TIGHT RATES FOR RANDOM RESHUFFLING: PHASE TRANSITIONS AND COST OF NONCONVEX COMPONENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Random Reshuffling (RR) processes all n components once per epoch in a fresh uniformly random order. We study the worst-case convergence of constant stepsize stochastic gradient descent under RR for a finite sum of smooth component functions whose average is strongly convex. Although recent works have established convergence guarantees for RR, most results require the number of epochs K to be sufficiently large, and a tight characterization across the full range of epoch counts has remained under-explored, both with and without component convexity. For convex components, by characterizing tight rates up to polylogarithmic factors, we establish a phase transition at $K \asymp \kappa$, where κ is the condition number: RR matches the convergence of with-replacement SGD below this threshold and achieves faster rates above it. Without component convexity, we prove that order $\max\{\kappa, \kappa^2/n\}$ epochs are necessary in the worst case to reduce the last-iterate squared distance below a constant fraction of its initial value. Moreover, beyond this threshold, the convergence rate of RR worsens by a factor of $\max\{1, \kappa/n\}$ relative to the convex-component rate. Together, these results characterize the cost of allowing nonconvex components, which takes different forms depending on the ratio κ/n.

1 INTRODUCTION

We study the finite-sum minimization problem:

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

When n is large, computing the full gradient at every iteration can be prohibitively expensive. Stochastic gradient methods (SGD) instead update using the gradient of a single component function. Classical with-replacement SGD (SGD) draws an independent index at each iteration, making the stochastic gradient conditionally unbiased. In practice, however, training data are commonly traversed without replacement: every component is used once in a random order before the next epoch begins.

A canonical method of this type is SGD with Random Reshuffling (RR). At the beginning of epoch k, RR draws an independent random permutation σ_k of $[n] := \{1, \dots, n\}$ and performs

$$\boldsymbol{x}_i^k = \boldsymbol{x}_{i-1}^k - \eta \nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k), \qquad i = 1, \dots, n, \tag{2}$$

with $\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_n^k$. Despite its simplicity, RR is substantially more difficult to analyze than SGD, since its next component gradient is biased conditional on the current iterate.

Nevertheless, a sequence of recent works (Ahn et al., 2020; Nguyen et al., 2021; Mishchenko et al., 2020; Liu & Zhou, 2024; Li, 2026) has shown that this dependence can be controlled to establish convergence guarantees. In particular, when the number of epochs K is sufficiently large, RR is known to achieve a convergence rate of $\mathcal{O}(1/nK^2)$ for smooth and strongly convex objectives, improving the $\mathcal{O}(1/nK)$ rate of SGD. In contrast, the behavior of RR when the number of epochs is not large enough remains less understood. While Safran & Shamir (2021) identified a phase transition in function-value convergence for a restricted quadratic class, existing analyses do not provide sharp last-iterate squared-distance guarantees under general smooth components (Mishchenko et al., 2020; Nagaraj et al., 2019), leaving open the corresponding comparison between RR and standard SGD in this small-epoch regime.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

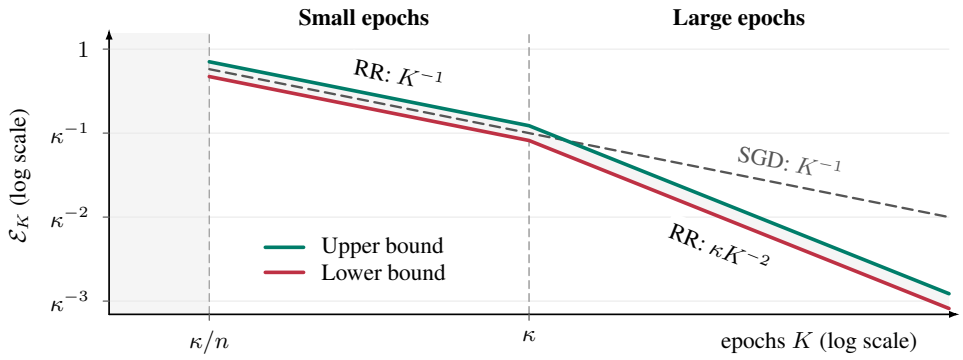


Figure 1: Visualization for the convex component case. Normalized squared distance $\mathcal{E}_K = (\mu^2 n/G_*^2)\mathbb{E}\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2$, with logarithmic factors suppressed. Matching RR upper (green) and lower (red) bounds have SGD scaling below κ and improve on it beyond this scale.

1.1 Summary of Our Contributions

In this work, we study the worst-case last-iterate squared-distance convergence of constant stepsize RR. We consider L-smooth component functions with a μ-strongly convex average, both with and without component convexity. Our results, up to polylogarithmic factors, establish a tight characterization of the worst-case convergence of RR across the full range of epoch counts. We summarize our main contributions below.

- **With Component Convexity.** Under component convexity, we establish a sharp phase transition of the convergence at $K \asymp \kappa$, where $\kappa := L/\mu$ denotes the condition number. Our new upper and lower bounds, together with the existing upper bound (Mishchenko et al., 2020) that applies for $K \gtrsim \kappa$, yield the tight rate $\widetilde{\Theta}\left(\frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}\right)$ for $K \gtrsim \kappa/n$ (Theorem 3.2). As a result, RR matches the rate of SGD for $\kappa/n \lesssim K \lesssim \kappa$ and improves on it for $K \gtrsim \kappa$. This extends the phase-transition picture beyond the quadratic setting of Safran & Shamir (2021).
- **Without Component Convexity.** We show that removing component convexity incurs a cost through a worse convergence rate or a longer initialization-error regime. We prove that RR can retain a constant fraction of its initial squared distance whenever $K \lesssim \max\{\kappa, \kappa^2/n\}$ (Theorem 4.2). Beyond this threshold, our matching upper and lower bounds establish the tight rate $\widetilde{\Theta}\left(\frac{LG_*^2}{\mu^3 nK^2}\max\left\{1, \frac{\kappa}{n}\right\}\right)$ (Theorem 4.1). For $\kappa \gtrsim n$, this introduces a slowdown by a factor of κ/n relative to the convex-component rate. For $n \gtrsim \kappa$, the eventual rate agrees with the convex-component case, making RR faster than SGD for $K \gtrsim \kappa$. However, the initialization-error regime $K \lesssim \kappa$ is wider compared to the $K \lesssim \kappa^2/n$ of SGD, and therefore RR can be slower than SGD in the intermediate range $\kappa^2/n \lesssim K \lesssim \kappa$. Overall, these results reveal that the effect of removing component convexity depends qualitatively on the ratio κ/n.

For these comparisons, we include SGD benchmarks under the same assumptions, constant-stepsize restriction, and last-iterate squared-distance criterion. See Figure 1 for an overview of the convex case, and Figure 2 for a more detailed summary overall.

2 Problem Setup and Related Work

Throughout, vectors are written in bold, and $\|\cdot\|$ denotes the Euclidean norm. We list some definitions and assumptions.

Definition 2.1 (Smoothness)**.** A differentiable function f is L-smooth, with $L > 0$, if

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

Definition 2.2 (Strong Convexity)**.** A differentiable function f is μ-strongly convex, with $\mu > 0$, if

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

Convexity corresponds to the case $\mu = 0$.

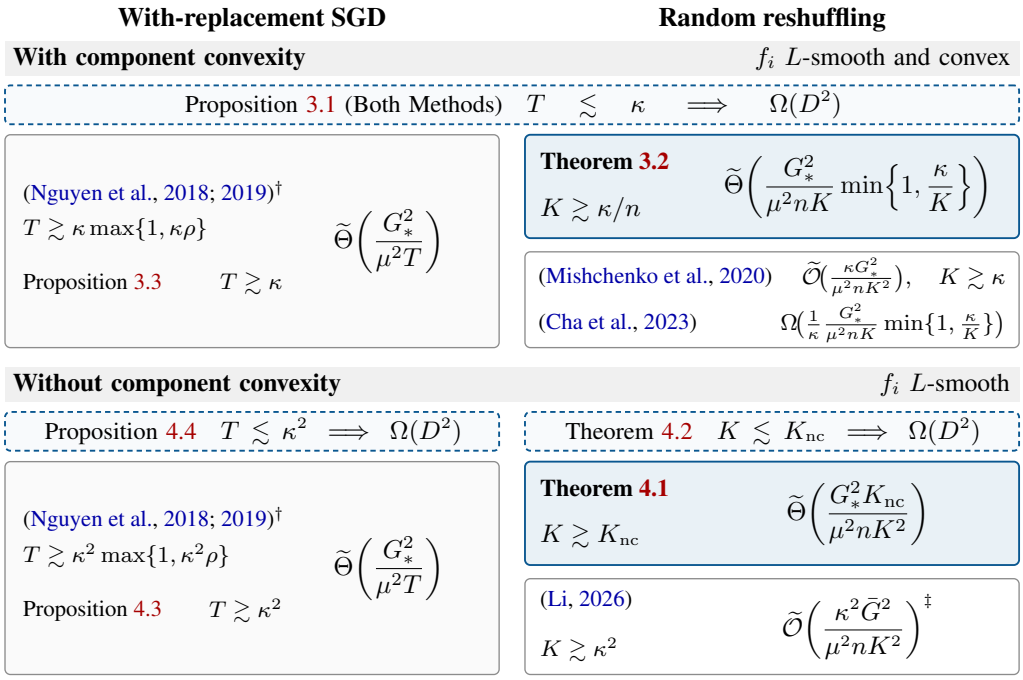


Figure 2: Overview of last-iterate squared-distance bounds after $T = nK$ updates and comparison with selected prior works. The SGD bounds in the propositions are included as benchmarks for comparison under the constant-stepsize setting. Here, $K_{\mathrm{nc}} := \max\{\kappa, \kappa^2/n\}$.

†UB uses diminishing stepsizes. Sufficient horizons use $\rho := \mu^2 \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 / G_*^2$.

‡$\bar{G} := G_{\max} + 2L \max\{D, G_{\max}/(2L)\}$ with $G_{\max} := \max_i \|\nabla f_i(\boldsymbol{x}^*)\|$.

We make the following assumptions.

Assumption 2.3 (Common Assumption)**.** Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, and F is μ-strongly convex, with $L \geq \mu > 0$.

Assumption 2.4 (Component Convexity)**.** Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex.

Under Assumption 2.3, F has a unique minimizer $\boldsymbol{x}^*$. We write $F^* := F(\boldsymbol{x}^*)$ for the minimum function value. When the overall function is strongly convex, both the squared distance $\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2$ and the function value gap $F(\boldsymbol{x}_n^K) - F^*$ are commonly used as the convergence metric. Under strong convexity and smoothness, the two metrics can be translated into each other up to constants depending on μ and L. We primarily focus on the squared distance metric throughout the paper.

Additional notations include $D := \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$ for the distance to the optimum at the initialization, and $\kappa := L/\mu$ for the condition number. Integers $n \geq 2$ and $K \geq 1$ denote the number of components and the number of epochs, respectively, where $T := nK$ denotes the total number of updates over K epochs. The component-gradient variability at the minimizer is denoted as

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

Expectations are taken over the sampled permutations or component indices. We write $a \gtrsim b$ to mean $a \geq Cb[\log(\mathrm{poly}(n, K, \mu, L, D/G_*))]^c$ for some numerical constant $C > 0$ and bounded c, and define $a \lesssim b$ analogously. We write $a \asymp b$ when both $a \gtrsim b$ and $a \lesssim b$ hold. Lastly, we define $\Lambda_T := 2\max\{2, \log(1 + \mu^2 T^2 \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 / G_*^2)\}$, and write Λ_{nK} when $T = nK$. The logarithmic term in Λ_T appears as a common factor in the step sizes for our upper bound theorems, which ensures a sufficient contraction of the initial squared-distance error $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2$.

2.1 Related Work

The convergence of without-replacement methods has been studied under a broad range of assumptions (Shamir, 2016; Haochen & Sra, 2019; Nagaraj et al., 2019; Mishchenko et al., 2020; Nguyen et al., 2021). Below, we focus on RR for smooth components with a strongly convex average, organizing the discussion by component convexity and the epoch ranges covered by existing guarantees.

With component convexity. Under Assumptions 2.3 and 2.4, Nagaraj et al. (2019) provide a function gap upper bound for a tail average of iterates with an additional global Lipschitz continuity of component functions. Relaxing the global Lipschitz continuity, Mishchenko et al. (2020, Theorem 2) establish the last-iterate squared-distance upper bound $\widetilde{\mathcal{O}}(LG_*^2/(\mu^3 nK^2))$ for $K \gtrsim \kappa$. Their Theorem 1 permits smaller epoch counts under strong convexity of every component, but does not provide a sharp small-epoch characterization. Liu & Zhou (2024) improve the large-epoch function-value upper bound, while Cha et al. (2023) provide matching function-value lower bounds. However, translating the latter bounds into squared distance via smoothness leaves a factor-κ gap.

For convex quadratic components with commuting Hessians, Safran & Shamir (2021) identify a phase transition: in the worst case, reshuffling improves over with-replacement SGD only beyond the condition-number scale. For general convex objectives, Liu (2026) establish RR guarantees that are competitive with those of SGD after any finite number of epochs, without assuming strong convexity of the average. Its within-epoch estimates also provide tools for our strongly convex analysis.

Without component convexity. Ahn et al. (2020) establish a $\widetilde{\mathcal{O}}(1/(nK^2))$ function-value rate for $K \gtrsim \kappa$, with problem-dependent prefactors suppressed. Their guarantee applies to the best epoch iterate and uses component-gradient bounds on the initial sublevel set; a corresponding last-iterate guarantee requires an additional bounded-iterates assumption. Nguyen et al. (2021) also analyze smooth, potentially nonconvex components with a strongly convex average, obtaining last-epoch guarantees and an improved dependence on n for random reshuffling under a generalized variance condition.

As a follow up, Li (2026, Corollary 3.7) obtains the last-iterate squared-distance bound $\widetilde{\mathcal{O}}(\kappa^2\bar{G}^2/(\mu^2 nK^2))$ for $K \gtrsim \kappa^2$, without a separate bounded-iterates assumption. Here, $\bar{G}$ depends on the initial distance and the largest component-gradient norm at the optimum, as specified in Figure 2. Thus, the coefficient of the decaying error term itself depends on initialization. Faster rates under quadratic structure or additional Hessian regularity are also available (Ahn et al., 2020; Li, 2026), but concern more restrictive function classes.

These upper bounds do not by themselves establish how many epochs RR needs to reduce initialization error or to outperform SGD. Although existing lower bounds for convex components remain valid in this setting, they neither match the available last-iterate squared-distance upper bounds in their dependence on the problem parameters nor establish the additional cost of removing component convexity. Our upper and lower bounds quantify this cost in terms of the convergence rate and the number of epochs needed to reduce the initialization error.

With-replacement SGD as a benchmark. Nguyen et al. (2018, Theorems 2 and 4) obtain an asymptotic last-iterate squared-distance rate of $\mathcal{O}(G_*^2/(\mu^2 T))$ with diminishing stepsizes, both with and without component convexity, without uniformly bounded stochastic gradients. Their initial stepsize scales are $1/L$ and μ/L^2, respectively, with an initialization-dependent transient. Nguyen et al. (2019) establish a corresponding asymptotic lower bound in the general stochastic setting under any deterministic varying stepsizes. For a direct comparison with RR, we adapt the analyses underlying Nguyen et al. (2018, Theorems 2 and 4) to constant stepsizes chosen for a prescribed horizon, and provide a simpler accompanying finite-sum lower-bound construction. These auxiliary results make the SGD benchmarks self-contained in our setting.

Without-replacement SGD. Without-replacement SGD processes each component exactly once per epoch, with different methods distinguished by how the component order is chosen. A classical alternative to RR is incremental gradient descent (IGD), which uses the same fixed order across epochs (Bertsekas, 2011). Cai & Diakonikolas (2025) establish last-iterate guarantees for IGD

with smooth convex components, while Koloskova et al. (2024) improve convergence guarantees for smooth nonconvex objectives. Kim et al. (2025) establish lower bounds showing that IGD can converge slowly on ill-conditioned strongly convex problems when the number of epochs is small relative to the condition number. The component order can also be chosen to improve convergence, as demonstrated by Rajput et al. (2022) for quadratic problems. Lu et al. (2022b) relate convergence to accumulated gradient errors, while Lu et al. (2022a) use gradient balancing to construct permutations with improved convergence guarantees over RR under smoothness and bounded gradient-deviation assumptions.

3 Convergence with Component Convexity

In this section, we study the convergence of RR when all component functions are convex. We compare RR with SGD after the same number of component-gradient evaluations, i.e., $T = nK$. To make the comparison precise, we consider the constant stepsize scheme for both methods, under the common criterion of the last-iterate squared-distance. Before presenting the main results, we first consider the regime where the computational budget is small relative to the condition number, i.e., $T \lesssim \kappa$ (equivalently, $K \lesssim \kappa/n$). The following proposition shows that neither sampling scheme can guarantee reducing the error below a constant fraction of its initial value.

Proposition 3.1 (Initialization LB for SGD and RR with convex components)**.** *For any $n \geq 2$, $\kappa \geq 4$, $T \lesssim \kappa$, $G_* \geq 0$, and $D \geq 0$, there exists a 3-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterates $\boldsymbol{x}_T$ and $\boldsymbol{x}_n^K$ obtained by with-replacement SGD and RR, respectively, satisfy*

$$\mathbb{E}\left\|\boldsymbol{x}_T - \boldsymbol{x}^*\right\|^2 \gtrsim D^2, \qquad \mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

The proof is given in Appendix B.1. Hence, $T \lesssim \kappa$ constitutes a common initialization-error regime for both SGD and RR, in which there are too few updates for either method to guarantee meaningful progress beyond the initialization scale D^2. To obtain convergence guarantees that improve upon the initialization scale D^2, we therefore focus on $T \gtrsim \kappa$, or equivalently, $K \gtrsim \kappa/n$, in the following subsections. We first present our characterization of RR in Section 3.1 and then compare it with the rate of SGD in Section 3.2. These results identify a phase transition at $K \asymp \kappa$: RR has the same worst-case rate scaling as SGD below this scale, but a faster rate above it.

3.1 Convergence of Random Reshuffling

We now present the tight rate of RR. Its dependence on K changes as the number of epochs passes the condition number κ. The following theorem characterizes this dependence in a single expression.

Theorem 3.2 (RR with convex components)**.** *Let $n \geq 2$, $\kappa \geq 4096$, $K \gtrsim \kappa/n$, and $G_* > 0$. Under Assumptions 2.3 and 2.4, the worst-case final-iterate error of RR with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}\right).$$

In particular, the upper bound is attained with $\eta = \Lambda_{nK}/\mu nK$ (see Theorem A.3). Provided that $D \geq G_*/(\mu\sqrt{nK})$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 4-dimensional finite-sum objective (see Theorem A.4).

The proofs are deferred to Appendices C.2 and C.3. For $K \gtrsim \kappa$, the minimum multiplier factor in the rate of the theorem selects the κ/K branch, and the corresponding upper bound follows from Mishchenko et al. (2020, Theorem 2) and Liu & Zhou (2024, Theorem 4.6). For the wider range of $K \gtrsim \kappa/n$, we establish the new upper bound $\widetilde{\mathcal{O}}(G_*^2/(\mu^2 nK))$. To our knowledge, this is the first result to establish such a rate for RR in this function class without requiring $K \gtrsim \kappa$. Our proof builds on Lemmas 4 and 5 of Liu (2026), adapting the analysis for convex objectives to retain the strong-convexity contraction in the squared-distance recursion. The two upper bounds match at $K \asymp \kappa$ up to polylogarithmic factors, and combining them yields the unified upper bound in Theorem A.3.

For the matching lower bound in Theorem A.4, our key construction uses a piecewise-quadratic function whose curvature differs on opposite sides of a hyperplane through the minimizer of the overall objective. This asymmetry induces a directional bias in the iterates, which we quantify to obtain the desired lower bound. Previously, Cha et al. (2023, Theorem 3.1) establish a function-value lower bound for the final iterate of RR, which translates into the squared-distance lower bound $\Omega(G_*^2/(\mu LnK)\min\{1, \kappa/K\})$ via L-smoothness. Our lower bound strengthens this rate by a factor of κ and exactly matches the upper bound up to polylogarithmic factors. Together, these results establish the tight worst-case last-iterate squared-distance rate in Theorem 3.2, including its dependence on the condition number.

3.2 Comparison with SGD

For comparison, the following proposition gives the SGD benchmark. It combines an upper bound, obtained by adapting Nguyen et al. (2018, Theorem 2) to a constant stepsize, with a matching lower bound up to polylogarithmic factors.

Proposition 3.3 (SGD with convex components)**.** *Let $n \geq 2$, $T \gtrsim \kappa$, and $G_* > 0$. Under Assumptions 2.3 and 2.4, the worst-case final-iterate error of with-replacement SGD with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{G_*^2}{\mu^2 T}\right).$$

In particular, the upper bound is attained with $\eta = \Lambda_T/\mu T$ (see Proposition A.1). Provided that $D \gtrsim G_*/(\mu\sqrt{T})$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 1-dimensional finite-sum objective (see Proposition A.2).

The proofs of the upper and lower bounds are given in Appendix B.1. With $T = nK$, the SGD benchmark is $\widetilde{\Theta}(G_*^2/(\mu^2 nK))$, and its iteration requirement $T \gtrsim \kappa$ is equivalent to the epoch requirement $K \gtrsim \kappa/n$ for RR.

We now compare Theorem 3.2 with the SGD benchmark in Proposition 3.3. When $\kappa/n \lesssim K \lesssim \kappa$, both methods have the worst-case rate $\widetilde{\Theta}(G_*^2/(\mu^2 nK))$. Thus, RR does not yet improve the polynomial dependence of the convergence rate. When $K \gtrsim \kappa$, the rate of RR becomes $\widetilde{\Theta}(LG_*^2/(\mu^3 nK^2))$, improving on the SGD rate by a factor of K/κ. The two branches meet at $K \asymp \kappa$, identifying the phase transition. Consequently, the number of epochs required for RR to achieve a better worst-case rate grows with the condition number.

We note that a similar phase transition for RR at $K \asymp \kappa$ was established by Safran & Shamir (2021). However, their analysis is restricted to convex quadratic components with commuting Hessians, which can essentially be reduced to a collection of 1-dimensional convex quadratics along different dimensions. In contrast, Theorem 3.2 applies to general convex and smooth components with a strongly convex average. Neither quadratic structure nor commutativity is required, and we do not impose any additional assumptions on the gradient noise away from the minimizer. Our results thus affirm that the requirement for sufficiently many epochs to improve over with-replacement sampling on ill-conditioned problems remains valid beyond the commuting quadratic class.

