$

for $K \gtrsim \kappa(\tau + 1)$.

The proof of Proposition 4.6 is in Appendix G. We show that the function class considered in Proposition 4.4 can be extended to $\mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ by following the proof step in Theorem 1 in Lu et al. (2022a) with slight modifications. The function class of the upper bound (Proposition 4.6) now includes the construction of the lower bound (Theorem 4.5). Therefore, when the objective is sufficiently ill-conditioned and a sufficiently many epochs are performed, our lower bound perfectly aligns with the upper bound in all factors, assuring that GraB is indeed the optimal permutation-based SGD algorithm.

4.3. Discussion of Existing Results

In this section, we take a deeper look at previous researches that address permutation-based SGD. We mainly discuss the dimension dependency hidden in the upper bounds.

Herding Bound. Bansal & Garg (2017) prove that there exists an efficient Herding algorithm that achieves Assumption 4.2 with $H = \mathcal{O}\left(\sqrt{d \log n}\right)$. Also, it is well known that $H = \Omega(\sqrt{d})$ (Behrend, 1954; Bárány, 2008). Thus, both Proposition 4.4 and Proposition 4.6 contain a dimension term in their convergence rates. Meanwhile, our lower bound results are based on fixed dimensional functions, so we can ignore the term H when we compare our lower bound results to the upper bound results. We also note that the assumption $n \geq H$ made in Proposition 4.6 is quite mild if the dimension of F is independent of n.

Comparison between Proposition 4.4 and Lu et al. (2022a), Theorem 1. In the original statement of Theorem 1 in Lu et al. (2022a), the authors use slightly different assumptions. Instead of smoothness with respect to the L_2 norm, they assume $L_{2,\infty}$-smoothness as follows:

$$\left\| \nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y}) \right\|_2 \leq L_{2,\infty} \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_\infty, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Lu et al. (2022a) also define the herding bound H with respect to different choices of norms. Specifically, the authors consider $\max_{k \in \{1,\dots,n\}} \left\| \sum_{i=1}^k \boldsymbol{z}_{\sigma(i)} \right\|_\infty \leq H_\infty$ with $\max_i \left\| \boldsymbol{z}_i \right\|_2 \leq 1$, and explain that combining the results from Harvey & Samadi (2014) and Alweiss et al. (2021) gives $H_\infty = \tilde{\mathcal{O}}(1)$. With these assumptions, the authors obtain the convergence rate of Algorithm 1 as the following:

$$F(\boldsymbol{x}_n^K) - F^* = \tilde{\mathcal{O}}\left(\frac{H_\infty^2 L_{2,\infty}^2 \nu^2}{\mu^3 n^2 K^2} \right). \tag{7}$$

However, we believe that Equation (7) is not also free from dimension dependency, since the term $L_{2,\infty}$ is likely to contain the dimension dependency in general (e.g., $L_{2,\infty} = \sqrt{d}L$ holds when $F(\boldsymbol{x}) = \frac{L}{2}\left\| \boldsymbol{x} \right\|^2$ for $\forall \boldsymbol{x} \in \mathbb{R}^d$). It is an open problem whether there exists a permutation-based SGD algorithm that gives a dimension-free upper bound while maintaining the same dependency on other factors.

Revisiting Rajput et al. (2022). We have discussed that the best-known upper bound of permutation-based SGD has dimension dependency. Earlier, we mentioned that our lower bound in Theorem 4.1 improves upon previous results from Theorem 2 of Rajput et al. (2022) by a factor of n. In fact, the construction of Rajput et al. (2022) is based on a $(2n+1)$-dimensional function, and applying the upper bounds for Algorithm 1 to this function results in a convergence rate of $\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$, due to the dimension dependency. More precisely, for the function constructed in Rajput et al. (2022), H is proportional to $\sqrt{n}$ and L is constant according to our L_2-norm-based notations, while we also have that H_∞ is constant and $L_{2,\infty}$ is proportional to $\sqrt{n}$ following the notations in Lu et al. (2022a). (Here we ignore log factors.) Thus, in terms of n dependency, we conclude that the actual gap between existing upper and lower bounds is n^2 rather than n, and that our results succeed in closing the gap completely.

5. Conclusion

We have shown convergence lower bounds for without-replacement SGD methods, focusing on matching the upper and lower bound in terms of the condition number κ. Our lower bounds for SGD-RR on weighted average iterates tightly match the corresponding upper bounds under both strong convexity and convexity assumptions. We also constructed lower bounds for permutation-based SGD with *and* without individual convexity assumptions, which tightly match the upper bounds for GraB in fixed-dimension settings, therefore implying that GraB achieves the optimal rate of convergence.

An immediate direction for future work is to investigate whether one can find lower bounds for arbitrary weighted average iterates of SGD-RR when $\eta = \Omega\left(\frac{1}{Ln}\right)$. In the discussion following Theorem 3.3 (Section 3.2), we outlined difficulties that arise in proving such a result for larger learning rates $\eta = \Omega\left(\frac{1}{Ln}\right)$.

We finally note that the power of general permutation-based SGD is not yet well-understood for the regime when the number of epochs is less than the condition number. Safran & Shamir (2021) show that SGD-RR does not enjoy faster convergence than *with-replacement* SGD in this regime, and it is still unclear whether the same restriction holds for permutation-based SGD as well.

Acknowledgements

This paper was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) funded by the Korea government (MSIT), two National Research Foundation of Korea (NRF) grants (No. NRF-2019R1A5A1028324, RS-2023-00211352) funded by the Korea government (MSIT), and a grant funded by Samsung Electronics Co., Ltd.

References

Ahn, K., Yun, C., and Sra, S. SGD with shuffling: Optimal rates without component convexity and large epoch requirements. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Alweiss, R., Liu, Y. P., and Sawhney, M. Discrepancy minimization via a self-balancing walk. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 14–20, 2021.

Bansal, N. and Garg, S. Algorithmic discrepancy beyond partial coloring. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 914–926, 2017.

Bárány, I. On the power of linear dependencies. In *Building bridges*, pp. 31–45. Springer, 2008.

Behrend, F. The steinitz-gross theorem on sums of vectors. *Canadian Journal of Mathematics*, 6:108–124, 1954.

Benaïm, M. Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, 33:1–68, 1999. URL http://www.numdam.org/item/SPS_1999__33__1_0/.

Bottou, L. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, 2009.

Bottou, L., Curtis, F. E., and Nocedal, J. 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.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. ISSN 1935-8237. doi: 10.1561/2200000050. URL http://dx.doi.org/10.1561/2200000050.

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.

Gürbüzbalaban, M., Ozdaglar, A. E., and Parrilo, P. A. Convergence rate of incremental gradient and incremental Newton methods. *SIAM J. Optim.*, 29(4):2542–2565, 2019. doi: 10.1137/17M1147846. URL https://doi.org/10.1137/17M1147846.

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. and Samadi, S. Near-optimal herding. In *Conference on Learning Theory*, pp. 1165–1182. PMLR, 2014.

Lu, Y., Guo, W., and Sa, C. D. GraB: Finding provably better data permutations than random reshuffling. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=nDemfqKHTpK.

Lu, Y., Meng, S. Y., and De Sa, C. a general analysis of example-selection for stochastic gradient descent. In *International Conference on Learning Representations*, 2022b.

Mishchenko, K., Khaled, A., and Richtarik, P. Random reshuffling: Simple analysis with vast improvements. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/c8cc6e90ccbff44c9cee23611711cdc4-Paper.pdf.

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.

Mortici, C. On Gospers formula for the gamma function. *Journal of Mathematical Inequalities*, 5, 12 2011. doi: 10.7153/jmi-05-53.

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.

Rajput, S., Gupta, A., and Papailiopoulos, D. S. Closing the convergence gap of SGD without replacement. *CoRR*, abs/2002.10400, 2020. URL https://arxiv.org/abs/2002.10400.

Rajput, S., Lee, K., and Papailiopoulos, D. Permutation-based SGD: Is random optimal? In *International Conference on Learning Representations*, 2022.

Recht, B. and Re, C. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. In Mannor, S., Srebro, N., and Williamson, R. C. (eds.), *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 11.1–11.24, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL https://proceedings.mlr.press/v23/recht12.html.

Recht, B. and Ré, C. Parallel stochastic gradient algorithms for large-scale matrix completion. *Math. Program. Comput.*, 5(2):201–226, 2013. doi: 10.1007/s12532-013-0053-8. URL https://doi.org/10.1007/s12532-013-0053-8.

Robbins, H. and Monro, S. 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.

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. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15151–15161. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/803ef56843860e4a48fc4cdb3065e8ce-Paper.pdf.

Tran, T. H., Scheinberg, K., and Nguyen, L. M. Nesterov accelerated shuffling gradient method for convex optimization. In *International Conference on Machine Learning*, pp. 21703–21732. PMLR, 2022.

Yun, C., Sra, S., and Jadbabaie, A. Open problem: Can Single-Shuffle SGD be better than Reshuffling SGD and GD? In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4653–4658. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/open-problem-yun21a.html.

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.

A. Comparison with Previous Results

Table 2 shows a detailed comparison of existing convergence rates and our results for permutation-based SGD. Note that the function class categories are divided with respect to the lower bound results— the selected upper bounds are the results with the best convergence rates among those of which the function class contains the constructed lower bounds. The upper bound results are colored white and the lower bound results are colored gray.

Similarly as in Table 1, the parameters L, μ, ν, and D are defined in Section 2. Algorithm outputs $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{x}}_{\text{tail}}$, and $\hat{\boldsymbol{x}}_{\text{avg}}$ are defined in Section 3. Function classes $\mathcal{F}$ and $\mathcal{F}_{\text{PŁ}}$ are defined in Sections 2 and 4, respectively. The herding bound H, which closely relates to the convergence rate of Algorithm 1, is defined in Section 4.

Table 2. A detailed comparison of existing convergence rates and our results for permutation-based SGD.

Random Reshuffling				
Function Class	Output	References	Convergence Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	Mishchenko et al. (2020)	$\tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 nK^2}\right)$	$K \gtrsim \kappa$
		Yun et al. (2022)	$\Omega\left(\frac{\nu^2}{\mu nK^2}\right)$	$\kappa \geq c$, $K \gtrsim \kappa$
		Ours, Theorem 3.1	$\Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$\kappa \geq c$, $K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}_{\text{tail}}$	Nagaraj et al. (2019)†	$\tilde{\mathcal{O}}\left(\frac{L^2\nu^2}{\mu^3 nK^2}\right)$	$K \gtrsim \kappa^2$
		Ours, Proposition 3.4	$\tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Ours, Theorem 3.3‡	$\Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right)$	$\kappa \geq c$, $K \gtrsim \kappa$
$\mathcal{F}(L, 0, 0, \nu)$	$\hat{\boldsymbol{x}}_{\text{avg}}$	Mishchenko et al. (2020)	$\mathcal{O}\left(\frac{L^{1/3}\nu^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$	$K \gtrsim \frac{L^2D^2n}{\nu^2}$
	$\hat{\boldsymbol{x}}$	Ours, Corollary 3.5‡	$\Omega\left(\frac{L^{1/3}\nu^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right)$	$K \gtrsim \max\{\frac{L^2D^2n}{\nu^2}, \frac{\nu}{\mu D n^{1/2}}\}$

Arbitrary Permutations				
Function Class	Output	References	Convergence Rate	Assumptions
$\mathcal{F}(L, \mu, 0, \nu)$	$\boldsymbol{x}_n^K$	Lu et al. (2022a) (GraB)	$\tilde{\mathcal{O}}\left(\frac{H^2L^2\nu^2}{\mu^3 n^2K^2}\right)$	$K \gtrsim \kappa$
	$\hat{\boldsymbol{x}}$	Ours, Theorem 4.1	$\Omega\left(\frac{L\nu^2}{\mu^2 n^2K^2}\right)$	-
$\mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$	$\boldsymbol{x}_n^K$	Ours, Proposition 4.6 (GraB)	$\tilde{\mathcal{O}}\left(\frac{H^2L^2\nu^2}{\mu^3 n^2K^2}\right)$	$n \geq H$, $K \gtrsim \kappa(\tau+1)$
	$\boldsymbol{x}_n^K$	Rajput et al. (2022)*	$\Omega\left(\frac{\nu^2}{Ln^3K^2}\right)$	$d = 2n+1$
	$\hat{\boldsymbol{x}}$	Ours, Theorem 4.5	$\Omega\left(\frac{L^2\nu^2}{\mu^3 n^2K^2}\right)$	$\tau = \kappa \geq 8n$, $K \geq \max\{\frac{\kappa^2}{n}, \kappa^{\frac{3}{2}}n^{\frac{1}{2}}\}$

† Assumes a stronger condition of $\|\nabla f_i(\boldsymbol{x})\| \leq \nu$ for all i and $\boldsymbol{x}$

‡ Additionally assumes $\eta \leq \frac{1}{c_2 Ln}$

* The constructed objective is a member of $\mathcal{F}\left(2L, \frac{n-1}{n}L, 1, \nu\right)$.

B. Proof of Theorem 3.1

Here we prove Theorem 3.1, restated below for the sake of readability.

Theorem 3.1. *For any $n \geq 2$ and $\kappa \geq c_1$, there exists a 3-dimensional function $F \in \mathcal{F}(L, \mu, 0, \nu)$ and an initialization point $\boldsymbol{x}_0$ such that for any constant step size η, the final iterate $\boldsymbol{x}_n^K$ of SGD-RR satisfies*

$$\mathbb{E}\left[F(\boldsymbol{x}_n^K) - F^*\right] = \begin{cases} \Omega\left(\frac{L\nu^2}{\mu^2 nK^2}\right), & \text{if } K \geq c_2\kappa, \\ \Omega\left(\frac{\nu^2}{\mu nK}\right), & \text{if } K < c_2\kappa, \end{cases}$$

for some universal constants c_1, c_2.

Proof. We prove the theorem statement for constants $c_1 = 2415$ and $c_2 = 161$.

As the convergence behavior of SGD heavily depends on the step size η, we consider three step-size regimes and use different objective functions with slow convergence rates in each case. Then we aggregate the three functions to obtain the final lower bound, which will be the minimum among the lower bounds from each regime. Throughout the proof, we will assume n is even. If n is odd, then we can use a similar technique with Theorem 1 of Safran & Shamir (2021), which is to set $n-1$ nontrivial components satisfying the statement, add a single zero component function, and scale by $\frac{n-1}{n}$.

To elaborate, we prove the following lower bounds for each of the following three regimes. Here we denote by F_j^* the minimizer of F_j for each $j = 1, 2, 3$. Note that the union of the three ranges completely covers the set of all positive learning rates, $\eta > 0$.

- If $\eta \in \left(0, \frac{1}{\mu n K}\right)$, there exists a 1-dimensional objective function $F_1(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with initialization $x_0^1 = D_0$ (for any D_0) satisfies
$$\mathbb{E}\left[F_1(x_n^K) - F_1^*\right] = \Omega\left(\mu D_0^2\right).$$
- If $\eta \in \left[\frac{1}{\mu n K}, \frac{1}{161 L n}\right]$, there exists a 1-dimensional objective function $F_2(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with initialization $x_0^1 = \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$ satisfies
$$\mathbb{E}\left[F_2(x_n^K) - F_2^*\right] = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right).$$
- If $\eta \geq \max\left\{\frac{1}{\mu n K}, \frac{1}{161 L n}\right\}$, there exists a 1-dimensional objective function $F_3(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with initialization $x_0^1 = 0$ satisfies
$$\mathbb{E}\left[F_3(x_n^K) - F_3^*\right] = \Omega\left(\frac{\nu^2}{\mu n K}\right).$$

Now we define the 3-dimensional function $F(x, y, z) = F_1(x) + F_2(y) + F_3(z)$, where F_1, F_2, and F_3 are chosen to satisfy the above lower bounds for ν replaced by $\frac{\nu}{\sqrt{3}}$. Note that scaling ν does not change the convergence rates above. We denote the components by $f_i(x, y, z) = f_{1,i}(x) + f_{2,i}(y) + f_{3,i}(z)$ for $i = 1, \dots, n$.

If H_1, H_2, and H_3 are L-smooth and μ-strongly convex, then $H(x, y, z) = H_1(x) + H_2(y) + H_3(z)$ satisfies
$$\mu \boldsymbol{I} \preceq \min\{\nabla^2 H_1(x), \nabla^2 H_2(y), \nabla^2 H_3(z)\} \preceq \nabla^2 H(x, y, z) \preceq \max\{\nabla^2 H_1(x), \nabla^2 H_2(y), \nabla^2 H_3(z)\} \preceq L\boldsymbol{I},$$
i.e., $H(\boldsymbol{x})$ must be L-smooth and μ-strongly convex.

Also, if H_1, H_2, and H_3 (each with n components $h_{1,i}$, $h_{2,i}$, and $h_{3,i}$) have bounded gradients (Assumption 2.4) for $\tau = 0$ and $\nu = \frac{\nu_0}{\sqrt{3}}$, then $H(x, y, z) = H_1(x) + H_2(y) + H_3(z)$ satisfies
$$\begin{aligned}
&\|\nabla h_i(x, y, z) - \nabla H(x, y, z)\|^2 \\
&= \|\nabla h_{1,i}(x) - \nabla H_1(x)\|^2 + \|\nabla h_{2,i}(y) - \nabla H_2(y)\|^2 + \|\nabla h_{3,i}(z) - \nabla H_3(z)\|^2 \\
&\leq \frac{\nu_0^2}{3} + \frac{\nu_0^2}{3} + \frac{\nu_0^2}{3} = \nu_0^2
\end{aligned}$$
for all $i = 1, \dots, n$, i.e., $H(x, y, z)$ satisfies Assumption 2.4 for $\tau = 0$ and $\nu = \nu_0$.

Since $F_1, F_2, F_3 \in \mathcal{F}(L, \mu, 0, \frac{\nu}{\sqrt{3}})$ by construction, we have $F \in \mathcal{F}(L, \mu, 0, \nu)$ from the above arguments.

Now suppose that we set $D_0 = \frac{\nu}{\mu}$ and initialize at the point $\left(\frac{\nu}{\mu}, \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}, 0\right)$.

If $K \geq 161\kappa$, then since $\frac{1}{\mu n K} \leq \frac{1}{161 L n}$ we can use the lower bound for $F_2(y)$. The lower bound for this case becomes
$$\mathbb{E}\left[F(x_n^K, y_n^K, z_n^K) - F^*\right] = \Omega\left(\min\left\{\frac{\nu^2}{\mu}, \frac{L\nu^2}{\mu^2 n K^2}, \frac{\nu^2}{\mu n K}\right\}\right) = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right).$$

If $K < 161\kappa$, then since $\frac{1}{\mu n K} > \frac{1}{161 L n}$ the middle step-size regime does not exist, i.e., we *cannot* use the lower bound for $F_2(y)$. Hence the lower bound for this case becomes

$$\mathbb{E}\left[F(x_n^K, y_n^K, z_n^K) - F^*\right] = \Omega\left(\min\left\{\frac{\nu^2}{\mu}, \frac{\nu^2}{\mu n K}\right\}\right) = \Omega\left(\frac{\nu^2}{\mu n K}\right),$$

which completes the proof. □

For the following subsections, we prove the lower bounds for F_1, F_2, and F_3 at the corresponding step size regimes. The proofs are similar to those of Yun et al. (2022), corresponding to the case $M = B = 1$ with slight modifications.

B.1. Lower Bound for $\eta \in \left(0, \frac{1}{\mu n K}\right)$

Here we show that there exists $F_1(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with $x_0^1 = D_0$ satisfies

$$\mathbb{E}\left[F_1(x_n^K) - F_1^*\right] = \Omega\left(\mu D_0^2\right).$$

Proof. We define $F_1(x) \in \mathcal{F}(\mu, \mu, 0, 0)$ by the following components:

$$f_i(x) = F_1(x) = \frac{\mu}{2}x^2.$$

Note that $\mathcal{F}(\mu, \mu, 0, 0) \subseteq \mathcal{F}(L, \mu, 0, \nu)$ and $F_1^* = 0$ at $x^* = 0$ by definition. Also, note that the components have no stochasticity, and hence we can drop the expectation notation, $\mathbb{E}[\cdot]$. Then we can easily compute per-epoch updates as:

$$x_0^{k+1} = (1 - \eta\mu)^n x_0^k, \quad \forall k = 1, \dots, K.$$

Since $x_0^1 = x_0 = D_0$ and $\eta \leq \frac{1}{\mu n K}$, for any $k = 1, \dots, K$ we have

$$x_0^{k+1} = (1 - \eta\mu)^{nk} \cdot D_0 \geq \left(1 - \frac{1}{nK}\right)^{nK} \cdot D_0 \geq \frac{D_0}{4}, \tag{8}$$

where in the last inequality we use $\left(1 - \frac{1}{m}\right)^m \geq \frac{1}{4}$ for all $m \geq 2$. Hence, for the final iterate we have $x_n^K \geq \frac{D_0}{4}$ and therefore

$$F_1(x_n^K) = \frac{\mu}{2}(x_n^K)^2 \geq \frac{\mu}{2}\left(\frac{D_0}{4}\right)^2 = \frac{\mu D^2}{32},$$

which concludes that $\mathbb{E}\left[F_1(x_n^K) - F_1^*\right] = \mathbb{E}\left[F_1(x_n^K)\right] = F_1(x_n^K) = \Omega\left(\mu D_0^2\right)$. □

B.2. Lower Bound for $\eta \in \left[\frac{1}{\mu n K}, \frac{1}{161 L n}\right]$

Here we show that there exists $F_2(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with $x_0^1 = \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$ satisfies

$$\mathbb{E}\left[F_2(x_n^K) - F_2^*\right] = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right).$$

Proof. We define $F_2(x) \in \mathcal{F}(L, \mu, 0, \nu)$ by the following components:

$$f_i(x) = \begin{cases} \left(L\mathbb{1}_{x<0} + \mu_0\mathbb{1}_{x\geq 0}\right)\frac{x^2}{2} + \nu x & \text{if } i \leq n/2, \\ \left(L\mathbb{1}_{x<0} + \mu_0\mathbb{1}_{x\geq 0}\right)\frac{x^2}{2} - \nu x & \text{otherwise,} \end{cases}$$

where we assume $\mu_0 \leq \frac{L}{2415}$ and later choose $\mu_0 = \frac{L}{2415}$. With this construction, the finite-sum objective becomes

$$F_2(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \left(L\mathbb{1}_{x<0} + \mu_0\mathbb{1}_{x\geq 0}\right)\frac{x^2}{2}.$$

Note that $F_2^* = 0$ at $x^* = 0$ by definition, and that μ_0 is different from μ. While $F_2(x) \in \mathcal{F}(L, \mu_0, 0, \nu)$ by construction, we can ensure that $\mathcal{F}(L, \mu_0, 0, \nu) \subset \mathcal{F}(L, \mu, 0, \nu)$ because the assumption $\kappa \geq 2415$ implies $\mu_0 = \frac{L}{2415} \geq \mu$.

First, we focus on a single epoch, and hence we write x_i instead of x_i^k, omitting the superscripts k for a while.

We use the following definition throughout the paper for notational simplicity.

Definition B.1. Define $\mathcal{S}_n$ as the set of all possible permutations of $\frac{n}{2}$ $+1$'s and $\frac{n}{2}$ -1's, where n is a positive, even integer. If SGD-RR samples a permutation σ in a certain epoch, we define the corresponding $s \in \mathcal{S}_n$ to satisfy $s_i = +1$ if $\sigma(i) \leq \frac{n}{2}$ and $s_i = -1$ if $\sigma(i) > \frac{n}{2}$.

Note that following Definition B.1, we can express the iterations for $i = 1, \dots, n$ via s as

$$x_i = x_{i-1} - \eta \nabla f_{\sigma(i)}(x) = x_{i-1} - \eta(L\mathbb{1}_{x_{i-1}<0} + \mu_0 \mathbb{1}_{x_{i-1}\geq 0})x_{i-1} - \eta\nu s_i.$$

Also, we can sum up the iterates to obtain

$$\begin{aligned} x_n &= x_{n-1} - \eta(L\mathbb{1}_{x_{n-1}<0} + \mu_0 \mathbb{1}_{x_{n-1}\geq 0})x_{n-1} - \eta\nu s_n \\ &= x_{n-2} - \eta(L\mathbb{1}_{x_{n-2}<0} + \mu_0 \mathbb{1}_{x_{n-2}\geq 0})x_{n-2} - \eta\nu s_{n-1} - \eta(L\mathbb{1}_{x_{n-1}<0} + \mu_0 \mathbb{1}_{x_{n-1}\geq 0})x_{n-1} - \eta\nu s_n \\ &\vdots \\ &= x_0 - \eta \sum_{i=0}^{n-1} (L\mathbb{1}_{x_i<0} + \mu_0 \mathbb{1}_{x_i\geq 0})x_i - \eta\nu \sum_{i=1}^{n} s_i \\ &= x_0 - \eta \sum_{i=0}^{n-1} (L\mathbb{1}_{x_i<0} + \mu_0 \mathbb{1}_{x_i\geq 0})x_i. \end{aligned}$$

Now we use the following three lemmas.

Lemma B.2. *For (fixed) $x_0 \geq 0$, $0 \leq i \leq \lfloor \frac{n}{2} \rfloor$, $\eta \leq \frac{1}{161Ln}$, and $\frac{L}{\mu_0} \geq 2415$,*

$$\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0 \mathbb{1}_{x_i\geq 0})x_i\right] \leq \frac{2}{3}Lx_0 - \frac{\eta L\nu}{480}\sqrt{i}.$$

Lemma B.3. *For (fixed) $x_0 \geq 0$, $0 \leq i \leq n-1$, and $\eta \leq \frac{1}{161Ln}$,*

$$\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0 \mathbb{1}_{x_i\geq 0})x_i\right] \leq \left(1 + \frac{161}{160} i\eta L\right)\mu_0 x_0 + \frac{161}{160}\eta\mu_0\nu\sqrt{i}.$$

Lemma B.4. *If $\eta \leq \frac{1}{161Ln}$, we have the followings.*

1. *For (fixed) $x_0 < 0$, we have*

$$\mathbb{E}\left[x_n\right] \geq \left(1 - \frac{160}{161}\eta Ln\right)x_0.$$

2. *If we initialize at $x_0^1 \geq 0$, then we always have $\mathbb{P}(x_n^k \geq 0) \geq \frac{1}{2}$ for future start-of-epoch iterates.*

See Appendix B.4 for the proofs of Lemmas B.2 to B.4.