4 Convergence without Component Convexity

Under component convexity, Section 3 shows that RR matches or improves upon SGD beyond their common initialization-error regime $T \lesssim \kappa$ in terms of worst-case convergence. We now examine how this relationship changes when component convexity is removed. Throughout this section, we retain Assumption 2.3: each component is L-smooth and the average objective is μ-strongly convex, but individual components may possibly be nonconvex. We first characterize the convergence of RR after sufficiently many epochs and establish an initialization lower bound for smaller epoch counts in Section 4.1. Although RR still achieves a faster rate than SGD beyond this threshold, dropping component convexity comes at a cost, which we examine in Section 4.2.

4.1 Convergence of Random Reshuffling

We first present the tight rate of RR when the number of epochs is sufficiently large. The following theorem characterizes its dependence on n, K, and κ without component convexity.

Theorem 4.1 (RR without component convexity)**.** *Let $n \geq 2$, $\kappa \geq 4096$, $K \gtrsim \max\{\kappa, \kappa^2/n\}$, and $G_* > 0$. Under Assumption 2.3, the worst-case final-iterate error of RR with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{LG_*^2}{\mu^3 nK^2} \max\left\{1, \frac{\kappa}{n}\right\}\right). \tag{3}$$

In particular, the upper bound is attained with $\eta = \Lambda_{nK}/\mu nK$ (see Theorem A.7). Provided that $D \geq \sqrt{\kappa} G_*/(\mu\sqrt{nK}) \max\{1, \sqrt{\kappa/n}\}$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 5-dimensional finite-sum objective (see Theorem A.8).

The proofs are deferred to Appendices D.2 and D.3. For the upper bound, we establish the per-epoch contraction inequality

$$\mathbb{E}\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\|^2 \leq \left(1 - \frac{\eta\mu n}{2}\right) \mathbb{E}\|\boldsymbol{x}_0^k - \boldsymbol{x}^*\|^2 + 4\eta^3 n^2 LG_*^2 + 64\eta^3 nL^2 G_*^2/\mu.$$

Compared with the convex-component analysis of Mishchenko et al. (2020), the contraction factor has the same form and the $\mathcal{O}(\eta^3 n^2 LG_*^2)$ noise term remains of the same order, while dropping component convexity introduces the additional $\mathcal{O}(\eta^3 nL^2 G_*^2/\mu)$ noise term. After unrolling the recursion for K epochs and substituting the stated stepsize, these two noise terms yield the 1 and κ/n branches in Equation (3), respectively. We note that the bound established in the recent work of Li (2026, Corollary 3.7) can be written as $\widetilde{\mathcal{O}}(L\bar{G}^2/(\mu^3 nK^2) \cdot \kappa)$ where $\bar{G} \geq G_*$ is the corresponding parameter in their analysis. Our upper bound therefore improves both the convergence rate and the epoch requirement.

For the matching lower bound, we design two key constructions corresponding to the two noise terms in the upper-bound analysis. The first is the piecewise-quadratic construction from the convex component case lower bound proof of Theorem A.4. For $K \gtrsim \kappa$, it yields $\Omega(LG_*^2/(\mu^3 nK^2))$, matching the 1 branch in Equation (3) and recovering the large-epoch rate under component convexity. The second uses 1-dimensional quadratic components with positive and negative curvatures and yields $\Omega(L^2 G_*^2/(\mu^4 n^2 K^2))$ for $K \gtrsim \kappa^2/n$, matching the additional κ/n branch. Combining these constructions on separate coordinate blocks yields a single fixed instance whose lower bound matches the upper bound up to polylogarithmic factors. In particular, the additional κ/n branch is unavoidable without component convexity.

A natural question is whether the rate in Equation (3) extends to epoch counts below the threshold $\max\{\kappa, \kappa^2/n\}$. Our next theorem shows that this is not possible in general.

Theorem 4.2 (Initialization LB for RR without convex components)**.** *For any $n \geq 1024$, $\kappa \geq 1024(1 + \log n)$, $K \lesssim \max\{\kappa, \kappa^2/n\}$, $G_* \geq 0$, and $D \geq 0$, there exists an $\mathcal{O}(\exp(n))$-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

The proof is given in Appendix D.1. The lower bound is established through two complementary constructions. The first construction covers the regime $\kappa \gtrsim n$, for which a constant fraction of the initialization error persists for up to κ^2/n epochs. When $n \gtrsim \kappa$, the second construction yields this behavior for up to κ epochs. Together, these constructions identify the threshold $\max\{\kappa, \kappa^2/n\}$, up to a remaining gap near the crossover regime $\kappa \asymp n$. In fact, our construction and the result remain valid for arbitrary without-replacement SGD. The same fixed objective and initialization satisfy $\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2 \gtrsim D^2$ simultaneously for every constant stepsize η and every sequence of epoch permutations $(\sigma_1, \ldots, \sigma_K)$, covering all $(n!)^K$ possible permutation sequences.

Theorem 4.2 shows that RR can retain a constant fraction of its initialization error for $K \lesssim \max\{\kappa, \kappa^2/n\}$, regardless of the choice of constant stepsize. In contrast, Theorem 4.1 establishes the tight worst-case rate for $K \gtrsim \max\{\kappa, \kappa^2/n\}$. Within the parameter range specified in Theorem 4.2, these results characterize the behavior of RR across the full range of epoch counts, up to

polylogarithmic factors. Also, compared with the convex-component result of RR in Theorem 3.2, the epoch requirement for RR to escape the initialization-error regime increases from $K \gtrsim \kappa/n$ to $K \gtrsim \max\{\kappa, \kappa^2/n\}$, and the large-epoch rate acquires an additional factor of $\max\{1, \kappa/n\}$. We defer a more detailed discussion of these differences and their implications for the comparison with SGD to Section 4.2.

Lastly, we make a remark on the comparison with Kim et al. (2025, Theorem 3.5), which considers the same function class as ours. For suitable initializations and $K \lesssim \kappa$, they establish an $\Omega(D^2)$ lower bound for Incremental Gradient Descent (IGD) under every constant stepsize using a construction involving concave components. Here, IGD processes the components in the same deterministic order $1, \dots, n$ at every epoch. Thus, a lower-bound instance need only be hard for this prescribed order, and their construction is tailored to exploit it. Establishing such a lower bound for RR is more challenging, as it requires a single fixed instance whose error remains large in expectation over the random permutations. Theorem 4.2 not only achieves this but also goes beyond it; Theorem 4.2 guarantees the same lower bound for every sequence of epoch permutations.

4.2 Comparison with With-replacement SGD

For comparison, we give two propositions describing the behavior of SGD after dropping the component-convexity assumption, across the full range of epoch counts. The first combines an upper bound, obtained by adapting Nguyen et al. (2018, Theorem 4) to a constant stepsize, with a lower bound matching up to polylogarithmic factors.

Proposition 4.3 (SGD without component convexity)**.** *Let $n \geq 2$, $T \gtrsim \kappa^2$, and $G_* > 0$. Under Assumption 2.3, the worst-case final-iterate error of with-replacement SGD with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{G_*^2}{\mu^2 T}\right).$$

In particular, the upper bound is attained with $\eta = \Lambda_T/\mu T$ (see Proposition A.5). Provided that $D \geq G_*/(\mu\sqrt{T})$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 1-dimensional finite-sum objective (see Proposition A.6).

The next proposition addresses the complementary small-budget regime.

Proposition 4.4 (Initialization LB for SGD without convex components)**.** *For any $n \geq 2$, $\kappa \geq 3$, $T \lesssim \kappa^2$, $G_* \geq 0$, and $D \geq 0$, there exists a 2-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD satisfies*

$$\mathbb{E}\,\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim D^2.$$

The proofs of Propositions 4.3 and 4.4 are given in Appendix B.2. Compared with Proposition 3.1, removing component convexity extends the regime in which SGD can retain a constant fraction of its initialization error D^2 from $T \lesssim \kappa$ to $T \lesssim \kappa^2$. The worst-case rate beyond this regime remains unchanged. With $T = nK$, the SGD benchmark translates to $\widetilde{\Theta}(G_*^2/(\mu^2 nK))$ for $K \gtrsim \kappa^2/n$.

To compare RR with SGD, write $K_{\mathrm{nc}} := \max\{\kappa, \kappa^2/n\}$. By Theorem 4.2, RR can escape the initialization-error regime only once $K \gtrsim K_{\mathrm{nc}}$, and beyond this threshold, its rate in Theorem 4.1 is $\widetilde{\Theta}(G_*^2/(\mu^2 nK) \cdot K_{\mathrm{nc}}/K)$. Since $K_{\mathrm{nc}} \geq \kappa^2/n$, the SGD benchmark also applies throughout this range. Thus, RR improves upon SGD by a factor of K/K_{nc}, up to polylogarithmic factors.

We now investigate the detailed cost of removing component convexity for RR. We examine three questions: **Q1:** whether RR can be slower than SGD, **Q2:** when RR achieves a faster rate than SGD, and **Q3:** how its large-epoch rate compares with the convex-component rate. The answers depend on the ratio κ/n. Throughout the comparison below, rates and epoch requirements are understood up to polylogarithmic factors.

Case 1: $\kappa \lesssim n$. Here, $K_{\mathrm{nc}} = \kappa$, so RR requires $K \gtrsim \kappa$ to escape the initialization-error regime, whereas the corresponding requirement for SGD is only $K \gtrsim \kappa^2/n$. Removing component convexity therefore extends this regime for both methods, but the extension is larger for RR. When these scales are sufficiently separated, there is an intermediate range $\kappa^2/n \lesssim K \lesssim \kappa$ in which Proposition 4.3

guarantees the upper bound $\widetilde{\mathcal{O}}(G_*^2/(\mu^2 nK))$ for SGD, while Theorem 4.2 gives an $\Omega(D^2)$ lower bound for RR. (**A1**) When the SGD upper bound is sufficiently below D^2, this yields instances on which RR is slower than SGD. Under component convexity, by contrast, there is no epoch range beyond their common initialization-error regime in which RR has a slower worst-case convergence rate than SGD.

For $K \gtrsim K_{\text{nc}} = \kappa$, however, RR improves upon the SGD benchmark by a factor of K/κ. (**A2**) The epoch scale at which RR achieves a faster rate is therefore unchanged from the convex-component case. (**A3**) Moreover, its rate in this range is $\widetilde{\Theta}(LG_*^2/(\mu^3 nK^2))$, which agrees with the convex-component rate in Theorem 3.2. Thus, removing component convexity can make RR slower than SGD before their comparison favors RR, even though neither the epoch scale for this improvement nor the eventual RR rate changes.

Case 2: $\kappa \gtrsim n$. Here, $K_{\text{nc}} = \kappa^2/n$, and both RR and SGD require $K \gtrsim \kappa^2/n$ to escape the regime in which a constant fraction of D^2 can persist. Unlike **Case 1**, there is no range of K for which SGD has already escaped the initialization-error regime while RR is still subject to its initialization lower bound. Outside their common initialization-error regime, RR has no worse worst-case rate scaling than SGD. (**A1**) Thus, removing component convexity does not introduce an analogous range in which RR is slower than SGD.

For $K \gtrsim K_{\text{nc}} = \kappa^2/n$, RR improves upon the SGD benchmark by a factor of nK/κ^2. (**A2**) However, the epoch scale required for this improvement increases from κ under component convexity to κ^2/n. (**A3**) At the same time, the large-epoch RR rate becomes $\widetilde{\Theta}(L^2 G_*^2/(\mu^4 n^2 K^2))$, which is worse than the convex-component rate in Theorem 3.2 by a factor of κ/n. Hence, RR still improves upon SGD, but this improvement requires more epochs and yields a worse rate than in the convex-component case.

Together, these cases show that the cost of removing component convexity takes different forms depending on the relative sizes of κ and n. For $\kappa \lesssim n$, the eventual RR rate and the epoch scale for improving upon SGD remain unchanged, but an intermediate regime emerges in which RR can be slower than SGD. For $\kappa \gtrsim n$, no such intermediate disadvantage arises outside the common initialization-error regime. Instead, the large-epoch RR rate worsens by a factor of κ/n, and the epoch scale required to improve upon SGD increases.

5 Conclusion

We characterize the worst-case last-iterate squared-distance convergence of constant stepsize RR for a finite sum of smooth functions with a strongly convex average. Our upper and lower bounds identify, up to polylogarithmic factors, how its comparison with SGD depends on component convexity and the number of epochs. With convex components, RR matches the SGD rate for $\kappa/n \lesssim K \lesssim \kappa$ and improves upon it for $K \gtrsim \kappa$, establishing a sharp phase transition beyond the quadratic setting.

Without component convexity, RR still improves upon SGD for $K \gtrsim \max\{\kappa, \kappa^2/n\}$, but removing this assumption has two distinct consequences. When n is sufficiently larger than κ, RR can be slower than SGD in an intermediate regime because its initialization error can persist longer, even though its eventual rate and the epoch scale for improving upon SGD remain unchanged. When κ is larger than n, the large-epoch rate worsens by a factor of κ/n compared to the convex component case, and the epoch scale for improving upon SGD increases from κ to κ^2/n.

Natural directions for future work include determining whether varying stepsizes or alternative permutation rules can change the comparison between without-replacement and with-replacement SGD.

AI use statement

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims and to assist in the writing of proofs. We did not use generative AI tools for any other applicable tasks requiring disclosure, and the remaining required-disclosure tasks were not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability.

We reviewed all AI-assisted work. In particular, all AI-assisted mathematical arguments and proofs were independently checked by the authors for correctness and consistency with the stated assumptions and results. All manuscript edits were reviewed and revised by the authors as appropriate.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This work is theoretical, and we do not anticipate any specific ethical concerns arising from its contributions.

Reproducibility statement

We state our assumptions in Section 2, the respective theoretical statements through Sections 3 and 4, and provide detailed proofs in Appendices.

References

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

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

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods as a model of forgetting. In *International Conference on Learning Representations*, volume 2025, pp. 102613–102647, 2025.

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In *International Conference on Machine Learning*, pp. 3855–3912. PMLR, 2023.

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

Jeff Haochen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In *International Conference on Machine Learning*, pp. 2624–2633. PMLR, 2019.

Yujun Kim, Jaeyoung Cha, and Chulhee Yun. Incremental gradient descent with small epoch counts is surprisingly slow on ill-conditioned problems. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=LiXD7mpjU0.

Anastasia Koloskova, Nikita Doikov, Sebastian U Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ZRMQX6aTUS.

Jiaxiang Li. Centered permutation prefixes for SGD with random reshuffling: Sharp rates, Hölder geometry, and composite proximal extensions. *arXiv preprint arXiv:2609.04578*, 2026.

Zijian Liu. Random reshuffling dominates stochastic gradient descent. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4859–4882. PMLR, 29 Jun–03 Jul 2026.

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=Xdy9bjwHDu.

Yucheng Lu, Wentao Guo, and Christopher De Sa. GraB: Finding provably better data permutations than random reshuffling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=nDemfqKHTpK.

Yucheng Lu, Si Yi Meng, and Christopher De Sa. A general analysis of example-selection for stochastic gradient descent. In *International Conference on Learning Representations*, 2022b.

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

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pp. 4703–4711. PMLR, 2019.

Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. SGD and Hogwild! Convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pp. 3750–3758. PMLR, 2018.

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

Phuong_Ha Nguyen, Lam Nguyen, and Marten van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in SGD. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

CONTENTS

1 Introduction **1**

1.1 Summary of Our Contributions 2

2 Problem Setup and Related Work **2**

2.1 Related Work 4

3 Convergence with Component Convexity **5**

3.1 Convergence of Random Reshuffling 5

3.2 Comparison with SGD 6

4 Convergence without Component Convexity **6**

4.1 Convergence of Random Reshuffling 7

4.2 Comparison with With-replacement SGD 8

5 Conclusion **9**

A Overview of Theoretical Results **12**

A.1 With Component Convexity 13

A.2 Without Component Convexity 14

B With-Replacement SGD Benchmarks **16**

B.1 With Component Convexity 16

B.2 Without Component Convexity 20

C Random Reshuffling with Component Convexity **22**

C.1 Initialization Lower Bound 22

C.2 Upper Bound 22

C.3 Lower Bound 27

D Random Reshuffling without Component Convexity **33**

D.1 Initialization Lower Bound 33

D.2 Upper Bound 46

D.3 Lower Bound 51

A OVERVIEW OF THEORETICAL RESULTS

In this section, we collect and organize the main propositions and theorems to provide a compact roadmap of the paper's theoretical results. For convenience, we restate each result together with its assumptions and provide direct references to the corresponding proofs in the subsequent appendices. This section is intended to serve as a quick index to the theoretical results developed throughout the paper. Refer Figure 3 for the theorem proof dependency.

Assumption 2.3 (Common Assumption)**.** Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, and F is μ-strongly convex, with $L \geq \mu > 0$.

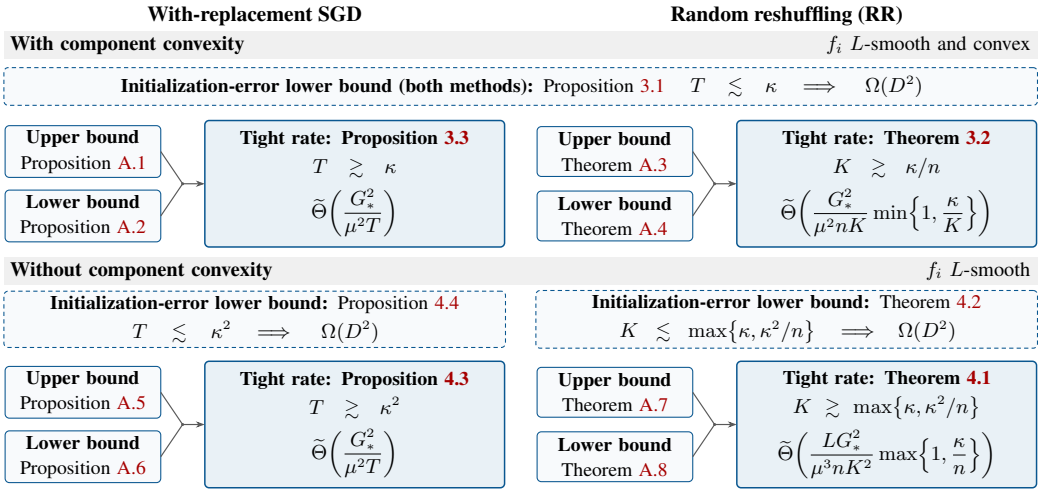


Figure 3: Overview of the results and their dependencies. Rows distinguish whether component convexity is assumed, and columns distinguish between SGD and RR. Rates concern the worst-case final-iterate squared distance; $T = nK$ for RR.

Assumption 2.4 (Component Convexity)**.** Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex.

A.1 With Component Convexity

Proposition 3.1 (Initialization LB for SGD and RR with convex components)**.** *For any $n \geq 2$, $\kappa \geq 4$, $T \lesssim \kappa$, $G_* \geq 0$, and $D \geq 0$, there exists a 3-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterates $\boldsymbol{x}_T$ and $\boldsymbol{x}_n^K$ obtained by with-replacement SGD and RR, respectively, satisfy*

$$\mathbb{E}\,\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim D^2, \qquad \mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

Proof. See Appendix B.1.

A.1.1 With-replacement SGD

Proposition 3.3 (SGD with convex components)**.** *Let $n \geq 2$, $T \gtrsim \kappa$, and $G_* > 0$. Under Assumptions 2.3 and 2.4, the worst-case final-iterate error of with-replacement SGD with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{G_*^2}{\mu^2 T}\right).$$

In particular, the upper bound is attained with $\eta = \Lambda_T/\mu T$ (see Proposition A.1). Provided that $D \gtrsim G_*/(\mu\sqrt{T})$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 1-dimensional finite-sum objective (see Proposition A.2).

Proof. The claim follows by combining the upper bound in Proposition A.1 with the matching lower bound in Proposition A.2. □

Proposition A.1 (SGD Upper Bound with convex components)**.** *Let $n \geq 2$, $T \gtrsim \kappa$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumptions 2.3 and 2.4. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD with a stepsize $\eta = \Lambda_T/\mu T$ satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \lesssim \frac{G_*^2}{\mu^2 T}.$$

Proof. See Appendix B.1.

Proposition A.2 (SGD Lower Bound with convex components). *For any $n \geq 2$, $T \geq 1$, $G_* \geq 0$, and $D \geq 0$, there exists a 1-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim \min\left\{D^2,\ \frac{G_*^2}{\mu^2 T}\right\}.$$

Proof. See Appendix B.1.

A.1.2 Random Reshuffling

Theorem 3.2 (RR with convex components). *Let $n \geq 2$, $\kappa \geq 4096$, $K \gtrsim \kappa/n$, and $G_* > 0$. Under Assumptions 2.3 and 2.4, the worst-case final-iterate error of RR with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{G_*^2}{\mu^2 nK}\min\left\{1,\ \frac{\kappa}{K}\right\}\right).$$

In particular, the upper bound is attained with $\eta = \Lambda_{nK}/\mu nK$ (see Theorem A.3). Provided that $D \geq G_*/(\mu\sqrt{nK})$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 4-dimensional finite-sum objective (see Theorem A.4).

Proof. The claim follows by combining the upper bound in Theorem A.3 with the matching lower bound in Theorem A.4. □

Theorem A.3 (RR Upper Bound with convex components). *Let $n \geq 2$, $K \gtrsim \kappa/n$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumptions 2.3 and 2.4. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR with a stepsize $\eta = \Lambda_{nK}/\mu nK$ satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \lesssim \frac{G_*^2}{\mu^2 nK}\min\left\{1,\ \frac{\kappa}{K}\right\}.$$

Proof. See Appendix C.2.

Theorem A.4 (RR Lower Bound with convex components). *For any even $n \geq 2$, $K \geq 1$, $\kappa \geq 4096$, $G_* \geq 0$, and $D \geq 0$, there exists a 4-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim \min\left\{D^2,\ \frac{G_*^2}{\mu^2 nK}\min\left\{1,\ \frac{\kappa}{K}\right\}\right\}.$$

Proof. See Appendix C.3.

A.2 Without Component Convexity

A.2.1 With-replacement SGD

Proposition 4.4 (Initialization LB for SGD without convex components). *For any $n \geq 2$, $\kappa \geq 3$, $T \lesssim \kappa^2$, $G_* \geq 0$, and $D \geq 0$, there exists a 2-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_T - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

Proof. See Appendix B.2.

Proposition 4.3 (SGD without component convexity). *Let $n \geq 2$, $T \gtrsim \kappa^2$, and $G_* > 0$. Under Assumption 2.3, the worst-case final-iterate error of with-replacement SGD with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{G_*^2}{\mu^2 T}\right).$$

In particular, the upper bound is attained with $\eta = \Lambda_T/\mu T$ (see Proposition A.5). Provided that $D \geq G_*/(\mu\sqrt{T})$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 1-dimensional finite-sum objective (see Proposition A.6).

Proof. The claim follows by combining the upper bound in Proposition A.5 with the matching lower bound in Proposition A.6. □

Proposition A.5 (SGD Upper Bound without component convexity)**.** *Let $n \geq 2$, $T \gtrsim \kappa^2$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumption 2.3. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD with a stepsize $\eta = \Lambda_T/\mu T$ satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \lesssim \frac{G_*^2}{\mu^2 T}.$$

Proof. See Appendix B.2.