If an epoch starts at (a fixed value) $x_0 \geq 0$, then from Lemmas B.2 and B.3 we have

$$\mathbb{E}\left[x_n - x_0\right] = -\eta \sum_{i=0}^{n-1} \mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0 \mathbb{1}_{x_i\geq 0})x_i\right]$$

$$
\begin{aligned}
&= -\eta \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0 \mathbb{1}_{x_i\ge 0})x_i\right] - \eta \sum_{i=\lfloor \frac{n}{2}\rfloor+1}^{n-1} \mathbb{E}\left[(L\mathbb{1}_{x_i<0}+\mu_0\mathbb{1}_{x_i\ge0})x_i\right] \\
&\ge -\eta \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \left(\frac{2}{3}Lx_0 - \frac{\eta L \nu}{480}\sqrt{i}\right) - \eta \sum_{i=\lfloor \frac{n}{2}\rfloor+1}^{n-1} \left(\left(1+\frac{161}{160}i\eta L\right)\mu_0 x_0 + \frac{161}{160}\eta\mu_0\nu\sqrt{i}\right) \\
&= -\eta \left(\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{2}{3}L + \sum_{i=\lfloor \frac{n}{2}\rfloor+1}^{n-1} \left(1+\frac{161}{160}i\eta L\right)\mu_0\right) x_0 - \eta \left(-\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{\eta L\nu}{480}\sqrt{i} + \sum_{i=\lfloor \frac{n}{2}\rfloor+1}^{n-1} \frac{161}{160}\eta\mu_0\nu\sqrt{i}\right).
\end{aligned}
$$

Now we can bound the coefficient of the x_0 term by the following inequality:

$$
\begin{aligned}
\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{2}{3}L + \sum_{i=\lfloor \frac{n}{2}\rfloor+1}^{n-1} \left(1+\frac{161}{160}i\eta L\right)\mu_0 &\le \left(\left\lfloor\frac{n}{2}\right\rfloor+1\right)\frac{2}{3}L + \left(n-\left\lfloor\frac{n}{2}\right\rfloor-1\right)\frac{L}{2415}\left(1+\frac{1}{160}\right) \\
&\le \frac{2}{3}Ln + \frac{Ln}{2400} \le \frac{3}{4}Ln,
\end{aligned}
$$

where we use $\mu_0 \le \frac{L}{2415}$ and $i\eta L \le \eta L n \le \frac{1}{161}$. Also, the constant term can be bounded as:

$$
\begin{aligned}
-\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{\eta L\nu}{480}\sqrt{i} + \sum_{i=\lfloor \frac{n}{2}\rfloor+1}^{n-1} \frac{161}{160}\eta\mu_0\nu\sqrt{i} &\le -\frac{\eta L\nu}{480}\int_0^{\lfloor \frac{n}{2}\rfloor} \sqrt{t}dt + \frac{161}{160}\eta\mu_0\nu \int_{\lfloor \frac{n}{2}\rfloor+1}^{n} \sqrt{t}dt \\
&\le -\frac{\eta L\nu}{480}\cdot\frac{2}{3}\left(\left\lfloor\frac{n}{2}\right\rfloor\right)^{3/2} + \frac{161}{160}\eta\mu_0\nu\cdot\frac{2}{3}\left(n^{3/2}-\left(\frac{n}{2}\right)^{3/2}\right) \\
&\le -\frac{\eta L\nu}{480}\cdot\frac{2}{3}\left(\frac{n}{3}\right)^{3/2} + \frac{161}{160}\eta\mu_0\nu\cdot\frac{2}{3}\cdot\frac{2\sqrt{2}-1}{2\sqrt{2}}n^{3/2} \\
&\le -\frac{\eta L\nu}{480}\cdot\frac{2}{9\sqrt{3}}n^{3/2} + \frac{161}{160}\eta\mu_0\nu\cdot\frac{1}{2}n^{3/2} \\
&\le -\eta L\nu n^{3/2}\left(\frac{2}{480\cdot 9\sqrt{3}} - \frac{161}{160\cdot 2\cdot 2415}\right) \le -\frac{\eta L \nu n^{3/2}}{18000},
\end{aligned}
$$

where we use $\mu_0 \le \frac{L}{2415}$, $\lfloor \frac{n}{2} \rfloor \ge \frac{n}{3}$ (for $n \ge 2$), and $\frac{2}{480\cdot 9\sqrt{3}} - \frac{161}{160\cdot 2\cdot 2415} > \frac{1}{18000}$. Hence we can conclude that

$$
\mathbb{E}\left[x_n - x_0\right] \ge -\eta\left(\frac{3}{4}Lnx_0 - \frac{\eta L\nu n^{3/2}}{18000}\right)
$$

and therefore

$$
\mathbb{E}\left[x_n\right] \ge \left(1-\frac{3}{4}\eta L n\right)x_0 + \frac{\eta^2 L\nu n^{3/2}}{18000}.
$$

If an epoch starts at (a fixed value) $x_0 < 0$, then from Lemma B.4 we have

$$
\mathbb{E}\left[x_n\right] \ge \left(1-\frac{160}{161}\eta L n\right)x_0 \ge \left(1-\frac{3}{4}\eta L n\right)x_0.
$$

From the second statement of Lemma B.4, we can observe that for all epochs we have $\mathbb{P}(x_0^k \ge 0) \ge \frac{1}{2}$ because we initialize at $x_0^1 \ge \frac{1}{27000}\cdot\frac{\nu}{\mu n^{1/2}K} \ge 0$. Therefore, taking expectations over x_0, we can conclude that each epoch must satisfy

$$
\begin{aligned}
\mathbb{E}\left[x_n\right] &= \mathbb{P}(x_0 \ge 0)\mathbb{E}[x_n | x_0 \ge 0] + \mathbb{P}(x_0 < 0)\mathbb{E}[x_n \mid x_0 < 0] \\
&\ge \mathbb{P}(x_0 \ge 0)\left(\left(1-\frac{3}{4}\eta L n\right)\mathbb{E}[x_0 \mid x_0 \ge 0] + \frac{\eta^2 L\nu n^{3/2}}{18000}\right) + \mathbb{P}(x_0 < 0)\left(\left(1-\frac{3}{4}\eta L n\right)\mathbb{E}[x_0 \mid x_0 < 0]\right)
\end{aligned}
$$

$$\geq \left(1 - \frac{3}{4}\eta L n\right) \mathbb{E}[x_0] + \frac{\eta^2 L \nu n^{3/2}}{36000}.$$

Since the above holds for all $\mu_0 \leq \frac{L}{2415}$, we may choose $\mu_0 = \frac{L}{2415}$, i.e., our function F_2 can be chosen as

$$F_2(x) = \left(L \mathbb{1}_{x<0} + \frac{L}{2415}\mathbb{1}_{x\geq 0}\right)\frac{x^2}{2}.$$

Note that since $\kappa \geq 2415$ is equivalent to $\frac{L}{2415} \geq \mu$, we have $F_2(x) \in \mathcal{F}(L, \frac{L}{2415}, 0, \nu) \subseteq \mathcal{F}(L, \mu, 0, \nu)$.

From here we focus on unrolling the per-epoch inequalities for all k, and hence we put the superscripts k back in our notation.

If the starting point of an epoch satisfies $\mathbb{E}[x_0^k] \geq \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$, then using $\eta \geq \frac{1}{\mu n K}$ we easily have

$$\begin{aligned}
\mathbb{E}\left[x_0^{k+1}\right] = \mathbb{E}\left[x_n^k\right] &\geq \left(1 - \frac{3}{4}\eta L n\right)\mathbb{E}[x_0^k] + \frac{\eta^2 L \nu n^{3/2}}{36000} \\
&\geq \left(1 - \frac{3}{4}\eta L n\right)\left(\frac{1}{27000}\cdot\frac{\nu}{\mu n^{1/2} K}\right) + \left(\frac{1}{\mu n K}\right)\frac{\eta L \nu n^{3/2}}{36000} \\
&\geq \frac{1}{27000}\cdot\frac{\nu}{\mu n^{1/2}K} - \frac{1}{36000}\cdot\frac{\eta L \nu n^{1/2}}{\mu K} + \frac{1}{36000}\cdot\frac{\eta L \nu n^{1/2}}{\mu K} = \frac{1}{27000}\cdot\frac{\nu}{\mu n^{1/2}K}. \qquad (9)
\end{aligned}$$

Therefore, if we set $x_0^1 \geq \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$, then the final iterate must also maintain $\mathbb{E}[x_n^K] \geq \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$. By Jensen's inequality, we can finally conclude that

$$\begin{aligned}
\mathbb{E}\left[F_2(x_n^K) - F_2^*\right] &= \mathbb{E}\left[F_2(x_n^K)\right] \\
&\geq \frac{L}{2\cdot 2415}\mathbb{E}\left[(x_n^K)^2\right] \\
&\geq \frac{L}{4830}\mathbb{E}\left[x_n^K\right]^2 \\
&\geq \frac{L}{4830}\cdot\left(\frac{1}{27000}\cdot\frac{\nu}{\mu n^{1/2}K}\right)^2 = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right).
\end{aligned}$$

□

B.3. Lower Bound for $\eta \geq \max\left\{\frac{1}{\mu n K}, \frac{1}{161 L n}\right\}$

Here we show that there exists $F_3(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with $x_0^1 = 0$ satisfies

$$\mathbb{E}\left[F_3(x_n^K) - F_3^*\right] = \Omega\left(\frac{\nu^2}{\mu n K}\right).$$

Proof. We define $F_3(x) \in \mathcal{F}(L, L, 0, \nu)$ by the following components:

$$f_i(x) = \begin{cases} \frac{Lx^2}{2} + \nu x & \text{if } i \leq n/2, \\ \frac{Lx^2}{2} - \nu x & \text{otherwise.} \end{cases}$$

With this construction, the finite-sum objective becomes

$$F_3(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{Lx^2}{2}.$$

Note that $\mathcal{F}(L, L, 0, \nu) \subseteq \mathcal{F}(L, \mu, 0, \nu)$ and $F_3^* = 0$ at $x^* = 0$ by definition.

First, we focus on a single epoch, and hence we write x_i instead of x_i^k, omitting the superscripts k for a while.

Similarly as in Appendix B.2, we can follow Definition B.1 to express the iterations for $i = 1, \dots, n$ via s as

$$x_i = x_{i-1} - \eta \nabla f_{\sigma(i)}(x) = x_{i-1} - \eta L x_{i-1} - \eta \nu s_i = (1 - \eta L) x_{i-1} - \eta \nu s_i.$$

Also, we can sum up the iterates to obtain

$$\begin{aligned} x_n &= (1 - \eta L) x_{n-1} - \eta \nu s_n \\ &= (1 - \eta L) \left((1 - \eta L) x_{n-2} - \eta \nu s_{n-1} \right) - \eta \nu s_n \\ &\vdots \\ &= (1 - \eta L)^n x_0 - \eta \nu \sum_{i=1}^{n} (1 - \eta L)^{n-i} s_i. \end{aligned}$$

Now we can square both terms and take expectations over $s \in \mathcal{S}_n$ to obtain

$$\begin{aligned} \mathbb{E}[x_n^2] &= \mathbb{E}\left[\left((1 - \eta L)^n x_0 - \eta \nu \sum_{i=1}^{n} (1 - \eta L)^{n-i} s_i \right)^2 \right] \\ &= (1 - \eta L)^{2n} x_0^2 - 2 (1 - \eta L)^n x_0 \cdot \eta \nu \mathbb{E}\left[\sum_{i=1}^{n} (1 - \eta L)^{n-i} s_i \right] + \eta^2 \nu^2 \mathbb{E}\left[\left(\sum_{i=1}^{n} (1 - \eta L)^{n-i} s_i \right)^2 \right] \\ &= (1 - \eta L)^{2n} x_0^2 + \eta^2 \nu^2 \mathbb{E}\left[\left(\sum_{i=1}^{n} (1 - \eta L)^{n-i} s_i \right)^2 \right], \end{aligned}$$

where the middle term is eliminated since $\mathbb{E}[s_i] = 0$ for all i. By Lemma 1 of Safran & Shamir (2020), we can bound

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} (1 - \eta L)^{n-i} s_i \right)^2 \right] \geq c \cdot \min \left\{ 1 + \frac{1}{\eta L}, \eta^2 L^2 n^3 \right\}$$

for some universal constant $c > 0$. Since $\eta \geq \frac{1}{161 L n}$, we can further lower bound the RHS by $\frac{c'}{\eta L}$ for some universal constant $c' > 0$. Then we have

$$\mathbb{E}[x_n^2] = (1 - \eta L)^{2n} x_0^2 + \eta^2 \nu^2 \mathbb{E}\left[\left(\sum_{i=1}^{n} (1 - \eta L)^{n-i} s_i \right)^2 \right] \geq (1 - \eta L)^{2n} x_0^2 + c' \frac{\eta \nu^2}{L}.$$

From here we focus on unrolling the per-epoch inequalities for all k, and hence we put the k's back in our notations.

Unrolling the inequalities, we obtain

$$\begin{aligned} \mathbb{E}[(x_n^K)^2] &\geq (1 - \eta L)^{2n} \mathbb{E}[(x_n^{K-1})^2] + c' \frac{\eta \nu^2}{L} \\ &\vdots \\ &\geq (1 - \eta L)^{2nK} (x_0^1)^2 + c' \frac{\eta \nu^2}{L} \sum_{k=0}^{K-1} (1 - \eta L)^{2nk} \geq c' \frac{\eta \nu^2}{L}, \end{aligned}$$

where we used $x_0^1 = 0$. Finally, from $\eta \geq \frac{1}{\mu n K}$ we can conclude that

$$\mathbb{E}[F_3(x_n^K) - F_3^*] = \mathbb{E}[F_3(x_n^K)] = \frac{L}{2} \mathbb{E}[(x_n^K)^2] \geq \frac{c'}{2} \eta \nu^2 \geq \frac{c'}{2} \frac{\nu^2}{\mu n K}.$$

□

B.4. Lemmas used in Theorem 3.1

In this subsection, we will prove the lemmas used in Theorem 3.1.

Lemma B.2. *For (fixed)* $x_0 \geq 0$*,* $0 \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor$*,* $\eta \leq \frac{1}{161Ln}$*, and* $\frac{L}{\mu_0} \geq 2415$*,*

$$\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i\right] \leq \frac{2}{3}Lx_0 - \frac{\eta L\nu}{480}\sqrt{i}.$$

Proof. For $i = 0$, the statement is trivial since $x_0 \geq 0$ and $\frac{L}{\mu_0} \geq 2415$ implies

$$\mathbb{E}\left[(L\mathbb{1}_{x_0<0} + \mu_0\mathbb{1}_{x_0\geq 0})x_0\right] = \mu_0 x_0 \leq \frac{1}{2415}Lx_0 \leq \frac{2}{3}Lx_0.$$

Hence we may assume that $1 \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor$.

Given $s = \{s_i\}_{i=1}^n \in \mathcal{S}_n$ (as in Definition B.1), let us denote the partial sums as $\mathcal{E}_i \triangleq \sum_{j=1}^i s_j$. We will use conditional expectations under $\mathcal{E}_i > 0$ and $\mathcal{E}_i \leq 0$, and then aggregate the results to obtain the final inequality.

First observe that

$$\begin{aligned}
\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i \middle| \mathcal{E}_i > 0\right] &\leq L\mathbb{E}\left[x_i \middle| \mathcal{E}_i > 0\right], \\
\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i \middle| \mathcal{E}_i \leq 0\right] &\leq \mu_0\mathbb{E}\left[x_i \middle| \mathcal{E}_i \leq 0\right],
\end{aligned}$$

since $(L\mathbb{1}_{x<0} + \mu_0\mathbb{1}_{x\geq 0}) \leq Lx$ and $(L\mathbb{1}_{x<0} + \mu_0\mathbb{1}_{x\geq 0}) \leq \mu_0 x$ for all $x \in \mathbb{R}$. By the law of total expectations, we have

$$\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i\right] \leq L\mathbb{P}(\mathcal{E}_i > 0)\mathbb{E}\left[x_i \middle| \mathcal{E}_i > 0\right] + \mu_0\mathbb{P}(\mathcal{E}_i \leq 0)\mathbb{E}\left[x_i \middle| \mathcal{E}_i \leq 0\right]. \tag{10}$$

First, we bound $\mathbb{E}\left[x_i \middle| \mathcal{E}_i > 0\right]$ for the former term. We can show that

$$\begin{aligned}
\mathbb{E}\left[x_i \middle| \mathcal{E}_i > 0\right] &= \mathbb{E}\left[x_0 - \eta\cdot\sum_{j=0}^{i-1}\left(\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\geq 0}\right)x_j + \nu s_{j+1}\right) \middle| \mathcal{E}_i > 0\right] \\
&= \mathbb{E}\left[x_0 - \eta\cdot\sum_{j=0}^{i-1}\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\geq 0}\right)\left(x_0 + (x_j - x_0)\right) - \eta\nu\mathcal{E}_i \middle| \mathcal{E}_i > 0\right] \\
&= x_0\mathbb{E}\left[1 - \eta\cdot\sum_{j=0}^{i-1}\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\geq 0}\right) \middle| \mathcal{E}_i > 0\right] \\
&\quad - \eta\mathbb{E}\left[\sum_{j=0}^{i-1}\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\geq 0}\right)(x_j - x_0) \middle| \mathcal{E}_i > 0\right] - \eta\nu\mathbb{E}\left[\mathcal{E}_i \middle| \mathcal{E}_i > 0\right] \\
&\leq x_0\mathbb{E}\left[1 - \eta\cdot\sum_{j=0}^{i-1}\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\geq 0}\right) \middle| \mathcal{E}_i > 0\right] \\
&\quad + \eta L\sum_{j=0}^{i-1}\mathbb{E}\left[|x_j - x_0| \middle| \mathcal{E}_i > 0\right] - \eta\nu\mathbb{E}\left[\mathcal{E}_i \middle| \mathcal{E}_i > 0\right].
\end{aligned} \tag{11}$$

Now we use the following lemmas.

Lemma B.5. *If* $n \geq 2$ *is an even number and* $0 \leq i \leq \frac{n}{2}$*, then* $\frac{\sqrt{i}}{10} \leq \mathbb{E}\left[|\mathcal{E}_i|\right] \leq \sqrt{i}$*.*

Lemma B.6 (Yun et al. (2022), Lemma 14)**.** *For all* $0 \leq i \leq n$*, we have* $\mathbb{P}(\mathcal{E}_i > 0) = \mathbb{P}(\mathcal{E}_i < 0) \geq \frac{1}{6}$*.*

We will prove Lemma B.5 later on.

Observe that the probability distribution of each $\mathcal{E}_i$ is symmetric by the definition of $\mathcal{S}_n$. Therefore we have

$$\begin{aligned}\mathbb{E}\left[|\mathcal{E}_i|\right] &= P(\mathcal{E}_i > 0)\mathbb{E}\left[|\mathcal{E}_i|\,|\mathcal{E}_i > 0\right] + P(\mathcal{E}_i = 0)\mathbb{E}\left[|\mathcal{E}_i|\,|\mathcal{E}_i = 0\right] + P(\mathcal{E}_i < 0)\mathbb{E}\left[|\mathcal{E}_i|\,|\mathcal{E}_i < 0\right]\\ &= P(\mathcal{E}_i > 0)\mathbb{E}\left[\mathcal{E}_i|\mathcal{E}_i > 0\right] + P(\mathcal{E}_i < 0)\mathbb{E}\left[-\mathcal{E}_i|\mathcal{E}_i < 0\right]\\ &= 2P(\mathcal{E}_i > 0)\mathbb{E}\left[\mathcal{E}_i|\mathcal{E}_i > 0\right].\end{aligned}$$

Using Lemmas B.5 and B.6, we can obtain

$$\frac{\sqrt{i}}{20} \le \frac{\mathbb{E}\left[|\mathcal{E}_i|\right]}{2} \le \mathbb{E}\left[\mathcal{E}_i|\mathcal{E}_i > 0\right] = \frac{\mathbb{E}\left[|\mathcal{E}_i|\right]}{2P(\mathcal{E}_i > 0)} \le 3\mathbb{E}\left[|\mathcal{E}_i|\right] \le 3\sqrt{i}. \tag{12}$$

We also use the following lemma, which we prove later on. This is a simple application of Lemmas B.5 and B.6.

Lemma B.7. *Suppose that $x_0 \ge 0$, $0 \le i \le n$, and $\eta \le \frac{1}{161Ln}$. Then we have*

$$\mathbb{E}\left[|x_i - x_0|\right] \le \frac{161}{160}\left(\eta L i x_0 + \eta\nu\sqrt{i}\right).$$

Now we bound the three terms of Equation (11) one by one. The first term can be bounded simply as

$$x_0\mathbb{E}\left[1 - \eta\cdot\sum_{j=0}^{i-1}\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\ge 0}\right)\middle|\,\mathcal{E}_i > 0\right] \le (1-\eta\mu_0 i)x_0. \tag{13}$$

For the second term of Equation (11), we use Lemma B.6 to obtain

$$\mathbb{E}\left[|x_i - x_0|\,|\,\mathcal{E}_i > 0\right] \le \frac{\mathbb{E}\left[|x_i - x_0|\right]}{\mathbb{P}(\mathcal{E}_i > 0)} \le 6\mathbb{E}\left[|x_i - x_0|\right],$$

and then use Lemma B.7 to obtain

$$\begin{aligned}\eta L\sum_{j=0}^{i-1}\mathbb{E}\left[|x_j - x_0|\,|\,\mathcal{E}_i > 0\right] &\le 6\eta L\sum_{j=0}^{i-1}\mathbb{E}\left[|x_j - x_0|\right]\\ &\le 6\eta L\cdot\frac{161}{160}\sum_{j=0}^{i-1}\left(\eta L j x_0 + \eta\nu\sqrt{j}\right)\\ &= \frac{483}{80}\left(\eta^2L^2x_0\sum_{j=0}^{i-1}j + \eta^2L\nu\sum_{j=0}^{i-1}\sqrt{j}\right)\\ &\le \frac{483}{80}\left(\eta^2L^2x_0\cdot\frac{1}{2}i^2 + \eta^2L\nu\cdot\frac{2}{3}i^{3/2}\right)\\ &\le \frac{483}{160}\eta^2L^2i^2x_0 + \frac{161}{40}\eta^2L\nu i^{3/2}.\end{aligned} \tag{14}$$

The last term of Equation (11) can be bounded using Equation (12) as

$$-\eta\nu\mathbb{E}\left[\mathcal{E}_i|\,\mathcal{E}_i > 0\right] \le -\eta\nu\frac{\sqrt{i}}{20}. \tag{15}$$

From Equations (13)-(15), we have

$$\begin{aligned}\mathbb{E}\left[x_i|\,\mathcal{E}_i > 0\right] &\le (1-\eta\mu_0 i)x_0 + \frac{483}{160}\eta^2L^2i^2x_0 + \frac{161}{40}\eta^2L\nu i^{3/2} - \eta\nu\frac{\sqrt{i}}{20}\\ &= \left(1 - \eta\mu_0 i + \frac{483}{160}\eta^2L^2i^2\right)x_0 - \left(\frac{1}{20} - \frac{161}{40}\eta L i\right)\eta\nu\sqrt{i}\end{aligned}$$

$$\leq \left(1 + \frac{3}{160 \cdot 161}\right) x_0 - \frac{\eta\nu\sqrt{i}}{40}, \tag{16}$$

where the last inequality comes from $\eta Li \leq \eta Ln \leq \frac{1}{161}$.

Next, we bound $\mathbb{E}\left[x_i \middle| \mathcal{E}_i \leq 0\right]$ for the former term. We can show that

$$\begin{aligned}
\mathbb{E}\left[x_i \middle| \mathcal{E}_i \leq 0\right] &\leq x_0 + \mathbb{E}\left[|x_i - x_0| \middle| \mathcal{E}_i \leq 0\right] \\
&\leq x_0 + \frac{\mathbb{E}\left[|x_i - x_0|\right]}{\mathbb{P}(\mathcal{E}_i \leq 0)} \\
&\leq x_0 + 6\mathbb{E}\left[|x_i - x_0|\right] && (\because \text{Lemma B.6}) \\
&\leq x_0 + 6 \cdot \frac{161}{160}\left(\eta Lix_0 + \eta\nu\sqrt{i}\right) && (\because \text{Lemma B.7}) \\
&= \left(1 + \frac{483}{80}\eta Li\right) x_0 + \frac{483}{80}\eta\nu\sqrt{i} \\
&= \left(1 + \frac{3}{80}\right) x_0 + \frac{483}{80}\eta\nu\sqrt{i}, && (17)
\end{aligned}$$

where the last inequality comes from $\eta Li \leq \eta Ln \leq \frac{1}{161}$.

Plugging in Equations (16) and (17) in (10), we have

$$\begin{aligned}
\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i\right] &\leq L\mathbb{P}(\mathcal{E}_i > 0)\mathbb{E}\left[x_i \middle| \mathcal{E}_i > 0\right] + \mu_0\mathbb{P}(\mathcal{E}_i \leq 0)\mathbb{E}\left[x_i \middle| \mathcal{E}_i \leq 0\right] \\
&\leq L\mathbb{P}(\mathcal{E}_i > 0)\left(\left(1 + \frac{3}{160 \cdot 161}\right) x_0 - \frac{\eta\nu\sqrt{i}}{40}\right) \\
&\quad + \mu_0\mathbb{P}(\mathcal{E}_i \leq 0)\left(\left(1 + \frac{3}{80}\right) x_0 + \frac{483}{80}\eta\nu\sqrt{i}\right) \\
&= \left(L\mathbb{P}(\mathcal{E}_i > 0)\left(1 + \frac{3}{160 \cdot 161}\right) + \mu_0\mathbb{P}(\mathcal{E}_i \leq 0)\left(1 + \frac{3}{80}\right)\right) x_0 \\
&\quad - \left(L\mathbb{P}(\mathcal{E}_i > 0) \cdot \frac{1}{40} - \mu_0\mathbb{P}(\mathcal{E}_i \leq 0) \cdot \frac{483}{80}\right)\eta\nu\sqrt{i}.
\end{aligned}$$

Since $\mathbb{P}(\mathcal{E}_i > 0) = \frac{1-\mathbb{P}(\mathcal{E}_i=0)}{2} \leq \frac{1}{2}$ by symmetry and $\mathbb{P}(\mathcal{E}_i \leq 0) = 1 - \mathbb{P}(\mathcal{E}_i > 0) \leq \frac{5}{6}$ by Lemma B.6, we have

$$L\mathbb{P}(\mathcal{E}_i > 0)\left(1 + \frac{3}{160 \cdot 161}\right) + \mu_0\mathbb{P}(\mathcal{E}_i \leq 0)\left(1 + \frac{3}{80}\right) \leq \left(\frac{1}{2}\left(1 + \frac{3}{160 \cdot 161}\right) + \frac{5}{6} \cdot \frac{1}{2415} \cdot \frac{83}{80}\right) L \leq \frac{2}{3}L,$$

where we use $\eta Li \leq \frac{1}{161}$, $\frac{L}{\mu} \geq 2415$, and $\frac{1}{2}\left(1 + \frac{3}{160 \cdot 161}\right) + \frac{5}{6} \cdot \frac{1}{2415} \cdot \frac{83}{80} \leq \frac{2}{3}$. Also, by Lemma B.6 we have

$$L\mathbb{P}(\mathcal{E}_i > 0) \cdot \frac{1}{40} - \mu_0\mathbb{P}(\mathcal{E}_i \leq 0) \cdot \frac{483}{80} \geq \left(\frac{1}{6} \cdot \frac{1}{40} - \frac{5}{6} \cdot \frac{1}{2415} \cdot \frac{483}{80}\right) L = \frac{1}{480}L,$$

where we use $\eta Li \leq \frac{1}{161}$ and $\frac{L}{\mu_0} \geq 2415$. Therefore we have

$$\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i\right] \leq \frac{2}{3}Lx_0 - \frac{\eta L\nu}{480}\sqrt{i}.$$

□

Lemma B.3. *For (fixed)* $x_0 \geq 0$, $0 \leq i \leq n-1$, *and* $\eta \leq \frac{1}{161Ln}$,

$$\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i\right] \leq \left(1 + \frac{161}{160}i\eta L\right)\mu_0 x_0 + \frac{161}{160}\eta\mu_0\nu\sqrt{i}.$$

Proof. Since $\eta \leq \frac{1}{161Ln}$, we can easily prove using Lemma B.7 as follows.

$$\begin{aligned}
\mathbb{E}\left[(L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i\right] &\leq \mu\mathbb{E}\left[\mu_0 x_i\right] \\
&\leq \mu_0 x_0 + \mu_0\mathbb{E}\left[|x_i - x_0|\right] \\
&\leq \mu_0 x_0 + \mu_0\frac{161}{160}\left(\eta L i x_0 + \eta\nu\sqrt{i}\right) \\
&\leq \left(1 + \frac{161}{160}i\eta L\right)\mu_0 x_0 + \frac{161}{160}\eta\mu_0\nu\sqrt{i}.
\end{aligned}$$

□

Lemma B.4. *If $\eta \leq \frac{1}{161Ln}$, we have the followings.*

1. *For (fixed) $x_0 < 0$, we have*

$$\mathbb{E}\left[x_n\right] \geq \left(1 - \frac{160}{161}\eta L n\right)x_0.$$

2. *If we initialize at $x_0^1 \geq 0$, then we always have $\mathbb{P}(x_n^k \geq 0) \geq \frac{1}{2}$ for future start-of-epoch iterates.*

Proof. We divide the proof into three parts. In the first part, we compare with the case of using a quadratic function instead, sharing the same permutation. In the second part, we assume $x_0 < 0$ and use the first part to prove the first result of the statement. In the third part, we assume $x_0 \geq 0$ and use the first part to prove the second result of the statement. Note that the statement in the first part holds for both $x_0 \geq 0$ or $x_0 < 0$.

Part 1. For comparison, we define and use the same function used in Appendix B.3:

$$h_i(x) = \begin{cases} \frac{Lx^2}{2} + \nu x & \text{if } i \leq n/2, \\ \frac{Lx^2}{2} - \nu x & \text{otherwise} \end{cases}$$

such that the finite-sum objective becomes

$$H(x) = \frac{1}{n}\sum_{i=1}^{n} h_i(x) = \frac{Lx^2}{2}.$$

Now let us think of SGD-RR run on the two functions $F_2(x)$ and $H(x)$, where both of the algorithms start from the same point x_0 and both share the same random permutation for all epochs. Let $x_{i,F}$ and $x_{i,H}$ be the output of the i-th iterate for SGD-RR on $F_2(x)$ and $H(x)$, respectively. Now we use mathematical induction on i to prove that $x_{i,F} \geq x_{i,H}$.

Base case. For $i = 0$, we have $x_{0,F} = x_{0,H} = x_0$.

Inductive Case. Let us assume that the induction hypothesis $x_{i,F} \geq x_{i,H}$ is true, and show that $x_{i+1,F} \geq x_{i+1,H}$ by considering the following three cases. Note that $f_i(x)$'s are the components of $F_2(x)$, s_i's are defined as in Definition B.1, and $\eta \leq \frac{1}{161Ln}$ implies $1 - \eta\mu \geq 1 - \eta L \geq 1 - \frac{1}{161n} \geq 0$.

- If $x_{i,F} \geq x_{i,H} \geq 0$, then we have

$$\begin{aligned}
x_{i+1,F} - x_{i+1,H} &= x_{i,F} - x_{i,H} - \eta\left(\nabla f_i(x_{i,F}) - \nabla h_i(x_{i,H})\right) \\
&= x_{i,F} - x_{i,H} - \eta\left(\mu x_{i,F} + \nu s_i - L x_{i,H} - \nu s_i\right) \\
&= (1 - \eta\mu)x_{i,F} - (1 - \eta L)x_{i,H} \geq 0,
\end{aligned}$$

  since $x_{i,F} \geq x_{i,H} \geq 0$ and $1 - \eta\mu \geq 1 - \eta L \geq 0$.

- If $x_{i,F} \geq 0 \geq x_{i,H}$, then we have

$$\begin{aligned} x_{i+1,F} - x_{i+1,H} &= x_{i,F} - x_{i,H} - \eta \left(\nabla f_i(x_{i,F}) - \nabla h_i(x_{i,H})\right) \\ &= x_{i,F} - x_{i,H} - \eta \left(\mu x_{i,F} + \nu s_i - L x_{i,H} - \nu s_i\right) \\ &= (1-\eta\mu) x_{i,F} - (1-\eta L) x_{i,H} \geq 0, \end{aligned}$$

  since $(1-\eta\mu)x_{i,F} \geq 0$ and $(1-\eta L)x_{i,H} \leq 0$.

- If $0 \geq x_{i,F} \geq x_{i,H}$, then we have

$$\begin{aligned} x_{i+1,F} - x_{i+1,H} &= x_{i,F} - x_{i,H} - \eta \left(\nabla f_i(x_{i,F}) - \nabla h_i(x_{i,H})\right) \\ &= x_{i,F} - x_{i,H} - \eta \left(L x_{i,F} + \nu s_i - L x_{i,H} - \nu s_i\right) \\ &= (1-\eta L) x_{i,F} - (1-\eta L) x_{i,H} \geq 0. \end{aligned}$$

Hence by induction, we have $x_{i+1,F} \geq x_{i+1,H}$ for all i.

From the above, we can observe that $\mathbb{E}[x_{n,F}] \geq \mathbb{E}[x_{n,H}] = (1-\eta L)^n x_0$.

Part 2. For the next step, let us assume $x_0 < 0$. Let us define

$$\varphi(z) = 1 - \frac{160}{161} nz - (1-z)^n.$$

Then for $z \in [0, 1 - (\frac{160}{161})^{\frac{1}{n-1}}]$, we have $\varphi'(z) = n((1-z)^{n-1} - \frac{160}{161}) \geq 0$ and hence $\varphi(z) \geq \varphi(0) = 0$.

Also, we can observe that for $n \geq 2$:

$$\left(1 - \frac{1}{161(n-1)}\right)^{n-1} \geq 1 - \frac{1}{161} \;\Rightarrow\; 1 - \left(\frac{160}{161}\right)^{\frac{1}{n-1}} \geq \frac{1}{161(n-1)} \geq \frac{1}{161n},$$

which implies that $\eta L \leq \frac{1}{161n} \leq 1 - \left(\frac{160}{161}\right)^{\frac{1}{n-1}}$. Hence we have $\varphi(\eta L) \geq 0$, or

$$(1-\eta L)^n \leq 1 - \frac{160}{161}\eta L n,$$

and for $x_0 < 0$ we have

$$\mathbb{E}[x_{n,H}] = (1-\eta L)^n x_0 \geq \left(1 - \frac{160}{161}\eta L n\right) x_0.$$

Applying ***Part 1***, we can conclude that $\mathbb{E}[x_{n,F}] \geq \mathbb{E}[x_{n,H}] \geq \left(1 - \frac{160}{161}\eta L n\right) x_0$.

Part 3. Now suppose that we initialize $x_0 \geq 0$. For $H(x)$ and some given permutation $s \in \mathcal{S}_n$, we have

$$x_{n,H} = (1-\eta L)^n x_0 - \eta\nu \sum_{i=1}^{n} (1-\eta L)^{n-i} s_i.$$

Now let us think of pairs of permutations $s, s' \in \mathcal{S}_n$ which satisfy $s_i = -s'_i$ for all i. By definition, the set $\mathcal{S}_n$ can be exactly partitioned into $\frac{1}{2}\binom{n}{n/2}$ disjoint pairs. Let us temporarily denote the final iterates obtained by choosing the permutations s and s' by $x^{s}_{n,H}$ and $x^{s'}_{n,H}$, respectively. Then we can observe that

$$\frac{1}{2}(x^{s}_{n,H} + x^{s'}_{n,H}) = (1-\eta L)^n x_0 - \eta\nu \sum_{i=1}^{n} (1-\eta L)^{n-i} \cdot \left(\frac{s_i + s'_i}{2}\right) = (1-\eta L)^n x_0,$$

which means that each pair of outcomes will be symmetric with respect to $(1-\eta L)^n x_0$. Hence the whole probability distribution of $(1-\eta L)^{-n} x_{n,H}$ will stay symmetric with respect to the initial point x_0.

Considering outputs after multiple epochs, we can sequentially apply the same logic to prove that the distribution of $(1-\eta L)^{-nk} x_{n,H}^k$ will always stay symmetric with respect to x_0^1 for all k. In other words, for each k, the distribution of outputs $x_{n,H}^k$ *conditioned only on the first epoch* x_0^1 will be symmetric with respect to $(1-\eta L)^{nk} x_0 \geq 0$. This automatically implies that we must have $\mathbb{P}(x_{n,H}^k \geq 0) \geq \mathbb{P}(x_{n,H}^k \geq (1-\eta L)^{nk} x_0) \geq \frac{1}{2}$ for any starting point $x_0^1 \geq 0$. Finally, since ***Part 1*** ensures $x_{n,F}^k \geq x_{n,H}^k$, we can conclude that $\mathbb{P}(x_{n,F}^k \geq 0) \geq \mathbb{P}(x_{n,H}^k \geq 0) \geq \frac{1}{2}$. □

Lemma B.7. *Suppose that* $x_0 \geq 0$, $0 \leq i \leq n$, *and* $\eta \leq \frac{1}{161Ln}$. *Then we have*

$$\mathbb{E}\left[|x_i - x_0|\right] \leq \frac{161}{160}\left(\eta L i x_0 + \eta\nu\sqrt{i}\right).$$

Proof. From $x_{i+1} = x_i - \eta\left((L\mathbb{1}_{x_i<0} + \mu_0\mathbb{1}_{x_i\geq 0})x_i + \nu s_{i+1}\right)$, we have for all $i = 1, \dots, n$:

$$\begin{aligned}
\mathbb{E}\left[|x_i - x_0|\right] &= \mathbb{E}\left[\left|-\eta \cdot \sum_{j=0}^{i-1}\left(\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\geq 0}\right)x_j + \nu s_{i+1}\right)\right|\right] \\
&\leq \eta\sum_{j=0}^{i-1}\mathbb{E}\left[\left|\left(L\mathbb{1}_{x_j<0} + \mu_0\mathbb{1}_{x_j\geq 0}\right)x_j\right|\right] + \eta\nu\mathbb{E}\left[\left|\sum_{j=1}^{i} s_j\right|\right] \\
&\leq \eta L\sum_{j=0}^{i-1}\mathbb{E}\left[|x_j|\right] + \eta\nu\mathbb{E}\left[|\mathcal{E}_i|\right] \\
&\leq \eta L i x_0 + \eta L\sum_{j=0}^{i-1}\mathbb{E}\left[|x_j - x_0|\right] + \eta\nu\sqrt{i}. && (\because \text{Lemma B.5})
\end{aligned}$$

Now let us think of a sequence $h(i)$ defined by $h(0) = 0$ and recursively as

$$h(i) = \eta L i x_0 + \eta L\sum_{j=0}^{i-1} h(j) + \eta\nu\sqrt{i}, \quad \text{for } i = 1, \dots, n.$$

Then obviously $h(i)$ monotonically increases since $h(i) - h(i-1) = \eta L x_0 + \eta L h(i-1) + \eta\nu(\sqrt{i} - \sqrt{i-1}) > 0$. We can plug in $h(j) \leq h(i)$ for all $j = 0, \dots, i-1$ to obtain $h(i) \leq \eta L i x_0 + \eta L i h(i) + \eta\nu\sqrt{i}$, and hence

$$h(i) \leq \frac{\eta L i x_0 + \eta\nu\sqrt{i}}{1 - \eta L i}.$$

Also, by induction, we have $\mathbb{E}\left[|x_i - x_0|\right] \leq h(i)$, since the sequence $\mathbb{E}\left[|x_i - x_0|\right]$ satisfies a recurrence of the same form but with an inequality. Hence, from $\eta L i \leq \eta L n \leq \frac{1}{161}$ we get

$$\mathbb{E}\left[|x_i - x_0|\right] \leq \frac{\eta L i x_0 + \eta\nu\sqrt{i}}{1 - \eta L i} \leq \frac{1}{1 - \eta L n}\left(\eta L i x_0 + \eta\nu\sqrt{i}\right) \leq \frac{161}{160}\left(\eta L i x_0 + \eta\nu\sqrt{i}\right).$$

□

Lemma B.5. *If* $n \geq 2$ *is an even number and* $0 \leq i \leq \frac{n}{2}$, *then* $\frac{\sqrt{i}}{10} \leq \mathbb{E}\left[|\mathcal{E}_i|\right] \leq \sqrt{i}$.

Proof. We assume $i \geq 1$ since the statement is vacuously true for $i = 0$.

For the upper bound, we use $\mathbb{E}\left[|\mathcal{E}_i|\right] \leq \sqrt{i}$ as in Lemma 12 of Rajput et al. (2020).

For the lower bound, we start from the following equation in Lemma 12 of Rajput et al. (2020):

$$\mathbb{E}\left[|\mathcal{E}_{i+1}|\right] = \left(1 - \frac{1}{n-i}\right)\mathbb{E}\left[|\mathcal{E}_i|\right] + \mathbb{P}(\mathcal{E}_i = 0).$$

We can explicitly compute for $i = 1, \dots, \frac{n}{2}$:

$$\mathbb{P}(\mathcal{E}_i = 0) = \mathbb{1}_{\{i \text{ is even}\}} \cdot \frac{\binom{i}{\frac{i}{2}}\binom{n-i}{\frac{n-i}{2}}}{\binom{n}{\frac{n}{2}}},$$

where $\mathcal{E}_i = 0$ has nonzero probability if and only if i is even. We also use the following lemma.

Lemma B.8. *For even, positive integers n, i with $n \geq 4$ and $2 \leq i \leq \lfloor \frac{n}{2} \rfloor$, we have*

$$\frac{\binom{i}{\frac{i}{2}}\binom{n-i}{\frac{n-i}{2}}}{\binom{n}{\frac{n}{2}}} \geq \frac{2}{5\sqrt{i}}.$$

This lemma yields $\mathbb{P}(\mathcal{E}_i = 0) \geq \frac{2}{5\sqrt{i}}$ for even i. We prove Lemma B.8 at the very end of Appendix B.4.

First, suppose that $i \geq 2$ is an *even* integer. Then since $i \leq \frac{n}{2}$, we have for $i \geq 4$:

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{E}_i|\right] &= \left(1 - \frac{1}{n-i+1}\right)\mathbb{E}\left[|\mathcal{E}_{i-1}|\right] + \mathbb{P}(\mathcal{E}_{i-1} = 0) \\
&\geq \left(1 - \frac{2}{n}\right)\mathbb{E}\left[|\mathcal{E}_{i-1}|\right] + \mathbb{1}_{i-1 \text{ is even}}\frac{2}{5\sqrt{i-1}} \\
&= \left(1 - \frac{2}{n}\right)\mathbb{E}\left[|\mathcal{E}_{i-1}|\right] \\
&\geq \left(1 - \frac{2}{n}\right)\left(\left(1 - \frac{2}{n}\right)\mathbb{E}\left[|\mathcal{E}_{i-2}|\right] + \mathbb{1}_{i-2 \text{ is even}}\frac{2}{5\sqrt{i-2}}\right) \\
&= \left(1 - \frac{2}{n}\right)^2\mathbb{E}\left[|\mathcal{E}_{i-2}|\right] + \left(1 - \frac{2}{n}\right)\frac{2}{5\sqrt{i-2}}. \qquad (18)
\end{aligned}$$

We can also explicitly compute the base case $i = 2$ as

$$\mathbb{E}\left[|\mathcal{E}_2|\right] = 2 \cdot \frac{2 \cdot \binom{n-2}{\frac{n}{2}}}{\binom{n}{\frac{n}{2}}} = \frac{4 \cdot (n-2)!(\frac{n}{2})!(\frac{n}{2})!}{(\frac{n}{2})!(\frac{n}{2}-2)!n!} = \frac{4(\frac{n}{2})(\frac{n}{2}-1)}{n(n-1)} = \frac{n-2}{n-1} = 1 - \frac{1}{n-1} \geq 1 - \frac{2}{n}, \qquad (19)$$

from the fact that $\mathcal{E}_2 = \pm 2$ each occurs $\binom{n-2}{\frac{n}{2}}$ times among a total of $\binom{n}{\frac{n}{2}}$ cases, and $\mathcal{E}_2 = 0$ otherwise. Also, note that we automatically have $\mathbb{E}\left[|\mathcal{E}_2|\right] \geq 1 - \frac{2}{n} \geq \frac{\sqrt{2}}{10}$, which proves the given statement for $i = 2$.

Now, unrolling the inequalities in (18), we have for $i \geq 4$:

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{E}_i|\right] &\geq \left(1 - \frac{2}{n}\right)^2\mathbb{E}\left[|\mathcal{E}_{i-2}|\right] + \left(1 - \frac{2}{n}\right)\frac{2}{5\sqrt{i-2}} \\
&\geq \left(1 - \frac{2}{n}\right)^2\left(\left(1 - \frac{2}{n}\right)^2\mathbb{E}\left[|\mathcal{E}_{i-4}|\right] + \left(1 - \frac{2}{n}\right)\frac{2}{5\sqrt{i-4}}\right) + \left(1 - \frac{2}{n}\right)\frac{2}{5\sqrt{i-2}} \\
&= \left(1 - \frac{2}{n}\right)^4\mathbb{E}\left[|\mathcal{E}_{i-4}|\right] + \left(1 - \frac{2}{n}\right)^3\frac{2}{5\sqrt{i-4}} + \left(1 - \frac{2}{n}\right)\frac{2}{5\sqrt{i-2}} \\
&\vdots \\
&\geq \left(1 - \frac{2}{n}\right)^{i-2}\mathbb{E}\left[|\mathcal{E}_2|\right] + \left(1 - \frac{2}{n}\right)\left(\sum_{p=0}^{\frac{i}{2}-2}\left(1 - \frac{2}{n}\right)^{2p}\frac{2}{5\sqrt{i-2-2p}}\right) \\
&\geq \left(1 - \frac{2}{n}\right)^{i-1} + \left(1 - \frac{2}{n}\right)\left(\sum_{p=0}^{\frac{i}{2}-2}\left(1 - \frac{2}{n}\right)^{2p}\frac{2}{5\sqrt{i-2-2p}}\right) \qquad (\because \text{Equation (19)})
\end{aligned}$$

$$\geq \left(1-\frac{2}{n}\right)^{i-1} + \left(1-\frac{2}{n}\right)\left(\sum_{p=0}^{\frac{i}{2}-2}\left(1-\frac{2}{n}\right)^{2p}\right)\frac{2}{5\sqrt{i-2}}$$

$$\geq \left(1-\frac{2}{n}\right)\left(\sum_{p=0}^{\frac{i}{2}-1}\left(1-\frac{2}{n}\right)^{2p}\right)\frac{2}{5\sqrt{i-2}}$$

$$= \left(1-\frac{2}{n}\right)\cdot\frac{1-\left(1-\frac{2}{n}\right)^{i}}{1-\left(1-\frac{2}{n}\right)^{2}}\cdot\frac{2}{5\sqrt{i-2}}$$

$$= \left(1-\frac{2}{n}\right)\cdot\frac{1}{\frac{4}{n}-\frac{4}{n^2}}\cdot\left(1-\left(1-\frac{2}{n}\right)^{i}\right)\cdot\frac{2}{5\sqrt{i-2}}$$

$$\geq \left(1-\frac{2}{n}\right)\cdot\frac{1}{\frac{4}{n}-\frac{4}{n^2}}\cdot\left(1-\frac{1}{1+\frac{2i}{n}}\right)\cdot\frac{2}{5\sqrt{i-2}} \tag{20}$$

$$= \frac{n(n-2)}{4(n-1)}\cdot\frac{2i}{n+2i}\cdot\frac{2}{5\sqrt{i-2}}$$

$$= \left(\frac{n-2}{n-1}\cdot\frac{\sqrt{i}}{\sqrt{i-2}}\right)\left(\frac{n}{n+2i}\cdot\frac{\sqrt{i}}{5}\right) \geq \frac{n}{n+2i}\cdot\frac{\sqrt{i}}{5} \geq \frac{\sqrt{i}}{10}. \tag{21}$$

In (20) we use $(1-x)^r \leq \frac{1}{1+rx}$ for all $0 \leq x \leq 1$ and $r \geq 0$. In (21) we use the fact that $2i \leq n$ and $\frac{n-2}{n-1}\cdot\frac{\sqrt{i}}{\sqrt{i-2}} \geq 1$, which is equivalent to

$$i(n-2)^2 \geq (i-2)(n-1)^2 \quad \Leftrightarrow \quad 2(n-1)^2 \geq i(2n-3),$$

which can be easily verified since $i(2n-3) \leq \frac{n}{2}(2n-3) = n^2 - \frac{3}{2}n \leq 2(n-1)^2$ for all $n \geq 2$. Also, note that we have to deal with the last iterate separately since Lemma B.8 applies only for $i \geq 2$.

Now suppose that $i \geq 1$ is an *odd* integer. Then we have

$$\mathbb{E}\left[|\mathcal{E}_{i+1}|\right] = \left(1-\frac{1}{n-i}\right)\mathbb{E}\left[|\mathcal{E}_i|\right] + \mathbb{1}_{i \text{ is even}}\frac{\binom{i}{\frac{i}{2}}\binom{n-i}{\frac{n-i}{2}}}{\binom{n}{\frac{n}{2}}} = \left(1-\frac{1}{n-i}\right)\mathbb{E}\left[|\mathcal{E}_i|\right] \qquad (\because i \text{ is odd})$$

and since $i+1$ is even, we can use the previous result as

$$\mathbb{E}\left[|\mathcal{E}_i|\right] = \frac{n-i}{n-i-1}\mathbb{E}\left[|\mathcal{E}_{i+1}|\right] \geq \frac{n-i}{n-i-1}\frac{\sqrt{i+1}}{10}.$$

Finally, since $\frac{n-i}{n-i-1} \geq 1 \geq \frac{\sqrt{i}}{\sqrt{i+1}}$, we can conclude that

$$\mathbb{E}\left[|\mathcal{E}_i|\right] \geq \frac{n-i}{n-i-1}\frac{\sqrt{i+1}}{10} \geq \frac{\sqrt{i}}{10}.$$

□

Lemma B.8. *For even, positive integers n, i with $n \geq 4$ and $2 \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor$, we have*

$$\frac{\binom{i}{\frac{i}{2}}\binom{n-i}{\frac{n-i}{2}}}{\binom{n}{\frac{n}{2}}} \geq \frac{2}{5\sqrt{i}}.$$

Proof. From Theorem 1 of Mortici (2011), for all $n \geq 1$ we have the expression

$$n! = \sqrt{\pi(2n+\alpha_n)}\cdot\frac{n^n}{e^n}, \quad \text{for some value } 0.333 \leq \alpha_n \leq 0.354.$$

Since $\frac{i}{2} \geq 1$, we have

$$\binom{i}{\frac{i}{2}} = \frac{i!}{((\frac{i}{2})!)^2} = \frac{i^i}{e^i} \cdot \frac{e^i}{(\frac{i}{2})^i} \frac{\sqrt{\pi(2i+\alpha_i)}}{\pi(i+\alpha_{i/2})} = 2^i \cdot \frac{\sqrt{(2i+\alpha_i)}}{\sqrt{\pi}(i+\alpha_{i/2})}.$$

Then we can compute

$$\begin{aligned}
\frac{\binom{i}{\frac{i}{2}}\binom{n-i}{\frac{n-i}{2}}}{\binom{n}{\frac{n}{2}}} &= \frac{\sqrt{(2i+\alpha_i)}}{\sqrt{\pi}(i+\alpha_{i/2})} \frac{\sqrt{(2(n-i)+\alpha_{n-i})}}{\sqrt{\pi}(n-i+\alpha_{(n-i)/2})} \frac{\sqrt{\pi}(n+\alpha_{n/2})}{\sqrt{(2n+\alpha_n)}} \\
&= \frac{1}{\sqrt{\pi}} \frac{\sqrt{(2i+\alpha_i)}}{(i+\alpha_{i/2})} \frac{\sqrt{(2(n-i)+\alpha_{n-i})}}{(n-i+\alpha_{(n-i)/2})} \frac{(n+\alpha_{n/2})}{\sqrt{(2n+\alpha_n)}} \\
&\geq \frac{1}{\sqrt{\pi}} \frac{\sqrt{(2i+0.33)}}{(i+0.354)} \frac{\sqrt{(2(n-i)+0.33)}}{(n-i+0.354)} \frac{(n+0.33)}{\sqrt{(2n+0.354)}} \\
&\geq \frac{1}{\sqrt{\pi}} \frac{\sqrt{2i}}{1.354i} \frac{\sqrt{(2(n-i))}}{1.354(n-i)} \frac{n}{\sqrt{2.354n}} = \frac{2}{1.354^2\sqrt{2.354\pi}} \frac{\sqrt{n}}{\sqrt{i(n-i)}} \\
&\geq \frac{2}{1.354^2\sqrt{2.354\pi}} \frac{1}{\sqrt{i}} \geq \frac{2}{5\sqrt{i}},
\end{aligned}$$

where $\frac{2}{1.354^2\sqrt{2.354\pi}} = 0.401157\cdots \geq 0.4 = \frac{2}{5}$. □

C. Proof of Theorem 3.3

Here we prove Theorem 3.3, restated below for the sake of readability.