Proposition A.6 (SGD Lower Bound without component convexity)**.** *For any $n \geq 2$, $T \geq 1$, $G_* \geq 0$, and $D \geq 0$, there exists a 1-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim \min\left\{D^2, \frac{G_*^2}{\mu^2 T}\right\}.$$

Proof. See Appendix B.2.

A.2.2 Random Reshuffling

Theorem 4.2 (Initialization LB for RR without convex components)**.** *For any $n \geq 1024$, $\kappa \geq 1024(1+\log n)$, $K \lesssim \max\{\kappa, \kappa^2/n\}$, $G_* \geq 0$, and $D \geq 0$, there exists an $\mathcal{O}(\exp(n))$-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

Proof. See Appendix D.1.

Theorem 4.1 (RR without component convexity)**.** *Let $n \geq 2$, $\kappa \geq 4096$, $K \gtrsim \max\{\kappa, \kappa^2/n\}$, and $G_* > 0$. Under Assumption 2.3, the worst-case final-iterate error of RR with a constant stepsize is*

$$\mathbb{E}\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2 = \widetilde{\Theta}\left(\frac{LG_*^2}{\mu^3 nK^2}\max\left\{1, \frac{\kappa}{n}\right\}\right). \tag{3}$$

In particular, the upper bound is attained with $\eta = \Lambda_{nK}/\mu nK$ (see Theorem A.7). Provided that $D \geq \sqrt{\kappa}G_*/(\mu\sqrt{n}K)\max\{1, \sqrt{\kappa/n}\}$, the lower bound holds for any constant stepsize $\eta > 0$ via a suitable 5-dimensional finite-sum objective (see Theorem A.8).

Proof. The claim follows by combining the upper bound in Theorem A.7 with the matching lower bound in Theorem A.8. □

Theorem A.7 (RR Upper Bound without component convexity)**.** *Let $n \geq 2$, $K \gtrsim \max\{\kappa, \kappa^2/n\}$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumption 2.3. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR with a stepsize $\eta = \Lambda_{nK}/\mu nK$ satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \lesssim \frac{LG_*^2}{\mu^3 nK^2}\max\left\{1, \frac{\kappa}{n}\right\}.$$

Proof. See Appendix D.2.

Theorem A.8 (RR Lower Bound without component convexity)**.** *For any even $n \geq 16$, $\kappa \geq 4096$, $K \geq \max\{\kappa, \kappa^2/n\}$, $G_* \geq 0$, and $D \geq 0$, there exists a 5-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim \min\left\{D^2, \frac{LG_*^2}{\mu^3 nK^2}\max\left\{1, \frac{\kappa}{n}\right\}\right\}.$$

Proof. See Appendix D.3.

B WITH-REPLACEMENT SGD BENCHMARKS

In this section, we collect the proofs for the SGD benchmarks used throughout the paper. These results provide the comparison baseline for our RR guarantees under the corresponding assumptions and convergence metric.

B.1 WITH COMPONENT CONVEXITY

Proposition 3.1 (Initialization LB for SGD and RR with convex components)**.** *For any $n \geq 2$, $\kappa \geq 4$, $T \lesssim \kappa$, $G_* \geq 0$, and $D \geq 0$, there exists a 3-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterates $\boldsymbol{x}_T$ and $\boldsymbol{x}_n^K$ obtained by with-replacement SGD and RR, respectively, satisfy*

$$\mathbb{E}\,\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim D^2, \qquad \mathbb{E}\,\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

We make the constants explicit and assume $T = nK \leq \kappa/4$. Fix any $n \geq 2$, $G_* \geq 0$, and $D > 0$, and consider the three-dimensional finite-sum objective

$$f_i(\boldsymbol{x}) = \frac{\mu}{2}u^2 + \frac{L}{2}v^2 + \frac{\mu}{2}\zeta^2 + a_i\zeta, \qquad \boldsymbol{x} = (u, v, \zeta)^\top \in \mathbb{R}^3, \qquad i \in [n],$$

where $a_1 = G_*\sqrt{n-1}$ and $a_i = -G_*/\sqrt{n-1}$ for $i = 2, \ldots, n$. Then $\frac{1}{n}\sum_{i=1}^n a_i = 0$ and $\frac{1}{n}\sum_{i=1}^n a_i^2 = G_*^2$. Each f_i has Hessian $\mathrm{diag}(\mu, L, \mu)$ and is therefore convex and L-smooth. Moreover, $F = \frac{1}{n}\sum_{i=1}^n f_i = \frac{\mu}{2}u^2 + \frac{L}{2}v^2 + \frac{\mu}{2}\zeta^2$ is μ-strongly convex with unique minimizer $\boldsymbol{x}^* = (0,0,0)^\top$. Thus, F satisfies Assumptions 2.3 and 2.4, and

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|^2 = \frac{1}{n}\sum_{i=1}^n a_i^2 = G_*^2.$$

Choose

$$\boldsymbol{x}_0 = \frac{D}{\sqrt{2}}(1, 1, 0)^\top, \qquad \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 = D^2.$$

The auxiliary coordinate ζ is introduced solely to realize the prescribed value of G_*. Its dynamics are decoupled from those of u and v, and its contribution to the squared distance is nonnegative, so we omit it from the lower-bound analysis. Since all component gradients in the (u, v) coordinates coincide, both with-replacement SGD and RR reduce to deterministic gradient descent in these coordinates, regardless of the sampled indices or permutations. In particular, after $T = nK$ component-gradient evaluations, the common (u, v) coordinates of their final iterates satisfy

$$\begin{pmatrix} u_T \\ v_T \end{pmatrix} = \frac{D}{\sqrt{2}}\begin{pmatrix} (1-\eta\mu)^T \\ (1-\eta L)^T \end{pmatrix}.$$

Consequently, for either final iterate $\boldsymbol{x}_{\mathrm{out}} \in \{\boldsymbol{x}_T, \boldsymbol{x}_n^K\}$,

$$\mathbb{E}\left[\|\boldsymbol{x}_{\mathrm{out}} - \boldsymbol{x}^*\|^2\right] \geq u_T^2 + v_T^2 = \frac{D^2}{2}\left((1-\eta\mu)^{2T} + (1-\eta L)^{2T}\right). \tag{4}$$

We bound this expression for every constant stepsize $\eta > 0$.

If $\eta > 2/L$, then $|1 - \eta L| > 1$, so the second term in Equation (4) gives

$$\frac{D^2}{2}(1-\eta L)^{2T} \geq \frac{D^2}{2}.$$

If $0 < \eta \leq 2/L$, then $\eta\mu \leq 2/\kappa$. Since $\kappa \geq 4$, we have $0 \leq \eta\mu \leq 1/2$ and Bernoulli's inequality therefore yields

$$(1-\eta\mu)^T \geq 1 - T\eta\mu \geq 1 - \frac{2T}{\kappa} \geq \frac{1}{2}.$$

Thus, the first term in Equation (4) is at least $D^2/8$. Combining the two cases, for every $\eta > 0$ we obtain

$$\mathbb{E}\left[\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2\right] \geq \frac{D^2}{8}, \qquad \mathbb{E}\left[\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\|^2\right] \geq \frac{D^2}{8},$$

which completes the proof.

Proposition A.1 (SGD Upper Bound with convex components)**.** *Let $n \geq 2$, $T \gtrsim \kappa$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumptions 2.3 and 2.4. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD with a stepsize $\eta = \Lambda_T/\mu T$ satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \lesssim \frac{G_*^2}{\mu^2 T}.$$

Proof. The condition $T \gtrsim \kappa$ in the proposition refers to $T \geq 2\kappa\Lambda_T$.

We first establish a second-moment bound for the stochastic gradient. Since each f_i is convex and L-smooth, we can use Garrigos & Gower (2023, Lemma 2.29):

$$\left\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\right\|^2 \leq 2L\Big(f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}^*) - \left\langle \nabla f_i(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\right\rangle\Big). \tag{5}$$

Averaging Equation (5) over $i \in [n]$ and using $\nabla F(\boldsymbol{x}^*) = 0$ gives

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\right\|^2 \leq 2L\big(F(\boldsymbol{x}) - F(\boldsymbol{x}^*)\big).$$

Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x})\|^2 &\leq \frac{2}{n}\sum_{i=1}^n \left\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\right\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|^2 \\ &\leq 4L\big(F(\boldsymbol{x}) - F(\boldsymbol{x}^*)\big) + 2G_*^2. \end{aligned} \tag{6}$$

Using the update rule and conditioning on $\boldsymbol{x}_t$, we have

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \mid \boldsymbol{x}_t\big] &= \mathbb{E}\Big[\left\|\boldsymbol{x}_t - \eta\nabla f_{i_{t+1}}(\boldsymbol{x}_t) - \boldsymbol{x}^*\right\|^2 \mid \boldsymbol{x}_t\Big] \\ &= \mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\left\langle \nabla f_{i_{t+1}}(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle + \eta^2\|\nabla f_{i_{t+1}}(\boldsymbol{x}_t)\|^2 \mid \boldsymbol{x}_t\big] \\ &= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\left\langle \mathbb{E}\big[\nabla f_{i_{t+1}}(\boldsymbol{x}_t) \mid \boldsymbol{x}_t\big], \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle \\ &\quad + \eta^2\mathbb{E}\big[\|\nabla f_{i_{t+1}}(\boldsymbol{x}_t)\|^2 \mid \boldsymbol{x}_t\big] \\ &= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\left\langle \frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle + \eta^2\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(\boldsymbol{x}_t)\|^2 \\ &= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\big\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\big\rangle + \frac{\eta^2}{n}\sum_{i=1}^n\|\nabla f_i(\boldsymbol{x}_t)\|^2. \end{aligned} \tag{7}$$

By the μ-strong convexity of F,

$$\big\langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\big\rangle \geq F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2.$$

Combining this inequality with Equation (6) and Equation (7) yields

$$\mathbb{E}\big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \mid \boldsymbol{x}_t\big] \leq (1 - \mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta(1 - 2L\eta)\big(F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)\big) + 2\eta^2 G_*^2. \tag{8}$$

Use the common logarithmic factor Λ_T defined in Section 2 and assume $T \geq 2\kappa\Lambda_T$. From $\eta = \frac{\Lambda_T}{\mu T}$, we have $\eta \leq \frac{1}{2L}$, and thus the function-gap term in Equation (8) is nonpositive and taking total expectation gives

$$\mathbb{E}\big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2\big] \leq (1 - \mu\eta)\mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2\big] + 2\eta^2 G_*^2. \tag{9}$$

Since $\mu\eta \leq \mu/(2L) \leq 1/2$, unrolling Equation (9) with $D_0 := \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$ gives

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2\big] &\leq (1 - \mu\eta)^T D_0^2 + 2\eta^2 G_*^2 \sum_{s=0}^{T-1}(1 - \mu\eta)^s \\ &\leq \exp(-\mu\eta T)D_0^2 + \frac{2\eta G_*^2}{\mu}\end{aligned}$$

$$= \exp(-\Lambda_T) D_0^2 + \frac{2G_*^2 \Lambda_T}{\mu^2 T}. \tag{10}$$

Let $r \coloneqq \mu^2 T D_0^2 / G_*^2$. Since $\Lambda_T \geq \log(1+r)$,

$$\exp(-\Lambda_T) D_0^2 \leq \frac{D_0^2}{1+r} = \frac{G_*^2}{\mu^2 T} \cdot \frac{r}{1+r} \leq \frac{G_*^2}{\mu^2 T}.$$

Substituting this into Equation (10) and using $\Lambda_T \geq 1$ gives

$$\mathbb{E}\left[\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2\right] \leq \frac{(1+2\Lambda_T) G_*^2}{\mu^2 T} \leq \frac{3 G_*^2 \Lambda_T}{\mu^2 T},$$

which completes the proof. □

Proposition A.2 (SGD Lower Bound with convex components)**.** *For any $n \geq 2$, $T \geq 1$, $G_* \geq 0$, and $D \geq 0$, there exists a 1-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim \min\left\{D^2, \frac{G_*^2}{\mu^2 T}\right\}.$$

Proof. Although the proposition is stated for a constant stepsize, we prove the stronger result for an arbitrary deterministic varying stepsize sequence $\{\eta_t\}_{t=1}^T \subset (0, \infty)$ fixed before the component indices are sampled. The result in the proposition follows by taking $\eta_t = \eta$ for all $t \in \{1, \ldots, T\}$.

For $x \in \mathbb{R}$, consider the component functions

$$f_1(x) \coloneqq \frac{\mu}{2} x^2 + G_* \sqrt{n-1}\, x,$$
$$f_i(x) \coloneqq \frac{\mu}{2} x^2 - \frac{G_*}{\sqrt{n-1}}\, x, \qquad i = 2, \ldots, n.$$

Averaging the component functions gives $F(x) = \frac{\mu}{2} x^2$ and the unique minimizer is $x^* = 0$. Moreover, for every $i \in [n]$ and $x \in \mathbb{R}$, $\nabla^2 f_i(x) = \mu$ holds. Thus, every component function is μ-strongly convex and μ-smooth. Since $\mu \leq L$, every component function is also L-smooth.

At the optimum, the component gradients satisfy

$$\nabla f_1(x^*) = G_* \sqrt{n-1},$$
$$\nabla f_i(x^*) = -\frac{G_*}{\sqrt{n-1}}, \qquad i = 2, \ldots, n.$$

It follows that

$$\frac{1}{n} \sum_{i=1}^n \left\|\nabla f_i(x^*)\right\|^2 = \frac{1}{n}\left(G_*^2 (n-1) + (n-1)\frac{G_*^2}{n-1}\right) = G_*^2.$$

For the prescribed $D > 0$, choose the initialization point $x_0 = D$, so that $|x_0 - x^*| = D$. The objective and initialization are fixed independently of the stepsize sequence. For every $t \in \{1, \ldots, T\}$, define

$$\xi_t \coloneqq \begin{cases} G_* \sqrt{n-1}, & i_t = 1, \\ -\dfrac{G_*}{\sqrt{n-1}}, & i_t \in \{2, \ldots, n\}. \end{cases}$$

Since the indices $i_1, \ldots, i_T$ are sampled independently and uniformly from $[n]$, we have

$$\mathbb{E}[\xi_t] = \frac{1}{n} G_* \sqrt{n-1} - \frac{n-1}{n} \frac{G_*}{\sqrt{n-1}} = 0,$$

and

$$\mathbb{E}[\xi_t^2] = \frac{1}{n} G_*^2 (n-1) + \frac{n-1}{n} \frac{G_*^2}{n-1} = G_*^2.$$

The random variables $\xi_1, \dots, \xi_T$ are therefore independent, centered, and have second moment G_*^2. Under the varying stepsize update

$$x_t = x_{t-1} - \eta_t \nabla f_{i_t}(x_{t-1}), \qquad t = 1, \dots, T,$$

the iterates satisfy

$$x_t = (1 - \mu\eta_t)x_{t-1} - \eta_t \xi_t.$$

For every $t \in \{0, \dots, T\}$, define

$$q_t \coloneqq \prod_{s=t+1}^{T} (1 - \mu\eta_s),$$

where an empty product is defined as one, so that $q_T = 1$. Iterating the recursion for x_t gives

$$x_T = q_0 x_0 - \sum_{t=1}^{T} \eta_t q_t \xi_t.$$

Because the stepsize sequence is deterministic and the random variables $\xi_1, \dots, \xi_T$ are independent and centered, all cross terms vanish after taking expectation. Therefore,

$$\begin{aligned}
\mathbb{E}[x_T^2] &= q_0^2 x_0^2 + \mathbb{E}\left[\left(\sum_{t=1}^{T} \eta_t q_t \xi_t\right)^2\right] \\
&= q_0^2 x_0^2 + \sum_{t=1}^{T} \eta_t^2 q_t^2 \mathbb{E}[\xi_t^2] \\
&= q_0^2 x_0^2 + G_*^2 \sum_{t=1}^{T} \eta_t^2 q_t^2.
\end{aligned}$$

By the definition of q_t, for every $t \in \{1, \dots, T\}$,

$$q_{t-1} = (1 - \mu\eta_t) q_t,$$

and hence

$$q_t - q_{t-1} = \mu\eta_t q_t.$$

Summing this identity over $t = 1, \dots, T$ yields

$$1 - q_0 = q_T - q_0 = \mu \sum_{t=1}^{T} \eta_t q_t.$$

Using the Cauchy–Schwarz inequality, we obtain

$$\sum_{t=1}^{T} \eta_t^2 q_t^2 \geq \frac{1}{T} \left(\sum_{t=1}^{T} \eta_t q_t\right)^2 = \frac{(1 - q_0)^2}{\mu^2 T}.$$

Substituting this inequality and $x_0^2 = D^2$ into the exact second-moment expression gives

$$\begin{aligned}
\mathbb{E}\left[|x_T - x^*|^2\right] &= \mathbb{E}[x_T^2] \\
&\geq D^2 q_0^2 + \frac{G_*^2}{\mu^2 T}(1 - q_0)^2 \\
&\geq \min\left\{D^2, \frac{G_*^2}{\mu^2 T}\right\} \left(q_0^2 + (1 - q_0)^2\right).
\end{aligned}$$

For every $q_0 \in \mathbb{R}$,

$$q_0^2 + (1 - q_0)^2 = 2\left(q_0 - \frac{1}{2}\right)^2 + \frac{1}{2} \geq \frac{1}{2}.$$

Consequently, we have

$$\mathbb{E}\left[|x_T - x^*|^2\right] \geq \frac{1}{2} \min\left\{D^2, \frac{G_*^2}{\mu^2 T}\right\}.$$

□

B.2 Without Component Convexity

Proposition 4.4 (Initialization LB for SGD without convex components)**.** *For any $n \geq 2$, $\kappa \geq 3$, $T \lesssim \kappa^2$, $G_* \geq 0$, and $D \geq 0$, there exists a 2-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD satisfies*

$$\mathbb{E}\,\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim D^2.$$

Proof. We make the constants explicit and assume $T \leq \kappa^2/6$. Fix any $n \geq 2$, $G_* \geq 0$, and $D \geq 0$, and let $m \coloneqq \lfloor n/2 \rfloor$. For $\boldsymbol{x} = (x, \zeta)^\top \in \mathbb{R}^2$, consider the two-dimensional component functions

$$f_i(\boldsymbol{x}) \coloneqq \frac{a_i}{2}x^2 + \frac{\mu}{2}\zeta^2 + b_i\zeta, \qquad a_i \coloneqq \begin{cases} L, & 1 \leq i \leq m, \\ 2\mu - L, & m+1 \leq i \leq 2m, \\ \mu, & 2m+1 \leq i \leq n. \end{cases}$$

Here $b_1 = G_*\sqrt{n-1}$ and $b_i = -G_*/\sqrt{n-1}$ for $i = 2, \ldots, n$, so that $\frac{1}{n}\sum_{i=1}^n b_i = 0$ and $\frac{1}{n}\sum_{i=1}^n b_i^2 = G_*^2$. Each f_i has Hessian $\operatorname{diag}(a_i, \mu)$. Since $\mu \leq L$, we have $|a_i| \leq L$ for every $i \in [n]$, so each f_i is L-smooth. Moreover,

$$\frac{1}{n}\sum_{i=1}^n a_i = \frac{mL + m(2\mu - L) + (n-2m)\mu}{n} = \mu.$$

Thus, $F(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x}) = \frac{\mu}{2}(x^2 + \zeta^2)$ is μ-strongly convex with unique minimizer $\boldsymbol{x}^* = (0,0)^\top$, and Assumption 2.3 holds. Moreover,

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|^2 = \frac{1}{n}\sum_{i=1}^n b_i^2 = G_*^2.$$

Choose $\boldsymbol{x}_0 = (D, 0)^\top$, so that $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$.

The auxiliary coordinate ζ is introduced solely to realize the prescribed value of G_*. Its dynamics are decoupled from those of x, and its contribution to the squared distance is nonnegative, so we omit it from the lower-bound analysis.

Let $i_1, \ldots, i_T$ be sampled independently and uniformly from $[n]$. For any constant stepsize $\eta > 0$, write the with-replacement SGD iterates as $\boldsymbol{x}_t = (x_t, \zeta_t)^\top$. Their first coordinates satisfy

$$x_t = x_{t-1} - \eta\partial_x f_{i_t}(\boldsymbol{x}_{t-1}) = (1 - \eta a_{i_t})x_{t-1}, \qquad t = 1, \ldots, T.$$

Define

$$A \coloneqq \frac{1}{n}\sum_{i=1}^n a_i^2 = \frac{mL^2 + m(2\mu - L)^2 + (n-2m)\mu^2}{n}.$$

Since $m/n \geq 1/3$ for every $n \geq 2$, we have $A \geq L^2/3$. Using the independence of i_t and x_{t-1} gives

$$\begin{aligned} \mathbb{E}\big[x_t^2 \mid x_{t-1}\big] &= \frac{1}{n}\sum_{i=1}^n (1 - \eta a_i)^2 x_{t-1}^2 \\ &= \left(1 - 2\mu\eta + A\eta^2\right) x_{t-1}^2. \end{aligned} \tag{11}$$

For every $\eta > 0$, we have

$$\begin{aligned} 1 - 2\mu\eta + A\eta^2 &= A\left(\eta - \frac{\mu}{A}\right)^2 + 1 - \frac{\mu^2}{A} \\ &\geq 1 - \frac{\mu^2}{A} \geq 1 - \frac{3}{\kappa^2}. \end{aligned}$$

Since $\kappa \geq 3$, we have $0 \leq 3/\kappa^2 \leq 1/2$. Taking total expectation in Equation (11), iterating the resulting recursion, and applying Bernoulli's inequality gives

$$\mathbb{E}\big[\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2\big] \geq \mathbb{E}\big[x_T^2\big] = \left(1 - 2\mu\eta + A\eta^2\right)^T D^2$$

$$\geq \left(1 - \frac{3}{\kappa^2}\right)^T D^2$$
$$\geq \left(1 - \frac{3T}{\kappa^2}\right) D^2 \geq \frac{D^2}{2}.$$

□

Proposition A.5 (SGD Upper Bound without component convexity)**.** *Let $n \geq 2$, $T \gtrsim \kappa^2$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumption 2.3. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD with a stepsize $\eta = \Lambda_T/\mu T$ satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \lesssim \frac{G_*^2}{\mu^2 T}.$$

Proof. The condition $T \gtrsim \kappa^2$ in the theorem refers to $T \geq 2\kappa^2\Lambda_T$.