Theorem 3.3. *For any $n \geq 2$ and $\kappa \geq c_1$, there exists a 2-dimensional function $F \in \mathcal{F}(L, \mu, 0, \nu)$ and an initialization point $\boldsymbol{x}_0$ such that for any constant step size $\eta \leq \frac{1}{c_2 L n}$, any weighted average iterate $\hat{\boldsymbol{x}}$ of SGD-RR of the form as in (4) satisfies*

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}) - F^*\right] = \begin{cases} \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right), & \text{if } K \geq c_2\kappa, \\ \Omega\left(\frac{\nu^2}{\mu}\right), & \text{if } K < c_2\kappa, \end{cases}$$

for the same universal constants c_1, c_2 as in Theorem 3.1.

Proof. We prove the theorem statement for the same constants $c_1 = 2415$ and $c_2 = 161$ as in Theorem 3.1.

Similarly as in Appendix B, we use different objective functions for two step-size regimes and aggregate the functions to obtain the final lower bound. Here we also assume n is even, where we can easily extend to odd n's by the same reasoning as in Appendix B.

Here we prove the following lower bounds for each regime. Here F_j^* is the minimizer of F_j for $j = 1, 2$.

- If $\eta \in \left(0, \frac{1}{\mu n K}\right)$, there exists a 1-dimensional objective function $F_1(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with initialization $x_0^1 = D_0$ (for any D_0) satisfies

$$\mathbb{E}\left[F_1(\hat{x}) - F_1^*\right] = \Omega\left(\mu D_0^2\right).$$

- If $\eta \in \left[\frac{1}{\mu n K}, \frac{1}{161 L n}\right]$, there exists a 1-dimensional objective function $F_2(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with initialization $x_0^1 = \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$ satisfies

$$\mathbb{E}\left[F_2(\hat{x}) - F_2^*\right] = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right).$$

Now we define the 2-dimensional function $F(x, y) = F_1(x) + F_2(y)$, where F_1 and F_2 are chosen to satisfy the above lower bounds for ν replaced by $\frac{\nu}{\sqrt{2}}$. Following the analyses in Appendix B, from $F_1, F_2 \in \mathcal{F}(L, \mu, 0, \frac{\nu}{\sqrt{2}})$ (by construction) we have $F \in \mathcal{F}(L, \mu, 0, \nu)$.

Now suppose that we set $D_0 = \frac{\nu}{\mu}$ and initialize at the point $\left(\frac{\nu}{\mu}, \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}\right)$.

If $K \geq 161\kappa$, then since $\frac{1}{\mu n K} \leq \frac{1}{161 L n}$ we can use the lower bound for $F_2(y)$. The lower bound for this case becomes

$$\mathbb{E}\left[F(\hat{x}, \hat{y}) - F^*\right] = \Omega\left(\min\left\{\frac{\nu^2}{\mu}, \frac{L\nu^2}{\mu^2 n K^2}\right\}\right) = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right).$$

If $K < 161\kappa$, then since $\frac{1}{\mu n K} > \frac{1}{161 L n}$ the latter step-size regime does not exist, i.e., we *cannot* use the lower bound for $F_2(y)$, and the lower bound for this case becomes

$$\mathbb{E}\left[F(\hat{x}, \hat{y}) - F^*\right] = \Omega\left(\frac{\nu^2}{\mu}\right),$$

which completes the proof. □

C.1. Lower Bound for $\eta \in \left(0, \frac{1}{\mu n K}\right)$

Here we show that there exists $F_1(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with $x_0^1 = D_0$ satisfies

$$\mathbb{E}\left[F_1(\hat{x}) - F_1^*\right] = \Omega\left(\mu D_0^2\right).$$

Proof. We define the same $F_1(x) \in \mathcal{F}(\mu, \mu, 0, 0)$ as in Appendix B.1 by the following components.

$$f_i(x) = F_1(x) = \frac{\mu x^2}{2}$$

Note that $\mathcal{F}(\mu, \mu, 0, 0) \subseteq \mathcal{F}(L, \mu, 0, \nu)$ and $F_1^* = 0$ at $x^* = 0$ by definition.

We start from Equation (8) in Appendix B, which gives

$$x_0^{k+1} = (1 - \eta\mu)^{nk} \cdot D_0 \geq \left(1 - \frac{1}{nK}\right)^{nK} \cdot D_0 \geq \frac{D_0}{4}$$

for all k. Then for any weighted average $\hat{x}$ we have

$$\hat{x} = \frac{\sum_{k=1}^{K+1} \alpha_k x_0^k}{\sum_{k=1}^{K+1} \alpha_k} \geq \frac{\sum_{k=1}^{K+1} \alpha_k \frac{D_0}{4}}{\sum_{k=1}^{K+1} \alpha_k} = \frac{D_0}{4}$$

and therefore

$$F_1(\hat{x}) \geq \frac{\mu}{2}\left(\frac{D_0}{4}\right)^2 = \frac{\mu D_0^2}{32},$$

which concludes that $\mathbb{E}\left[F_1(\hat{x}) - F_1^*\right] = \mathbb{E}\left[F_1(\hat{x})\right] = F_1(\hat{x}) = \Omega\left(\mu D_0^2\right)$. □

C.2. Lower Bound for $\eta \in \left[\frac{1}{\mu n K}, \frac{1}{161 L n}\right]$

Here we show that there exists $F_2(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that SGD-RR with $x_0^1 = \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$ satisfies

$$\mathbb{E}\left[F_2(\hat{x}) - F_2^*\right] = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right).$$

Proof. We define the $F_2(x) \in \mathcal{F}(L, \mu, 0, \nu)$ as in Appendix B.2 by the following components:

$$f_i(x) = \begin{cases} (L\mathbb{1}_{x<0} + \mu_0 \mathbb{1}_{x\geq 0}) \frac{x^2}{2} + \nu x & \text{if } i \leq n/2, \\ (L\mathbb{1}_{x<0} + \mu_0 \mathbb{1}_{x\geq 0}) \frac{x^2}{2} - \nu x & \text{otherwise,} \end{cases}$$

where we assume $\mu_0 \leq \frac{L}{2415}$ and later choose $\mu_0 = \frac{L}{2415}$. With this construction, the finite-sum objective becomes

$$F_2(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = (L\mathbb{1}_{x<0} + \mu_0 \mathbb{1}_{x\geq 0}) \frac{x^2}{2}.$$

Note that $F_2^* = 0$ at $x^* = 0$ by definition, and that μ_0 is different from μ. While $F_2(x) \in \mathcal{F}(L, \mu_0, 0, \nu)$ by construction, we can ensure that $\mathcal{F}(L, \mu_0, 0, \nu) \subset \mathcal{F}(L, \mu, 0, \nu)$ because the assumption $\kappa \geq 2415$ implies $\mu_0 = \frac{L}{2415} \geq \mu$.

We start from Equation (9) in Appendix B, which gives

$$\mathbb{E}\left[x_0^{k+1}\right] = \mathbb{E}\left[x_n^k\right] \geq \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$$

for all k. If we set $x_0 \geq \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$ then all end-of-epoch iterates must maintain $\mathbb{E}[x_n^k] \geq \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}$. This implies that for any weighted average $\hat{x}$ we have

$$\mathbb{E}[\hat{x}] = \mathbb{E}\left[\frac{\sum_{k=1}^{K+1} \alpha_k x_0^k}{\sum_{k=1}^{K+1} \alpha_k}\right] = \frac{\sum_{k=1}^{K+1} \alpha_k \mathbb{E}\left[x_0^k\right]}{\sum_{k=1}^{K+1} \alpha_k} \geq \frac{\sum_{k=1}^{K+1} \alpha_k \left(\frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}\right)}{\sum_{k=1}^{K+1} \alpha_k} = \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}.$$

Finally, by Jensen's inequality, we have

$$\begin{aligned} \mathbb{E}\left[F_2(\hat{x}) - F_2^*\right] &= \mathbb{E}\left[F_2(\hat{x})\right] \\ &\geq \frac{L}{2 \cdot 2415} \mathbb{E}\left[\hat{x}^2\right] \\ &\geq \frac{L}{4830} \mathbb{E}\left[\hat{x}\right]^2 \\ &\geq \frac{L}{4830} \cdot \left(\frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K}\right)^2 = \Omega\left(\frac{L\nu^2}{\mu^2 n K^2}\right). \end{aligned}$$

□

C.3. Proof of Corollary 3.5

Here we prove Corollary 3.5, restated below for the sake of readability.

Corollary 3.5. *For any $n \geq 2$, there exists a 2-dimensional function $F \in \mathcal{F}(L, 0, 0, \nu)$ such that if*

$$K \geq c_3 \max\left\{\frac{L^2 D^2 n}{\nu^2}, \frac{\nu}{\mu D n^{1/2}}\right\}, \tag{5}$$

then for any constant step size $\eta \leq \frac{1}{c_2 L n}$, any weighted average iterate $\hat{\boldsymbol{x}}$ of SGD-RR of the form as in (4) satisfies

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}) - F^*\right] = \Omega\left(\frac{L^{1/3}\nu^{2/3} D^{4/3}}{n^{1/3} K^{2/3}}\right),$$

for some universal constants c_2 and c_3.

Proof. Suppose that c_1, c_2 are the same constants as in Theorem 3.3, and let $c_3 = \max\{c_1^{3/2}, c_2^3\}$. We use results of Theorem 3.3 in Appendix C, but we view the initialization D_0 for the first coordinate as a separate constant instead of setting to a certain value like $D_0 = \frac{\nu}{\mu}$.

Let us choose $\mu = \frac{L^{1/3}\nu^{2/3}}{D^{2/3}n^{1/3}K^{2/3}}$. First, we can check that our choice of μ and $K \geq c_3 \frac{\nu}{\mu D n^{1/2}} \geq c_3 \frac{\nu}{L D n^{1/2}}$ implies

$$\kappa = \frac{L^{2/3}D^{2/3}n^{1/3}K^{2/3}}{\nu^{2/3}} \geq c_3^{2/3} \geq c_1,$$

and $K \geq c_3 \frac{L^2D^2n}{\nu^2}$ implies

$$K \geq c_3^{1/3} \cdot \frac{L^{2/3}D^{2/3}n^{1/3}K^{2/3}}{\nu^{2/3}} = c_3^{1/3}\kappa \geq c_2\kappa.$$

Since $\kappa \geq c_1$ and $K \geq c_2\kappa$, by Theorem 3.3 there exists some $F_\mu \in \mathcal{F}(L, \mu, 0, \nu)$ satisfying

$$\mathbb{E}\left[F_\mu(\hat{\boldsymbol{x}}) - F_\mu^*\right] = \Omega\left(\min\left\{\mu D_0^2, \frac{L\nu^2}{\mu^2 n K^2}\right\}\right).$$

for initialization $(D_0, \frac{1}{27000} \cdot \frac{\nu}{\mu n^{1/2} K})$. Note that we have $D^2 = D_0^2 + \frac{1}{27000^2} \cdot \frac{\nu^2}{\mu^2 n K^2}$. Since $K \geq c_3 \frac{\nu}{\mu D n^{1/2}}$, or equivalently $D \geq c_3 \frac{\nu}{\mu n^{1/2} K}$, we have $D_0 = \Omega(\frac{\nu}{\mu n^{1/2} K})$ and therefore

$$\mathbb{E}\left[F_\mu(\hat{\boldsymbol{x}}) - F_\mu^*\right] = \Omega\left(\min\left\{\mu D^2, \frac{L\nu^2}{\mu^2 n K^2}\right\}\right).$$

Letting $F = F_\mu$ and plugging in $\mu = \frac{L^{1/3}\nu^{2/3}}{D^{2/3}n^{1/3}K^{2/3}}$, we can conclude that

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}) - F^*\right] = \Omega\left(\frac{L^{1/3}\nu^{2/3}D^{4/3}}{n^{1/3}K^{2/3}}\right).$$

Finally we can check that $F \in \mathcal{F}(L, \mu, 0, \nu) \subseteq \mathcal{F}(L, 0, 0, \nu)$ for all $\mu > 0$, which completes the proof. □

D. Proof of Proposition 3.4

Here we prove Proposition 3.4, restated below for the sake of readability.

Proposition 3.4. *Suppose that $F \in \mathcal{F}(L, \mu, 0, \nu)$, and that we choose η as*

$$\eta = \min\left\{\frac{1}{\sqrt{2}Ln}, \frac{9}{\mu n K}\max\left\{1, \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)\right\}\right\}.$$

Then, for SGD-RR with constant step size η and $K \geq 5$, the tail average iterate $\hat{\boldsymbol{x}}_{\text{tail}}$ satisfies:

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] = \tilde{\mathcal{O}}\left(\frac{LD^2}{K}e^{-\frac{1}{9\sqrt{2}}\frac{K}{\kappa}} + \frac{L\nu^2}{\mu^2 n K^2}\right).$$

Proof. We start from the following lemma from Mishchenko et al. (2020).

Lemma D.1 (Mishchenko et al. (2020), Lemma 3)**.** *Assume that functions $f_1, \dots, f_n$ are convex and $F, f_1, \dots, f_n$ are L-smooth. Suppose that σ_*^2 is defined as in Equation (3). Then SGD-RR with step size $\eta \leq \frac{1}{\sqrt{2}Ln}$ satisfies*

$$\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}_0^k - \boldsymbol{x}^*\|^2\right] - 2\eta n \mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] + \frac{1}{2}\eta^3 L \sigma_*^2 n^2.$$

Note that the assumptions in Definition 2.5 of $\mathcal{F}(L, \mu, 0, \nu)$ includes all the required conditions above, *plus* an additional condition that F is μ-strongly convex. Also, note that the σ_*^2 term from the original paper can be replaced with ν^2, which is safe by the same reasoning as what we mentioned in Proposition 3.2. Hence we can use the following inequality:

$$\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}_0^k - \boldsymbol{x}^*\|^2\right] - 2\eta n \mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] + \frac{1}{2}\eta^3 L \nu^2 n^2.$$

From strong convexity, for all k we have

$$\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] \geq \frac{\mu}{2}\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right]. \tag{22}$$

Now we apply (22) to *only exactly half* of the term involving $\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right]$ to obtain

$$\left(1 + \frac{\eta\mu n}{2}\right)\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}_0^{k} - \boldsymbol{x}^*\|^2\right] - \eta n\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] + \frac{1}{2}\eta^3 L\nu^2 n^2.$$

Since $0 \leq \eta\mu n \leq \eta L n \leq \frac{1}{\sqrt{2}} \leq 1$ implies $\frac{1}{1+\frac{\eta\mu n}{2}} \leq 1 - \frac{\eta\mu n}{3}$ and $\frac{2}{3} \leq \frac{1}{1+\frac{\eta\mu n}{2}} \leq 1$, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] &\leq \frac{1}{1+\frac{\eta\mu n}{2}}\left(\mathbb{E}\left[\|\boldsymbol{x}_0^{k} - \boldsymbol{x}^*\|^2\right] - \eta n\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] + \frac{1}{2}\eta^3 L\nu^2 n^2\right)\\ &\leq \left(1 - \frac{\eta\mu n}{3}\right)\mathbb{E}\left[\|\boldsymbol{x}_0^{k} - \boldsymbol{x}^*\|^2\right] - \frac{2}{3}\eta n\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] + \frac{1}{2}\eta^3 L\nu^2 n^2. \qquad (23)\end{aligned}$$

We derive two different types of weaker inequalities from (23), as:

$$\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] \leq \left(1 - \frac{\eta\mu n}{3}\right)\mathbb{E}\left[\|\boldsymbol{x}_0^{k} - \boldsymbol{x}^*\|^2\right] + \frac{1}{2}\eta^3 L\nu^2 n^2, \tag{24}$$

$$\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}_0^{k} - \boldsymbol{x}^*\|^2\right] - \frac{2}{3}\eta n\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] + \frac{1}{2}\eta^3 L\nu^2 n^2. \tag{25}$$

From (24), we can unroll the inequality to obtain

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] &\leq \left(1 - \frac{\eta\mu n}{3}\right)^k\mathbb{E}\left[\|\boldsymbol{x}_0^{1} - \boldsymbol{x}^*\|^2\right] + \frac{1}{2}\eta^3 L\nu^2 n^2\sum_{j=0}^{K-1}\left(1 - \frac{\eta\mu n}{3}\right)^j\\ &\leq \left(1 - \frac{\eta\mu n}{3}\right)^k\mathbb{E}\left[\|\boldsymbol{x}_0^{1} - \boldsymbol{x}^*\|^2\right] + \frac{1}{2}\eta^3 L\nu^2 n^2\sum_{j=0}^{\infty}\left(1 - \frac{\eta\mu n}{3}\right)^j\\ &= \left(1 - \frac{\eta\mu n}{3}\right)^k\mathbb{E}\left[\|\boldsymbol{x}_0^{1} - \boldsymbol{x}^*\|^2\right] + \frac{1}{2}\eta^3 L\nu^2 n^2\frac{1}{1-\left(1-\frac{\eta\mu n}{3}\right)}\\ &= \left(1 - \frac{\eta\mu n}{3}\right)^k D^2 + \frac{3}{2}\cdot\frac{\eta^2 L\nu^2 n}{\mu} \qquad (D := \|\boldsymbol{x}_0^1 - \boldsymbol{x}^*\|)\\ &\leq e^{-\frac{1}{3}\eta\mu n k}D^2 + \frac{3}{2}\cdot\frac{\eta^2 L\nu^2 n}{\mu} \qquad (26)\end{aligned}$$

which holds for all k.

From (25), we can rearrange terms as

$$\eta n\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] \leq \frac{3}{2}\mathbb{E}\left[\|\boldsymbol{x}_0^{k} - \boldsymbol{x}^*\|^2\right] - \frac{3}{2}\mathbb{E}\left[\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2\right] + \frac{3}{4}\eta^3 L\nu^2 n^2$$

and average the inequality from $k = \lceil\frac{K}{2}\rceil$ to K to obtain

$$\begin{aligned}&\frac{\eta n}{K - \lceil\frac{K}{2}\rceil + 1}\sum_{k=\lceil\frac{K}{2}\rceil}^{K}\mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right]\\ &\leq \frac{1}{K - \lceil\frac{K}{2}\rceil + 1}\left(\frac{3}{2}\mathbb{E}\left[\|\boldsymbol{x}_0^{\lceil\frac{K}{2}\rceil} - \boldsymbol{x}^*\|^2\right] - \frac{3}{2}\mathbb{E}\left[\|\boldsymbol{x}_0^{K+1} - \boldsymbol{x}^*\|^2\right]\right) + \frac{3}{4}\eta^3 L\nu^2 n^2\\ &\leq \frac{3/2}{K - \lceil\frac{K}{2}\rceil + 1}\mathbb{E}\left[\|\boldsymbol{x}_0^{\lceil\frac{K}{2}\rceil} - \boldsymbol{x}^*\|^2\right] + \frac{3}{4}\eta^3 L\nu^2 n^2. \qquad (27)\end{aligned}$$

Therefore we have

$$\begin{aligned}
\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] &= \mathbb{E}\left[F\left(\frac{1}{K - \lceil \frac{K}{2} \rceil + 1} \sum_{k=\lceil \frac{K}{2} \rceil}^{K} \boldsymbol{x}_n^k\right) - F^*\right] \\
&= \mathbb{E}\left[F\left(\frac{1}{K - \lceil \frac{K}{2} \rceil + 1} \sum_{k=\lceil \frac{K}{2} \rceil}^{K} \boldsymbol{x}_0^{k+1}\right) - F^*\right] \\
&\leq \frac{1}{K - \lceil \frac{K}{2} \rceil + 1} \sum_{k=\lceil \frac{K}{2} \rceil}^{K} \mathbb{E}\left[F(\boldsymbol{x}_0^{k+1}) - F(\boldsymbol{x}^*)\right] && (\because \text{Jensen's inequality}) \\
&\leq \frac{3/2}{K - \lceil \frac{K}{2} \rceil + 1} \cdot \frac{1}{\eta n} \mathbb{E}\left[\|\boldsymbol{x}_0^{\lceil \frac{K}{2} \rceil} - \boldsymbol{x}^*\|^2\right] + \frac{3}{4}\eta^2 L \nu^2 n && (\because \text{By (27)}) \\
&\leq \frac{3}{\eta n K} \mathbb{E}\left[\|\boldsymbol{x}_0^{\lceil \frac{K}{2} \rceil} - \boldsymbol{x}^*\|^2\right] + \frac{3}{4}\eta^2 L \nu^2 n && (\because \lceil \tfrac{K}{2} \rceil \leq \tfrac{K}{2} + 1) \\
&\leq \frac{3}{\eta n K}\left(e^{-\frac{1}{3}\eta\mu n (\lceil \frac{K}{2} \rceil - 1)} D^2 + \frac{3}{2} \cdot \frac{\eta^2 L \nu^2 n}{\mu}\right) + \frac{1}{2}\eta^2 L \nu^2 n && (\because \text{By (26), for } k = \lceil \tfrac{K}{2} \rceil) \\
&= \frac{3D^2}{\eta n K} e^{-\frac{1}{3}\eta\mu n (\lceil \frac{K}{2} \rceil - 1)} + \frac{9}{2} \cdot \frac{\eta L \nu^2}{\mu K} + \frac{1}{2}\eta^2 L \nu^2 n \\
&\leq \frac{3D^2}{\eta n K} e^{-\frac{1}{9}\eta\mu n K} + \frac{9}{2} \cdot \frac{\eta L \nu^2}{\mu K} + \frac{1}{2}\eta^2 L \nu^2 n. \qquad (28)
\end{aligned}$$

Note that in the last inequality, $K \geq 5$ implies $\lceil \frac{K}{2} \rceil - 1 \geq \frac{K}{3}$.

Now we will divide into four possible cases according to how we choose η, and then derive that desired upper bound holds in each case from Equation (28). Note that we have $\max\left\{1, \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)\right\} = 1$ if and only if $K \leq \frac{e^{1/2} L^{1/2} \nu}{\mu^{3/2} n^{1/2} D}$, which is again equivalent to $\mu D^2 \leq \frac{e L \nu^2}{\mu^2 n K^2}$.

Case (a) Suppose that $\eta = \frac{1}{\sqrt{2}Ln} \leq \frac{9}{\mu n K} \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)$, where $\max\left\{1, \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)\right\} = \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)$.

From Equation (28) we have

$$\begin{aligned}
\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] &\leq \frac{3D^2}{\eta n K} e^{-\frac{1}{9}\eta\mu n K} + \frac{9}{2} \cdot \frac{\eta L \nu^2}{\mu K} + \frac{1}{2}\eta^2 L \nu^2 n \\
&\leq \frac{3\sqrt{2}LD^2}{K} e^{-\frac{1}{9\sqrt{2}} \frac{K}{L/\mu}} + \frac{81}{2} \frac{L\nu^2}{\mu^2 n K^2} \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right) + \frac{81}{2} \frac{L\nu^2}{\mu^2 n K^2} \log^2\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right) \\
&= \tilde{\mathcal{O}}\left(\frac{LD^2}{K} e^{-\frac{1}{9\sqrt{2}} \frac{K}{L/\mu}} + \frac{L\nu^2}{\mu^2 n K^2}\right).
\end{aligned}$$

Case (b) Suppose that $\eta = \frac{1}{\sqrt{2}Ln} \leq \frac{9}{\mu n K}$, where $\max\left\{1, \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)\right\} = 1$.

From Equation (28) we have

$$\begin{aligned}
\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] &\leq \frac{3D^2}{\eta n K} e^{-\frac{1}{9}\eta\mu n K} + \frac{9}{2} \cdot \frac{\eta L \nu^2}{\mu K} + \frac{1}{2}\eta^2 L \nu^2 n \\
&\leq \frac{3\sqrt{2}LD^2}{K} e^{-\frac{1}{9\sqrt{2}} \frac{K}{L/\mu}} + \frac{81}{2} \frac{L\nu^2}{\mu^2 n K^2} + \frac{81}{2} \frac{L\nu^2}{\mu^2 n K^2} \\
&= \tilde{\mathcal{O}}\left(\frac{LD^2}{K} e^{-\frac{1}{9\sqrt{2}} \frac{K}{L/\mu}} + \frac{L\nu^2}{\mu^2 n K^2}\right).
\end{aligned}$$

Case (c) Suppose that $\eta = \frac{9}{\mu n K} \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right) \leq \frac{1}{\sqrt{2}Ln}$, where $\max\left\{1, \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)\right\} = \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)$.

From Equation (28) we have

$$
\begin{aligned}
\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] &\le \frac{3D^2}{\eta n K} e^{-\frac{1}{9}\eta\mu n K} + \frac{9}{2}\cdot\frac{\eta L\nu^2}{\mu K} + \frac{1}{2}\eta^2 L\nu^2 n \\
&= \frac{\mu D^2}{3\log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)}\cdot\frac{L\nu^2}{\mu^3 n D^2 K^2} + \frac{81}{2}\frac{L\nu^2}{\mu^2 n K^2}\log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right) + \frac{81}{2}\frac{L\nu^2}{\mu^2 n K^2}\log^2\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right) \\
&= \frac{L\nu^2}{\mu^2 n K^2}\left(\frac{1}{3\log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)} + \frac{81}{2}\log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right) + \frac{81}{2}\log^2\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)\right) \\
&= \tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 n K^2}\right).
\end{aligned}
$$

Case (d) Suppose that $\eta = \frac{9}{\mu n K} \le \frac{1}{\sqrt{2}Ln}$, where $\max\left\{1, \log\left(\frac{\mu^3 n D^2 K^2}{L\nu^2}\right)\right\} = 1$.