We adapt the argument of Nguyen et al. (2018, Theorem 4) to a constant stepsize. We first establish a second-moment bound for the stochastic gradient, following their Lemma 2. Since each f_i is L-smooth,

$$\left\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\right\|^2 \leq L^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2.$$

Averaging this inequality over $i \in [n]$ and using the μ-strong convexity of F gives

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\right\|^2 \leq L^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2$$
$$\leq 2L\kappa\big(F(\boldsymbol{x}) - F(\boldsymbol{x}^*)\big).$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x})\|^2 \leq \frac{2}{n}\sum_{i=1}^n \left\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}^*)\right\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|^2$$
$$\leq 4L\kappa\big(F(\boldsymbol{x}) - F(\boldsymbol{x}^*)\big) + 2G_*^2. \tag{12}$$

Since i_{t+1} is sampled independently and uniformly from $[n]$, using the update rule and conditioning on $\boldsymbol{x}_t$, we have

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \mid \boldsymbol{x}_t\big] &= \mathbb{E}\Big[\left\|\boldsymbol{x}_t - \eta\nabla f_{i_{t+1}}(\boldsymbol{x}_t) - \boldsymbol{x}^*\right\|^2 \mid \boldsymbol{x}_t\Big] \\
&= \mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\left\langle\nabla f_{i_{t+1}}(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle + \eta^2\|\nabla f_{i_{t+1}}(\boldsymbol{x}_t)\|^2 \mid \boldsymbol{x}_t\big] \\
&= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\left\langle\mathbb{E}\big[\nabla f_{i_{t+1}}(\boldsymbol{x}_t) \mid \boldsymbol{x}_t\big], \boldsymbol{x}_t - \boldsymbol{x}^*\right\rangle \\
&\quad + \eta^2\mathbb{E}\big[\|\nabla f_{i_{t+1}}(\boldsymbol{x}_t)\|^2 \mid \boldsymbol{x}_t\big] \\
&= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\big\langle\nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\big\rangle + \frac{\eta^2}{n}\sum_{i=1}^n\|\nabla f_i(\boldsymbol{x}_t)\|^2.
\end{aligned} \tag{13}$$

By the μ-strong convexity of F,

$$\big\langle\nabla F(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^*\big\rangle \geq F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2.$$

Combining this inequality with Equation (12) and Equation (13) yields

$$\mathbb{E}\big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \mid \boldsymbol{x}_t\big] \leq (1 - \mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta(1 - 2L\kappa\eta)\big(F(\boldsymbol{x}_t) - F(\boldsymbol{x}^*)\big) + 2\eta^2 G_*^2. \tag{14}$$

Let $D_0 := \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$ and recall that Λ_T is the common logarithmic factor defined in Section 2. Assume $T \geq 2\kappa^2\Lambda_T$, which makes the condition $T \gtrsim \kappa^2$ precise. From $\eta = \frac{\Lambda_T}{\mu T}$, we have $\eta \leq \frac{1}{2L\kappa}$, and thus the function-gap term in Equation (14) is nonpositive and taking total expectation gives

$$\mathbb{E}\big[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2\big] \leq (1 - \mu\eta)\mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2\big] + 2\eta^2 G_*^2. \tag{15}$$

Since $\mu\eta \le \mu/(2L\kappa) = 1/(2\kappa^2) \le 1/2$, unrolling Equation (15) gives

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2\right] &\le (1-\mu\eta)^T D_0^2 + 2\eta^2 G_*^2 \sum_{s=0}^{T-1}(1-\mu\eta)^s \\ &\le \exp(-\mu\eta T) D_0^2 + \frac{2\eta G_*^2}{\mu} \\ &= \exp(-\Lambda_T) D_0^2 + \frac{2G_*^2 \Lambda_T}{\mu^2 T}. \end{aligned} \tag{16}$$

Let $r \coloneqq \mu^2 T D_0^2 / G_*^2$. Since $\Lambda_T \ge \log(1+r)$,

$$\exp(-\Lambda_T) D_0^2 \le \frac{D_0^2}{1+r} = \frac{G_*^2}{\mu^2 T} \cdot \frac{r}{1+r} \le \frac{G_*^2}{\mu^2 T}.$$

Substituting this into Equation (16) and using $\Lambda_T \ge 1$ gives

$$\mathbb{E}\left[\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2\right] \le \frac{(1+2\Lambda_T)G_*^2}{\mu^2 T} \le \frac{3G_*^2\Lambda_T}{\mu^2 T},$$

which completes the proof. □

Proposition A.6 (SGD Lower Bound without component convexity)**.** *For any $n \ge 2$, $T \ge 1$, $G_* \ge 0$, and $D \ge 0$, there exists a 1-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_T$ obtained by with-replacement SGD satisfies*

$$\mathbb{E}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \gtrsim \min\left\{D^2, \frac{G_*^2}{\mu^2 T}\right\}.$$

Proof. The one-dimensional instance constructed in the proof of Proposition A.2 satisfies Assumptions 2.3 and 2.4, and is therefore also admissible under Assumption 2.3 alone. Using the same instance and initialization, Proposition A.2 gives the claimed lower bound. □

C Random Reshuffling with Component Convexity

In this section, we provide the proofs for RR under component convexity.

C.1 Initialization Lower Bound

Proposition 3.1 applies to both SGD and RR. Since its proof is already given in Appendix B.1 as part of the SGD benchmarks, we only restate the proposition here for completeness and refer the reader to Appendix B.1 for the proof.

Proposition 3.1 (Initialization LB for SGD and RR with convex components)**.** *For any $n \ge 2$, $\kappa \ge 4$, $T \lesssim \kappa$, $G_* \ge 0$, and $D \ge 0$, there exists a 3-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterates $\boldsymbol{x}_T$ and $\boldsymbol{x}_n^K$ obtained by with-replacement SGD and RR, respectively, satisfy*

$$\mathbb{E}\left\|\boldsymbol{x}_T - \boldsymbol{x}^*\right\|^2 \gtrsim D^2, \qquad \mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

C.2 Upper Bound

Theorem A.3 (RR Upper Bound with convex components)**.** *Let $n \ge 2$, $K \gtrsim \kappa/n$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumptions 2.3 and 2.4. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR with a stepsize $\eta = \Lambda_{nK}/\mu nK$ satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \lesssim \frac{G_*^2}{\mu^2 nK} \min\left\{1, \frac{\kappa}{K}\right\}.$$

Proof. The condition $K \gtrsim \kappa/n$ in the theorem refers to $K \geq \max\{2\Lambda_{nK}, 16\kappa\Lambda_{nK}/n\}$.

We first give a new proof of the bound $\mathbb{E}\left[\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2\right] \lesssim G_*^2/(\mu^2 nK)$ that applies for any $K \gtrsim \kappa/n$. For $K \geq \sqrt{2}\kappa\Lambda_{nK}$, we use Theorem 2 of Mishchenko et al. (2020) to obtain the sharper bound $\kappa G_*^2/(\mu^2 nK^2)$.

We denote the Bregman divergence of the component functions as

$$D_i := D_{f_i}, \quad \text{where } D_f(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle .$$

Our analysis relies on inequalities from Lemmas 4 and 5 and Equation (7) of Liu (2026), which we state here for completeness:

First, the RR update satisfies the exact identity

$$\begin{aligned}\left\|\boldsymbol{x}_i^k - \boldsymbol{x}^*\right\|^2 = \left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2 - 2\eta\big(f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k) - f_{\sigma_k(i)}(\boldsymbol{x}^*)\big) \\ + \eta^2\left\|\nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right\|^2 - 2\eta D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k).\end{aligned} \tag{17}$$

Second, when $\eta \leq 1/(6L)$, the expected sampled component value can be converted to the full objective:

$$\begin{aligned}\mathbb{E}F(\boldsymbol{x}_{i-1}^k) \leq \mathbb{E}f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k) + \eta\mathbb{E}\left\|\nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right\|^2 \\ + \frac{4\eta}{3n}\sum_{j<i}\mathbb{E}\left\|\nabla f_{\sigma_k(j)}(\boldsymbol{x}_{j-1}^k)\right\|^2.\end{aligned} \tag{18}$$

Finally Equation (7) of Liu (2026) gives the convenient gradient estimate

$$\mathbb{E}\left\|\nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right\|^2 \leq 3L\mathbb{E}D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k) + 18G_*^2. \tag{19}$$

These relations retain the within-epoch dependence instead of replacing it with an unbiased-noise surrogate.

We begin by taking expectations in Equation (17):

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}_i^k - \boldsymbol{x}^*\right\|^2\right] =& \mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] - 2\eta(\mathbb{E}\left[f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right] - F(\boldsymbol{x}^*)) \\ &+ \eta^2\mathbb{E}\left[\left\|\nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right\|^2\right] - 2\eta\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] \\ \overset{(a)}{\leq}& \mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] - 2\eta(\mathbb{E}\left[F(\boldsymbol{x}_{i-1}^k)\right] - F(\boldsymbol{x}^*)) \\ &+ 3\eta^2\mathbb{E}\left[\left\|\nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right\|^2\right] + \frac{8\eta^2}{3n}\sum_{j<i}\mathbb{E}\left[\left\|\nabla f_{\sigma_k(j)}(\boldsymbol{x}_{j-1}^k)\right\|^2\right] \\ &- 2\eta\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] \\ \overset{(b)}{\leq}& \mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] - 2\eta(\mathbb{E}\left[F(\boldsymbol{x}_{i-1}^k)\right] - F(\boldsymbol{x}^*)) \\ &- \eta(2 - 9\eta L)\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] + \frac{8\eta^2 L}{n}\sum_{j<i}\mathbb{E}\left[D_{\sigma_k(j)}(\boldsymbol{x}^*, \boldsymbol{x}_{j-1}^k)\right] \\ &+ 102\eta^2 G_*^2 \\ \overset{(c)}{\leq}& (1 - \eta\mu)\mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] \\ &- \eta\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] + \frac{\eta}{2n}\sum_{j<i}\mathbb{E}\left[D_{\sigma_k(j)}(\boldsymbol{x}^*, \boldsymbol{x}_{j-1}^k)\right] \\ &+ 102\eta^2 G_*^2,\end{aligned}$$

where (a) follows from Equation (18), (b) follows from Equation (19), and (c) uses the strong convexity of F and $16\eta L \leq 1$.

Now suppose $\eta \leq 1/(2\mu n)$. Then $(1 - \eta\mu)^n \geq 1/2$. Over an epoch, contraction therefore cannot shrink a negative Bregman term by more than a factor of two, so it absorbs all of the corresponding

positive history terms. In particular, the coefficient of $\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right]$ over the epoch recursion is

$$-(1-\eta\mu)^{n-i}\eta + \frac{\eta}{2n}\sum_{j=i+1}^{n}(1-\eta\mu)^{n-j},$$

which is upper bounded by $-(1-\eta\mu)^n\eta + n\cdot\frac{\eta}{2n} \le 0$ for all $i \in [n]$.

As a result, after n iterations within an epoch,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}^*\right\|^2\right] &= \mathbb{E}\left[\left\|\boldsymbol{x}_n^{k} - \boldsymbol{x}^*\right\|^2\right] \\ &\le (1-\eta\mu)^n\mathbb{E}\left[\left\|\boldsymbol{x}_0^{k} - \boldsymbol{x}^*\right\|^2\right] + 102\eta^2 G_*^2\sum_{i=0}^{n-1}(1-\eta\mu)^i.\end{aligned}$$

Over the epoch recursion, using constant stepsize η for all $k \in [K]$, we have

$$\mathbb{E}\left[\left\|\boldsymbol{x}_n^{K} - \boldsymbol{x}^*\right\|^2\right] \le (1-\eta\mu)^{nK}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 + \frac{102\eta G_*^2}{\mu}.$$

In particular, take $\eta = \Lambda_{nK}/(\mu nK)$. The stepsize restrictions above hold when

$$K \ge 2\Lambda_{nK}, \qquad nK \ge 16\kappa\Lambda_{nK},$$

which makes the condition $K \gtrsim \kappa/n$ precise under our convention that it also includes $K \gtrsim 1$. By the definition of Λ_{nK},

$$(1-\eta\mu)^{nK}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 \le \exp(-\Lambda_{nK})\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 \le \frac{G_*^2}{\mu^2(nK)^2}.$$

This gives

$$\mathbb{E}\left[\left\|\boldsymbol{x}_n^{K} - \boldsymbol{x}^*\right\|^2\right] \le \frac{(1+102\Lambda_{nK})G_*^2}{\mu^2 nK} \lesssim \frac{G_*^2}{\mu^2 nK}.$$

It remains to obtain the sharper bound for $K \gtrsim \kappa$. Under Assumptions 2.3 and 2.4, Theorem 2 of Mishchenko et al. (2020) applies with their $\gamma = \eta$, $T = K$, and $\sigma_*^2 = G_*^2$. In particular, if $\eta \le 1/(\sqrt{2}Ln)$, their result gives

$$\mathbb{E}\left[\left\|\boldsymbol{x}_n^{K} - \boldsymbol{x}^*\right\|^2\right] \le \left(1 - \frac{\eta\mu n}{2}\right)^K\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 + \eta^2\kappa n G_*^2.$$

For the stepsize $\eta = \Lambda_{nK}/(\mu nK)$, the condition holds whenever $K \ge \sqrt{2}\kappa\Lambda_{nK}$. Consequently,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}_n^{K} - \boldsymbol{x}^*\right\|^2\right] &\le \exp(-\Lambda_{nK}/2)\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 + \frac{\kappa G_*^2\Lambda_{nK}^2}{\mu^2 nK^2} \\ &\le \frac{G_*^2}{\mu^2(nK)^2} + \frac{\kappa G_*^2\Lambda_{nK}^2}{\mu^2 nK^2} \\ &\le \frac{(1+\Lambda_{nK}^2)\kappa G_*^2}{\mu^2 nK^2},\end{aligned}$$

where the last inequality uses $n\kappa \ge 1$.

Finally, we combine the two bounds by dividing the cases:

- If $K < \sqrt{2}\kappa\Lambda_{nK}$, then $1 \le \sqrt{2}\Lambda_{nK}\min\{1, \kappa/K\}$. The first bound therefore gives
$$\mathbb{E}\left[\left\|\boldsymbol{x}_n^{K} - \boldsymbol{x}^*\right\|^2\right] \le \sqrt{2}\Lambda_{nK}(1+102\Lambda_{nK})\frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}.$$
- If $K \ge \sqrt{2}\kappa\Lambda_{nK}$, then $K \ge \kappa$, and the second bound gives
$$\mathbb{E}\left[\left\|\boldsymbol{x}_n^{K} - \boldsymbol{x}^*\right\|^2\right] \le (1+\Lambda_{nK}^2)\frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}.$$

Since Λ_{nK} is logarithmic, both cases conclude $\mathbb{E}\left[\left\|\boldsymbol{x}_n^{K} - \boldsymbol{x}^*\right\|^2\right] \lesssim \frac{G_*^2}{\mu^2 nK}\min\{1, \kappa/K\}$. □

C.2.1 Upper Bound for Function Value

As an additional result, we obtain a function-value guarantee by averaging the within-epoch iterates with weights concentrated toward the end of the run. Nagaraj et al. (2019, Theorem 2) also study function-value convergence covering our epoch range, but assume a uniform bound $\|\nabla f_i(\boldsymbol{x})\| \le G$ (i.e. Lipschitz continuity) for every component and every point in their feasible set. Our bound instead depends on G_*^2 defined as the noise at the optimum. In the overlapping epoch range $\kappa/n \lesssim K \lesssim \kappa$, the rate below matches, up to logarithmic factors, with $\Omega(G_*^2/(\mu nK))$ last-iterate lower-bound scaling of Cha et al. (2023, Theorem 3.1).

Define

$$\bar{\boldsymbol{x}} := \frac{1}{W}\sum_{k=1}^{K}\sum_{i=1}^{n}(1-\eta\mu/2)^{n(K-k)+n-i}\boldsymbol{x}_{i-1}^k, \quad W := \sum_{k=1}^{K}\sum_{i=1}^{n}(1-\eta\mu/2)^{n(K-k)+n-i}. \tag{20}$$

Proposition C.1 (RR Average Function-Value Upper Bound with convex components)**.** *Let $n \ge 2$, $K \gtrsim \kappa/n$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumptions 2.3 and 2.4. For any initialization point $\boldsymbol{x}_0$, the weighted average $\bar{\boldsymbol{x}}$ in Equation (20) obtained by RR with a stepsize $\eta = \Lambda_{nK}/(\mu nK)$ satisfies*

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^*)\right] \lesssim \frac{G_*^2}{\mu nK}.$$

Proof. The condition $K \gtrsim \kappa/n$ in the proposition refers to $K \ge \max\{2\Lambda_{nK}, 16\kappa\Lambda_{nK}/n\}$. Since $\Lambda_{nK} \ge 4$, the stepsize $\eta = \Lambda_{nK}/(\mu nK)$ then satisfies

$$\frac{2}{\mu nK} \le \eta \le \min\left\{\frac{1}{2\mu n}, \frac{1}{16L}\right\}.$$

We follow the first part of the proof of Theorem A.3, retaining one half of the function-gap term to control the weighted average. We use the same Bregman divergences $D_i = D_{f_i}$ and the estimates in Equations (17) to (19).

We begin by taking expectations in Equation (17):

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{x}_i^k - \boldsymbol{x}^*\right\|^2\right] =& \mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] - 2\eta(\mathbb{E}\left[f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right] - F(\boldsymbol{x}^*)) \\
&+ \eta^2\mathbb{E}\left[\left\|\nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right\|^2\right] - 2\eta\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] \\
\overset{(a)}{\le}& \mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] - 2\eta(\mathbb{E}\left[F(\boldsymbol{x}_{i-1}^k)\right] - F(\boldsymbol{x}^*)) \\
&+ 3\eta^2\mathbb{E}\left[\left\|\nabla f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k)\right\|^2\right] + \frac{8\eta^2}{3n}\sum_{j<i}\mathbb{E}\left[\left\|\nabla f_{\sigma_k(j)}(\boldsymbol{x}_{j-1}^k)\right\|^2\right] \\
&- 2\eta\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] \\
\overset{(b)}{\le}& \mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] - 2\eta(\mathbb{E}\left[F(\boldsymbol{x}_{i-1}^k)\right] - F(\boldsymbol{x}^*)) \\
&- \eta(2 - 9\eta L)\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] + \frac{8\eta^2 L}{n}\sum_{j<i}\mathbb{E}\left[D_{\sigma_k(j)}(\boldsymbol{x}^*, \boldsymbol{x}_{j-1}^k)\right] \\
&+ 102\eta^2 G_*^2 \\
\overset{(c)}{\le}& (1-\eta\mu/2)\mathbb{E}\left[\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2\right] - \eta(\mathbb{E}\left[F(\boldsymbol{x}_{i-1}^k)\right] - F(\boldsymbol{x}^*)) \\
&- \eta\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right] + \frac{\eta}{2n}\sum_{j<i}\mathbb{E}\left[D_{\sigma_k(j)}(\boldsymbol{x}^*, \boldsymbol{x}_{j-1}^k)\right] \\
&+ 102\eta^2 G_*^2,
\end{aligned}$$

where (a) follows from Equation (18), (b) follows from Equation (19), and (c) uses $16\eta L \le 1$ and the μ-strong convexity of F. In (c), we use $F(\boldsymbol{x}_{i-1}^k) - F(\boldsymbol{x}^*) \ge (\mu/2)\left\|\boldsymbol{x}_{i-1}^k - \boldsymbol{x}^*\right\|^2$ for only one half of the function-gap term.

Since $\eta \le 1/(2\mu n)$, Bernoulli's inequality gives

$$(1-\eta\mu/2)^n \ge 1-\eta\mu n/2 \ge \frac{3}{4} \ge \frac{1}{2}.$$

As in the proof of Theorem A.3, each negative Bregman term therefore absorbs all of its positive history contributions within the epoch. Specifically, for every $i \in [n]$, the coefficient of $\mathbb{E}\left[D_{\sigma_k(i)}(\boldsymbol{x}^*, \boldsymbol{x}_{i-1}^k)\right]$ after unrolling the recursion over the epoch is bounded by

$$-\eta(1-\eta\mu/2)^{n-i} + \frac{\eta}{2n}\sum_{j=i+1}^{n}(1-\eta\mu/2)^{n-j} \le -\eta(1-\eta\mu/2)^n + \frac{\eta}{2} \le 0.$$

The component convexity ensures that these Bregman divergences are nonnegative. Consequently,

$$\begin{aligned}
&\eta\sum_{i=1}^{n}(1-\eta\mu/2)^{n-i}\mathbb{E}\left[F(\boldsymbol{x}_{i-1}^k) - F(\boldsymbol{x}^*)\right] \\
&\quad\le (1-\eta\mu/2)^n\mathbb{E}\left[\left\|\boldsymbol{x}_0^k - \boldsymbol{x}^*\right\|^2\right] - \mathbb{E}\left[\left\|\boldsymbol{x}_n^k - \boldsymbol{x}^*\right\|^2\right] + 102\eta^2 G_*^2\sum_{i=0}^{n-1}(1-\eta\mu/2)^i.
\end{aligned}$$

Multiplying the epoch-k inequality by $(1-\eta\mu/2)^{n(K-k)}$ and summing over $k \in [K]$, the squared-distance terms telescope because $\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_n^k$. Dropping the final nonpositive squared-distance term yields

$$\begin{aligned}
&\eta\sum_{k=1}^{K}\sum_{i=1}^{n}(1-\eta\mu/2)^{n(K-k)+n-i}\mathbb{E}\left[F(\boldsymbol{x}_{i-1}^k) - F(\boldsymbol{x}^*)\right] \\
&\quad\le (1-\eta\mu/2)^{nK}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 + 102\eta^2 G_*^2\sum_{t=0}^{nK-1}(1-\eta\mu/2)^t \\
&\quad\le (1-\eta\mu/2)^{nK}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 + \frac{204\eta G_*^2}{\mu}.
\end{aligned}$$

The normalizing weight in Equation (20) satisfies

$$W = \sum_{t=0}^{nK-1}(1-\eta\mu/2)^t = \frac{1-(1-\eta\mu/2)^{nK}}{\eta\mu/2} \ge \frac{1}{\eta\mu}, \tag{21}$$

since $\eta \ge 2/(\mu nK)$ gives $(1-\eta\mu/2)^{nK} \le \exp(-\eta\mu nK/2) \le e^{-1} < 1/2$. Dividing by ηW and using the convexity of F, we obtain

$$\mathbb{E}\left[F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^*)\right] \le \mu(1-\eta\mu/2)^{nK}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 + 204\eta G_*^2.$$

Finally, recall the common logarithmic factor Λ_{nK} defined in Section 2. For $\eta = \Lambda_{nK}/(\mu nK)$, its definition gives

$$\begin{aligned}
(1-\eta\mu/2)^{nK}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 &\le \exp(-\Lambda_{nK}/2)\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 \\
&\le \frac{\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2}{1+\mu^2(nK)^2\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2/G_*^2} \le \frac{G_*^2}{\mu^2(nK)^2}.
\end{aligned}$$

Substituting these bounds and using $nK \ge 1$ yields

$$\begin{aligned}
\mathbb{E}\left[F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^*)\right] &\le \frac{G_*^2}{\mu(nK)^2} + \frac{204\Lambda_{nK}G_*^2}{\mu nK} \\
&\le \frac{(1+204\Lambda_{nK})G_*^2}{\mu nK} \lesssim \frac{G_*^2}{\mu nK},
\end{aligned}$$

which completes the proof. □

C.3 Lower Bound

Theorem A.4 (RR Lower Bound with convex components)**.** *For any even $n \geq 2$, $K \geq 1$, $\kappa \geq 4096$, $G_* \geq 0$, and $D \geq 0$, there exists a 4-dimensional finite-sum objective F satisfying Assumptions 2.3 and 2.4, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim \min\left\{D^2, \frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}\right\}.$$

Proof. We construct two one-dimensional finite sums and one two-dimensional finite sum, each covering a different range of step sizes. The three functions are then combined across coordinates so that a single objective gives the desired lower bound for every step size. Write $\eta_0 := 1/(2\mu nK)$. Our constructions F_1, F_2, and F_3, with component functions $f_{1,i}$, $f_{2,i}$, and $f_{3,i}$, respectively, will have the following properties:

- **Small step size regime.** Starting from $u_0 = D/\sqrt{2}$, for every $0 \leq \eta < \eta_0$, the final iterate on F_1 satisfies
$$|u_n^K|^2 \geq \frac{D^2}{8}.$$
- **Moderate step size regime.** Starting from $\boldsymbol{y}_0 = \mathbf{0} \in \mathbb{R}^2$, for every $\eta_0 \leq \eta < 2/L$, the final iterate on F_2 satisfies
$$\mathbb{E}\left\|\boldsymbol{y}_n^K\right\|^2 \geq 2^{-32}\frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}.$$
- **Large step size regime.** Starting from $v_0 = D/\sqrt{2}$, for every $\eta \geq 2/L$, the final iterate on F_3 satisfies
$$|v_n^K|^2 \geq \frac{D^2}{2}.$$

If $\eta_0 \geq 2/L$, the moderate regime is empty, and the small and large regimes already cover every nonnegative step size. Each F_j is μ-strongly convex and minimized at zero, and all component functions are convex and L-smooth. The component gradients of F_1 and F_3 vanish at zero, whereas those of F_2 have empirical second moment G_*^2. We verify these properties below.