From Equation (28) we have

$$
\begin{aligned}
\mathbb{E}\left[F(\hat{\boldsymbol{x}}_{\text{tail}}) - F^*\right] &\le \frac{3D^2}{\eta n K} e^{-\frac{1}{9}\eta\mu n K} + \frac{9}{2}\cdot\frac{\eta L\nu^2}{\mu K} + \frac{1}{2}\eta^2 L\nu^2 n \\
&= \frac{\mu D^2}{3e} + \frac{81}{2}\frac{L\nu^2}{\mu^2 n K^2} + \frac{81}{2}\frac{L\nu^2}{\mu^2 n K^2} \\
&\le \frac{L\nu^2}{\mu^2 n K^2}\cdot\left(\frac{1}{3e} + 81\right) \\
&= \tilde{\mathcal{O}}\left(\frac{L\nu^2}{\mu^2 n K^2}\right).
\end{aligned}
$$

Therefore Proposition 3.4 holds for all cases, which completes the proof. □

E. Proof of Theorem 4.1

Here we prove Theorem 4.1, restated below for the sake of readability.

Theorem 4.1. *For any $n \ge 2$ and $\kappa \ge 4$, there exists a 4-dimensional function $F \in \mathcal{F}(L, \mu, 0, \nu)$ and an initialization point $\boldsymbol{x}_0$ such that for any permutation-based SGD with any constant step size η, any weighted average iterate $\hat{\boldsymbol{x}}$ of the form as in Equation (4) satisfies*

$$
F(\hat{\boldsymbol{x}}) - F^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).
$$

Proof. Similarly as in Appendix B, we define objective functions for three step-size regimes and aggregate the functions to obtain the final lower bound. Here we also assume n is even, where we can easily extend to odd n's by the same reasoning as in Appendix B.

We will prove the following lower bounds for each regime. Here F_j^* is the minimizer of F_j for $j = 1, 2, 3$.

- If $\eta \in \left(0, \frac{1}{\mu n K}\right)$, there exists a 1-dimensional objective function $F_1(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that any permutation-based SGD with initialization $x_0^1 = \frac{L^{1/2}\nu}{\mu^{3/2} n K}$ satisfies

$$
F_1(\hat{x}) - F_1^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).
$$

- If $\eta \in \left[\frac{1}{\mu n K}, \frac{1}{L}\right)$, there exists a 2-dimensional objective function $F_2(y, z) \in \mathcal{F}\left(L, \mu, 0, \sqrt{2}\nu\right)$ such that any permutation-based SGD with initialization $(y_0^1, z_0^1) = \left(\frac{\nu}{2L}, 0\right)$ satisfies

$$
F_2(\hat{y}, \hat{z}) - F_2^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).
$$

Note that $\hat{y}$ and $\hat{z}$ share same weights $\{\alpha_k\}_{k=1}^{K+1}$.

- If $\eta \geq \frac{1}{L}$, there exists a 1-dimensional objective function $F_3(w) \in \mathcal{F}(2L, \mu, 0, \nu)$ such that any permutation-based SGD with initialization $w_0^1 = \frac{\nu}{\mu n K}$ satisfies

$$F_3(\hat{w}) - F_3^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).$$

Now we define the 4-dimensional function $F(\boldsymbol{x}) = F(x, y, z, w) = F_1(x) + F_2(y, z) + F_3(w)$, where F_1, F_2, and F_3 are chosen to satisfy the above lower bounds.

Following the analyses in Appendix B, we have $F \in \mathcal{F}(2L, \mu, 0, 2\nu)$, which allows us to directly apply the convergence rates in the lower bounds of F_1, F_2 and F_3 to the aggregated function F.

When $K \leq \frac{\kappa}{n}$, the second step-size regime becomes invalid. In this case, we define $F(x, y, z, w) = F_1(x) + F_1(y) + F_1(z) + F_3(w) \in \mathcal{F}(2L, \mu, 0, 2\nu)$. The final lower bound is then the minimum of the lower bounds obtained for the remaining two regimes, which is $\Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right)$.

Note that we assumed $\kappa \geq 4$ and our constructed function is $2L$-*smooth* and μ-*strongly convex*. Thus, $\kappa \geq 4$ is equivalent to $\mu \leq \frac{L}{2}$ throughout the proof.

Finally, rescaling L and ν will give us the function $F \in \mathcal{F}(L, \mu, 0, \nu)$ satisfying $F(\hat{\boldsymbol{x}}) - F^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right)$. □

For the following subsections, we prove the lower bounds for F_1, F_2, and F_3 at the corresponding step size regimes.

E.1. Lower Bound for $\eta \in \left(0, \frac{1}{\mu n K}\right)$

Here we show that there exists $F_1(x) \in \mathcal{F}(L, \mu, 0, \nu)$ such that any permutation-based SGD with $x_0^1 = \frac{L^{1/2}\nu}{\mu^{3/2} n K}$ satisfies

$$F_1(\hat{x}) - F_1^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).$$

Proof. We define $F_1(x) \in \mathcal{F}(\mu, \mu, 0, 0)$ by the following components:

$$f_i(x) = F(x) = \frac{\mu}{2}x^2.$$

Note that $\mathcal{F}(\mu, \mu, 0, 0) \subseteq \mathcal{F}(L, \mu, 0, \nu)$ and $F_1^* = 0$ at $x^* = 0$ by definition.

In this regime, we will see that the step size is too small so that $\left\{x_n^k\right\}_{k=1}^{K}$ cannot even reach near the optimal point. We start from $x_0^1 = \frac{L^{1/2}\nu}{\mu^{3/2} n K}$. Since the gradient of all component functions evaluated at point x is fixed deterministically to μx, regardless of the permutation-based SGD algorithm we use, we have

$$\begin{aligned} x_n^k = x_0^1 (1 - \eta\mu)^{nk} &\geq \frac{L^{1/2}\nu}{\mu^{3/2} n K}\left(1 - \frac{1}{nK}\right)^{nk} \geq \frac{L^{1/2}\nu}{\mu^{3/2} n K}\left(1 - \frac{1}{nK}\right)^{nK} \\ &\overset{(a)}{>} \frac{L^{1/2}\nu}{\mu^{3/2} n K}\frac{1}{e}\left(1 - \frac{1}{nK}\right) \overset{(b)}{\geq} \frac{L^{1/2}\nu}{\mu^{3/2} n K}\frac{1}{2e}, \end{aligned}$$

where (a) comes from Lemma E.2 and (b) comes from the assumption that $n \geq 2$. Therefore, we have $\hat{x} = \Omega\left(\frac{L^{1/2}\nu}{\mu^{3/2} n K}\right)$ for any nonnegative weights $\{\alpha_k\}_{k=1}^{K+1}$. With this $\hat{x}$, we have

$$F_1(\hat{x}) - F_1^* = \frac{\mu}{2}\hat{x}^2 = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).$$

□

E.2. Lower Bound for $\eta \in \left[\frac{1}{\mu n K}, \frac{1}{L}\right)$

Here we show that there exists $F_2(y,z) \in \mathcal{F}(L, \mu, 0, \sqrt{2}\nu)$ such that any permutation-based SGD with $(y_0^1, z_0^1) = \left(\frac{\nu}{2L}, 0\right)$ satisfies

$$F_2(\hat{y}, \hat{z}) - F_2^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).$$

Proof. Let us define the function g_{+1}, g_{-1} as follows.

$$g_{+1}(x) = \left(L\mathbb{1}_{x<0} + \frac{L}{2}\mathbb{1}_{x\geq 0}\right)\frac{x^2}{2} + \nu x,$$
$$g_{-1}(x) = \left(L\mathbb{1}_{x<0} + \frac{L}{2}\mathbb{1}_{x\geq 0}\right)\frac{x^2}{2} - \nu x.$$

Note that g_{+1} and g_{-1} are μ-*strongly convex* since $\mu \leq \frac{L}{2}$. We define $F_2(x) \in \mathcal{F}(L, \mu, 0, \sqrt{2}\nu)$ by the following components:

$$f_i(y,z) = \begin{cases} g_{+1}(y) + g_{-1}(z) & \text{if } i \leq \frac{n}{2}, \\ g_{-1}(y) + g_{+1}(z) & \text{otherwise.} \end{cases}$$

With this construction, the finite-sum objective becomes

$$F_2(y,z) = \frac{1}{n}\sum_{i=1}^{n} f_i(y,z) = \left(L\mathbb{1}_{y<0} + \frac{L}{2}\mathbb{1}_{y\geq 0}\right)\frac{y^2}{2} + \left(L\mathbb{1}_{z<0} + \frac{L}{2}\mathbb{1}_{z\geq 0}\right)\frac{z^2}{2}.$$

Note that $F_2^* = 0$ at $(y^*, z^*) = (0,0)$ by definition.

We start at $(y_0^1, z_0^1) = \left(\frac{\nu}{2L}, 0\right)$. We now use the following lemma to find the lower bound of $\left\{y_n^k + z_n^k\right\}_{k=1}^{K}$ that holds for every permutation.

Lemma E.1. *Consider the optimization process whose setting is given as E.2 with $\eta < \frac{1}{L}$. For any $0 \leq t \leq \frac{n}{2} - 1$ and any $k \in \{1, \cdots, K\}$, if $y_{2t}^k + z_{2t}^k \geq 0$ holds, then*

$$y_{2t+2}^k + z_{2t+2}^k \geq \left(1 - \frac{\eta L}{2}\right)(1 - \eta L)\left(y_{2t}^k + z_{2t}^k\right) + \frac{\eta^2 L \nu}{2}$$

holds regardless of which functions are used at the $(2t+1)$-th and the $(2t+2)$-th iterations of the k-th epoch. Consequently, if $y_0^k + z_0^k \geq \frac{\eta\nu}{3-\eta L}$, $y_n^k + z_n^k \geq \frac{\eta\nu}{3-\eta L}$ holds regardless of the permutation σ_k.

The proof of the lemma is in Appendix E.4. In our setting, $y_0^1 + z_0^1 = \frac{\nu}{2L} \geq \frac{\eta\nu}{3-\eta L}$ since $\eta < \frac{1}{L}$. Thus, we have $y_n^k + z_n^k \geq \frac{\eta\nu}{3-\eta L}$ for every $k \in [K]$.

For $\hat{y} + \hat{z}$, we get

$$\begin{aligned}\hat{y} + \hat{z} &= \frac{\sum_{k=1}^{K+1}\alpha_k\left(y_0^k + z_0^k\right)}{\sum_{k=1}^{K+1}\alpha_k} \geq \frac{\eta\nu}{3-\eta L}\cdot\frac{\sum_{k=1}^{K+1}\alpha_k}{\sum_{k=1}^{K+1}\alpha_k} \\ &= \frac{\eta\nu}{3-\eta L} > \frac{\eta\nu}{3} = \Omega\left(\frac{\nu}{\mu n K}\right),\end{aligned}$$

and using the inequality $2(a^2+b^2) \geq (a+b)^2$,

$$\begin{aligned}F_2(\hat{y}, \hat{z}) - F_2^* &= \left(L\mathbb{1}_{y<0} + \frac{L}{2}\mathbb{1}_{y\geq 0}\right)\frac{\hat{y}^2}{2} + \left(L\mathbb{1}_{z<0} + \frac{L}{2}\mathbb{1}_{z\geq 0}\right)\frac{\hat{z}^2}{2} \\ &\geq \frac{L}{4}\left(\hat{y}^2 + \hat{z}^2\right)\end{aligned}$$

$$\begin{aligned}&\geq \frac{L}{8}\left(\hat{y}+\hat{z}\right)^2\\&=\Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).\end{aligned}$$

□

E.3. Lower Bound for $\eta > \frac{1}{L}$

Here we show that there exists $F_3(w) \in \mathcal{F}(2L, \mu, 0, \nu)$ such that any permutation-based SGD with $w_0^1 = \frac{\nu}{\mu n K}$ satisfies

$$F_3(\hat{w}) - F_3^* = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right).$$

Proof. We define $F_3(w) \in \mathcal{F}(2L, 2L, 0, 0)$ by the following components:

$$f_i(w) = F_3(w) = Lw^2.$$

Note that $\mathcal{F}(2L, 2L, 0, 0) \subseteq \mathcal{F}(2L, \mu, 0, \nu)$ and $F_3^* = 0$ at $w^* = 0$ by definition.

In this regime, we will see that the step size is so large that $\left\{w_n^k\right\}_{k=1}^K$ diverges. We start from $w_0^1 = \frac{\nu}{\mu n K}$. Since the gradient of all component functions at point w is fixed deterministically to $2Lw$, we have for every $k \in [K]$,

$$w_n^k = \left(1 - 2\eta L\right)^{nk} w_0^1 \geq 1^{nk}\frac{\nu}{\mu n K} = \Omega\left(\frac{\nu}{\mu n K}\right),$$

where we used the fact that n is even in the second step. Thus, regardless of the permutation-based SGD we use, we have $\hat{w} = \Omega\left(\frac{\nu}{\mu n K}\right)$ and $F_3(\hat{w}) - F_3^* = L\hat{w}^2 = \Omega\left(\frac{L\nu^2}{\mu^2 n^2 K^2}\right)$. □

E.4. Lemmas used in Theorem 4.1

In this subsection, we will prove the lemmas used in Theorem 4.1.

Lemma E.1. *Consider the optimization process whose setting is given as E.2 with $\eta < \frac{1}{L}$. For any $0 \leq t \leq \frac{n}{2} - 1$ and any $k \in \{1, \cdots, K\}$, if $y_{2t}^k + z_{2t}^k \geq 0$ holds, then*

$$y_{2t+2}^k + z_{2t+2}^k \geq \left(1 - \frac{\eta L}{2}\right)\left(1 - \eta L\right)\left(y_{2t}^k + z_{2t}^k\right) + \frac{\eta^2 L \nu}{2}$$

holds regardless of which functions are used at the $(2t+1)$-th and the $(2t+2)$-th iterations of the k-th epoch. Consequently, if $y_0^k + z_0^k \geq \frac{\eta\nu}{3-\eta L}$, $y_n^k + z_n^k \geq \frac{\eta\nu}{3-\eta L}$ holds regardless of the permutation σ_k.

Proof. Without loss of generality, assume $y_{2t}^k \geq z_{2t}^k$. Since we assumed $y_{2t}^k + z_{2t}^k$ is nonnegative, $y_{2t}^k \geq 0$ holds. Depending on which function is used at $(2t+1)$-th iteration of k-th epoch, we consider following two cases:

(a) $y_{2t+1}^k = y_{2t}^k - \eta\nabla g_{-1}(y_{2t}^k)$ and $z_{2t+1}^k = z_{2t}^k - \eta\nabla g_{+1}(z_{2t}^k)$,

(b) $y_{2t+1}^k = y_{2t}^k - \eta\nabla g_{+1}(y_{2t}^k)$ and $z_{2t+1}^k = z_{2t}^k - \eta\nabla g_{-1}(z_{2t}^k)$.

Note that $y_{2t+2}^k + z_{2t+2}^k$ is independent of which function is used at $(2t+2)$-th iteration of the k-th epoch, because $y_{2t+2}^k = \left(1 - \eta\left(L\mathbb{1}_{y_{2t+1}^k < 0} + \frac{L}{2}\mathbb{1}_{y_{2t+1}^k \geq 0}\right)\right) y_{2t+1}^k \pm \eta\nu$ and $z_{2t+2}^k = \left(1 - \eta\left(L\mathbb{1}_{z_{2t+1}^k < 0} + \frac{L}{2}\mathbb{1}_{z_{2t+1}^k \geq 0}\right)\right) z_{2t+1}^k \mp \eta\nu$ so that summation of y and z results in the canceling of $\eta\nu$ terms.

Case (a) For Case (a), $y_{2t+1}^k = (1-\frac{\eta L}{2})y_{2t}^k + \eta\nu > 0$ holds since $y_{2t}^k \geq 0$, but the signs of z_{2t}^k and z_{2t+1}^k are undetermined. Thereby, we split the cases by the signs of z_{2t}^k and z_{2t+1}^k.

First, assume $z_{2t}^k < 0$ and $z_{2t+1}^k < 0$. In this setting, $y_{2t}^k \geq 0$, $y_{2t+1}^k \geq 0$, $z_{2t}^k < 0$, $z_{2t+1}^k < 0$. Then,

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t+1}^k + (1-\eta L) z_{2t+1}^k \\
&= \left(1 - \frac{\eta L}{2}\right)\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k + \eta\nu\right) + (1-\eta L)\left((1-\eta L) z_{2t}^k - \eta\nu\right) \\
&= \left(1 - \frac{\eta L}{2}\right)^2 y_{2t}^k + (1-\eta L)^2 z_{2t}^k + \frac{\eta^2 L\nu}{2} \\
&= \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2}\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - (1-\eta L) z_{2t}^k\right) \\
&\geq \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2},
\end{aligned}$$

where the last inequality holds because $y_{2t}^k \geq 0$ and $z_{2t}^k < 0$.

Next, assume $z_{2t}^k \geq 0$ and $z_{2t+1}^k < 0$. In this setting, $y_{2t}^k \geq 0$, $y_{2t+1}^k \geq 0$, $z_{2t}^k \geq 0$, $z_{2t+1}^k < 0$. Similarly,

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t+1}^k + (1-\eta L) z_{2t+1}^k \\
&= \left(1 - \frac{\eta L}{2}\right)\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k + \eta\nu\right) + (1-\eta L)\left(\left(1 - \frac{\eta L}{2}\right) z_{2t}^k - \eta\nu\right) \\
&= \left(1 - \frac{\eta L}{2}\right)^2 y_{2t}^k + (1-\eta L)\left(1 - \frac{\eta L}{2}\right) z_{2t}^k + \frac{\eta^2 L\nu}{2} \\
&= \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2}\left(1 - \frac{\eta L}{2}\right) y_{2t}^k \\
&\geq \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2}.
\end{aligned}$$

Finally, assume both $z_{2t}^k \geq 0$ and $z_{2t+1}^k \geq 0$. In this setting, $y_{2t}^k \geq 0$, $y_{2t+1}^k \geq 0$, $z_{2t}^k \geq 0$, $z_{2t+1}^k \geq 0$. Since $0 \leq z_{2t+1}^k = \left(1 - \frac{\eta L}{2}\right) z_{2t}^k - \eta\nu$, $z_{2t}^k \geq \frac{\eta\nu}{1-\eta L/2}$ holds. Then,

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t+1}^k + \left(1 - \frac{\eta L}{2}\right) z_{2t+1}^k \\
&= \left(1 - \frac{\eta L}{2}\right)\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k + \eta\nu\right) + \left(1 - \frac{\eta L}{2}\right)\left(\left(1 - \frac{\eta L}{2}\right) z_{2t}^k - \eta\nu\right) \\
&= \left(1 - \frac{\eta L}{2}\right)^2 y_{2t}^k + \left(1 - \frac{\eta L}{2}\right)^2 z_{2t}^k \\
&= \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2}\left(1 - \frac{\eta L}{2}\right)(y_{2t}^k + z_{2t}^k) - \frac{\eta^2 L\nu}{2} \\
&\geq \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2}.
\end{aligned}$$

In the last inequality, we used the fact that $y_{2t}^k \geq 0$ and $z_{2t}^k \geq \frac{\eta\nu}{1-\eta L/2}$.

We don't have to consider the case when $z_{2t}^k < 0$ and $z_{2t+1}^k \geq 0$ hold, because $z_{2t+1}^k = (1-\eta L)z_{2t}^k - \eta\nu < 0$ if $z_{2t}^k < 0$. Thus, we have proven the first inequality of the lemma for Case (a).

Case (b) For Case (b), we consider three cases depending on the signs of y_{2t+1}^k and z_{2t+1}^k.

First, assume $y_{2t+1}^k \geq 0$ and $z_{2t+1}^k < 0$ hold. In this case, if $z_{2t}^k \geq 0$, then $z_{2t+1}^k = \left(1 - \frac{\eta L}{2}\right) z_{2t}^k + \eta\nu > 0$; therefore, $z_{2t}^k < 0$ should hold. So in this setting, we have $y_{2t}^k \geq 0$, $y_{2t+1}^k \geq 0$, $z_{2t}^k < 0$, and $z_{2t+1}^k < 0$. Using this fact,

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t+1}^k + (1-\eta L) z_{2t+1}^k \\
&= \left(1 - \frac{\eta L}{2}\right) \left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu\right) + (1-\eta L)\left((1-\eta L) z_{2t}^k + \eta\nu\right) \\
&= \left(1 - \frac{\eta L}{2}\right)^2 y_{2t}^k + (1-\eta L)^2 z_{2t}^k - \frac{\eta^2 L\nu}{2} \\
&= \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2}\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - (1-\eta L) z_{2t}^k - 2\eta\nu\right) \\
&= \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2}\left(y_{2t+1}^k - z_{2t+1}^k\right) \\
&\geq \left(1 - \frac{\eta L}{2}\right)(1-\eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2}.
\end{aligned}$$

Second, assume $y_{2t+1}^k < 0$ and $z_{2t+1}^k \geq 0$. If $z_{2t}^k < 0$, the setting becomes $y_{2t}^k \geq 0$, $y_{2t+1}^k < 0$, $z_{2t}^k < 0$, $z_{2t+1}^k \geq 0$ so that

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= (1-\eta L) y_{2t+1}^k + \left(1 - \frac{\eta L}{2}\right) z_{2t+1}^k \\
&= (1-\eta L)\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu\right) + \left(1 - \frac{\eta L}{2}\right)\left((1-\eta L) z_{2t}^k + \eta\nu\right) \\
&= (1-\eta L)\left(1 - \frac{\eta L}{2}\right)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2}.
\end{aligned}$$

If $z_{2t}^k \geq 0$, the setting becomes $y_{2t}^k \geq 0$, $y_{2t+1}^k < 0$, $z_{2t}^k \geq 0$, $z_{2t+1}^k \geq 0$ so that

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= (1-\eta L) y_{2t+1}^k + \left(1 - \frac{\eta L}{2}\right) z_{2t+1}^k \\
&= (1-\eta L)\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu\right) + \left(1 - \frac{\eta L}{2}\right)\left(\left(1 - \frac{\eta L}{2}\right) z_{2t}^k + \eta\nu\right) \\
&= (1-\eta L)\left(1 - \frac{\eta L}{2}\right)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2}\left(1 - \frac{\eta L}{2}\right) z_{2t}^k \\
&\geq (1-\eta L)\left(1 - \frac{\eta L}{2}\right)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2}.
\end{aligned}$$

Lastly, assume $y_{2t+1}^k \geq 0$ and $z_{2t+1}^k \geq 0$. If $z_{2t}^k < 0$, the setting becomes $y_{2t}^k \geq 0$, $y_{2t+1}^k \geq 0$, $z_{2t}^k < 0$, $z_{2t+1}^k \geq 0$ so that

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t+1}^k + \left(1 - \frac{\eta L}{2}\right) z_{2t+1}^k \\
&= \left(1 - \frac{\eta L}{2}\right)\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu\right) + \left(1 - \frac{\eta L}{2}\right)\left((1-\eta L) z_{2t}^k + \eta\nu\right) \\
&= (1-\eta L)\left(1 - \frac{\eta L}{2}\right)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2}\left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu\right) \\
&= (1-\eta L)\left(1 - \frac{\eta L}{2}\right)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2} + \frac{\eta L}{2} y_{2t+1}^k \\
&\geq (1-\eta L)\left(1 - \frac{\eta L}{2}\right)(y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L\nu}{2}.
\end{aligned}$$

If $z_{2t}^k \geq 0$, the setting becomes $y_{2t}^k \geq 0$, $y_{2t+1}^k \geq 0$, $z_{2t}^k \geq 0$, $z_{2t+1}^k \geq 0$ so that

$$\begin{aligned}
y_{2t+2}^k + z_{2t+2}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t+1}^k + \left(1 - \frac{\eta L}{2}\right) z_{2t+1}^k \\
&= \left(1 - \frac{\eta L}{2}\right) \left(\left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu\right) + \left(1 - \frac{\eta L}{2}\right) \left(\left(1 - \frac{\eta L}{2}\right) z_{2t}^k + \eta\nu\right) \\
&= (1 - \eta L) \left(1 - \frac{\eta L}{2}\right) (y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L \nu}{2} + \frac{\eta L}{2} \left(\left(1 - \frac{\eta L}{2}\right) (y_{2t}^k + z_{2t}^k) - \eta\nu\right) \\
&= (1 - \eta L) \left(1 - \frac{\eta L}{2}\right) (y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L \nu}{2} + \frac{\eta L}{2} \left(y_{2t+1}^k + \left(1 - \frac{\eta L}{2}\right) z_{2t}^k\right) \\
&\geq (1 - \eta L) \left(1 - \frac{\eta L}{2}\right) (y_{2t}^k + z_{2t}^k) + \frac{\eta^2 L \nu}{2}.
\end{aligned}$$

We do not have to consider the case when y_{2t+1}^k and z_{2t+1}^k are both less than 0, because $y_{2t+1}^k + z_{2t+1}^k \geq 0$ always holds. This can be shown by case analysis on the sign of z_{2t}^k: if $z_{2t}^k \geq 0$, then

$$\begin{aligned}
y_{2t+1}^k + z_{2t+1}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu + \left(1 - \frac{\eta L}{2}\right) z_{2t}^k + \eta\nu \\
&= \left(1 - \frac{\eta L}{2}\right) (y_{2t}^k + z_{2t}^k) \geq 0,
\end{aligned}$$

and if $z_{2t}^k < 0$, then

$$\begin{aligned}
y_{2t+1}^k + z_{2t+1}^k &= \left(1 - \frac{\eta L}{2}\right) y_{2t}^k - \eta\nu + (1 - \eta L) z_{2t}^k + \eta\nu \\
&= (1 - \eta L)(y_{2t}^k + z_{2t}^k) + \frac{\eta L}{2} y_{2t}^k \geq 0.
\end{aligned}$$

Therefore, we have proven the first inequality of the lemma for Case (b).