We aggregate the functions by defining

$$\begin{aligned} F(u, \boldsymbol{y}, v) &:= F_1(u) + F_2(\boldsymbol{y}) + F_3(v), \\ f_i(u, \boldsymbol{y}, v) &:= f_{1,i}(u) + f_{2,i}(\boldsymbol{y}) + f_{3,i}(v), \qquad i \in [n], \\ \boldsymbol{x}_0 &:= \left(\frac{D}{\sqrt{2}}, 0, 0, \frac{D}{\sqrt{2}}\right)^\top. \end{aligned}$$

Block diagonality preserves component convexity, component L-smoothness, and average μ-strong convexity. The minimizer is $\boldsymbol{x}^* = \mathbf{0} \in \mathbb{R}^4$, and the initialization has distance exactly D. Moreover,

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_{2,i}(\mathbf{0})\|^2 = G_*^2.$$

Under any realized sequence of permutations, each block follows its own updates with the same permutations. Hence,

$$\begin{aligned} \mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 &= |u_n^K|^2 + \mathbb{E}\left\|\boldsymbol{y}_n^K\right\|^2 + |v_n^K|^2 \\ &\geq \min\left\{\frac{D^2}{8}, 2^{-32}\frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}, \frac{D^2}{2}\right\} \\ &\geq 2^{-32}\min\left\{D^2, \frac{G_*^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\}\right\}. \end{aligned}$$

It remains to give the three constructions and establish their claimed bounds.

Construction of F_1. Let

$$F_1(u) = \frac{\mu}{2}u^2, \qquad f_{1,i}(u) = F_1(u), \quad i \in [n].$$

Every component is convex and L-smooth, the average is μ-strongly convex, and all component gradients vanish at the minimizer $u^* = 0$. Since the components are identical, the iterate is independent of the permutations. With $u_0 = D/\sqrt{2}$ and $0 \le \eta < \eta_0$, Bernoulli's inequality gives

$$u_n^K = (1-\eta\mu)^{nK}\frac{D}{\sqrt{2}} \ge (1-\eta\mu nK)\frac{D}{\sqrt{2}} \ge \frac{D}{2\sqrt{2}}.$$

Consequently, $|u_n^K|^2 \ge D^2/8$.

Construction of F_2. Define

$$\boldsymbol{\nu} := \begin{pmatrix} \sqrt{8\mu/L} \\ \sqrt{1-8\mu/L} \end{pmatrix}, \qquad \boldsymbol{B} := \operatorname{diag}\left(\mu, \frac{L}{8}\right).$$

Since $\kappa \ge 4096$, we have $\|\boldsymbol{\nu}\| = 1$, $1 - 8\mu/L \ge 3/4$, and $\boldsymbol{B} \succeq \mu \boldsymbol{I}_2$. Since n is even, choose slopes $\xi_1, \dots, \xi_n$ to consist of $n/2$ copies of each of $+1$ and -1. Then,

$$\sum_{i=1}^n \xi_i = 0, \qquad \frac{1}{n}\sum_{i=1}^n \xi_i^2 = 1,$$

and the slope multiset is invariant under sign reversal.

For $\boldsymbol{y} = (x, y)^\top \in \mathbb{R}^2$, define

$$f_{2,i}(\boldsymbol{y}) := \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{B}\boldsymbol{y} + \frac{7L}{16}\big(\min\{\boldsymbol{\nu}^\top \boldsymbol{y}, 0\}\big)^2 + G_* \xi_i \boldsymbol{\nu}^\top \boldsymbol{y}.$$

The one-sided quadratic term is convex, and its gradient is

$$\frac{7L}{8}\min\{\boldsymbol{\nu}^\top \boldsymbol{y}, 0\}\boldsymbol{\nu},$$

which is $7L/8$-Lipschitz. Since $\|\boldsymbol{B}\| = L/8$, every component is convex and L-smooth. The affine terms cancel in the average, giving

$$F_2(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{B}\boldsymbol{y} + \frac{7L}{16}\big(\min\{\boldsymbol{\nu}^\top \boldsymbol{y}, 0\}\big)^2.$$

Thus F_2 is μ-strongly convex and uniquely minimized at $\mathbf{0}$. Also,

$$\nabla f_{2,i}(\mathbf{0}) = G_* \xi_i \boldsymbol{\nu}, \qquad \frac{1}{n}\sum_{i=1}^n \|\nabla f_{2,i}(\mathbf{0})\|^2 = G_*^2.$$

We use the initialization $\boldsymbol{y}_0 = \mathbf{0}$.

Fix $\eta_0 \le \eta < 2/L$ and index the updates consecutively by $t = 0, \dots, nK-1$, so that $\boldsymbol{y}_{nK} = \boldsymbol{y}_n^K$. Let S_t denote the selected slope at update t. For each $k = 0, \dots, K-1$, the vector $(S_{kn}, \dots, S_{kn+n-1})$ is a uniform ordering of the slope multiset, independently across epochs. Set

$$\boldsymbol{A} := \boldsymbol{I}_2 - \eta\boldsymbol{B} = \operatorname{diag}\left(1-\eta\mu, 1-\frac{\eta L}{8}\right),$$
$$q_t := \max\{-\boldsymbol{\nu}^\top \boldsymbol{y}_t, 0\}.$$

The exact update is

$$\boldsymbol{y}_{t+1} = \boldsymbol{A}\boldsymbol{y}_t - \eta G_* S_t \boldsymbol{\nu} + \frac{7\eta L}{8} q_t \boldsymbol{\nu}.$$

Here $0 < 1 - \eta L/8 < 1 - \eta\mu < 1$. Define the terminal weights and their weighted sum by

$$\gamma_t := (1-\eta\mu)^{nK-1-t}, \qquad Q := \sum_{t=0}^{nK-1} \gamma_t q_t.$$

Since $x_0 = 0$ and $\mathbb{E}S_t = 0$ at every position, unrolling the first coordinate gives

$$\mathbb{E}x_{nK} = \frac{7\eta L}{8}\sqrt{\frac{8\mu}{L}}\,\mathbb{E}Q. \tag{22}$$

We will lower-bound $\mathbb{E}Q$ by comparing the iterates with the linear recursion obtained by removing the one-sided quadratic term.

Using the same slopes, define

$$\widetilde{\boldsymbol{y}}_0 := \mathbf{0}, \qquad \widetilde{\boldsymbol{y}}_{t+1} := \boldsymbol{A}\widetilde{\boldsymbol{y}}_t - \eta G_* S_t \boldsymbol{\nu}, \qquad \widetilde{q}_t := \max\{-\boldsymbol{\nu}^\top \widetilde{\boldsymbol{y}}_t, 0\}.$$

For $m \geq 0$, write

$$\delta_m := \boldsymbol{\nu}^\top \boldsymbol{A}^m \boldsymbol{\nu} = \frac{8\mu}{L}(1-\eta\mu)^m + \left(1 - \frac{8\mu}{L}\right)\left(1 - \frac{\eta L}{8}\right)^m \geq 0.$$

Subtracting the two recursions and unrolling yields

$$\boldsymbol{\nu}^\top \boldsymbol{y}_t = \boldsymbol{\nu}^\top \widetilde{\boldsymbol{y}}_t + \frac{7\eta L}{8}\sum_{j<t} \delta_{t-1-j} q_j.$$

The sum on the right is nonnegative. Since the map $s \mapsto \max\{s, 0\}$ is 1-Lipschitz, it follows that

$$\widetilde{q}_t \leq q_t + \frac{7\eta L}{8}\sum_{j<t} \delta_{t-1-j} q_j.$$

We retain a subset of positions near the end of the run. Let

$$H := \left\lfloor \frac{1}{4\eta\mu} \right\rfloor.$$

Since $\eta\mu \leq 2/\kappa \leq 1/2048$ and $\eta \geq \eta_0$, we have

$$\frac{1}{8\eta\mu} \leq H \leq \frac{1}{4\eta\mu} \leq \frac{nK}{2}.$$

Let $\mathcal{I}$ consist of the times $t \in \{nK - H, \ldots, nK - 1\}$ whose within-epoch position $p := t \bmod n$ satisfies

$$\begin{cases} \lceil n/4 \rceil \leq p \leq \lfloor 3n/4 \rfloor, & \eta L n \leq 512, \\ p \geq \lceil n/4 \rceil, & \eta L n > 512. \end{cases}$$

We claim that

$$|\mathcal{I}| \geq \frac{H}{4}, \qquad \gamma_t \geq \frac{3}{4} \quad (t \in \mathcal{I}).$$

If $\eta L n \leq 512$, then

$$\frac{1}{4\eta\mu} = \frac{\kappa n}{4\eta L n} \geq 2n,$$

so $H \geq 2n$. The terminal window ends at an epoch boundary and therefore contains $\lfloor H/n \rfloor$ complete epochs. Each complete epoch contains at least $n/2$ positions between $\lceil n/4 \rceil$ and $\lfloor 3n/4 \rfloor$. Consequently,

$$|\mathcal{I}| \geq \frac{n}{2}\left\lfloor \frac{H}{n} \right\rfloor \geq \frac{H}{4},$$

where the second inequality uses $H/n \geq 2$. If $\eta L n > 512$, at least half of the positions in each complete epoch, and in the possible leading suffix of an epoch, satisfy $p \geq \lceil n/4 \rceil$. Thus $|\mathcal{I}| \geq H/2$ in this case. Finally, for every $t \in \mathcal{I}$, Bernoulli's inequality gives

$$\gamma_t \geq (1-\eta\mu)^{H-1} \geq 1 - (H-1)\eta\mu \geq \frac{3}{4}.$$

We next sum the comparison between q_t and $\widetilde{q}_t$ over $\mathcal{I}$. For $m = t-1-j$, we have $\gamma_t/\gamma_j = (1-\eta\mu)^{-m-1}$. Hence, for every $j \in \{0, \dots, nK-1\}$,

$$\begin{aligned}\sum_{\substack{t\in\mathcal{I}\\ t>j}} \gamma_t \delta_{t-1-j} &\le \gamma_j \left(\frac{8\mu H}{L(1-\eta\mu)} + \frac{1-8\mu/L}{1-\eta\mu} \sum_{m=0}^{\infty} \left(\frac{1-\eta L/8}{1-\eta\mu} \right)^m \right) \\ &= \gamma_j \left(\frac{8\mu H}{L(1-\eta\mu)} + \frac{8}{\eta L} \right) \\ &\le \frac{32}{3\eta L} \gamma_j.\end{aligned}$$

The last inequality uses $H \le 1/(4\eta\mu)$ and $1-\eta\mu \ge 15/16$. Multiplying the pointwise comparison by γ_t and interchanging the nonnegative sums now gives

$$\sum_{t\in\mathcal{I}} \gamma_t \widetilde{q}_t \le Q + \frac{7\eta L}{8} \sum_{j=0}^{nK-1} q_j \sum_{\substack{t\in\mathcal{I}\\ t>j}} \gamma_t \delta_{t-1-j} \le \frac{31}{3} Q.$$

Equivalently,

$$Q \ge \frac{3}{31} \sum_{t\in\mathcal{I}} \gamma_t \widetilde{q}_t. \tag{23}$$

It remains to lower-bound the fluctuations of the linear recursion at each $t \in \mathcal{I}$. Fix such a time and write $t = kn + p$, where $0 \le k < K$ and $0 \le p < n$. Define

$$c_i := \begin{cases} \delta_{p-1-i}, & 0 \le i < p, \\ 0, & p \le i < n, \end{cases} \qquad Y := \sum_{i=0}^{n-1} c_i S_{kn+i}.$$

Unrolling from the start of the current epoch gives

$$-\boldsymbol{\nu}^\top \widetilde{\boldsymbol{y}}_t = -\boldsymbol{\nu}^\top \boldsymbol{A}^p \widetilde{\boldsymbol{y}}_{kn} + \eta G_* Y.$$

The first term is centered and independent of the current epoch permutation, whereas Y has a symmetric distribution because the slope multiset is invariant under sign reversal. Conditional Jensen's inequality and symmetry therefore imply

$$\mathbb{E}\widetilde{q}_t \ge \frac{\eta G_*}{2} \mathbb{E}|Y|.$$

To bound $\mathbb{E}|Y|$, write

$$\overline{c} := \frac{1}{n} \sum_{i=0}^{n-1} c_i, \qquad C := \sum_{i=0}^{n-1} (c_i - \overline{c})^2.$$

Exchangeability and $\sum_{i=0}^{n-1} S_{kn+i} = 0$ give

$$\mathbb{E}S_{kn+i}^2 = 1, \qquad \mathbb{E}[S_{kn+i} S_{kn+j}] = -\frac{1}{n-1} \quad (i \ne j).$$

It follows that

$$\mathbb{E}Y^2 = \sum_{i=0}^{n-1} c_i^2 - \frac{1}{n-1} \sum_{i\ne j} c_i c_j = \frac{n}{n-1} C.$$

We also obtain a fourth-moment bound directly from a pairing representation of the shuffled slopes. Choose a uniform perfect matching of the positions and independently assign opposite signs to the two positions in each pair. This procedure is invariant under permutations of the positions and hence generates the uniform ordering of the prescribed slope multiset. Conditional on the matching $\mathcal{M}$, we may write

$$Y = \sum_{\{i,j\}\in\mathcal{M}} \epsilon_{\{i,j\}} (c_i - c_j),$$

where the $\epsilon_{\{i,j\}}$ are independent Rademacher variables. Since each position belongs to exactly one pair,

$$\sum_{\{i,j\}\in\mathcal{M}} (c_i - c_j)^2 \le \sum_{\{i,j\}\in\mathcal{M}} \left(2(c_i-\overline{c})^2 + 2(c_j-\overline{c})^2\right) = 2C.$$

Expanding the fourth moment over the independent signs gives

$$\mathbb{E}_\epsilon Y^4 \le 3\left(\sum_{\{i,j\}\in\mathcal{M}} (c_i - c_j)^2\right)^2 \le 12C^2.$$

Averaging over the matching yields $\mathbb{E}Y^4 \le 12C^2$. Since $1 \le p < n$, the coefficient vector contains both 1 and 0, so $C > 0$. Hölder's inequality now gives

$$\mathbb{E}|Y| \ge \frac{(\mathbb{E}Y^2)^{3/2}}{(\mathbb{E}Y^4)^{1/2}} \ge \frac{\sqrt{C}}{\sqrt{12}}.$$

Consequently,

$$\mathbb{E}\widetilde{q}_t \ge \frac{\eta G_*}{2\sqrt{12}}\sqrt{C}. \tag{24}$$

We next lower-bound C for the coefficients at each retained position. Write

$$M := \min\left\{n, \frac{1}{\eta L}\right\}.$$

Suppose first that $\eta L n \le 512$. Then $n/4 \le p \le 3n/4$ by the definition of $\mathcal{I}$. Let

$$r := \min\left\{p, \left\lfloor \frac{2}{\eta L}\right\rfloor\right\}.$$

Since $\eta L < 2$, we have $\lfloor 2/(\eta L)\rfloor \ge 1/(\eta L)$, and hence $r \ge M/4$. For $0 \le m < r$, Bernoulli's inequality gives

$$\delta_m \ge \left(1 - \frac{8\mu}{L}\right)\left(1 - \frac{\eta L}{8}\right)^m \ge \frac{3}{4}\left(1 - \frac{m\eta L}{8}\right) \ge \frac{9}{16}.$$

Thus at least r coefficients are at least $9/16$, while the $n-p$ suffix coefficients vanish. Using the pairwise representation of centered squared deviations, we obtain

$$C = \frac{1}{n}\sum_{0\le i<j<n} (c_i - c_j)^2 \ge \frac{r(n-p)}{n}\left(\frac{9}{16}\right)^2 \ge \frac{81}{4096}M \ge \frac{M}{64}.$$

Suppose next that $\eta L n > 512$. Then $p \ge n/4$, so $\eta L p/8 > 16$. In particular,

$$\left(1 - \frac{\eta L}{8}\right)^{2p} \le e^{-\eta L p/4} \le \frac{1}{2}, \qquad 1 - \left(1 - \frac{\eta L}{8}\right)^2 \le \frac{\eta L}{4}.$$

Therefore,

$$\sum_{i=0}^{n-1} c_i^2 = \sum_{m=0}^{p-1} \delta_m^2 \ge \frac{9}{16}\sum_{m=0}^{p-1}\left(1 - \frac{\eta L}{8}\right)^{2m} \ge \frac{9}{8\eta L}.$$

Also,

$$\sum_{i=0}^{n-1} c_i \le \sum_{m=0}^{\infty} \delta_m = \frac{8\mu/L}{\eta\mu} + \frac{1 - 8\mu/L}{\eta L/8} \le \frac{16}{\eta L}.$$

Since $\eta L n > 512$ and $M = 1/(\eta L)$ in this case,

$$\begin{aligned} C &= \sum_{i=0}^{n-1} c_i^2 - \frac{1}{n}\left(\sum_{i=0}^{n-1} c_i\right)^2 \\ &\ge \frac{9}{8\eta L} - \frac{256}{n\eta^2 L^2} \ge \frac{1}{2\eta L} \ge \frac{M}{64}. \end{aligned}$$

Thus $C \ge M/64$ in both cases. Substituting into Equation (24) gives, for every $t \in \mathcal{I}$,

$$\mathbb{E}\widetilde{q}_t \ge \frac{\eta G_*}{16\sqrt{12}}\sqrt{M} \ge \frac{\eta G_*}{128}\sqrt{M}.$$

Using Equation (23), together with the bounds on $|\mathcal{I}|$, γ_t, and H, we obtain

$$\begin{aligned} \mathbb{E}Q &\ge \frac{3}{31}\sum_{t\in\mathcal{I}} \gamma_t \mathbb{E}\widetilde{q}_t \\ &\ge \frac{3}{31}\cdot\frac{3}{4}\cdot\frac{H}{4}\cdot\frac{\eta G_*}{128}\sqrt{M} \ge \frac{G_*}{2^{16}\mu}\sqrt{M}. \end{aligned}$$

Substituting into Equation (22) yields

$$\begin{aligned} \mathbb{E}x_{nK} &\ge \frac{7\eta L}{8}\sqrt{\frac{8\mu}{L}}\frac{G_*}{2^{16}\mu}\sqrt{M} \\ &\ge \frac{\eta G_*}{2^{15}}\sqrt{\frac{LM}{\mu}}. \end{aligned}$$

Jensen's inequality and the definition of M therefore imply

$$\mathbb{E}\left\|\boldsymbol{y}_{nK}\right\|^2 \ge (\mathbb{E}x_{nK})^2 \ge 2^{-30}\frac{G_*^2}{\mu}\min\{\eta^2 L n, \eta\}.$$

Both arguments of the minimum are nondecreasing in $\eta \ge 0$. Since $\eta \ge \eta_0 = 1/(2\mu n K)$,

$$\begin{aligned} \min\{\eta^2 L n, \eta\} &\ge \min\{\eta_0^2 L n, \eta_0\} \\ &= \frac{1}{4\mu n K}\min\left\{\frac{\kappa}{K}, 2\right\} \\ &\ge \frac{1}{4\mu n K}\min\left\{1, \frac{\kappa}{K}\right\}. \end{aligned}$$

Consequently,

$$\mathbb{E}\left\|\boldsymbol{y}_n^K\right\|^2 \ge 2^{-32}\frac{G_*^2}{\mu^2 n K}\min\left\{1, \frac{\kappa}{K}\right\}.$$

Construction of F_3. Let

$$F_3(v) = \frac{L}{2}v^2, \qquad f_{3,i}(v) = F_3(v), \quad i \in [n].$$

Every component is convex and L-smooth, the average is μ-strongly convex, and all component gradients vanish at the minimizer $v^* = 0$. With $v_0 = D/\sqrt{2}$, for every $\eta \ge 2/L$ we have

$$|v_n^K| = |1 - \eta L|^{nK}\frac{D}{\sqrt{2}} \ge \frac{D}{\sqrt{2}},$$

since $|1 - \eta L| \ge 1$. Thus $|v_n^K|^2 \ge D^2/2$. All component functions and the initialization are fixed independently of the step size, completing the proof. □

D Random Reshuffling without Component Convexity

In this section, we provide the proofs for random reshuffling without assuming component convexity.