Putting the results of Case (a) and (b) together, we have

$$y_{2t+2}^k + z_{2t+2}^k \geq \left(1 - \frac{\eta L}{2}\right) (1 - \eta L) \left(y_{2t}^k + z_{2t}^k\right) + \frac{\eta^2 L \nu}{2}$$

for any $0 \leq t \leq \frac{n}{2} - 1$ and any $k \in \{1, \cdots, K\}$, proving the first part of the lemma.

It now remains to prove the second part, namely that

$$y_n^k + z_n^k \geq \frac{\eta\nu}{3 - \eta L}$$

holds if $y_0^k + z_0^k \geq \frac{\eta\nu}{3 - \eta L}$. From the first part of the lemma, we can see that the updates over a single epoch can be bounded as

$$\begin{aligned}
y_n^k + z_n^k &\geq \left(1 - \frac{\eta L}{2}\right) (1 - \eta L) \left(y_{n-2}^k + z_{n-2}^k\right) + \frac{\eta^2 L \nu}{2} \\
&\vdots \\
&\geq \left(1 - \frac{\eta L}{2}\right)^{\frac{n}{2}} (1 - \eta L)^{\frac{n}{2}} \left(y_0^k + z_0^k\right) + \frac{\eta^2 L \nu}{2} \cdot \sum_{i=0}^{\frac{n}{2}-1} \left(1 - \frac{\eta L}{2}\right)^i (1 - \eta L)^i \\
&= \left(1 - \frac{\eta L}{2}\right)^{\frac{n}{2}} (1 - \eta L)^{\frac{n}{2}} \left(y_0^k + z_0^k\right) + \frac{\eta^2 L \nu}{2} \cdot \frac{1 - \left(1 - \frac{\eta L}{2}\right)^{\frac{n}{2}} (1 - \eta L)^{\frac{n}{2}}}{1 - \left(1 - \frac{\eta L}{2}\right) (1 - \eta L)}
\end{aligned}$$

$$
\begin{aligned}
&= \left(1-\frac{\eta L}{2}\right)^{\frac{n}{2}} (1-\eta L)^{\frac{n}{2}} \left(y_0^k + z_0^k\right) + \frac{\eta \nu}{3-\eta L}\left(1-\left(1-\frac{\eta L}{2}\right)^{\frac{n}{2}}(1-\eta L)^{\frac{n}{2}}\right) \\
&= \frac{\eta \nu}{3-\eta L} + \left(1-\frac{\eta L}{2}\right)^{\frac{n}{2}} (1-\eta L)^{\frac{n}{2}} \left(y_0^k + z_0^k - \frac{\eta \nu}{3-\eta L}\right) \\
&\geq \frac{\eta \nu}{3-\eta L}.
\end{aligned}
$$

This ends the proof of the lemma. □

Lemma E.2. *For any $t \geq 2$, the following inequality holds:*

$$\left(1-\frac{1}{t}\right)^t > \frac{1}{e}\left(1-\frac{1}{t}\right).$$

Proof.

$$
\begin{aligned}
\left(1+\frac{1}{t-1}\right)^{t-1} < e &\iff \left(\frac{t}{t-1}\right)^{t-1} < e \\
&\iff \left(\frac{t-1}{t}\right)^{t-1} > \frac{1}{e} \iff \left(1-\frac{1}{t}\right)^t > \frac{1}{e}\left(1-\frac{1}{t}\right).
\end{aligned}
$$

□

F. Proof of Theorem 4.5

Here we prove Theorem 4.5, restated below for the sake of readability.

Theorem 4.5. *For any $n \geq 104$, L and μ satisfying $\kappa \geq 8n$, and $K \geq \max\left\{\frac{\kappa^2}{n}, \kappa^{3/2} n^{1/2}\right\}$, there exists a 4-dimensional function $F \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$ and an initialization point $\boldsymbol{x}_0$ such that for any permutation-based SGD with any constant step size η, any weighted average iterate $\hat{\boldsymbol{x}}$ of the form as in Equation (4) satisfies*

$$F(\hat{\boldsymbol{x}}) - F^* = \Omega\left(\frac{L^2 \nu^2}{\mu^3 n^2 K^2}\right).$$

Proof. Similarly as in Appendix B, we define objective functions for four step-size regimes and aggregate the functions to obtain the final lower bound. Here we also assume n is even, where we can easily extend to odd n's by the same reasoning as in Appendix B.

We will prove the following lower bounds for each regime. Here F_j^* is the minimizer of F_j for $j = 1, 2, 3, 4$.

- If $\eta \in \left(0, \frac{1}{2\mu n K}\right)$, there exists a 1-dimensional objective function $F_1(x) \in \mathcal{F}_{\text{PŁ}}(L, \mu, 0, \nu)$ such that any permutation-based SGD with initialization $x_0^1 = \frac{L\nu}{\mu^2 n K}$ satisfies

$$F_1(\hat{x}) - F_1^* = \Omega\left(\frac{L^2 \nu^2}{\mu^3 n^2 K^2}\right).$$

- If $\eta \in \left[\frac{1}{2\mu n K}, \frac{2}{nL}\right]$, there exists a 1-dimensional objective function $F_2(y) \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$ such that any permutation-based SGD with initialization $y_0^1 = \frac{\nu}{60L}$ satisfies

$$F_2(\hat{y}) - F_2^* = \Omega\left(\frac{L^2 \nu^2}{\mu^3 n^2 K^2}\right).$$

- If $\eta \in \left[\frac{2}{nL}, \frac{1}{L}\right]$, there exists a 1-dimensional objective function $F_3(z) \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$ such that any permutation-based SGD with initialization $z_0^1 = \frac{3\nu}{8nL}$ satisfies
$$F_3(\hat{z}) - F_3^* = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right).$$
- If $\eta > \frac{1}{L}$, there exists a 1-dimensional objective function $F_4(w) \in \mathcal{F}_{\text{PŁ}}(2L, \mu, 0, \nu)$ such that any permutation-based SGD with initialization $w_0^1 = \frac{L^{1/2}\nu}{\mu^{3/2}nK}$ satisfies
$$F_4(\hat{w}) - F_4^* = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right).$$

Now we define the 4-dimensional function $F(\boldsymbol{x}) = F(x, y, z, w) = F_1(x) + F_2(y) + F_3(z) + F_4(w)$, where F_1, F_2, F_3, and F_4 are chosen to satisfy the above lower bounds.

In fact, our constructed F_1, F_2, F_3, and F_4 are strongly convex; however, for simplicity of exposition, we stated as a member of a larger class $\mathcal{F}_{\text{PŁ}}$. Following the analyses in Appendix B, F is a $2L$-*smooth* and μ-*strongly convex* function.

Also, if four functions H_1, H_2, H_3, and H_4 (each with n components $h_{1,i}$, $h_{2,i}$, $h_{3,i}$, and $h_{4,i}$) satisfies Assumption 2.4 for $\tau = \tau_0$ and $\nu = \nu_0$, then $H(\boldsymbol{x}) = H_1(x) + H_2(y) + H_3(z) + H_4(w)$ satisfies

$$\begin{aligned}
\|\nabla h_i(\boldsymbol{x}) - \nabla H(\boldsymbol{x})\|^2 &= \|\nabla h_{1,i}(x) - \nabla H_1(x)\|^2 + \|\nabla h_{2,i}(y) - \nabla H_2(y)\|^2 \\
&\quad + \|\nabla h_{3,i}(z) - \nabla H_3(z)\|^2 + \|\nabla h_{4,i}(w) - \nabla H_4(w)\|^2 \\
&\leq (\tau_0 \|H_1(x)\| + \nu_0)^2 + (\tau_0 \|H_2(y)\| + \nu_0)^2 + (\tau_0 \|H_3(z)\| + \nu_0)^2 + (\tau_0 \|H_4(w)\| + \nu_0)^2 \\
&\leq (2\tau_0^2 \|H_1(x)\|^2 + 2\nu_0^2) + (2\tau_0^2 \|H_2(y)\|^2 + 2\nu_0^2) \\
&\quad + (2\tau_0^2 \|H_3(z)\|^2 + 2\nu_0^2) + (2\tau_0^2 \|H_4(w)\|^2 + 2\nu_0^2) \\
&\leq 2\tau_0^2 \left(\|H_1(x)\|^2 + \|H_2(y)\|^2 + \|H_3(z)\|^2 + \|H_4(w)\|^2\right) + 8\nu_0^2 \\
&= 2\tau_0^2 \|H(\boldsymbol{x})\|^2 + 8\nu_0^2 < (2\tau_0 \|H(\boldsymbol{x})\| + 3\nu_0)^2.
\end{aligned}$$

for all $i = 1, \dots, n$, i.e., $H(\boldsymbol{x})$ satisfies Assumption 2.4 for $\tau = 2\tau_0$ and $\nu = 3\nu_0$. Combining these results, we obtain $F \in \mathcal{F}_{\text{PŁ}}\left(2L, \mu, \frac{2L}{\mu}, 3\nu\right)$, which allows us to directly apply the convergence rates in the lower bounds of F_1, F_2, F_3, and F_4 to the aggregated function F.

Note that we assumed $\kappa \geq 8n$ and our constructed function is $2L$-*smooth* and μ-*strongly convex*. Thus, $\kappa \geq 8n$ is equivalent to $\frac{L}{\mu} \geq 4n$ throughout the proof. Also, combining $K \geq \max\left\{\frac{\kappa^2}{n}, \kappa^{3/2}n^{1/2}\right\}$ and $\kappa \geq 8n$, we have $K \geq \frac{\kappa^2}{n} \geq \kappa$ and thus $\frac{1}{2\mu nK} \leq \frac{2}{nL}$ holds so all step size regimes are valid.

Finally, rescaling L and ν will give us the function $F \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$ satisfying $F(\hat{\boldsymbol{x}}) - F^* = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right)$. □

For the following subsections, we prove the lower bounds for F_1, F_2, F_3, and F_4 at the corresponding step size regime.

F.1. Lower Bound for $\eta \in \left(0, \frac{1}{2\mu nK}\right)$

Here we show that there exists $F_1(x) \in \mathcal{F}_{\text{PŁ}}(L, \mu, 0, \nu)$ such that any permutation-based SGD with $x_0^1 = \frac{L\nu}{\mu^2 nK}$ satisfies
$$F_1(\hat{x}) - F_1^* = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right).$$

Proof. We define $F_1(x) \in \mathcal{F}_{\text{PŁ}}(\mu, \mu, 0, 0)$ by the following components:
$$f_i(x) = F_1(x) = \frac{\mu}{2}x^2.$$

Note that $\mathcal{F}_{\text{PŁ}}(\mu, \mu, 0, 0) \subseteq \mathcal{F}_{\text{PŁ}}(L, \mu, 0, \nu)$ and $F_1^* = 0$ at $x^* = 0$ by definition.

In this regime, we will see that the step size is too small so that $\left\{x_n^k\right\}_{k=1}^K$ cannot even reach near the optimal point. We start from $x_0^1 = \frac{L\nu}{\mu^2 nK}$. Since the gradient of all component functions evaluated at point x is fixed deterministically to μx, regardless of the permutation-based SGD algorithm we use, we have

$$\begin{aligned} x_n^k = x_0^1(1-\eta\mu)^{nk} &\geq \frac{L\nu}{\mu^2 nK}\left(1 - \frac{1}{2nK}\right)^{nk} \geq \frac{L\nu}{\mu^2 nK}\left(1 - \frac{1}{2nK}\right)^{nK} \\ &> \frac{L\nu}{\mu^2 nK}\left(1 - \frac{1}{nK}\right)^{nK} \overset{(a)}{>} \frac{L\nu}{\mu^2 nK}\frac{1}{e}\left(1 - \frac{1}{nK}\right) \overset{(b)}{\geq} \frac{L\nu}{\mu^2 nK}\frac{1}{2e}. \end{aligned}$$

where (a) comes from Lemma E.2 and (b) comes from assumption that $n \geq 2$. Therefore, we have $\hat{x} = \Omega\left(\frac{L\nu}{\mu^2 nK}\right)$ for any nonnegative weights $\{\alpha_k\}_{k=1}^{K+1}$. With this $\hat{x}$, we have

$$F_1(\hat{x}) - F_1^* = \frac{\mu}{2}\hat{x}^2 = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right).$$

□

F.2. Lower Bound for $\eta \in \left[\frac{1}{2\mu nK}, \frac{2}{nL}\right]$

Here we show that there exists $F_2(y) \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$ such that any permutation-based SGD with $y_0^1 = \frac{\nu}{60L}$ satisfies

$$F_2(\hat{y}) - F_2^* = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right).$$

Proof. We define $F_2(y)$ by the following components:

$$f_i(y) = \begin{cases} g_1(y) & \text{if } i \leq \frac{n}{2}, \\ g_2(y) & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned} g_1(y) &= \frac{L}{2}y^2 - \nu y, \\ g_2(y) &= -\frac{L}{2}\left(1 - \frac{2\mu}{L}\right)y^2 + \nu y. \end{aligned}$$

With this construction, the finite-sum objective becomes

$$F_2(y) = \frac{1}{n}\sum_{i=1}^{n} f_i(y) = \frac{\mu}{2}y^2.$$

Note that all the components are *L-smooth* and F is *μ-strongly convex*. Moreover,

$$\begin{aligned} \|\nabla f_1(y) - \nabla F_2(y)\| &= \|(Ly - \nu) - \mu y\| \leq \|(L-\mu)y\| + \nu \\ &\leq L\|y\| + \nu = \frac{L}{\mu}\|\nabla F_2(y)\| + \nu, \\ \left\|\nabla f_{\frac{n}{2}+1}(y) - \nabla F_2(y)\right\| &= \left\|\left(-L\left(1 - \frac{2\mu}{L}\right)y + \nu\right) - \mu y\right\| \leq \|(L-\mu)y\| + \nu \\ &\leq L\|y\| + \nu = \frac{L}{\mu}\|\nabla F_2(y)\| + \nu, \end{aligned}$$

and thereby $F_2 \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$. Also, we can easily verify $F_2^* = 0$ at $y^* = 0$.

To simplify notation, we will write $\nabla f_i(y) = a_i y - b_i$ temporarily. Then, $a_i \in \left\{L, -L\left(1-\frac{2\mu}{L}\right)\right\}$, $b_i \in \{\nu, -\nu\}$ holds and we can write y_n^k as

$$\begin{aligned}
y_1^k &= y_0^k - \eta \nabla f_{\sigma_k(1)}\left(y_0^k\right) = \left(1-\eta a_{\sigma_k(1)}\right) y_0^k + \eta b_{\sigma_k(1)}, \\
y_2^k &= y_1^k - \eta \nabla f_{\sigma_k(2)}\left(y_1^k\right) = \left(1-\eta a_{\sigma_k(2)}\right) y_1^k + \eta b_{\sigma_k(2)} \\
&= \left(1-\eta a_{\sigma_k(2)}\right)\left(1-\eta a_{\sigma_k(1)}\right) y_0^k + \eta b_{\sigma_k(2)} + \eta b_{\sigma_k(1)}\left(1-\eta a_{\sigma_k(2)}\right), \\
&\vdots \\
y_n^k &= y_{n-1}^k - \eta \nabla f_{\sigma_k(n)}\left(y_{n-1}^k\right) = \prod_{i=1}^n \left(1-\eta a_{\sigma_k(i)}\right) y_0^k + \eta \sum_{i=1}^n b_{\sigma_k(i)} \prod_{j=i+1}^n \left(1-\eta a_{\sigma_k(j)}\right). \qquad (29)
\end{aligned}$$

Define $S := \prod_{i=1}^n \left(1-\eta a_{\sigma_k(i)}\right) = \prod_{i=1}^n (1-\eta a_i)$ and $A_\sigma := \eta \sum_{i=1}^n b_{\sigma(i)} \prod_{j=i+1}^n \left(1-\eta a_{\sigma(j)}\right)$. Then, we can write Equation (29) as $y_n^k = S y_0^k + A_\sigma$. Note that S is independent of the choice of σ_k and A_σ is the term that we can control using permutation-based SGD.

We now consider which permutation σ minimizes A_σ. Choose an arbitrary σ and assume there exists $t \in \{1, \cdots, n-1\}$ such that $f_{\sigma(t)} = g_2$ and $f_{\sigma(t+1)} = g_1$. Then, define another permutation σ' by $\sigma'(t) = \sigma(t+1)$, $\sigma'(t+1) = \sigma(t)$ and $\sigma'(i) = \sigma(i)$ for $i \in \{1, \cdots, n\} \setminus \{t, t+1\}$.

Let y_σ and $y_{\sigma'}$ as the value of y_n^k generated by σ and σ' starting from the same y_0^k, respectively. Since $b_{\sigma(i)} \prod_{j=i+1}^n \left(1-\eta a_{\sigma(j)}\right) = b_{\sigma'(i)} \prod_{j=i+1}^n \left(1-\eta a_{\sigma'(j)}\right)$ for $i \in \{1, \cdots, n\} \setminus \{t, t+1\}$, we have

$$\begin{aligned}
y_\sigma - y_{\sigma'} &= \prod_{i=1}^n \left(1-\eta a_{\sigma(i)}\right) y_0^k + \eta \sum_{i=1}^n b_{\sigma(i)} \prod_{j=i+1}^n \left(1-\eta a_{\sigma(j)}\right) \\
&\quad - \prod_{i=1}^n \left(1-\eta a_{\sigma'(i)}\right) y_0^k + \eta \sum_{i=1}^n b_{\sigma'(i)} \prod_{j=i+1}^n \left(1-\eta a_{\sigma'(j)}\right) \\
&= \eta \Bigg(b_{\sigma(t)} \prod_{j=t+1}^n \left(1-\eta a_{\sigma(j)}\right) + b_{\sigma(t+1)} \prod_{j=t+2}^n \left(1-\eta a_{\sigma(j)}\right) \\
&\qquad - b_{\sigma'(t)} \prod_{j=t+1}^n \left(1-\eta a_{\sigma'(j)}\right) - b_{\sigma'(t+1)} \prod_{j=t+2}^n \left(1-\eta a_{\sigma'(j)}\right) \Bigg) \\
&= \left(\eta \prod_{j=t+2}^n \left(1-\eta a_{\sigma(j)}\right)\right) \cdot \left(-\nu(1-\eta L) + \nu - \nu\left(1+\eta L\left(1-\frac{2\mu}{L}\right)\right) - (-\nu)\right) \\
&= \left(\eta \prod_{j=t+2}^n \left(1-\eta a_{\sigma(j)}\right)\right) \cdot 2\eta\mu\nu > 0. \qquad (30)
\end{aligned}$$

Thereby, we can conclude that the permutation σ that minimizes A_σ should satisfy $\sigma(i) \le n/2$ for $i \le n/2$ and $\sigma(i) > n/2$ for $i > n/2$, i.e., $f_{\sigma(i)} = g_1$ for $i \le n/2$ and $f_{\sigma(i)} = g_2$ for $i > n/2$. Let σ^* denote such σ.

With this permutation σ^*, A_{σ^*} becomes

$$A_{\sigma^*} = \eta\nu \cdot \left(1+\eta L\left(1-\frac{2\mu}{L}\right)\right)^{\frac{n}{2}} \sum_{i=0}^{\frac{n}{2}-1} (1-\eta L)^i - \eta\nu \cdot \sum_{i=0}^{\frac{n}{2}-1} \left(1+\eta L\left(1-\frac{2\mu}{L}\right)\right)^i.$$

Here, we introduce $\beta := 1 - \frac{2\mu}{L}$ and $m := \frac{n}{2}$ to simplify notation a bit. Note that $\beta \ge 1 - \frac{1}{m}$ holds since we assumed $\frac{L}{\mu} > n$. Then A_{σ^*} can be rearranged as

$$A_{\sigma^*} = \eta\nu \cdot (1+\eta L\beta)^m \frac{1-(1-\eta L)^m}{\eta L} - \eta\nu \cdot \frac{(1+\eta L\beta)^m - 1}{\eta L\beta}$$

$$= \frac{\nu}{L\beta} \cdot \left((1+\eta L\beta)^m(\beta-1) - \beta(1+\eta L\beta)^m(1-\eta L)^m + 1\right).$$

Using Lemma F.1 (substituting ηL to x), we have $\frac{\nu}{L\beta} \cdot \left((1+\eta L\beta)^m(\beta-1) - \beta(1+\eta L\beta)^m(1-\eta L)^m + 1\right) \geq \frac{\eta^2 m L\nu}{30}$. We now show a lower bound for S.

$$\begin{aligned}
S &= \prod_{i=1}^{n}(1-\eta a_i) \\
&= (1-\eta L)^m\,(1+\eta L\beta)^m \\
&= (1-\eta L(1-\beta) - \eta^2L^2\beta)^m \\
&= \left(1 - \eta L \cdot \frac{2\mu}{L} - \eta^2 L^2\left(1 - \frac{2\mu}{L}\right)\right)^m \\
&> (1 - 2\eta\mu - \eta^2L^2)^m \\
&\geq \begin{cases} (1-4\eta\mu)^m > 1 - 4\eta m\mu, & \text{(if } \frac{1}{2\mu nK} \leq \eta < \frac{2\mu}{L^2}\text{)} \\ (1-2\eta^2L^2)^m > 1 - 2\eta^2 m L^2. & \text{(if } \frac{2\mu}{L^2} \leq \eta \leq \frac{2}{nL}\text{)} \end{cases}
\end{aligned}$$

We start at $y_0^1 = \frac{\nu}{60L}$. Being aware of $\kappa = \frac{2L}{\mu}$ in the construction, we first verify that

$$\frac{\nu}{60L} = \frac{\nu}{60L} \cdot \frac{K}{K} \geq \frac{\nu}{60L} \cdot \frac{\kappa^2}{nK} = \frac{4L\nu}{60\mu^2} \cdot \frac{1}{nK} > \frac{L\nu}{240\mu^2 nK}.$$

For the case when $\frac{1}{2\mu nK} \leq \eta < \frac{2\mu}{L^2}$,

$$\begin{aligned}
y_n^1 &= Sy_0^1 + A_{\sigma_1} \\
&\geq (1-4\eta m\mu)\frac{L\nu}{240\mu^2 nK} + \frac{\eta^2 m L\nu}{30} \\
&= \frac{L\nu}{240\mu^2 nK} - \frac{\eta L\nu}{120\mu K} + \frac{\eta^2 nL\nu}{60} \qquad (\because n = 2m) \\
&= \frac{L\nu}{240\mu^2 nK} - \frac{\eta nL\nu}{60}\left(\frac{1}{2\mu nK} - \eta\right) \\
&\geq \frac{L\nu}{240\mu^2 nK}.
\end{aligned}$$

Applying this process in a chain, we then gain $y_n^k \geq \frac{L\nu}{240\mu^2 nK}$ for all $k \in \{1, \cdots, K\}$. Therefore, regardless of the choice of $\{\alpha_k\}_{k=1}^{K+1}$, $\hat{y} = \Omega\left(\frac{L\nu}{\mu^2 nK}\right)$ holds and $F_2(\hat{y}) - F_2^* = \frac{\mu}{2}\hat{y}^2 = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right)$.

For the case when $\frac{2\mu}{L^2} \leq \eta \leq \frac{2}{nL}$, we have

$$\begin{aligned}
y_n^1 &= Sy_0^1 + A_{\sigma_1} \\
&\geq (1 - 2\eta^2 m L^2)\frac{\nu}{60L} + \frac{\eta^2 m L\nu}{30} \\
&= \frac{\nu}{60L}.
\end{aligned}$$

Applying this process in a chain, we then gain $y_n^k \geq \frac{\nu}{60L}$ for all $k \in \{1, \cdots, K\}$. Therefore, regardless of the choice of $\{\alpha_k\}_{k=1}^{K+1}$, $\hat{y} = \Omega\left(\frac{\nu}{L}\right)$ holds and $F_2(\hat{y}) - F_2^* = \frac{\mu}{2}\hat{y}^2 = \Omega\left(\frac{\mu\nu^2}{L^2}\right) = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right)$, where we used $K \geq \frac{\kappa^2}{n}$ in the last step. □

F.3. Lower Bound for $\eta \in \left[\frac{2}{nL}, \frac{1}{L}\right]$

Here we show that there exists $F_3(z) \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$ such that any permutation-based SGD with $z_0^1 = \frac{3\nu}{8nL}$ satisfies

$$F_3(\hat{z}) - F_3^* = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right).$$

Proof. We define $F_3(z)$ by the following components:

$$f_i(z) = \begin{cases} \frac{L}{2}z^2 - \nu z & \text{if } i = 1, \\ -\frac{L}{4(n-1)}z^2 + \frac{\nu}{n-1}z & \text{otherwise.} \end{cases}$$

With this construction, the finite-sum objective becomes

$$F_3(z) = \frac{1}{n}\sum_{i=1}^{n} f_i(z) = \frac{L}{4n}z^2.$$

Note that all the components are *L-smooth* and F is *μ-strongly convex* since we assumed $\frac{L}{4n} \geq \mu$. Moreover,

$$\begin{aligned} \|\nabla f_1(z) - \nabla F_3(z)\| &= \left\|(Lz - \nu) - \frac{Lz}{2n}\right\| \leq \left\|\left(1 - \frac{1}{2n}\right)Lz\right\| + \nu \\ &\leq \|Lz\| + \nu = 2n\|\nabla F_3(z)\| + \nu \leq \frac{L}{\mu}\|\nabla F_3(z)\| + \nu, \\ \|\nabla f_2(z) - \nabla F_3(z)\| &= \left\|\left(-\frac{Lz}{2(n-1)} + \frac{\nu}{n-1}\right) - \frac{Lz}{2n}\right\| \\ &< \|Lz\| + \nu = 2n\|\nabla F_3(z)\| + \nu \leq \frac{L}{\mu}\|\nabla F_3(z)\| + \nu, \end{aligned}$$

and thereby $F_3 \in \mathcal{F}_{\text{PŁ}}\left(L, \mu, \frac{L}{\mu}, \nu\right)$. Also, we can easily verify $F_3^* = 0$ at $z^* = 0$.

Similarly as in (29), we temporarily write $\nabla f_i(y) = a_i y - b_i$ where $a_i \in \left\{L, -\frac{L}{2(n-1)}\right\}$, $b_i \in \left\{\nu, -\frac{\nu}{n-1}\right\}$ holds. We then write y_n^k as $Sy_0^k + A_{\sigma_k}$, where $S := \prod_{i=1}^n \left(1 - \eta a_{\sigma_k(i)}\right) = \prod_{i=1}^n \left(1 - \eta a_i\right)$ is independent of the choice of σ_k, and $A_\sigma := \eta\sum_{i=1}^n b_{\sigma(i)}\prod_{j=i+1}^n \left(1 - \eta a_{\sigma(j)}\right)$ is the term that we can control using permutation-based SGD.

We will first find what permutation σ leads to the smallest A_σ. Choose arbitrary σ and assume that $\sigma(1) \neq 1$. Define $t := \sigma^{-1}(1)$. We then define another permutation σ' by $\sigma'(t-1) = 1$, $\sigma'(t) = \sigma(t-1)$ and $\sigma'(i) = \sigma(i)$ for $i \in \{1, \cdots, n\} \setminus \{t-1, t\}$.

Let z_σ and $z_{\sigma'}$ as the value of z_n^k generated by σ and σ' starting from the same z_0^k, respectively. We will show that $z_{\sigma_1} > z_{\sigma_1'}$. In a similar manner as (30),

$$\begin{aligned} z_\sigma - z_{\sigma'} &= Sz_0^k + A_\sigma - Sz_0^k - A_{\sigma'} \\ &= \left(\eta\prod_{j=t+1}^n\left(1 + \frac{\eta L}{2(n-1)}\right)\right) \cdot \left(\left(-\frac{\nu}{n-1}(1 - \eta L) + \nu\right) - \left(\nu\left(1 + \frac{\eta L}{2(n-1)}\right) - \frac{\nu}{n-1}\right)\right) \\ &= \left(\eta\prod_{j=t+1}^n\left(1 + \frac{\eta L}{2(n-1)}\right)\right) \cdot \left(\frac{\eta L\nu}{2(n-1)}\right) > 0 \end{aligned}$$

holds. Thus, we can conclude that the permutation σ satisfying $\sigma(1) = 1$ is the permutation that minimizes A_σ. Let σ^* denote such σ.

With this permutation σ^*, A_{σ^*} becomes

$$A_{\sigma^*} = \eta\nu\left(1 + \frac{\eta L}{2(n-1)}\right)^{n-1} - \frac{\eta\nu}{n-1}\sum_{i=0}^{n-2}\left(1 + \frac{\eta L}{2(n-1)}\right)^i$$

$$\begin{aligned}
&= \eta\nu\left(1+\frac{\eta L}{2(n-1)}\right)^{n-1} - \frac{\eta\nu}{n-1}\cdot\frac{\left(1+\eta L/(2(n-1))\right)^{n-1}-1}{\eta L/(2(n-1))}\\
&= \frac{2\nu}{L} - \left(1+\frac{\eta L}{2(n-1)}\right)^{n-1}\left(\frac{2\nu}{L}-\eta\nu\right). \qquad (31)
\end{aligned}$$

Note that $\eta L \le 1$, so $\frac{2\nu}{L} - \eta\nu$ is nonnegative. Using Lemma F.4, we have

$$\left(1+\frac{\eta L}{2(n-1)}\right)^{n-1} < e^{\frac{\eta L}{2}} < 1 + \frac{\eta L}{2} + \frac{5\eta^2L^2}{32}. \qquad (32)$$

Substituting (32) to (31) results

$$\begin{aligned}
A_{\sigma^*} &> \frac{2\nu}{L} - \left(1+\frac{\eta L}{2}+\frac{5\eta^2L^2}{32}\right)\left(\frac{2\nu}{L}-\eta\nu\right)\\
&= \frac{3\eta^2L\nu}{16} + \frac{5\eta^3L^2\nu}{32}\\
&> \frac{3\eta^2L\nu}{16}.
\end{aligned}$$

We start at $z_0^1 = \frac{3\nu}{8nL}$. Using $S = (1-\eta L)\left(1+\frac{\eta L}{2(n-1)}\right)^{n-1} > 1-\eta L$, we have

$$\begin{aligned}
z_n^1 &= Sz_0^1 + A_{\sigma^*}\\
&\ge (1-\eta L)z_0^1 + \frac{3\eta^2L\nu}{16}\\
&= (1-\eta L)\frac{3\nu}{8nL} + \frac{3\eta^2L\nu}{16}\\
&= \frac{3\nu}{8nL} - \frac{3\eta\nu}{8n} + \frac{3\eta^2L\nu}{16}\\
&= \frac{3\nu}{8nL} + \frac{3\eta L\nu}{16}\left(\eta - \frac{2}{nL}\right)\\
&\ge \frac{3\nu}{8nL}.
\end{aligned}$$

Applying this process in a chain, we then gain $z_n^k \ge \frac{3\nu}{8nL}$ for all $k \in \{1,\cdots,K\}$. Therefore, regardless of the choice of $\{\alpha_k\}_{k=1}^{K+1}$, $\hat{z} = \Omega\left(\frac{\nu}{nL}\right)$ holds and $F_3(\hat{z}) - F_3^* = \frac{L}{4n}\hat{z}^2 = \Omega\left(\frac{\nu^2}{n^3L}\right) = \Omega\left(\frac{L^2\nu^2}{\mu^3n^2K^2}\right)$, where we used $K \ge \kappa^{3/2}n^{1/2}$ in the last step. □

F.4. Lower Bound for $\eta > \frac{1}{L}$

Here we show that there exists $F_4(w) \in \mathcal{F}_{\text{PŁ}}(2L,\mu,0,\nu)$ such that any permutation-based SGD with $w_0^1 = \frac{L^{1/2}\nu}{\mu^{3/2}nK}$ satisfies

$$F_4(\hat{w}) - F_4^* = \Omega\left(\frac{L^2\nu^2}{\mu^3n^2K^2}\right).$$

Proof. We define $F_4(w) \in \mathcal{F}_{\text{PŁ}}(2L,2L,0,0)$ by the following components:

$$f_i(w) = Lw^2.$$

Note that $\mathcal{F}_{\text{PŁ}}(2L,2L,0,0) \subseteq \mathcal{F}_{\text{PŁ}}(2L,\mu,0,\nu)$ and $F_4^* = 0$ at $w^* = 0$ by definition.

In this regime, we will see that the step size is too large so that $\left\{w_n^k\right\}_{k=1}^K$ diverges. We start from $w_0^1 = \frac{L^{1/2}\nu}{\mu^{3/2}nK}$. Since the gradient of all component functions evaluated at point w is fixed deterministically to $2Lw$, we have for every $k \in [K]$,

$$w_n^k = (1-2\eta L)^{nk}\, w_0^1 \ge 1^{nk}\frac{L^{1/2}\nu}{\mu^{3/2}nK} = \Omega\left(\frac{L^{1/2}\nu}{\mu^{3/2}nK}\right),$$

where we used the fact that n is even in the second step. Thus, regardless of the permutation-based SGD algorithm we use, we have $\hat{w} = \Omega\left(\frac{L^{1/2}\nu}{\mu^{3/2}nK}\right)$ and $F_4(\hat{w}) - F_4^* = L\hat{w}^2 = \Omega\left(\frac{L^2\nu^2}{\mu^3 n^2 K^2}\right)$. □

F.5. Lemmas used in Theorem 4.5

In this subsection, we will prove the lemmas used in Theorem 4.5.

Lemma F.1. *For any even $n \geq 104$, any $0 < x \leq \frac{2}{n}$ and any $1 - \frac{2}{n} \leq \beta < 1$, let $m = \frac{n}{2}$. Then, the following inequality holds:*

$$(1+\beta x)^m(\beta - 1) - \beta(1+\beta x)^m(1-x)^m + 1 \geq \frac{mx^2}{30}. \tag{33}$$

Proof. To prove the lemma, we focus on the coefficients of x^k for $0 \leq k \leq 2m$.

Define a_k as the absolute value of x^k's coefficient in $(1+\beta x)^m(\beta-1)$. Using the fact that $1 - \frac{1}{m} \leq \beta < 1$, $a_k = \left|\binom{m}{k}\beta^k(\beta-1)\right| \leq \frac{m^k}{k!} \cdot \frac{1}{m} = \frac{m^{k-1}}{k!}$. Note that for $k \geq m+1$, a_k is 0.

Let b_k be x^k's coefficient in $\beta(1+\beta x)^m(1-x)^m$. While the sequence of coefficients $\{b_k\}$ have alternating signs, we can define a *positive* sequence c_k which upper bounds the sequence $|b_k|$. Since $(1+\beta x)^m(1-x)^m = (1-(1-\beta)x - \beta x^2)^m$,

$$\begin{aligned}
|b_k| &= \beta \cdot \left| \sum_{t=\max\{0,k-m\}}^{\lfloor \frac{k}{2} \rfloor} (-\beta)^t(-(1-\beta))^{k-2t} \frac{m!}{t!(k-2t)!(m-k+t)!} \right| \\
&\leq 1 \cdot \sum_{t=\max\{0,k-m\}}^{\lfloor \frac{k}{2} \rfloor} \beta^t (1-\beta)^{k-2t} \frac{m!}{t!(k-2t)!(m-k+t)!} \\
&\triangleq c_k
\end{aligned}$$

Then x^k's coefficient in LHS of Equation (33) is lower bounded by $-(a_k + c_k)$. For even $k < m$, we have

$$\begin{aligned}
\frac{c_{k+1}}{c_k} &\leq (1-\beta) \max_{t \leq \lfloor \frac{k}{2} \rfloor} \frac{m!/(t!(k+1-2t)!(m-k-1+t)!)}{m!/(t!(k-2t)!(m-k+t)!)} \\
&\leq \frac{1}{m} \max_{t \leq \lfloor \frac{k}{2} \rfloor} \frac{m-k+t}{k+1-2t} \\
&\leq \frac{1}{m} \max_{t \leq \lfloor \frac{k}{2} \rfloor} (m-k+t) \\
&\leq \frac{1}{m} \cdot m \\
&= 1.
\end{aligned} \tag{34}$$

For odd $k < m$, we have

$$\begin{aligned}
c_{k+1} &= \beta^{\frac{k+1}{2}} \frac{m!}{(\frac{k+1}{2})!(m - \frac{k+1}{2})!} + \sum_{t=0}^{\frac{k-1}{2}} \beta^t (1-\beta)^{k+1-2t} \frac{m!}{t!(k+1-2t)!(m-k-1+t)!} \\
&< 1^{\frac{k+1}{2}} \cdot \frac{m^{\frac{k+1}{2}}}{(\frac{k+1}{2})!} + c_k \cdot (1-\beta) \max_{t \leq \lfloor \frac{k}{2} \rfloor} \frac{m!/(t!(k+1-2t)!(m-k-1+t)!)}{m!/(t!(k-2t)!(m-k+t)!)} \\
&\leq \frac{m^{\frac{k+1}{2}}}{(\frac{k+1}{2})!} + c_k.
\end{aligned} \tag{35}$$

For $k \geq m$, we have

$$\frac{c_{k+1}}{c_k} \leq (1-\beta) \max_{t \leq \lfloor \frac{k}{2} \rfloor} \frac{m!/(t!(k+1-2t)!(m-k-1+t)!)}{m!/(t!(k-2t)!(m-k+t)!)}$$

$$\begin{aligned}
&\leq \frac{1}{m} \max_{t \leq \lfloor \frac{k}{2} \rfloor} \frac{m-k+t}{k+1-2t} \\
&\leq \frac{1}{m} \max_{t \leq \lfloor \frac{k}{2} \rfloor} (m-k+t) \\
&\leq \frac{1}{m} \cdot \left(m - \frac{k}{2}\right) \\
&\leq \frac{1}{m} \cdot \frac{m}{2} \\
&= \frac{1}{2}. \qquad (36)
\end{aligned}$$

Using (34), (35) and (36), we will show $c_k \leq \frac{m^{k-1}}{k!}$ for $4 \leq k \leq 2m$.

Note that $c_1 = (1-\beta) \cdot \frac{m!}{(m-1)!} \leq 1$. Also, we can easily prove $\sum_{i=0}^{p} \frac{m^i}{i!} \leq \frac{m^p}{(p-1)!}$ for $\forall m \geq 3$, $\forall 2 \leq p \leq m-1$ using mathematical induction. Therefore, for $k \leq m$,

$$\begin{cases} c_k \leq \sum_{i=0}^{\frac{k-1}{2}} \frac{m^i}{i!} \leq \frac{m^{\frac{k-1}{2}}}{(\frac{k-1}{2}-1)!} & \text{if } k \text{ is odd,} \\ c_k \leq \sum_{i=0}^{\frac{k}{2}} \frac{m^i}{i!} \leq \frac{m^{\frac{k}{2}}}{(\frac{k}{2}-1)!} & \text{if } k \text{ is even,} \end{cases}$$

and applying Lemma F.2 and Lemma F.3, we finally get $c_k \leq \frac{m^{k-1}}{k!}$.

For $k > m$,

$$\begin{aligned}
c_k &\leq c_m \cdot \left(\frac{1}{2}\right)^{k-m} \\
&< \frac{m^{m-1}}{m!} \cdot \frac{m}{m+1} \cdot \frac{m}{m+2} \cdot \dots \cdot \frac{m}{k} \\
&= \frac{m^{k-1}}{k!}.
\end{aligned}$$

Thus, we have proven $c_k \leq \frac{m^{k-1}}{k!}$ for $4 \leq k \leq 2m$. Since we also have $a_k \leq \frac{m^{k-1}}{k!}$, we can conclude that $a_k + c_k \leq \frac{2 \cdot m^{k-1}}{k!}$, i.e., the absolute value of the x^k's coefficient of LHS of our statement is upper bounded by $\frac{2 \cdot m^{k-1}}{k!}$ when $4 \leq k \leq 2m$.

We now consider the coefficient of x^k when $k < 4$. For $k = 0$, the coefficient is

$$(\beta - 1) - \beta \cdot 1 + 1 = 0.$$

For $k = 1$, the coefficient is

$$\beta m(\beta - 1) - \beta(\beta m - m) = 0.$$

For $k = 2$, the coefficient is

$$\beta^2 \cdot \binom{m}{2} \cdot (\beta - 1) - \beta \cdot \left((1-\beta)^2 \cdot \binom{m}{2} - \beta \cdot m\right) = \beta^2 \cdot \frac{m(m+1)}{2} - \beta \cdot \frac{m(m-1)}{2}.$$

For fixed m, RHS is a quadratic with respect to β, and it is minimized when β is $1 - \frac{1}{m}$. Hence the above equation can be lower bounded by

$$\begin{aligned}
&\left(1 - \frac{1}{m}\right)^2 \cdot \frac{m(m+1)}{2} - \left(1 - \frac{1}{m}\right) \cdot \frac{m(m-1)}{2} \\
&= \frac{m}{2} - 1 + \frac{1}{2m} \\
&\geq \frac{2m}{5}. \qquad (m \geq 10). \qquad (37)
\end{aligned}$$

For $k = 3$, the coefficient is

$$\begin{aligned}
&\beta^3 \cdot \binom{m}{3} \cdot (\beta - 1) - \beta \cdot \left(-\binom{m}{3} + m \cdot \beta \cdot \binom{m}{2} - \binom{m}{2} \cdot \beta^2 \cdot m + \binom{m}{3} \cdot \beta^3 \right) \\
&= \frac{\beta}{6} \cdot m(m-1)(m-2) - \frac{\beta^2}{2} \cdot m^2(m-1) + \frac{\beta^3}{3} \cdot (m+1)m(m-1). \qquad (38)
\end{aligned}$$

For fixed m, (38) is a cubic function with respect to β. Differentiating this function, we get

$$\begin{aligned}
&\beta^2(m+1)m(m-1) - \beta m^2(m-1) + \frac{m(m-1)(m-2)}{6} \\
&= m(m-1) \cdot \left(\beta^2(m+1) - \beta m + \frac{m-2}{6} \right) \\
&= m(m-1) \cdot \left(\beta m(\beta - 1) + \beta^2 + \frac{m-2}{6} \right) \\
&\geq m(m-1) \cdot \left(-\beta + \beta^2 + \frac{m-2}{6} \right) && (\because \beta - 1 \geq -\frac{1}{m}) \\
&= m(m-1) \cdot \left(\beta(\beta - 1) + \frac{m-2}{6} \right) \\
&\geq m(m-1) \cdot \left(-\frac{1}{m} + \frac{m-2}{6} \right) && (\because \beta \leq 1 \ \& \ \beta - 1 \geq -\frac{1}{m}) \\
&> 0. && (\because m \geq 4)
\end{aligned}$$

Thereby, (38) is minimized when β is $1 - \frac{1}{m}$, and substituting such β to (38) results

$$\begin{aligned}
&\frac{1 - \frac{1}{m}}{6} \cdot m(m-1)(m-2) - \frac{\left(1 - \frac{1}{m}\right)^2}{2} \cdot m^2(m-1) + \frac{\left(1 - \frac{1}{m}\right)^3}{3} \cdot (m+1)m(m-1) \\
&= -\frac{m^2}{6} - \frac{1}{m} + \frac{1}{3m^2} + \frac{5}{6} \\
&\geq -\frac{m^2}{6}. \qquad (39)
\end{aligned}$$

Remind that x^k's coefficient in LHS of Equation (33) is lower bounded by $-(a_k + c_k)$. Summing up (37), (39), and the fact that $a_k + c_k \leq \frac{2 \cdot m^{k-1}}{k!}$ for $k \geq 4$, we obtain

$$\begin{aligned}
&(1 + \beta x)^m(\beta - 1) - \beta(1 + \beta x)^m(1 - x)^m + 1 \\
&\geq \frac{2m}{5}x^2 - \frac{m^2}{6}x^3 - \sum_{k=4}^{2m} x^k \cdot \frac{2 \cdot m^{k-1}}{k!} \\
&> \frac{2}{5}mx^2 - \frac{mx}{6}mx^2 - \sum_{k=4}^{\infty} x^k \cdot \frac{2 \cdot m^{k-1}}{k!} \\
&\geq \frac{2}{5}mx^2 - \frac{1}{6}mx^2 - mx^2 \cdot \frac{2}{m^2x^2} \cdot \sum_{k=4}^{\infty} \frac{(mx)^k}{k!} && (\because mx \leq 1).
\end{aligned}$$

For the last term, $\frac{1}{m^2x^2} \cdot \sum_{k=4}^{\infty} \frac{(mx)^k}{k!}$ is an increasing function of mx so it is maximized when mx is 1. Thereby we can further extend the above inequality as:

$$\begin{aligned}
&\frac{2}{5}mx^2 - \frac{1}{6}mx^2 - mx^2 \cdot 2\left(e - 1 - \frac{1}{1!} - \frac{1}{2!} - \frac{1}{3!} \right) \\
&\geq \frac{2}{5}mx^2 - \frac{1}{6}mx^2 - \frac{1}{5}mx^2 \\
&= \frac{1}{30}mx^2.
\end{aligned}$$

□

Lemma F.2. *For $m \geq 52$ and even $4 \leq k \leq m$, $\frac{m^{k-1}}{k!} > \frac{m^{\frac{k}{2}}}{\left(\frac{k}{2}-1\right)!}$ holds.*

Proof. We first consider the case when $k \geq 14$. Since $m \geq k$, it is sufficient to show $m^{\frac{k}{2}-2} > \frac{(k-1)!}{\left(\frac{k}{2}-1\right)!}$. Taking log on both sides, this inequality becomes

$$\left(\frac{k}{2}-2\right)\log m > \sum_{i=\frac{k}{2}}^{k-1} \log i.$$

Using $\sum_{i=\frac{k}{2}}^{k-1} \log i < \int_{\frac{k}{2}}^{k} \log x \, dx$, we will instead prove following inequality when $k \geq 16$:

$$\log m > \frac{\int_{\frac{k}{2}}^{k} \log x \, dx}{\frac{k}{2}-2} = \frac{k \log k - \frac{k}{2}\log\frac{k}{2} - \frac{k}{2}}{\frac{k}{2}-2}.$$

Define $f(X) := \frac{2X\log(2X) - X\log X - X}{X-2} = \frac{X\log X + 2X\log 2 - X}{X-2}$. Then,

$$\begin{aligned} f'(X) &= \left(\frac{X\log X + 2X\log 2 - X}{X-2}\right)' \\ &= \frac{X - 2\log X - 4\log 2}{(X-2)^2}. \end{aligned}$$

We can numerically check that $f'(\frac{k}{2}) > 0$ holds for $k \geq 14$. Therefore, for fixed m, $\arg\max_{k\geq 14} f\left(\frac{k}{2}\right) = 2\lfloor\frac{m}{2}\rfloor$. We now have to prove $\log m > f\left(\lfloor\frac{m}{2}\rfloor\right)$. Let $s = \lfloor\frac{m}{2}\rfloor$. Then $f\left(\lfloor\frac{m}{2}\rfloor\right)$ becomes

$$f(s) = \frac{s\log s + 2s\log 2 - s}{s-2}.$$

Combining $\log m \geq \log(2s)$ and

$$\begin{aligned} &\log(2s) \geq \frac{s\log s + 2s\log 2 - s}{s-2} \\ &\iff (s-2)\log(2s) \geq s\log s + 2s\log 2 - s \\ &\iff s \geq (s+2)\log 2 + 2\log s \\ &\impliedby s \geq 26 \iff m \geq 52, \end{aligned}$$

we have proven the statement.

Now, we are left to prove the lemma for $k < 14$. Exchanging $m^{\frac{k}{2}}$ and $k!$ in the statement of the lemma, we have

$$m^{\frac{k}{2}-1} > \frac{k!}{\left(\frac{k}{2}-1\right)!}. \tag{40}$$

We can numerically check that

- for $k = 4$, $m \geq 25$ is sufficient,
- for $k = 6$, $m \geq 19$ is sufficient,
- for $k = 8$, $m \geq 19$ is sufficient,
- for $k = 10$, $m \geq 20$ is sufficient,
- for $k = 12$, $m \geq 21$ is sufficient,

for (40) to hold. This ends the proof of the lemma. □

Lemma F.3. *For $m \geq 52$ and odd $4 \leq k \leq m$, $\frac{m^{k-1}}{k!} > \frac{m^{\frac{k-1}{2}}}{(\frac{k-1}{2}-1)!}$*

Proof.

$$\frac{m^{k-1}}{k!} = \frac{m}{k} \cdot \frac{m^{k-2}}{(k-1)!} > \frac{m}{k} \cdot \frac{m^{\frac{k-1}{2}}}{\left(\frac{k-1}{2}-1\right)!} \geq \frac{m^{\frac{k-1}{2}}}{\left(\frac{k-1}{2}-1\right)!},$$

where we used Lemma F.2 in the first inequality. This ends the proof. □

Lemma F.4. *For $x \leq 1$, the following inequality holds:*

$$e^{\frac{x}{2}} < 1 + \frac{x}{2} + \frac{5x^2}{32}.$$

Proof. Using Taylor expansion,

$$\begin{aligned} e^{\frac{x}{2}} &= 1 + \frac{x}{2} + \frac{x^2}{8} + \sum_{i=3}^{\infty} \frac{1}{i!} \cdot \frac{x^i}{2^i} \\ &= 1 + \frac{x}{2} + \frac{x^2}{8} + x^2 \sum_{i=3}^{\infty} \frac{1}{i!} \cdot \frac{x^{i-2}}{2^i} \\ &\leq 1 + \frac{x}{2} + \frac{x^2}{8} + x^2 \sum_{i=3}^{\infty} \frac{1}{i!} \cdot \frac{1}{2^i} \\ &= 1 + \frac{x}{2} + \frac{x^2}{8} + x^2 \left(e^{\frac{1}{2}} - 1 - \frac{1}{1! \cdot 2} - \frac{1}{2! \cdot 2^2} \right) \\ &\leq 1 + \frac{x}{2} + \frac{x^2}{8} + \frac{x^2}{32} \\ &= 1 + \frac{x}{2} + \frac{5x^2}{32}. \end{aligned}$$

□

G. Proof of Proposition 4.6

Here we prove Proposition 4.6, restated below for the sake of readability.

Proposition 4.6 (Extended version of Lu et al. (2022a), Theorem 1)**.** *Suppose that $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ and $n \geq H$. Under Assumption 4.2, with constant step size η as*

$$\eta = \frac{2}{\mu n K} W_0 \left(\frac{\left(F(\boldsymbol{x}_0^1) - F^* + \nu^2/L\right) \mu^3 n^2 K^2}{192 H^2 L^2 \nu^2} \right),$$

where W_0 denotes the Lambert W function, Algorithm 1 converges at the rate

$$F(\boldsymbol{x}_n^K) - F^* = \tilde{\mathcal{O}}\left(\frac{H^2 L^2 \nu^2}{\mu^3 n^2 K^2} \right)$$

for $K \gtrsim \kappa(\tau + 1)$.

Proof of Proposition 4.6. While Lu et al. (2022a) gained convergence rate for $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, 0, \nu)$, we found out that their result can easily be extended to $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ with a slight adjustment. We basically follow up the proof step in Theorem 1 of Lu et al. (2022a). We first state 2 lemmas that will help us prove the proposition.

Lemma G.1 (Extended version of Lu et al. (2022a), Lemma 2)**.** *Applying offline GraB to a function $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ with $\eta n L < 1$ results*

$$F(\boldsymbol{x}_n^K) - F^* \leq \rho^K (F(\boldsymbol{x}_0^1) - F^*) + \frac{\eta n L^2}{2} \sum_{k=1}^{K} \rho^{K-k} \Delta_k^2 - \frac{\eta n}{4} \sum_{k=1}^{K} \rho^{K-k} \left\| \nabla F(\boldsymbol{x}_0^k) \right\|^2,$$

where $\rho = 1 - \frac{\eta n \mu}{2}$ and $\Delta_k = \max_{m=1,\cdots,n} \left\| \boldsymbol{x}_m^k - \boldsymbol{x}_0^k \right\|$ for all $k \in [K]$.

Lemma G.2 (Extended version of Lu et al. (2022a), Lemma 3)**.** *Applying offline GraB to a function $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ with $\eta n L \leq \frac{1}{2}$ results*

$$\begin{aligned}
&\Delta_1 \leq 2\eta n \nu + 2\eta n(\tau + 1) \cdot \left\| \nabla F(\boldsymbol{x}_0^1) \right\|, \ \textit{and} \\
&\Delta_k \leq 2\eta H \nu + (2\eta H \tau + 2\eta n) \cdot \left\| \nabla F \left(\boldsymbol{x}_0^k \right) \right\| + (4\eta H L(\tau + 1) + 8\eta n L) \cdot \Delta_{k-1}
\end{aligned}$$

for $k \in [K] \setminus \{1\}$.