D.1 Initialization Lower Bound

Theorem 4.2 (Initialization LB for RR without convex components)**.** *For any $n \geq 1024$, $\kappa \geq 1024(1+\log n)$, $K \lesssim \max\{\kappa, \kappa^2/n\}$, $G_* \geq 0$, and $D \geq 0$, there exists an $\mathcal{O}(\exp(n))$-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim D^2.$$

Proof. We prove the initialization lower bound separately in two regimes, according to the relative sizes of κ and n. When $\kappa \gtrsim n$, we construct an instance for which the initialization error remains of order D^2 for up to $K \lesssim \kappa^2/n$ epochs. When $\kappa \lesssim n$, we construct a different instance that retains an order D^2 error for up to $K \lesssim \kappa$ epochs. Together, these two constructions yield the threshold

$$K \lesssim \max\{\kappa, \kappa^2/n\},$$

corresponding to the two branches of Theorem 4.2. The two regimes are treated in Appendices D.1.1 and D.1.2, respectively. □

Before delving into the main proofs for the two regimes, we note two limitations of the current constructions. First, they leave a gap around the crossover regime $\kappa \asymp n$, where neither current parameter range directly applies; specifically, our current result does not cover $n < \kappa < 1024n\log_2 n$. Second, the construction used in the regime $\kappa \lesssim n$ has dimension exponential in n. Closing the crossover gap and obtaining the same lower bound with a fixed dimensional construction remain open. We now proceed to the main proofs.

D.1.1 Regime I: $\kappa \gtrsim n$

Proof. We make the constants and the condition on κ explicit: assume $n \geq 2$, $\kappa \geq 1024n\log_2 n$, and $1 \leq K \leq \kappa^2/(16384n)$. Fix $G_* > 0$ and $D > 0$, and set $\eta_0 := 4096\mu/L^2$. We divide the range of constant step sizes into three regimes and construct a one-dimensional finite sum for each regime. The functions are then combined across coordinates so that a single objective gives the lower bound for every step size and every sequence of epoch permutations. Our constructions F_1, F_2, and F_3, with component functions $f_{1,i}$, $f_{2,i}$, and $f_{3,i}$, respectively, will have the following properties:

- **Small step size regime.** Starting from $u_0 = D/\sqrt{3}$, for every $0 \leq \eta < \eta_0$ and every sequence of permutations, the scalar iterate on F_1 satisfies

$$|u_n^K|^2 \geq \frac{3D^2}{16}.$$

- **Moderate step size regime.** Starting from $z_0 = D/\sqrt{3}$, for every $\eta_0 \leq \eta < 2/L$ and every sequence of permutations, the scalar iterate on F_2 satisfies

$$|z_n^K|^2 \geq \frac{3D^2}{16}.$$

- **Large step size regime.** Starting from $v_0 = D/\sqrt{3}$, for every $\eta \geq 2/L$ and every sequence of permutations, the scalar iterate on F_3 satisfies

$$|v_n^K|^2 \geq \frac{D^2}{3}.$$

Their averages are

$$F_1(u) = \frac{\mu}{2}u^2, \qquad F_2(z) = \frac{\mu}{2}z^2, \qquad F_3(v) = \frac{L}{2}v^2.$$

Every component is L-smooth, and its gradient vanishes at zero. We verify these assertions and the three lower bounds in the constructions below.

We aggregate the three constructions across separate coordinates. Their component gradients vanish at zero, so we add one scalar coordinate ζ to obtain the prescribed gradient second moment G_*^2 at the minimizer. This additional coordinate is initialized at zero and does not change the updates in the three construction coordinates. Set

$$b_1^{\mathrm{noise}} = G_*\sqrt{\frac{n}{2}}, \qquad b_2^{\mathrm{noise}} = -G_*\sqrt{\frac{n}{2}}, \qquad b_i^{\mathrm{noise}} = 0 \quad (i \geq 3),$$

and define, for $\boldsymbol{x} = (u, z, v, \zeta)^\top \in \mathbb{R}^4$,

$$f_i(\boldsymbol{x}) := f_{1,i}(u) + f_{2,i}(z) + f_{3,i}(v) + \frac{\mu}{2}\zeta^2 + b_i^{\mathrm{noise}}\zeta,$$
$$\boldsymbol{x}_0 := \frac{D}{\sqrt{3}}(1, 1, 1, 0)^\top.$$

The two nonzero noise slopes cancel, while their squared magnitudes give

$$\sum_{i=1}^n b_i^{\mathrm{noise}} = 0,$$
$$\frac{1}{n}\sum_{i=1}^n |b_i^{\mathrm{noise}}|^2 = \frac{1}{n}\left(\frac{n}{2}G_*^2 + \frac{n}{2}G_*^2\right) = G_*^2.$$

Consequently, the average objective is

$$F(\boldsymbol{x}) = \frac{\mu}{2}(u^2 + z^2 + \zeta^2) + \frac{L}{2}v^2.$$

Thus $\boldsymbol{x}^* = \mathbf{0}$ and the strong-convexity modulus is exactly μ. Block diagonality preserves L-smoothness of the components, and the v-coordinate makes their gradient Lipschitz constants exactly L. Thus Assumption 2.3 holds, and

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 = \frac{D^2}{3} + \frac{D^2}{3} + \frac{D^2}{3} = D^2,$$
$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|^2 = \frac{1}{n}\sum_{i=1}^n |b_i^{\mathrm{noise}}|^2 = G_*^2.$$

The construction and initialization are fixed before choosing either η or the permutation sequence. For any fixed sequence of permutations, each coordinate follows its own component updates. For a given step size, at least one of the three regime bounds applies, and the other coordinates contribute nonnegative squared values. Therefore,

$$\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 = |u_n^K|^2 + |z_n^K|^2 + |v_n^K|^2 + |\zeta_n^K|^2$$
$$\geq \min\left\{\frac{3D^2}{16}, \frac{3D^2}{16}, \frac{D^2}{3}\right\} = \frac{3D^2}{16},$$

once the three coordinate bounds are established. This bound holds for every sequence of epoch permutations, including those chosen adaptively by a without-replacement SGD method. Taking expectations also gives the claimed lower bound for RR. We now give the constructions and prove their claimed properties.

Construction of F_1. Let

$$f_{1,i}(u) = F_1(u) := \frac{\mu}{2}u^2, \qquad i \in [n],$$

and initialize $u_0 = D/\sqrt{3}$. For $0 \le \eta < \eta_0$, the horizon condition gives

$$\eta\mu nK \le \eta_0 \mu nK = \frac{4096nK}{\kappa^2} \le \frac{1}{4}.$$

Since all component gradients coincide, Bernoulli's inequality yields

$$u_n^K = (1-\eta\mu)^{nK}\frac{D}{\sqrt{3}} \ge (1-\eta\mu nK)\frac{D}{\sqrt{3}} \ge \frac{3D}{4\sqrt{3}}.$$

Thus $|u_n^K|^2 \ge 3D^2/16$.

Construction of F_2. Let $J := \lfloor \log_2 n \rfloor$. For $r = 0, \dots, J-1$, define

$$m_r := 2^{J-r}, \qquad c_r := \frac{D}{\sqrt{3}} - \sum_{j=0}^{r-1} w_j,$$

$$w_r := \begin{cases} D/(8\sqrt{3}), & r = 0, \\ 128nD/(\sqrt{3}\kappa), & r \ge 1, \end{cases}$$

$$h_r(z) := \frac{L}{4}\max\{0, \min\{z - c_r + w_r, c_r - z\}\}.$$

The support of h_r is $[c_r - w_r, c_r]$. These intervals are adjacent and have disjoint interiors, and

$$\sum_{r=0}^{J-1} w_r = \frac{D}{\sqrt{3}}\left(\frac{1}{8} + \frac{128n(J-1)}{\kappa}\right) \le \frac{D}{4\sqrt{3}},$$

where we used $J - 1 \le \log_2 n$ and $\kappa \ge 1024n \log_2 n$. Hence all supports lie in $[3D/(4\sqrt{3}), D/\sqrt{3}]$. For each r, choose signs $s_{1r}, \dots, s_{nr}$ consisting of $m_r/2$ copies of each of $+1$ and -1, and $n - m_r$ zeros. Define

$$f_{2,i}(z) := \frac{\mu}{2}z^2 + \sum_{r=0}^{J-1} s_{ir}\int_0^z h_r(t)\,dt, \qquad F_2(z) = \frac{\mu}{2}z^2.$$

The formula for F_2 follows from $\sum_{i=1}^n s_{ir} = 0$ for every r. The signed sum of tents is continuous and piecewise linear with slopes of magnitude at most $L/4$, since their interiors are disjoint. Thus each $f'_{2,i}$ is globally $(\mu + L/4)$-Lipschitz, so each component is L-smooth. Also $f'_{2,i}(0) = 0$, and some components are nonconvex: on the upper half of a tent with sign $+1$, their curvature is $\mu - L/4 < 0$. We initialize $z_0 = D/\sqrt{3}$.

Fix $\eta_0 \le \eta < 2/L$. Since $n/2 < m_0 \le n$, $m_{r-1} = 2m_r$, and $m_{J-1} = 2$, the intervals

$$\left[\eta_0, \frac{4}{m_0 L}\right], \qquad \left[\frac{2}{m_r L}, \frac{4}{m_r L}\right] \quad (r = 1, \dots, J-1)$$

cover this regime without gaps. Indeed, consecutive endpoints agree, the last upper endpoint is $2/L$, and $\eta_0 m_0 L/4 = 1024 m_0/\kappa \le 1024n/\kappa \le 1$. Choose an interval containing η and retain its index r. Then

$$\frac{m_r \eta L}{4} \le 1, \qquad w_r \ge \frac{256n\mu D}{\sqrt{3}m_r\eta L^2} \ge \frac{256n\mu c_r}{m_r\eta L^2}. \tag{25}$$

To verify the width bound when $r = 0$, use $\eta \ge \eta_0$ and $m_0 > n/2$ to obtain

$$\frac{256n\mu D}{\sqrt{3}m_0\eta L^2} \le \frac{nD}{16\sqrt{3}m_0} \le \frac{D}{8\sqrt{3}} = w_0.$$

For $r \ge 1$, the lower endpoint $\eta \ge 2/(m_r L)$ instead gives

$$\frac{256n\mu D}{\sqrt{3}m_r\eta L^2} \le \frac{128n\mu D}{\sqrt{3}L} = \frac{128nD}{\sqrt{3}\kappa} = w_r.$$

The last inequality in Equation (25) follows from $c_r \le D/\sqrt{3}$.

We show that each epoch preserves $[c_r - w_r/4, \infty)$. Every scalar update $z \mapsto z - \eta f'_{2,i}(z)$ is nondecreasing: for $z' \geq z$, the difference between its values is at least $[1 - \eta(\mu + L/4)](z' - z) \geq 0$, because $\eta(\mu + L/4) \leq 2/\kappa + 1/2 \leq 1$. Its composition over an epoch is therefore nondecreasing as well. It suffices to show that an epoch starting at the boundary $c_r - w_r/4$ ends at or above that boundary.

Fix any permutation σ and set $\alpha := \eta L/4$. We introduce e_j to represent the distance of the jth test iterate to the left of the selected tent's right endpoint c_r, so that the iterate is written as $c_r - e_j$. The starting point $c_r - w_r/4$ corresponds to $e_0 = w_r/4$, and the desired endpoint bound is equivalent to $e_n \leq e_0$. This change of variable is useful because, on the upper half of the tent, $h_r(c_r - e) = (L/4)e$ for $0 \leq e \leq w_r/2$. The update in this distance coordinate therefore takes the following affine form. We first define e_j by this recursion, then verify that $0 \leq e_j \leq w_r/2$ throughout the epoch, so it indeed describes the actual test trajectory:

$$e_0 = \frac{w_r}{4}, \qquad e_j = (1 - \eta\mu + s_{\sigma(j),r}\alpha)e_{j-1} + \eta\mu c_r, \quad j = 1, \ldots, n. \tag{26}$$

All multipliers are nonnegative. In any contiguous product, at most $m_r/2$ factors exceed one, and each is at most $1 + \alpha$. Thus every such product is bounded by

$$(1 + \alpha)^{m_r/2} \leq \exp(m_r\alpha/2) \leq e^{1/2} < \frac{5}{3}.$$

Here $m_r\alpha \leq 1$ by Equation (25), and $\alpha \leq 1/2$. Multiplying the width bound by $m_r\eta^2L^2/256$ gives

$$n\eta\mu c_r \leq \frac{w_r m_r \alpha^2}{16} \leq \frac{w_r}{32}.$$

Unrolling Equation (26), the coefficient of the initial value and each added term is a contiguous product of the above multipliers. Consequently, for every $j = 0, \ldots, n$,

$$0 \leq e_j \leq \frac{5}{3}\left(\frac{w_r}{4} + j\eta\mu c_r\right) \leq \frac{5}{3}\left(\frac{w_r}{4} + \frac{w_r}{32}\right) = \frac{15w_r}{32} < \frac{w_r}{2}.$$

Hence $c_r - e_j$ stays in the upper half of the selected tent, where $h_r(c_r - e_j) = (L/4)e_j$ and all other tents vanish. Substituting $z = c_r - e_{j-1}$ into the actual update gives precisely Equation (26) for the new distance from c_r. Starting from $e_0 = w_r/4$, induction therefore identifies $c_r - e_j$ with the actual test trajectory.

Because the signs are balanced, the full-epoch multiplier is

$$\begin{aligned} P &:= \prod_{j=1}^{n}(1 - \eta\mu + s_{\sigma(j),r}\alpha) \\ &= (1 - \eta\mu)^{n-m_r}\left((1 - \eta\mu)^2 - \alpha^2\right)^{m_r/2} \leq (1 - \alpha^2)^{m_r/2} \leq e^{-m_r\alpha^2/2}. \end{aligned}$$

Since $m_r\alpha^2/2 \leq 1/4$ and $1 - e^{-t} \geq t - t^2/2 \geq 7t/8$ for $0 \leq t \leq 1/4$, we have $1 - P \geq 7m_r\alpha^2/16$. Using the same bound $5/3$ for all suffix products in the unrolled recursion gives

$$\begin{aligned} e_n &\leq P\frac{w_r}{4} + \frac{5}{3}n\eta\mu c_r \\ &\leq \frac{w_r}{4} - \frac{7w_r m_r \alpha^2}{64} + \frac{5w_r m_r \alpha^2}{48} = \frac{w_r}{4} - \frac{w_r m_r \alpha^2}{192} \leq \frac{w_r}{4}. \end{aligned}$$

Thus the test epoch ends at $c_r - e_n \geq c_r - w_r/4$. By monotonicity, any epoch starting above this boundary also ends above it, regardless of which tents its intermediate iterates visit. Since $z_0 = D/\sqrt{3} \geq c_r - w_r/4 \geq 3D/(4\sqrt{3})$, induction over epochs yields

$$z_n^k \geq c_r - \frac{w_r}{4} \geq \frac{3D}{4\sqrt{3}}, \qquad k = 1, \ldots, K,$$

for every permutation sequence. In particular, $|z_n^K|^2 \geq 3D^2/16$ throughout the moderate regime.

Construction of F_3. Let

$$f_{3,i}(v) = F_3(v) := \frac{L}{2}v^2, \qquad i \in [n].$$

These components are convex and L-smooth, and their gradients vanish at zero. Starting from $v_0 = D/\sqrt{3}$, every component update satisfies

$$v_i^k = (1 - \eta L)v_{i-1}^k.$$

Thus, for every sequence of permutations,

$$v_n^K = (1 - \eta L)^{nK} \frac{D}{\sqrt{3}}.$$

For $\eta \geq 2/L$, we have $|1 - \eta L| \geq 1$, and hence

$$|v_n^K|^2 = (1 - \eta L)^{2nK} \frac{D^2}{3} \geq \frac{D^2}{3}.$$

□

D.1.2 REGIME II: $\kappa \lesssim n$

Proof. We make the condition on K explicit and assume

$$\kappa \leq n, \qquad 1 \leq K \leq \frac{\kappa}{1024(1 + \log n)}.$$

We divide the range of step sizes into three regimes and construct a finite sum for each regime. The functions are then combined across coordinates so that a single objective gives the lower bound for every step size. Write

$$A := 128(1 + \log n), \qquad \eta_0 := \frac{A}{nL}.$$

Since $K \geq 1$, we have $\kappa \geq 1024(1 + \log n)$. Together with $\kappa \leq n$, this gives

$$n \geq \kappa \geq 1024(1 + \log n), \qquad \frac{K}{n} \leq \frac{1}{1024(1 + \log n)}, \qquad \eta_0 L = \frac{A}{n} \leq \frac{1}{8}.$$

In particular, $n \geq 1024$ and $K \leq n$.

Our constructions F_1, F_2, and F_3, with component functions $f_{1,i}$, $f_{2,i}$, and $f_{3,i}$, respectively, will have the following properties:

- **Small step size regime.** Starting from $u_0 = D/\sqrt{3}$, for every $0 \leq \eta < \eta_0$, the scalar iterate on F_1 satisfies
$$|u_n^K|^2 \geq \frac{D^2}{4}.$$
- **Moderate step size regime.** There is an initialization $\boldsymbol{w}_0$ with $\|\boldsymbol{w}_0\|^2 = D^2/3$ such that, for every $\eta_0 \leq \eta < 2/L$ and every sequence of permutations, the iterate on F_2 satisfies
$$\|\boldsymbol{w}_n^K\|^2 \geq D^2.$$
- **Large step size regime.** Starting from $v_0 = D/\sqrt{3}$, for every $\eta \geq 2/L$, the scalar iterate on F_3 satisfies
$$|v_n^K|^2 \geq \frac{D^2}{3}.$$

The functions F_1 and F_3 are one-dimensional, whereas F_2 is finite-dimensional. Their averages are

$$F_1(u) = \frac{\mu}{2}u^2, \qquad F_2(\boldsymbol{w}) = \frac{\mu}{2}\|\boldsymbol{w}\|^2, \qquad F_3(v) = \frac{L}{2}v^2.$$

Every component is twice differentiable and L-smooth, and its gradient vanishes at zero. We verify these assertions and the three lower bounds in the constructions below.

We aggregate the three constructions across separate coordinates. Their component gradients vanish at zero, so we add one scalar coordinate ζ to obtain the prescribed gradient second moment G_*^2 at the minimizer. This additional coordinate is initialized at zero and does not change the updates in the three construction blocks. Set

$$b_1 := G_*\sqrt{n-1}, \qquad b_i := -G_*/\sqrt{n-1}\,(i \geq 2)$$

and define, for $\boldsymbol{x} = (u, \boldsymbol{w}, v, \zeta)$,

$$f_i(\boldsymbol{x}) := f_{1,i}(u) + f_{2,i}(\boldsymbol{w}) + f_{3,i}(v) + \frac{\mu}{2}\zeta^2 + b_i\zeta,$$
$$\boldsymbol{x}_0 := \left(\frac{D}{\sqrt{3}}, \boldsymbol{w}_0, \frac{D}{\sqrt{3}}, 0\right).$$

The noise slopes cancel, while their squared magnitudes give

$$\sum_{i=1}^n b_i = 0, \quad \frac{1}{n}\sum_{i=1}^n b_i^2 = G_*^2.$$

Consequently, the average objective is

$$F(\boldsymbol{x}) = \frac{\mu}{2}\left(u^2 + \|\boldsymbol{w}\|^2 + \zeta^2\right) + \frac{L}{2}v^2.$$

Thus $\boldsymbol{x}^* = 0$ and the strong-convexity modulus is exactly μ. Block diagonality preserves L-smoothness of the components, and the v-coordinate makes their gradient Lipschitz constants exactly L. Also,

$$\|\boldsymbol{x}_0\|^2 = \frac{D^2}{3} + \|\boldsymbol{w}_0\|^2 + \frac{D^2}{3} = D^2,$$
$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(0)\|^2 = \frac{1}{n}\sum_{i=1}^n b_i^2 = G_*^2.$$

For any fixed sequence of permutations, each coordinate block follows its own component updates. For a given step size, at least one of the three regime bounds applies, and the other coordinates contribute nonnegative squared norms. Therefore,

$$\|\boldsymbol{x}_n^K\|^2 = |u_n^K|^2 + \|\boldsymbol{w}_n^K\|^2 + |v_n^K|^2 + |\zeta_n^K|^2$$
$$\geq \min\left\{\frac{D^2}{4}, D^2, \frac{D^2}{3}\right\} = \frac{D^2}{4},$$

once the three coordinate bounds are established. We now give the constructions and prove their claimed properties.

Construction of F_1. Let

$$F_1(u) = \frac{\mu}{2}u^2, \qquad f_{1,i}(u) = F_1(u), \quad i \in [n].$$

The components are twice differentiable and L-smooth, and their gradients vanish at zero. With $u_0 = D/\sqrt{3}$, all permutations give

$$u_n^K = (1-\eta\mu)^{nK}\frac{D}{\sqrt{3}}.$$

For $0 \leq \eta \leq \eta_0$, we have $\eta\mu \leq \eta_0 L \leq 1/8$. The total contraction over nK updates is controlled by

$$\eta\mu nK \leq \eta_0\mu nK = \frac{AK}{\kappa} \leq \frac{1}{8}.$$

Bernoulli's inequality therefore gives

$$|u_n^K|^2 = \frac{D^2}{3}(1-\eta\mu)^{2nK} \geq \frac{D^2}{3}(1-2\eta\mu nK) \geq \frac{D^2}{4}.$$

Construction of F_2. Our goal is to construct a single finite sum for which $\|\boldsymbol{w}_n^K\| \geq D$ for every $\eta \in [\eta_0, 2/L)$ and every sequence of permutations. We construct a collection of coordinate blocks, one for each of finitely many narrow step size intervals covering $[\eta_0, 2/L]$. We refer to these intervals as bands. For any chosen step size, it will suffice to find one band whose block produces the required lower bound. All band blocks are included in the objective in advance, so this choice in the analysis does not make the objective depend on the step size.

Each band is an $(n + 2 + N)$-dimensional block with coordinates

$$(p, q, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^N, \qquad N = \binom{n}{m}, \quad m = \lfloor n/2 \rfloor.$$

These coordinates are organized into three parts: a clock, an encoder, and a bank of branch coordinates. The scalar p serves as the clock, the pair $(q, \boldsymbol{y}) \in \mathbb{R}^{n+1}$ serves as the encoder, and $\boldsymbol{z} = (z_S)_{S \in \mathcal{S}} \in \mathbb{R}^N$ contains one branch coordinate for every possible set S of m component indices.

The clock is designed to detect the midpoint of the final epoch. By that time, the encoder records which component indices have appeared during the first m updates at the final epoch. The branch coordinates are indexed by all possible such sets S. Once the clock and encoder identify the observed prefix set S, the corresponding coordinate z_S is seeded, and the construction is designed so that the remaining updates in the epoch amplify this coordinate.