We defer the proofs of the lemmas to Appendix G.1. We start by finding the upper bound of $\sum_{k=1}^{K} \rho^{K-k} \Delta_k^2$. From Lemma G.2, we have

$$\Delta_k \leq 2\eta H \nu + (2\eta H \tau + 2\eta n) \cdot \left\| \nabla F \left(\boldsymbol{x}_0^k \right) \right\| + (4\eta H L(\tau + 1) + 8\eta n L) \cdot \Delta_{k-1}$$

for $k \in [K] \setminus \{1\}$. Taking square on both sides and applying the inequality $3\left(a^2 + b^2 + c^2\right) \geq (a + b + c)^2$, we get

$$\Delta_k^2 \leq 3\eta^2 \left(4HL(\tau + 1) + 8nL\right)^2 \Delta_{k-1}^2 + 12\eta^2 H^2 \nu^2 + 12\eta^2 (H\tau + n)^2 \left\| \nabla F \left(\boldsymbol{x}_0^k \right) \right\|^2.$$

Similarly, for $k = 1$, we have

$$\Delta_1^2 \leq 8\eta^2 n^2 (\tau + 1)^2 \left\| \nabla F(\boldsymbol{x}_0^1) \right\|^2 + 8\eta^2 n^2 \nu^2.$$

Hence,

$$\begin{aligned}
&\sum_{k=1}^{K} \rho^{K-k} \Delta_k^2 \\
&= \sum_{k=2}^{K} \rho^{K-k} \Delta_k^2 + \rho^{K-1} \Delta_1^2 \\
&\leq \sum_{k=2}^{K} \rho^{K-k} \left(3\eta^2 \left(4HL(\tau + 1) + 8nL\right)^2 \Delta_{k-1}^2 + 12\eta^2 H^2 \nu^2 + 12\eta^2 (H\tau + n)^2 \left\| \nabla F(\boldsymbol{x}_0^k) \right\|^2 \right) \\
&\quad + \rho^{K-1} \left(8\eta^2 n^2 (\tau + 1)^2 \left\| \nabla F(\boldsymbol{x}_0^1) \right\|^2 + 8\eta^2 n^2 \nu^2 \right) \\
&\leq 3\rho^{-1} \eta^2 \left(4HL(\tau + 1) + 8nL\right)^2 \sum_{k=2}^{K} \rho^{K-(k-1)} \Delta_{k-1}^2 + \frac{12\eta^2 H^2 \nu^2}{1 - \rho} + 8\rho^{K-1} \eta^2 n^2 \nu^2 \\
&\quad + 12\eta^2 (H\tau + n)^2 \sum_{k=2}^{K} \rho^{K-k} \left\| \nabla F(\boldsymbol{x}_0^k) \right\|^2 + 8\eta^2 n^2 (\tau + 1)^2 \rho^{K-1} \left\| \nabla F(\boldsymbol{x}_0^1) \right\|^2.
\end{aligned}$$

From the assumption that $H \leq n$, $H\tau + n \leq n(\tau + 1)$ holds. Then, we get

$$\begin{aligned}
\sum_{k=1}^{K} \rho^{K-k} \Delta_k^2 \leq 3\rho^{-1} \eta^2 \left(4HL(\tau + 1) + 8nL\right)^2 \sum_{k=1}^{K} \rho^{K-k} \Delta_k^2 + \frac{12\eta^2 H^2 \nu^2}{1 - \rho} \\
+ 8\rho^{K-1} \eta^2 n^2 \nu^2 + 12\eta^2 n^2 (\tau + 1)^2 \sum_{k=1}^{K} \rho^{K-k} \left\| \nabla F(\boldsymbol{x}_0^k) \right\|^2. \qquad (41)
\end{aligned}$$

We now define our step size as:

$$\eta = \min\left(\frac{1}{64nL(\tau+1)}, \frac{2}{\mu n K} W_0\left(\frac{\left(F(\boldsymbol{x}_0^1) - F^* + \nu^2/L\right)\mu^3 n^2 K^2}{192 H^2 L^2 \nu^2}\right)\right).$$

We first focus on $\eta \leq \frac{1}{64nL(\tau+1)}$. With this step size range, $\rho = 1 - \frac{\eta n \mu}{2} \geq 1 - \frac{\eta n L}{2} > \frac{1}{2}$ and

$$\begin{aligned} \eta(4HL(\tau+1) + 8nL) &\leq \frac{4HL(\tau+1)}{64nL(\tau+1)} + \frac{8nL}{64nL(\tau+1)} \\ &\leq \frac{H}{16n} + \frac{1}{8(\tau+1)} < \frac{1}{4} \end{aligned} \tag{42}$$

holds. Thereby,

$$3\rho^{-1}\eta^2\left(4HL(\tau+1) + 8nL\right)^2 < 3 \cdot 2 \cdot \frac{1}{16} < \frac{1}{2}$$

holds and (41) becomes

$$\sum_{k=1}^{K} \rho^{K-k} \Delta_k^2 \leq \frac{24\eta^2 H^2 \nu^2}{1-\rho} + 32\rho^K \eta^2 n^2 \nu^2 + 24\eta^2 n^2 (\tau+1)^2 \sum_{k=1}^{K} \rho^{K-k} \left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2.$$

Substituting this inequality to Lemma G.1, we obtain

$$\begin{aligned} F(\boldsymbol{x}_n^K) - F^* &\leq \rho^K(F(\boldsymbol{x}_0^1) - F^*) + \frac{\eta n L^2}{2} \sum_{k=1}^{K} \rho^{K-k} \Delta_k^2 - \frac{\eta n}{4} \sum_{k=1}^{K} \rho^{K-k} \left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2 \\ &\leq \rho^K(F(\boldsymbol{x}_0^1) - F^*) + \frac{12\eta^3 n L^2 H^2 \nu^2}{1-\rho} + 16\rho^K \eta^3 n^3 L^2 \nu^2 \\ &\quad + 12\eta^3 n^3 L^2 (\tau+1)^2 \sum_{k=1}^{K} \rho^{K-k} \left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2 - \frac{\eta n}{4} \sum_{k=1}^{K} \rho^{K-k} \left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2 \\ &\leq \rho^K(F(\boldsymbol{x}_0^1) - F^*) + \frac{24\eta^2 L^2 H^2 \nu^2}{\mu} + 16\rho^K \eta^3 n^3 L^2 \nu^2, \end{aligned} \tag{43}$$

where the last inequality holds because

$$\begin{aligned} 12\eta^3 n^3 L^2 (\tau+1)^2 &= \frac{\eta n}{4} \cdot 48\eta^2 n^2 L^2 (\tau+1)^2 \\ &\leq \frac{\eta n}{4} \cdot \frac{48}{64^2} \qquad \left(\because \eta \leq \frac{1}{64nL(\tau+1)}\right) \\ &< \frac{\eta n}{4}. \end{aligned}$$

The RHS of (43) can further be extended as

$$\begin{aligned} &\left(1 - \frac{\eta n \mu}{2}\right)^K \left((F(\boldsymbol{x}_0^1) - F^*) + 16\eta^3 n^3 L^2 \nu^2\right) + \frac{24\eta^2 L^2 H^2 \nu^2}{\mu} \\ &< e^{-\frac{\eta n \mu K}{2}} \left(F(\boldsymbol{x}_0^1) - F^* + \nu^2/L\right) + \frac{24\eta^2 L^2 H^2 \nu^2}{\mu}. \end{aligned} \tag{44}$$

Taking derivative of (44) with respect to η, we can obtain η that minimizes (44) is

$$\eta = \frac{2}{\mu n K} W_0\left(\frac{\left(F(\boldsymbol{x}_0^1) - F^* + \nu^2/L\right)\mu^3 n^2 K^2}{192 H^2 L^2 \nu^2}\right),$$

where W_0 denotes the Lambert W function. By substituting this η to (44), we finally obtain

$$F(\boldsymbol{x}_n^K) - F^* = \tilde{\mathcal{O}}\left(\frac{H^2L^2\nu^2}{\mu^3n^2K^2}\right). \tag{45}$$

In addition, to make use of such η to obtain (45), the following condition

$$\frac{2}{\mu nK}W_0\left(\frac{\left(F(\boldsymbol{x}_0^1) - F^* + \nu^2/L\right)\mu^3n^2K^2}{192H^2L^2\nu^2}\right) \leq \frac{1}{64nL(\tau+1)}$$

must hold. Thus, we require

$$K \gtrsim \kappa(\tau+1)$$

to guarantee the convergence rate. □

G.1. Lemmas used in Proposition 4.6

G.1.1. Proof for Lemma G.1

Lemma G.1 (Extended version of Lu et al. (2022a), Lemma 2)**.** *Applying offline GraB to a function $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ with $\eta nL < 1$ results*

$$F(\boldsymbol{x}_n^K) - F^* \leq \rho^K(F(\boldsymbol{x}_0^1) - F^*) + \frac{\eta nL^2}{2}\sum_{k=1}^K \rho^{K-k}\Delta_k^2 - \frac{\eta n}{4}\sum_{k=1}^K \rho^{K-k}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2,$$

where $\rho = 1 - \frac{\eta n\mu}{2}$ and $\Delta_k = \max_{m=1,\cdots,n}\left\|\boldsymbol{x}_m^k - \boldsymbol{x}_0^k\right\|$ for all $k \in [K]$.

Proof of Lemma G.1. The update process within a k-th epoch can be written as:

$$x_0^{k+1} = x_0^k - \eta n \cdot \frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right).$$

Using smoothness and $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = -\frac{1}{2}\|\boldsymbol{a}\|^2 - \frac{1}{2}\|\boldsymbol{b}\|^2 + \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$, we get

$$\begin{aligned}
F(\boldsymbol{x}_0^{k+1}) &\leq F(\boldsymbol{x}_0^k) - \eta n\left\langle \nabla F(\boldsymbol{x}_0^k), \frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\rangle + \frac{\eta^2n^2L}{2}\left\|\frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2 \\
&= F(\boldsymbol{x}_0^k) - \frac{\eta n}{2}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2 - \frac{\eta n}{2}\left\|\frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2 \\
&\quad + \frac{\eta n}{2}\left\|\nabla F(\boldsymbol{x}_0^k) - \frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2 + \frac{\eta^2n^2L}{2}\left\|\frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2 \\
&\leq F(\boldsymbol{x}_0^k) - \frac{\eta n}{2}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2 + \frac{\eta n}{2}\left\|\nabla F(\boldsymbol{x}_0^k) - \frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2,
\end{aligned}$$

where we used $\eta nL < 1$ in the last inequality. In addition, we can expand the last term as

$$\begin{aligned}
\left\|\nabla F(\boldsymbol{x}_0^k) - \frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2 &= \left\|\frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}(\boldsymbol{x}_0^k) - \frac{1}{n}\sum_{t=1}^n \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2 \\
&\leq \frac{1}{n}\sum_{t=1}^n \left\|\nabla f_{\sigma_k(t)}(\boldsymbol{x}_0^k) - \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right)\right\|^2 \qquad (\because \text{Jensen's Inequality})
\end{aligned}$$

$$\begin{aligned} &\le \frac{L^2}{n}\sum_{t=1}^{n}\left\|\boldsymbol{x}_0^k - \boldsymbol{x}_{t-1}^k\right\|^2 && (\because \text{ smoothness}) \\ &\le L^2\Delta_k^2. && (\because \Delta_k = \max_{m=1,\cdots,n}\|\boldsymbol{x}_m^k - \boldsymbol{x}_0^k\|) \end{aligned}$$

Combining these two results, we get

$$F(\boldsymbol{x}_0^{k+1}) \le F(\boldsymbol{x}_0^k) + \frac{\eta n L^2 \Delta_k^2}{2} - \frac{\eta n}{2}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2.$$

Using the PŁ inequality, this inequality becomes

$$\begin{aligned} F(\boldsymbol{x}_0^{k+1}) &\le F(\boldsymbol{x}_0^k) + \frac{\eta n L^2 \Delta_k^2}{2} - \frac{\eta n}{4}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2 - \frac{\eta n}{4}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2 \\ &\le F(\boldsymbol{x}_0^k) + \frac{\eta n L^2 \Delta_k^2}{2} - \frac{\eta n \mu}{2}(F(\boldsymbol{x}_0^k) - F^*) - \frac{\eta n}{4}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2. \end{aligned}$$

Define $\rho := 1 - \frac{\eta n \mu}{2}$. Subtracting F^* on both sides, we get

$$F(\boldsymbol{x}_0^{k+1}) - F^* \le \rho(F(\boldsymbol{x}_0^k) - F^*) + \frac{\eta n L^2 \Delta_k^2}{2} - \frac{\eta n}{4}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2.$$

This inequality holds for all $k \in \{1, \cdots, K\}$. Unrolling for entire epochs gives

$$F(\boldsymbol{x}_0^{K+1}) - F^* \le \rho^K(F(\boldsymbol{x}_0^1) - F^*) + \frac{\eta n L^2}{2}\sum_{k=1}^{K}\rho^{K-k}\Delta_k^2 - \frac{\eta n}{4}\sum_{k=1}^{K}\rho^{K-k}\left\|\nabla F(\boldsymbol{x}_0^k)\right\|^2.$$

This ends the proof of the lemma. □

G.1.2. Proof for Lemma G.2

Lemma G.2 (Extended version of Lu et al. (2022a), Lemma 3)**.** *Applying offline GraB to a function $F \in \mathcal{F}_{\text{PŁ}}(L, \mu, \tau, \nu)$ with $\eta n L \le \frac{1}{2}$ results*

$$\begin{aligned} &\Delta_1 \le 2\eta n\nu + 2\eta n(\tau+1)\cdot\left\|\nabla F(\boldsymbol{x}_0^1)\right\|, \text{ and} \\ &\Delta_k \le 2\eta H\nu + (2\eta H\tau + 2\eta n)\cdot\left\|\nabla F\left(\boldsymbol{x}_0^k\right)\right\| + (4\eta HL(\tau+1) + 8\eta nL)\cdot\Delta_{k-1} \end{aligned}$$

for $k \in [K] \setminus \{1\}$.

Proof of Lemma G.2. We first consider the situation after the first epoch. For $m \in [n]$ and $k \in [K] \setminus \{1\}$, proper additions and subtractions give us

$$\begin{aligned} \boldsymbol{x}_m^k &= \boldsymbol{x}_0^k - \eta\sum_{t=1}^{m}\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right) \\ &= \boldsymbol{x}_0^k - \eta\sum_{t=1}^{m}\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}^{k-1}\right) \\ &\quad - \eta\sum_{t=1}^{m}\left(\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^k\right) - \nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}^{k-1}\right)\right) \\ &= \boldsymbol{x}_0^k - \eta\sum_{t=1}^{m}\left(\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}^{k-1}\right) - \frac{1}{n}\sum_{s=1}^{n}\nabla f_{\sigma_{k-1}(s)}\left(\boldsymbol{x}_{s-1}^{k-1}\right)\right) \\ &\quad - \frac{\eta m}{n}\sum_{s=1}^{n}\nabla f_{\sigma_{k-1}(s)}\left(\boldsymbol{x}_{s-1}^{k-1}\right) \end{aligned}$$

$$
\begin{aligned}
&\quad - \eta \sum_{t=1}^{m} \left(\nabla f_{\sigma_k(t)} \left(\boldsymbol{x}_{t-1}^{k} \right) - \nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) \right) \\
&= \boldsymbol{x}_0^k - \eta \sum_{t=1}^{m} \left(\nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{s-1}^{k-1} \right) \right) \\
&\quad - \eta m \nabla F \left(\boldsymbol{x}_n^{k-1} \right) \\
&\quad - \frac{\eta m}{n} \sum_{s=1}^{n} \left(\nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{s-1}^{k-1} \right) - \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{n}^{k-1} \right) \right) \\
&\quad - \eta \sum_{t=1}^{m} \left(\nabla f_{\sigma_k(t)} \left(\boldsymbol{x}_{t-1}^{k} \right) - \nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) \right).
\end{aligned}
$$

Here, $\sigma_{k-1}^{-1}(t)$ indicates in which iteration is the t-th sample used at the $(k-1)$-th epoch and $\nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right)$ indicates the gradient with respect to the same sample used in the t-th iteration of the k-th epoch, but which was computed previously in the $(k-1)$-th epoch. Using the triangle inequality, we gain

$$
\begin{aligned}
\left\| \boldsymbol{x}_m^k - \boldsymbol{x}_0^k \right\| \leq \eta & \left\| \sum_{t=1}^{m} \left(\nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{s-1}^{k-1} \right) \right) \right\| \\
&+ \eta m \left\| \nabla F \left(\boldsymbol{x}_n^{k-1} \right) \right\| \\
&+ \frac{\eta m}{n} \left\| \sum_{s=1}^{n} \left(\nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{s-1}^{k-1} \right) - \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{n}^{k-1} \right) \right) \right\| \\
&+ \eta \left\| \sum_{t=1}^{m} \left(\nabla f_{\sigma_k(t)} \left(\boldsymbol{x}_{t-1}^{k} \right) - \nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) \right) \right\|. \qquad (46)
\end{aligned}
$$

Here, the first term in (46) is the term in which *Herding* intervenes and it enables us to gain the upper bound. To do so, we first upper bound the norm of each component as

$$
\begin{aligned}
&\left\| \nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{s-1}^{k-1} \right) \right\| \\
\leq & \left\| \nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) \right\| \\
&+ \left\| \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{s-1}^{k-1} \right) \right\| \\
\leq & \left(\nu + \tau \left\| \nabla F \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) \right\| \right) + \frac{L}{n} \sum_{s=1}^{n} \left\| \boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} - \boldsymbol{x}_{s-1}^{k-1} \right\| \\
\leq & \, \nu + \tau \left(\left\| \nabla F \left(\boldsymbol{x}_0^k \right) \right\| + \left\| \nabla F \left(\boldsymbol{x}_0^{k-1} \right) - \nabla F \left(\boldsymbol{x}_0^k \right) \right\| + \left\| \nabla F \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) - \nabla F \left(\boldsymbol{x}_0^{k-1} \right) \right\| \right) \\
&+ \frac{L}{n} \sum_{s=1}^{n} \left(\left\| \boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} - \boldsymbol{x}_0^{k-1} \right\| + \left\| \boldsymbol{x}_0^{k-1} - \boldsymbol{x}_{s-1}^{k-1} \right\| \right) \\
\leq & \, \nu + \tau \left(\left\| \nabla F \left(\boldsymbol{x}_0^k \right) \right\| + 2L\Delta_{k-1} \right) + 2L\Delta_{k-1} \\
= & \, \nu + \tau \cdot \left\| \nabla F \left(\boldsymbol{x}_0^k \right) \right\| + 2L \left(\tau + 1 \right) \cdot \Delta_{k-1}.
\end{aligned}
$$

Now, define z_t^k as

$$
z_t^k := \frac{\nabla f_{\sigma_k(t)} \left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1} \right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_{k-1}(s)} \left(\boldsymbol{x}_{s-1}^{k-1} \right)}{\nu + \tau \cdot \left\| \nabla F \left(\boldsymbol{x}_0^k \right) \right\| + 2L \left(\tau + 1 \right) \cdot \Delta_{k-1}}
$$

for $t \in [n]$. Then, $\left\|z_t^k\right\| \leq 1$ holds.
We now apply Herding algorithm to upper bound the first term of (46). Since $\|z_t^k\| \leq 1$, we then get following inequality for all $m \in [n]$:

$$\begin{aligned}\left\|\sum_{t=1}^{m}\left(\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}\right)-\frac{1}{n}\sum_{s=1}^{n}\nabla f_{\sigma_{k-1}(s)}\left(\boldsymbol{x}_{s-1}^{k-1}\right)\right)\right\| \\ \leq H\left(\nu+\tau\cdot\left\|\nabla F\left(\boldsymbol{x}_0^k\right)\right\|+2L\left(\tau+1\right)\cdot\Delta_{k-1}\right).\end{aligned} \tag{47}$$

For the remaining terms in (46), we can upper bound each of them by

$$\begin{aligned}\left\|\sum_{s=1}^{n}\left(\nabla f_{\sigma_{k-1}(s)}\left(\boldsymbol{x}_{s-1}^{k-1}\right)-\nabla f_{\sigma_{k-1}(s)}\left(\boldsymbol{x}_{n}^{k-1}\right)\right)\right\| &\leq \sum_{s=1}^{n}\left\|\nabla f_{\sigma_{k-1}(s)}\left(\boldsymbol{x}_{s-1}^{k-1}\right)-\nabla f_{\sigma_{k-1}(s)}\left(\boldsymbol{x}_{n}^{k-1}\right)\right\| \\ &\leq L\sum_{s=1}^{n}\left\|\boldsymbol{x}_{s-1}^{k-1}-\boldsymbol{x}_{n}^{k-1}\right\| \\ &\leq L\sum_{s=1}^{n}\left(\left\|\boldsymbol{x}_{s-1}^{k-1}-\boldsymbol{x}_{0}^{k-1}\right\|+\left\|\boldsymbol{x}_{0}^{k-1}-\boldsymbol{x}_{n}^{k-1}\right\|\right) \\ &\leq 2nL\Delta_{k-1}\end{aligned} \tag{48}$$

and

$$\begin{aligned}&\left\|\sum_{t=1}^{m}\left(\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^{k}\right)-\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}\right)\right)\right\| \\ &\leq \sum_{t=1}^{m}\left\|\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}_{t-1}^{k}\right)-\nabla f_{\sigma_k(t)}\left(\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}\right)\right\| \\ &\leq L\sum_{t=1}^{m}\left\|\boldsymbol{x}_{t-1}^{k}-\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}\right\| \\ &\leq L\sum_{t=1}^{m}\left(\left\|\boldsymbol{x}_{t-1}^{k}-\boldsymbol{x}_{0}^{k}\right\|+\left\|\boldsymbol{x}_{0}^{k}-\boldsymbol{x}_{0}^{k-1}\right\|+\left\|\boldsymbol{x}_{0}^{k-1}-\boldsymbol{x}^{k-1}_{\sigma_{k-1}^{-1}(\sigma_k(t))-1}\right\|\right) \\ &\leq mL\left(\Delta_k+2\Delta_{k-1}\right).\end{aligned} \tag{49}$$

By summing up (47)-(49) and taking a max over $m \in \{1, \cdots, n\}$ on both side of (46),

$$\begin{aligned}\Delta_k &\leq \eta H\left(\nu+\tau\cdot\left\|\nabla F\left(\boldsymbol{x}_0^k\right)\right\|+2L\left(\tau+1\right)\cdot\Delta_{k-1}\right) \\ &\quad+\eta n\left\|\nabla F\left(\boldsymbol{x}_0^k\right)\right\|+\frac{\eta n}{n}\cdot 2nL\Delta_{k-1}+\eta nL\left(\Delta_k+2\Delta_{k-1}\right) \\ &\leq \eta H\nu+\left(\eta H\tau+\eta n\right)\cdot\left\|\nabla F\left(\boldsymbol{x}_0^k\right)\right\|+\left(2\eta HL(\tau+1)+4\eta nL\right)\cdot\Delta_{k-1}+\eta nL\Delta_k.\end{aligned}$$

Using $\eta nL \leq \frac{1}{2}$, we finally get

$$\Delta_k \leq 2\eta H\nu+\left(2\eta H\tau+2\eta n\right)\cdot\left\|\nabla F\left(\boldsymbol{x}_0^k\right)\right\|+\left(4\eta HL(\tau+1)+8\eta nL\right)\cdot\Delta_{k-1}.$$

We now move on to the first epoch case. By properly decomposing the term, we gain

$$\begin{aligned}\boldsymbol{x}_m^1 &= \boldsymbol{x}_0^1-\eta\sum_{t=1}^{m}\nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_{t-1}^1\right) \\ &= \boldsymbol{x}_0^1-\eta\sum_{t=1}^{m}\frac{1}{n}\sum_{s=1}^{n}\nabla f_{\sigma_1(s)}\left(\boldsymbol{x}_0^1\right) \\ &\quad-\eta\sum_{t=1}^{m}\left(\nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_0^1\right)-\frac{1}{n}\sum_{s=1}^{n}\nabla f_{\sigma_1(s)}\left(\boldsymbol{x}_0^1\right)\right)\end{aligned}$$

$$-\eta \sum_{t=1}^{m} \left(\nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_{t-1}^1\right) - \nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_0^1\right)\right).$$

In a similar way to the technique we used above, we have

$$\begin{aligned}
\left\|\boldsymbol{x}_m^1 - \boldsymbol{x}_0^1\right\| \leq \eta & \left\| \sum_{t=1}^{m} \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_1(s)}\left(\boldsymbol{x}_0^1\right) \right\| \\
& + \eta \left\| \sum_{t=1}^{m} \left(\nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_0^1\right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_1(s)}\left(\boldsymbol{x}_0^1\right) \right) \right\| \\
& + \eta \left\| \sum_{t=1}^{m} \left(\nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_{t-1}^1\right) - \nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_0^1\right)\right) \right\| \\
\leq & \, \eta \sum_{t=1}^{m} \left\| \nabla F\left(\boldsymbol{x}_0^1\right) \right\| \\
& + \eta \sum_{t=1}^{m} \left\| \nabla f_{\sigma_1(t)}\left(\boldsymbol{x}_0^1\right) - \frac{1}{n} \sum_{s=1}^{n} \nabla f_{\sigma_1(s)}\left(\boldsymbol{x}_0^1\right) \right\| \\
& + \eta \sum_{t=1}^{m} L \left\| \boldsymbol{x}_{t-1}^1 - \boldsymbol{x}_0^1 \right\|.
\end{aligned}$$

Taking a max over $m \in \{1, \cdots, n\}$ in both sides, we gain

$$\Delta_1 \leq \eta n \left\| \nabla F\left(\boldsymbol{x}_0^1\right) \right\| + \eta n \left(\nu + \tau \cdot \left\| \nabla F\left(\boldsymbol{x}_0^1\right) \right\|\right) + \eta n L \Delta_1$$

and using the fact that $\eta n L \leq \frac{1}{2}$, we finally obtain

$$\Delta_1 \leq 2\eta n \nu + 2\eta n(\tau + 1) \cdot \left\| \nabla F(\boldsymbol{x}_0^1) \right\|.$$

□