We first partition the moderate step size range $[\eta_0, 2/L)$ into finitely many narrow intervals, which we call bands. Each band is constructed around its left endpoint, viewed as a nominal step size. Set

$$\varepsilon := n^{-2}, \qquad \alpha := Ln^{-4}.$$

Let M be the smallest positive integer such that $(1 + \varepsilon)^M \eta_0 \geq 2/L$, and define

$$\begin{aligned} \eta_j &:= (1 + \varepsilon)^j \eta_0, \qquad j = 0, \ldots, M, \\ I_j &:= [\eta_j, \min\{\eta_{j+1}, 2/L\}], \qquad j = 0, \ldots, M - 1. \end{aligned}$$

These intervals cover $[\eta_0, 2/L]$. Moreover,

$$\begin{aligned} M &\leq 1 + \frac{\log(2n/A)}{\log(1 + \varepsilon)} \\ &\leq 1 + 2n^2 \log(2n) \leq 1 + \frac{n^3}{2} \leq n^3, \end{aligned}$$

where we used $\log(1 + \varepsilon) \geq \varepsilon/2$ and $\log(2n) \leq n/4$ for $n \geq 16$. All bands are included in the construction before η is chosen. The bound $M \leq n^3$ will control both the dimension and the amount of initial distance assigned to each block. Intuitively, we choose the bands sufficiently narrow so that, for any stepsize within a band, the clock can accurately detect the exact midpoint of the final epoch, namely, the time immediately after its first m updates.

We next define the vectors that represent component indices and subsets of these indices. Let

$$\mathcal{S} := \{S \subseteq [n] : |S| = m\}, \quad \boldsymbol{a}_i := \boldsymbol{e}_i - \frac{1}{n}\mathbf{1}, \quad \boldsymbol{a}_S := \mathbb{1}_S - \frac{m}{n}\mathbf{1}, \quad a_{S,i} := (\boldsymbol{a}_S)_i = \mathbb{1}_{\{i \in S\}} - \frac{m}{n}.$$

Here $\boldsymbol{e}_i$ is the i-th standard basis vector, $\mathbf{1}$ is the all-ones vector, and $\mathbb{1}_S$ is the indicator vector of S. In particular,

$$\sum_{i \in S} \boldsymbol{a}_i = \mathbb{1}_S - \frac{m}{n}\mathbf{1} = \boldsymbol{a}_S,$$

$$a_{S,i} = \begin{cases} 1 - m/n, & i \in S, \\ -m/n, & i \notin S. \end{cases}$$

These quantities also satisfy

$$\sum_{i=1}^n \boldsymbol{a}_i = 0, \qquad \sum_{i=1}^n a_{S,i} = 0, \qquad \|\boldsymbol{a}_i\| \leq 1, \qquad \|\boldsymbol{a}_S\| \leq \sqrt{n}, \qquad \frac{1}{3} \leq \frac{m}{n} \leq \frac{1}{2}, \qquad |a_{S,i}| \leq \frac{2}{3}.$$

Based on this definition, we now explain how each band is designed. As explained earlier, each band has coordinates $(p, q, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}^{n+2+N}$, where $p, q \in \mathbb{R}$, $\boldsymbol{y} \in \mathbb{R}^n$, and $\boldsymbol{z} = (z_S)_{S \in \mathcal{S}}$. The scalar p is

the clock, and the pair $(q, \boldsymbol{y})$ records the component indices through a quadratic coupling. We call each z_S a branch coordinate, since it is associated with one possible observed prefix S. We initialize each band at $(s, s, 0, 0)$, where

$$s := \frac{D}{\sqrt{6M}}, \qquad t_* := (K-1)n + m.$$

The index t counts completed component updates, so t_* is the time immediately after the first m updates of the final epoch. Only p and q are initially nonzero, and the choice of s ensures that $M(s^2 + s^2) = D^2/3$.

Fix a band j and temporarily write $\bar{\eta} := \eta_j$. Define

$$\bar{\rho} := 1 - \bar{\eta}\mu, \qquad \bar{\beta} := \frac{\bar{\eta}\alpha}{\bar{\rho}}, \qquad \bar{p} := s\bar{\rho}^{t_*},$$
$$\delta_p := \frac{\bar{\eta}\mu\bar{p}}{8}, \qquad \delta_y := \frac{\bar{\beta}\bar{p}}{8}, \qquad \lambda := \frac{L\delta_y}{256}.$$

These parameters depend on the fixed band, not on the chosen step size. The number $\bar{\rho}$ is the contraction factor of the clock at the nominal step size $\bar{\eta}$, and $\bar{\beta}$ is the corresponding normalized coupling strength of the index recorder. The value $\bar{p}$ is the nominal clock position at time t_*. The width δ_p specifies how close the clock must be to $\bar{p}$ for a seed update to occur, while δ_y is the tolerance for recognizing the encoded set of component indices. For each S, the vector $p\bar{\beta}\boldsymbol{a}_S$ is the target location of $\boldsymbol{y}$ when the clock equals p and the observed prefix is S. Finally, λ controls the size of the seed potential and is chosen proportional to $L\delta_y$ so that this potential remains L-smooth up to a fixed constant. We implement these requirements using the following smooth functions:

$$\phi(t) := (1 - t^2)_+^3,$$
$$\psi(\boldsymbol{y}) := (1 - \|\boldsymbol{y}\|^2)_+^3,$$
$$\theta(z) := \delta_y \tanh(z/\delta_y),$$

where $(t)_+ := \max\{t, 0\}$. The functions ϕ and ψ vanish outside $[-1, 1]$ and the unit ball, respectively. The bounded function θ satisfies $\theta(0) = 0$ and $\theta'(0) = 1$, which will allow a branch to be seeded without changing the clock or the index recorder at that update. Define

$$\chi_S(p, \boldsymbol{y}) := \phi\left(\frac{p - \bar{p}}{\delta_p}\right) \psi\left(\frac{\boldsymbol{y} - p\bar{\beta}\boldsymbol{a}_S}{\delta_y}\right),$$
$$Q_i(p, \boldsymbol{y}, \boldsymbol{z}) := \lambda \sum_{S \in \mathcal{S}} a_{S,i}\theta(z_S)\chi_S(p, \boldsymbol{y}).$$

The function χ_S can be nonzero only when both $|p - \bar{p}| < \delta_p$ and $\|\boldsymbol{y} - p\bar{\beta}\boldsymbol{a}_S\| < \delta_y$. It therefore requires the clock to be near the intended time and the recorded vector to be near the representation of S. The function Q_i combines these possible seed terms with the centered coefficients $a_{S,i}$. Define the i-th component on this band by

$$H_i(p, q, \boldsymbol{y}, \boldsymbol{z}) := \frac{\mu}{2}\left(p^2 + q^2 + \|\boldsymbol{y}\|^2 + \sum_{S \in \mathcal{S}} z_S^2\right) - \alpha q\langle \boldsymbol{a}_i, \boldsymbol{y}\rangle$$
$$+ \frac{L}{2}\sum_{S \in \mathcal{S}} a_{S,i} z_S^2 + Q_i(p, \boldsymbol{y}, \boldsymbol{z}). \tag{27}$$

The first term in Equation (27) supplies the common strongly convex quadratic average. The coupling $-\alpha q\langle \boldsymbol{a}_i, \boldsymbol{y}\rangle$ records the component index i, the term $(L/2)a_{S,i}z_S^2$ determines the curvature of branch S, and Q_i supplies the seed. All three terms that depend on the component index have zero average over i. Restoring the band index, denote this component by $H_{i,j}$ and set

$$f_{2,i}(\boldsymbol{w}) := \sum_{j=0}^{M-1} H_{i,j}(p_j, q_j, \boldsymbol{y}_j, \boldsymbol{z}_j),$$
$$\boldsymbol{w}_0 := \big((s, s, 0, 0)\big)_{j=0}^{M-1}.$$

The initialization satisfies $\|\boldsymbol{w}_0\|^2 = 2Ms^2 = D^2/3$. Also, $\sum_i \boldsymbol{a}_i = 0$ and $\sum_i a_{S,i} = 0$ imply

$$\frac{1}{n}\sum_{i=1}^{n} Q_i = 0, \qquad \frac{1}{n}\sum_{i=1}^{n} H_i = \frac{\mu}{2}\left(p^2 + q^2 + \|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2\right).$$

Consequently, $F_2(\boldsymbol{w}) = \mu\|\boldsymbol{w}\|^2/2$. It remains to verify two properties: each component is twice differentiable and L-smooth, and at least one branch reaches magnitude D for every step size in the moderate regime.

We first verify global regularity of the components. The main point is that, despite the large number of subsets, at most one seed term can have nonzero derivatives at a given point. Since $\bar{\eta} \le 2/L$, $\kappa \ge 16$, and $\mu \ge L/n$,

$$\bar{\rho} \ge \frac{1}{2}, \qquad 0 < \bar{\beta} \le 4n^{-4},$$
$$\frac{\delta_p}{\delta_y} = \frac{\mu\bar{\rho}}{\alpha} \ge \frac{n^3}{2} > 1, \qquad \frac{\delta_p}{\bar{p}} = \frac{\bar{\eta}\mu}{8} \le \frac{1}{16}, \qquad \|\bar{\beta}\boldsymbol{a}_S\| \le \bar{\beta}\sqrt{n} \le 1.$$

Whenever the clock factor is nonzero, $|p - \bar{p}| < \delta_p \le \bar{p}/16$, and hence $p \ge \bar{p}/2$. For distinct $S, S' \in \mathcal{S}$, the indicator vectors differ in at least two coordinates because both subsets have size m. Thus $\|\boldsymbol{a}_S - \boldsymbol{a}_{S'}\| \ge \sqrt{2}$, and the centers of their recognition regions satisfy

$$\|p\bar{\beta}\boldsymbol{a}_S - p\bar{\beta}\boldsymbol{a}_{S'}\| \ge \frac{\bar{\beta}\bar{p}}{\sqrt{2}} > 2\delta_y.$$

Thus the supports of the functions χ_S have disjoint interiors. This prevents the number N of subset coordinates from entering the smoothness bound.

The functions ϕ and ψ are twice differentiable: their values and derivatives through order two match zero at the support boundaries. Inside their supports, direct differentiation gives

$$\begin{aligned}
\phi'(t) &= -6t(1-t^2)^2, \\
\phi''(t) &= -6(1-t^2)^2 + 24t^2(1-t^2), \\
\nabla\psi(\boldsymbol{y}) &= -6(1-\|\boldsymbol{y}\|^2)^2\boldsymbol{y}, \\
\nabla^2\psi(\boldsymbol{y}) &= -6(1-\|\boldsymbol{y}\|^2)^2 I + 24(1-\|\boldsymbol{y}\|^2)\boldsymbol{y}\boldsymbol{y}^\top.
\end{aligned}$$

These formulas imply

$$|\phi'| \le 2, \qquad |\phi''| \le 6, \qquad \|\nabla\psi\| \le 2, \qquad \|\nabla^2\psi\| \le 6.$$

For the last bound, the radial and tangential eigenvalues are $\phi''(\|\boldsymbol{y}\|)$ and $-6(1-\|\boldsymbol{y}\|^2)^2$, respectively, inside the unit ball. The linear map $(p, \boldsymbol{y}) \mapsto \boldsymbol{y} - p\bar{\beta}\boldsymbol{a}_S$ has operator norm $\sqrt{1 + \|\bar{\beta}\boldsymbol{a}_S\|^2} \le \sqrt{2}$. Using $\delta_p \ge \delta_y$ and the product rule, we obtain

$$\begin{aligned}
\|\nabla\chi_S\| &\le \frac{2}{\delta_p} + \frac{2\sqrt{2}}{\delta_y} \le \frac{2+2\sqrt{2}}{\delta_y} < \frac{5}{\delta_y}, \\
\|\nabla^2\chi_S\| &\le \frac{6}{\delta_p^2} + \frac{12}{\delta_y^2} + \frac{8\sqrt{2}}{\delta_p\delta_y} \le \frac{18+8\sqrt{2}}{\delta_y^2} < \frac{30}{\delta_y^2}.
\end{aligned}$$

For the remaining factor $\theta(z) = \delta_y \tanh(z/\delta_y)$, we have

$$\begin{aligned}
\theta'(z) &= 1 - \tanh^2(z/\delta_y), \\
\theta''(z) &= -\frac{2}{\delta_y}\tanh(z/\delta_y)\big(1 - \tanh^2(z/\delta_y)\big).
\end{aligned}$$

Thus $|\theta| \le \delta_y$, $|\theta'| \le 1$, and $|\theta''| \le 2/\delta_y$. To see how these estimates control a seed term, write its Hessian in the variables $(p, \boldsymbol{y}, z_S)$ as

$$\nabla^2\big(\lambda\theta(z_S)\chi_S\big) = \lambda\begin{pmatrix} \theta(z_S)\nabla^2\chi_S & \theta'(z_S)\nabla\chi_S \\ \theta'(z_S)(\nabla\chi_S)^\top & \theta''(z_S)\chi_S \end{pmatrix}.$$

Here the derivatives of χ_S are with respect to $(p, \boldsymbol{y})$. The upper-left block, symmetric mixed block, and lower-right block have norms at most $30\lambda/\delta_y$, $5\lambda/\delta_y$, and $2\lambda/\delta_y$, respectively. Adding these three bounds gives

$$\|\nabla^2(\lambda\theta(z_S)\chi_S)\| \le (30+5+2)\frac{\lambda}{\delta_y} = \frac{37L}{256} < \frac{L}{6}.$$

The symmetric mixed block has norm $\lambda|\theta'|\|\nabla\chi_S\|$. At any point, at most one summand in Q_i has nonzero derivatives; all derivatives through order two vanish on the support boundaries. Since $|a_{S,i}| \le 1$, the same bound holds for $\|\nabla^2 Q_i\|$ globally.

For the quadratic part of Equation (27), the clock curvature is μ, the $(q, \boldsymbol{y})$ block has norm at most $\mu + \alpha\|\boldsymbol{a}_i\|$, and each branch curvature is $\mu + La_{S,i}$. Since $\|\boldsymbol{a}_i\| \le 1$, $|a_{S,i}| \le 2/3$, and $\alpha \le L/2$, the quadratic Hessian has norm at most $\mu + 2L/3$. Adding the seed Hessian bound therefore gives

$$\|\nabla^2 H_i\| \le \mu + \frac{2L}{3} + \frac{L}{6} < L,$$

where we used $\mu \le L/16$. Block diagonality across bands proves that each $f_{2,i}$ is twice differentiable and L-smooth. Since $\delta_p < \bar{p}$, every clock bump vanishes in a neighborhood of $p = 0$. Thus $\nabla f_{2,i}(0) = 0$ for every i.

We now establish the lower bound in the moderate step size regime. Fix $\eta \in [\eta_0, 2/L)$ and an arbitrary sequence of permutations. Choose a band I_j containing η and suppress its index for the rest of the analysis. This only selects a block already present in F_2. Since the bands use separate coordinates, a lower bound on one branch in this block is also a lower bound on $\|\boldsymbol{w}_n^K\|$. We will first locate the intended seed time, then identify the prefix set stored in $\boldsymbol{y}$, and finally bound the amplification of its branch. At a high level, the dynamics up to the midpoint $t^* = (K-1)n + m$ are designed so that, if S denotes the set of component indices appearing in the first m updates of the final epoch, then

$$p_{t^*} \approx \bar{p}, \qquad q_{t^*} \approx p_{t^*}, \qquad y_{t^*} \approx p_{t^*}\bar{\beta}a_S, \qquad z_{S',t^*} = 0 \quad \text{for all } S' \in \mathcal{S}.$$

Thus, p identifies the midpoint of the final epoch, while (q, y) encode the observed prefix set S; the branch coordinates remain inactive until this information is used to seed the corresponding branch z_S.

Write

$$\rho := 1 - \eta\mu, \qquad \beta := \frac{\eta\alpha}{\rho}.$$

Then $\rho \ge 1/2$, $0 < \beta \le 4n^{-4}$, and $\bar{\eta} \le \eta \le (1+\varepsilon)\bar{\eta}$.

First consider the clock and index recorder with the seed potential Q_i omitted. This gives an explicit reference trajectory that we can use to locate the supports of the seed terms. After locating those supports, we will verify that Q_i does not alter the clock or the index recorder along the actual trajectory. Let i_t be the index of the component used at update t, with $t = 1, \dots, nK$. Starting from $(p_0, q_0, \boldsymbol{y}_0) = (s, s, 0)$, the quadratic part of Equation (27) gives

$$\begin{aligned} p_{t+1} &= \rho p_t, \\ q_{t+1} &= \rho q_t + \eta\alpha\langle \boldsymbol{a}_{i_{t+1}}, \boldsymbol{y}_t\rangle, \\ \boldsymbol{y}_{t+1} &= \rho \boldsymbol{y}_t + \eta\alpha q_t \boldsymbol{a}_{i_{t+1}}. \end{aligned}$$

In particular, $p_t = s\rho^t$.

We first show that the clock factor can be active only at $t = t_*$. Recall that the factor $\phi((p-\bar{p})/\delta_p)$ vanishes unless $|p - \bar{p}| < \delta_p$. At the nominal step size $\bar{\eta}$, the clock reaches $\bar{p}$ exactly at time t_*, and the band width ensures that replacing $\bar{\eta}$ by η changes this position by much less than δ_p. Indeed, $\bar{\rho} \ge 1/2$ and $\bar{\eta} \le \eta \le (1+\varepsilon)\bar{\eta}$ imply

$$0 \le \frac{(\eta - \bar{\eta})\mu}{\bar{\rho}} \le 2\varepsilon\bar{\eta}\mu.$$

Since $t_* \le nK$, Bernoulli's inequality gives

$$\frac{p_{t_*}}{\bar{p}} = \frac{s\rho^{t_*}}{s\bar{\rho}^{t_*}} = \left(1 - \frac{(\eta-\bar{\eta})\mu}{\bar{\rho}}\right)^{t_*}$$

$$\geq 1 - t_* \frac{(\eta - \bar{\eta})\mu}{\bar{\rho}} \geq 1 - 2nK\varepsilon\bar{\eta}\mu.$$

We also have $p_{t_*}/\bar{p} \leq 1$, because $\rho \leq \bar{\rho}$. It follows that

$$\frac{|p_{t_*} - \bar{p}|}{\delta_p} \leq \frac{2nK\varepsilon\bar{\eta}\mu\bar{p}}{\bar{\eta}\mu\bar{p}/8} = 16nK\varepsilon$$
$$= \frac{16K}{n} \leq \frac{1}{64} < \frac{1}{8}.$$

To exclude all other update times, it suffices to check the two neighboring clock positions, since $p_t = s\rho^t$ is strictly decreasing. For these neighboring times, we have

$$\frac{p_{t_*+1}}{\bar{p}} \leq \rho \leq 1 - \bar{\eta}\mu,$$
$$\frac{p_{t_*-1}}{\bar{p}} \geq \frac{1 - 2nK\varepsilon\bar{\eta}\mu}{1 - \bar{\eta}\mu} = 1 + \frac{\bar{\eta}\mu(1 - 2nK\varepsilon)}{1 - \bar{\eta}\mu} \geq 1 + \frac{\bar{\eta}\mu}{2}.$$

Here $2nK\varepsilon = 2K/n < 1/2$, which justifies the last lower bound for p_{t_*-1}. Both gaps exceed $\delta_p/\bar{p} = \bar{\eta}\mu/8$. Monotonicity of the clock therefore proves

$$|p_{t_*} - \bar{p}| \leq \frac{\delta_p}{8},$$
$$|p_t - \bar{p}| > \delta_p, \qquad t \in \{0, \ldots, nK\} \setminus \{t_*\}. \tag{28}$$

Next, we quantify the information about component indices stored in $(q, \boldsymbol{y})$. Dividing out the common contraction factor ρ isolates the small changes caused by the choice of component. Define

$$E_i := \begin{pmatrix} 0 & \boldsymbol{a}_i^\top \\ \boldsymbol{a}_i & 0 \end{pmatrix}, \qquad \|E_i\| = \|\boldsymbol{a}_i\| \leq 1.$$

The two coupled recursions can be written as

$$\begin{pmatrix} q_{t+1} \\ \boldsymbol{y}_{t+1} \end{pmatrix} = \rho \begin{pmatrix} 1 & \beta\boldsymbol{a}_{i_{t+1}}^\top \\ \beta\boldsymbol{a}_{i_{t+1}} & I \end{pmatrix} \begin{pmatrix} q_t \\ \boldsymbol{y}_t \end{pmatrix}$$
$$= \rho(I + \beta E_{i_{t+1}}) \begin{pmatrix} q_t \\ \boldsymbol{y}_t \end{pmatrix}.$$

Applying this identity repeatedly from $(q_0, \boldsymbol{y}_0) = (s, 0)$ gives the exact product formula

$$\rho^{-t} \begin{pmatrix} q_t \\ \boldsymbol{y}_t \end{pmatrix} = (I + \beta E_{i_t}) \cdots (I + \beta E_{i_1}) \begin{pmatrix} s \\ 0 \end{pmatrix}.$$

Expanding the product, the constant term is $(s, 0)$. Since $E_i(s, 0) = (0, s\boldsymbol{a}_i)$, the first-order term is $(0, \beta s \sum_{\ell=1}^t \boldsymbol{a}_{i_\ell})$. For each $\ell \geq 2$, there are $\binom{t}{\ell}$ products containing ℓ matrices, and each product has operator norm at most one. Thus the total higher-order contribution is bounded independently of the order of the component indices by

$$s \sum_{\ell=2}^{t} \binom{t}{\ell} \beta^\ell \leq s \sum_{\ell=2}^{\infty} \frac{(t\beta)^\ell}{\ell!} = s(e^{t\beta} - 1 - t\beta)$$
$$\leq s\beta^2 t^2 \leq s\beta^2 n^2 K^2.$$

Here $\beta nK \leq 4n^{-2} \leq 1$, and we used $e^u - 1 - u \leq u^2$ for $0 \leq u \leq 1$. Taking the $\boldsymbol{y}$-coordinates proves

$$\frac{\boldsymbol{y}_t}{s\rho^t} = \beta \sum_{\ell=1}^{t} \boldsymbol{a}_{i_\ell} + \boldsymbol{r}_t, \qquad \|\boldsymbol{r}_t\| \leq \beta^2 n^2 K^2. \tag{29}$$

Let S be the set of component indices used in the first m updates of the final epoch. Each completed epoch uses every component index exactly once, so its contribution to the sum in Equation (29) is zero. At time t_*, the remaining prefix is precisely S, and hence

$$\sum_{\ell=1}^{t_*} \boldsymbol{a}_{i_\ell} = (K - 1) \sum_{i=1}^{n} \boldsymbol{a}_i + \sum_{i \in S} \boldsymbol{a}_i = \boldsymbol{a}_S.$$

We must compare this actual encoding, which uses β, with the nominal encoding $\bar{\beta}\boldsymbol{a}_S$. The difference between the two coupling strengths satisfies

$$\frac{\beta}{\bar{\beta}} - 1 = \frac{\eta\bar{\rho} - \bar{\eta}\rho}{\bar{\eta}\rho} = \frac{\eta - \bar{\eta}}{\bar{\eta}\rho},$$
$$0 \le \frac{\beta}{\bar{\beta}} - 1 \le \frac{\varepsilon}{\rho} \le 2\varepsilon.$$

Thus $\bar{\beta} \le \beta \le (1+2\varepsilon)\bar{\beta} \le 2\bar{\beta}$, and at time t_* we have

$$\boldsymbol{y}_{t_*} - p_{t_*}\bar{\beta}\boldsymbol{a}_S = p_{t_*}\left((\beta - \bar{\beta})\boldsymbol{a}_S + \boldsymbol{r}_{t_*}\right).$$

There are two sources of error: replacing β by the nominal value $\bar{\beta}$, and the higher-order remainder $\boldsymbol{r}_{t_*}$. Using $\|\boldsymbol{a}_S\| \le \sqrt{n}$, $\|\boldsymbol{r}_{t_*}\| \le \beta^2 n^2 K^2$, and $\beta \le 2\bar{\beta}$, we obtain

$$\begin{aligned}\|\boldsymbol{y}_{t_*} - p_{t_*}\bar{\beta}\boldsymbol{a}_S\| &\le p_{t_*}\left(|\beta - \bar{\beta}|\sqrt{n} + \beta^2 n^2 K^2\right) \\ &\le \bar{p}\left(2\varepsilon\bar{\beta}\sqrt{n} + 4\bar{\beta}^2 n^2 K^2\right).\end{aligned}$$

Dividing by the recognition width $\delta_y = \bar{\beta}\bar{p}/8$ gives

$$\begin{aligned}\frac{\|\boldsymbol{y}_{t_*} - p_{t_*}\bar{\beta}\boldsymbol{a}_S\|}{\delta_y} &\le 8\left(2\varepsilon\sqrt{n} + 4\bar{\beta}n^2K^2\right) \\ &\le 16n^{-3/2} + 128\left(\frac{K}{n}\right)^2 < \frac{1}{8},\end{aligned}$$

where we used $n \ge 1024$ and $K/n \le 1/1024$. Thus the encoded vector lies within $\delta_y/8$ of the center for the actual prefix S. Together with Equation (28), this bounds the arguments of ϕ and ψ by $1/8$ in absolute value and norm, respectively, and therefore

$$\chi_S(p_{t_*}, \boldsymbol{y}_{t_*}) \ge \left(1 - \frac{1}{64}\right)^6 > \frac{3}{4}.$$

By disjointness of the supports, every other $\chi_{S'}$ vanishes there.

We now verify that the clock and index recorder formulas above hold for the actual component gradient trajectory. Before time t_*, the clock lies outside every seed support by Equation (28). All branch coordinates start at zero, and their homogeneous quadratic updates preserve zero. Inductively, the actual clock and index recorder therefore follow the preceding recursions up to time t_*. At that time every $z_{S'}$ is still zero. Differentiating the seed potential with respect to the clock and the index recorder gives

$$\nabla_{(p,\boldsymbol{y})} Q_i = \lambda \sum_{S' \in \mathcal{S}} a_{S',i}\theta(z_{S'})\nabla\chi_{S'} = 0,$$
$$\partial_q Q_i = 0,$$

because $\theta(0) = 0$ and q does not enter Q_i. On the other hand, $\theta'(0) = 1$ gives

$$\partial_{z_{S'}} Q_i = a_{S',i}\lambda\theta'(0)\chi_{S'} = a_{S',i}\lambda\chi_{S'}.$$

At the current point, only χ_S is nonzero, so only branch z_S receives a nonzero seed gradient. This explains why the seed changes one branch while preserving the clock and recursions for the index recorder.

The next component index i_{t_*+1} has not appeared in the final epoch's first m positions, so it lies outside S and $a_{S,i_{t_*+1}} = -m/n$. Since $z_{S,t_*} = 0$, the quadratic part contributes no branch gradient at this update. The component gradient step is therefore

$$\begin{aligned}z_{S,t_*+1} &= z_{S,t_*} - \eta\left[(\mu + La_{S,i_{t_*+1}})z_{S,t_*} + a_{S,i_{t_*+1}}\lambda\chi_S\right] \\ &= \eta\frac{m}{n}\lambda\chi_S,\end{aligned}$$

where χ_S is evaluated at $(p_{t_*}, \boldsymbol{y}_{t_*})$. Using $m/n \ge 1/3$ and $\chi_S \ge 3/4$ gives

$$z_{S,t_*+1} = \eta\frac{m}{n}\lambda\chi_S(p_{t_*}, \boldsymbol{y}_{t_*}) \ge \frac{\eta\lambda}{4} > 0. \tag{30}$$

The clock update is still $p_{t_*+1} = \rho p_{t_*}$ because its seed derivative is zero. After this update, Equation (28) keeps the clock strictly below the support. Induction then shows that no seed gradient acts again. Thus the clock and index recorder recursions describe the actual trajectory throughout the run, and exactly one branch coordinate is seeded in this band.

It remains to show that the positive seed grows enough during the rest of the final epoch. Every component index remaining in that epoch lies outside S, so its coefficient in branch S is always $a_{S,i} = -m/n$. Its curvature in the selected coordinate is

$$\mu + La_{S,i} = \mu - L\frac{m}{n} \le \frac{L}{16} - \frac{L}{3} \le -\frac{L}{4}.$$

Since the seed terms vanish afterwards, the remaining updates satisfy

$$\begin{aligned} z_{S,t+1} &= \left(1 - \eta\left(\mu - L\frac{m}{n}\right)\right) z_{S,t} \\ &\ge \left(1 + \frac{\eta L}{4}\right) z_{S,t}, \qquad t = t_* + 1, \dots, nK - 1. \end{aligned}$$

The seed is positive and the multiplier exceeds one, so the branch remains positive. There are $nK - (t_* + 1) = n - m - 1 \ge 2n/5$ such updates, since $n \ge 10$. Moreover,

$$\log\left(1 + \frac{\eta L}{4}\right) \ge \frac{\eta L}{4 + \eta L} \ge \frac{\eta L}{6},$$

where $\eta L \le 2$. Therefore,

$$\begin{aligned} z_{S,nK} &\ge z_{S,t_*+1}\left(1 + \frac{\eta L}{4}\right)^{n-m-1} \\ &\ge z_{S,t_*+1} \exp\left(\frac{2n}{5} \cdot \frac{\eta L}{6}\right) = z_{S,t_*+1} \exp\left(\frac{\eta n L}{15}\right). \end{aligned}$$

To compare this final value with D, we must also account for the size of the seed in Equation (30). Recall that λ contains $\bar{\beta}$, which measures the small coupling, and $\bar{p}$, which is the clock value after almost nK contractions. These band parameters satisfy

$$\begin{aligned} \bar{\beta} &= \frac{\bar{\eta}\alpha}{\bar{\rho}} \ge \bar{\eta}\alpha \ge \frac{\eta\alpha}{2} = \frac{\eta L}{2n^4}, \\ \bar{p} &= s(1 - \bar{\eta}\mu)^{t_*} \ge s\exp(-2\bar{\eta}\mu n K) \ge s\exp\left(-\frac{2\eta n L K}{\kappa}\right). \end{aligned}$$

Here $\bar{\eta} \ge \eta/(1 + \varepsilon) \ge \eta/2$, and the bound on $\bar{p}$ uses $\log(1 - \bar{\eta}\mu) \ge -2\bar{\eta}\mu$. Combining these inequalities with Equation (30) and $\lambda = L\bar{\beta}\bar{p}/(8 \cdot 256)$ gives

$$\begin{aligned} z_{S,nK} &\ge \frac{(\eta L)^2}{128^2 n^4} s \exp\left(\frac{\eta n L}{15} - \frac{2\eta n L K}{\kappa}\right) \\ &\ge s n^{-6} \exp\left(\frac{\eta n L}{16}\right). \end{aligned}$$

For the last inequality, the step size lower bound gives

$$\frac{(\eta L)^2}{128^2 n^4} \ge \frac{A^2}{128^2 n^6} \ge n^{-6},$$

and the condition on K gives

$$\frac{1}{15} - \frac{2K}{\kappa} \ge \frac{1}{15} - \frac{1}{512} \ge \frac{1}{16}.$$

Thus the amplification during the final epoch dominates the decay of the clock over the preceding updates. Finally, $\eta n L \ge A$ implies

$$\exp\left(\frac{\eta n L}{16}\right) \ge \exp\left(\frac{A}{16}\right) = e^8 n^8 \ge n^8.$$

It follows that

$$z_{S,nK} \geq sn^2 = \frac{Dn^2}{\sqrt{6M}} \geq D\sqrt{\frac{n}{6}} \geq D,$$

where $M \leq n^3$ and $n \geq 1024$ were used. In particular,

$$\|\boldsymbol{w}_n^K\|^2 \geq z_{S,nK}^2 \geq D^2.$$

This proves the moderate step size bound for every permutation sequence.

Each band uses two scalar coordinates p, q, an n-dimensional vector $\boldsymbol{y}$, and N scalar branches. Thus the dimension of F_2 is exactly $M(n+2+N)$ and can be bounded as

$$M(n+2+N) \leq n^3 \left(n+2+\binom{n}{\lfloor n/2 \rfloor}\right).$$

Construction of F_3. Let

$$F_3(v) = \frac{L}{2}v^2, \qquad f_{3,i}(v) = F_3(v), \quad i \in [n].$$

Every component is twice differentiable and L-smooth, and its gradient vanishes at zero. Since all components are identical, every permutation gives the same scalar update $v^+ = (1-\eta L)v$. For $v_0 = D/\sqrt{3}$ and every $\eta \geq 2/L$,

$$|v_n^K|^2 = |1-\eta L|^{2nK}\frac{D^2}{3} \geq \frac{D^2}{3},$$

since $|1-\eta L| \geq 1$. □

D.2 Upper Bound

Theorem A.7 (RR Upper Bound without component convexity)**.** *Let $n \geq 2$, $K \gtrsim \max\{\kappa, \kappa^2/n\}$, and $G_* > 0$. Suppose F is a finite-sum objective satisfying Assumption 2.3. For any initialization point $\boldsymbol{x}_0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR with a stepsize $\eta = \Lambda_{nK}/\mu nK$ satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \lesssim \frac{LG_*^2}{\mu^3 nK^2}\max\left\{1, \frac{\kappa}{n}\right\}.$$

Proof. The condition $K \gtrsim \max\left\{\kappa, \kappa^2/n\right\}$ in the theorem refers to $K \geq \Lambda_{nK}\max\left\{4\kappa, 32\kappa^2/n\right\}$.

We first prove a bound for any constant stepsize satisfying

$$0 < \eta \leq \min\left\{\frac{1}{4nL}, \frac{\mu}{32L^2}\right\}. \tag{31}$$

We then substitute $\eta = \Lambda_{nK}/(\mu nK)$, which satisfies the above η bound under the stated epoch condition.

Shifted iterates and within-epoch distances. Fix an epoch k and condition on its starting point $\boldsymbol{x}_0^k$. Write $\sigma = \sigma_k$ and suppress the epoch superscript k until the epoch recursion is established. All expectations in this part are conditional on that starting point. For $j \in [n]$ and $0 \leq i \leq n$, define

$$\mathbf{h}_j(\boldsymbol{x}) := \nabla f_j(\boldsymbol{x}) - \nabla f_j(\boldsymbol{x}^*), \qquad \mathbf{s}_i := \sum_{r=1}^{i}\nabla f_{\sigma(r)}(\boldsymbol{x}^*), \qquad \widetilde{\boldsymbol{x}}_i := \boldsymbol{x}_i + \eta\mathbf{s}_i.$$

Here $\mathbf{s}_0 = \mathbf{0}$, each $\mathbf{h}_j$ is L-Lipschitz, $\|\mathbf{h}_j(\boldsymbol{x})\| \leq L\|\boldsymbol{x}-\boldsymbol{x}^*\|$, and $n^{-1}\sum_j \mathbf{h}_j(\boldsymbol{x}) = \nabla F(\boldsymbol{x})$. Since $\sum_j \nabla f_j(\boldsymbol{x}^*) = \mathbf{0}$, the shifted updates

$$\widetilde{\boldsymbol{x}}_{i+1} := \widetilde{\boldsymbol{x}}_i - \eta\mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i), \qquad \widetilde{\boldsymbol{x}}_0 = \boldsymbol{x}_0, \tag{32}$$

satisfy $\widetilde{\boldsymbol{x}}_n = \boldsymbol{x}_n$.

For distinct positions $r, t \in [n]$, uniform sampling without replacement gives

$$\mathbb{E}\left\langle \nabla f_{\sigma(r)}(\boldsymbol{x}^*), \nabla f_{\sigma(t)}(\boldsymbol{x}^*)\right\rangle = \frac{\sum_{a\neq b}\langle \nabla f_a(\boldsymbol{x}^*), \nabla f_b(\boldsymbol{x}^*)\rangle}{n(n-1)} = -\frac{G_*^2}{n-1}.$$

Expanding the squared prefix sum therefore yields

$$\mathbb{E}\left\|\mathbf{s}_i\right\|^2 = \frac{i(n-i)}{n-1}G_*^2 \le nG_*^2, \qquad \sum_{i=0}^{n-1}\mathbb{E}\left\|\mathbf{s}_i\right\|^2 = \frac{n(n+1)}{6}G_*^2 \le \frac{n^2}{4}G_*^2. \tag{33}$$

Write

$$d_i := (\mathbb{E}\left\|\boldsymbol{x}_i - \boldsymbol{x}^*\right\|^2)^{1/2}, \qquad W := \frac{1}{n}\sum_{i=0}^{n-1} d_i^2.$$

In this paragraph, we first aim to bound d_0 in terms of W, η, n, and G_*^2. Unrolling Equation (32) and using $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \eta\mathbf{s}_i$ gives

$$\boldsymbol{x}_i - \boldsymbol{x}_0 = -\eta\sum_{j=0}^{i-1}\mathbf{h}_{\sigma(j+1)}(\boldsymbol{x}_j) - \eta\mathbf{s}_i.$$

Applying Minkowski's inequality then gives

$$\begin{aligned}(\mathbb{E}\left\|\boldsymbol{x}_i - \boldsymbol{x}_0\right\|^2)^{1/2} &\le \eta\sum_{j<i}\left(\mathbb{E}\left\|\mathbf{h}_{\sigma(j+1)}(\boldsymbol{x}_j)\right\|^2\right)^{1/2} + \eta(\mathbb{E}\left\|\mathbf{s}_i\right\|^2)^{1/2}\\ &\le \eta L\sum_{j<i} d_j + \eta(\mathbb{E}\left\|\mathbf{s}_i\right\|^2)^{1/2}.\end{aligned}$$

For each $0 \le i < n$, the triangle inequality in L^2 and the preceding bound give

$$\begin{aligned}d_0 &\le d_i + (\mathbb{E}\left\|\boldsymbol{x}_i - \boldsymbol{x}_0\right\|^2)^{1/2}\\ &\le d_i + \eta L\sum_{j<i} d_j + \eta(\mathbb{E}\left\|\mathbf{s}_i\right\|^2)^{1/2}.\end{aligned}$$

By Cauchy–Schwarz,

$$\sum_{j=0}^{n-1} d_j \le \left(n\sum_{j=0}^{n-1} d_j^2\right)^{1/2} = n\sqrt{W},$$

so $n^{-1}\sum_{i=0}^{n-1} d_i \le \sqrt{W}$ and $\sum_{j<i} d_j \le n\sqrt{W}$ for every i. Averaging the bound for d_0 over $i = 0, \dots, n-1$ and applying Cauchy–Schwarz to the scalar sequence $(\mathbb{E}\left\|\mathbf{s}_i\right\|^2)^{1/2}$ yields

$$\begin{aligned}d_0 &\le \frac{1}{n}\sum_{i=0}^{n-1} d_i + \frac{\eta L}{n}\sum_{i=0}^{n-1}\sum_{j<i} d_j + \frac{\eta}{n}\sum_{i=0}^{n-1}(\mathbb{E}\left\|\mathbf{s}_i\right\|^2)^{1/2}\\ &\le (1+\eta nL)\sqrt{W} + \eta\left(\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left\|\mathbf{s}_i\right\|^2\right)^{1/2}\\ &= (1+\eta nL)\sqrt{W} + \eta\sqrt{\frac{n+1}{6}}G_*\\ &\le \frac{5}{4}\sqrt{W} + \frac{\eta\sqrt{n}}{2}G_*.\end{aligned}$$

Here the equality uses Equation (33), and the last inequality uses $\eta nL \le 1/4$ and $(n+1)/6 \le n/4$ for $n \ge 2$.

Squaring with $(a+b)^2 \le \frac{5}{4}a^2 + 5b^2$ gives

$$d_0^2 \le \frac{125}{64}W + \frac{5}{4}\eta^2 nG_*^2 \le 2W + 2\eta^2 nG_*^2. \tag{34}$$

Controlling the sampling bias by a permutation swap. The component gradient under RR is biased. We control the bias here.

Fix $i < n$, draw R uniformly from $[n]$ independently of σ, and set $J := \sigma(R)$. Let σ' exchange positions $i+1$ and R in σ, and let $\boldsymbol{x}'_t, \mathbf{s}'_t, \widetilde{\boldsymbol{x}}'_t$ be the corresponding quantities from the same starting point. For each fixed $R = r$, exchanging two positions is a bijection on the set of permutations. Hence σ' is uniform and $\sigma'(i+1) = J$, which gives

$$\mathbb{E}_{\sigma,R} \langle \mathbf{h}_J(\boldsymbol{x}'_i), \widetilde{\boldsymbol{x}}'_i - \boldsymbol{x}^* \rangle = \mathbb{E}_\sigma \left\langle \mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^* \right\rangle. \tag{35}$$

Conditional on σ, the original trajectory $\boldsymbol{x}_i$ is fixed and $J = \sigma(R)$ is uniform on $[n]$. Since $\nabla F(\boldsymbol{x}^*) = 0$, we have

$$\mathbb{E}_R[\mathbf{h}_J(\boldsymbol{x}_i) \mid \sigma] = \frac{1}{n} \sum_{a=1}^{n} \mathbf{h}_a(\boldsymbol{x}_i) = \nabla F(\boldsymbol{x}_i), \tag{36}$$

$$\mathbb{E}_{\sigma,R} \langle \mathbf{h}_J(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^* \rangle = \mathbb{E}_\sigma \langle \nabla F(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^* \rangle. \tag{37}$$

Subtracting Equation (37) from Equation (35) and writing $B_i := \mathbb{E} \left\langle \mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i) - \nabla F(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^* \right\rangle$, we obtain

$$B_i = \mathbb{E}_{\sigma,R} \big[\langle \mathbf{h}_J(\boldsymbol{x}'_i), \widetilde{\boldsymbol{x}}'_i - \boldsymbol{x}^* \rangle - \langle \mathbf{h}_J(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^* \rangle \big]. \tag{38}$$

The role of B_i becomes clear in the squared-distance recursion: expanding the squared norm in Equation (32) and decomposing $\mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i)$ into $\nabla F(\boldsymbol{x}_i)$ and its residual splits the expected cross term into the full-gradient contribution and $-2\eta B_i$. The full-gradient term is controlled by strong convexity and smoothness in Equation (43), while a bound on $\left| \sum_{i=0}^{n-1} B_i \right|$ controls the accumulated sampling bias in the epoch recursion.

If $R > i$, both trajectories agree through step i and the difference is zero. For a fixed $R = r \le i$, put $q_t := (\mathbb{E}_\sigma \|\widetilde{\boldsymbol{x}}'_t - \widetilde{\boldsymbol{x}}_t\|^2)^{1/2}$. The trajectories agree through step $r-1$. At step r, the shifted update gives

$$\widetilde{\boldsymbol{x}}'_r - \widetilde{\boldsymbol{x}}_r = -\eta \big[\mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_{r-1}) - \mathbf{h}_{\sigma(r)}(\boldsymbol{x}_{r-1}) \big].$$

Since $\|\mathbf{h}_a(\boldsymbol{x})\| \le L \|\boldsymbol{x} - \boldsymbol{x}^*\|$ for every a, the triangle inequality gives $q_r \le 2\eta L d_{r-1}$. For $r \le t \le i$, the prefix-sum difference is constant in t (but depends on σ):

$$\mathbf{s}'_t - \mathbf{s}_t = \nabla f_{\sigma(i+1)}(\boldsymbol{x}^*) - \nabla f_{\sigma(r)}(\boldsymbol{x}^*).$$

Each of $\sigma(i+1)$ and $\sigma(r)$ is marginally uniform on $[n]$, so each gradient on the right has L^2 norm G_*. Thus, without requiring independence,

$$(\mathbb{E}_\sigma \|\mathbf{s}'_t - \mathbf{s}_t\|^2)^{1/2} \le 2G_*, \qquad (\mathbb{E}_\sigma \|\boldsymbol{x}'_t - \boldsymbol{x}_t\|^2)^{1/2} \le q_t + 2\eta G_*.$$

The second bound follows from $\boldsymbol{x}'_t - \boldsymbol{x}_t = \widetilde{\boldsymbol{x}}'_t - \widetilde{\boldsymbol{x}}_t - \eta(\mathbf{s}'_t - \mathbf{s}_t)$. For $r \le t < i$, both runs use the same component at step $t+1$, and hence

$$\begin{aligned} \widetilde{\boldsymbol{x}}'_{t+1} - \widetilde{\boldsymbol{x}}_{t+1} &= \widetilde{\boldsymbol{x}}'_t - \widetilde{\boldsymbol{x}}_t - \eta \big[\mathbf{h}_{\sigma(t+1)}(\boldsymbol{x}'_t) - \mathbf{h}_{\sigma(t+1)}(\boldsymbol{x}_t) \big], \\ q_{t+1} &\le q_t + \eta L (\mathbb{E}_\sigma \|\boldsymbol{x}'_t - \boldsymbol{x}_t\|^2)^{1/2} \\ &\le (1 + \eta L) q_t + 2\eta^2 L G_*. \end{aligned}$$

Here the first inequality uses the L^2 triangle inequality and Lipschitzness.

Unrolling this recursion and using $(1+\eta L)^n \le e^{1/4} < 4/3$ yields

$$q_i \le 2\eta L (1+\eta L)^{i-r} d_{r-1} + 2\eta \big((1+\eta L)^{i-r} - 1 \big) G_* \le \frac{8}{3} \eta L d_{r-1} + \frac{2}{3} \eta G_*,$$

$$(\mathbb{E}_\sigma \|\boldsymbol{x}'_i - \boldsymbol{x}_i\|^2)^{1/2} \le q_i + 2\eta G_* \le \frac{8}{3} \eta (L d_{r-1} + G_*). \tag{39}$$

For each fixed $R = r$, swapping positions r and $i+1$ is a bijection on the set of permutations, so σ' is uniform. Since both trajectories use the same update rule and starting point, $\widetilde{\boldsymbol{x}}'_i$ and $\widetilde{\boldsymbol{x}}_i$ have the same distribution under $\mathbb{E}_\sigma$. Using $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \eta \mathbf{s}_i$ and the L^2 triangle inequality, we obtain

$$(\mathbb{E}_\sigma \|\widetilde{\boldsymbol{x}}'_i - \boldsymbol{x}^*\|^2)^{1/2} = (\mathbb{E}_\sigma \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^*\|^2)^{1/2}$$

$$
\begin{aligned}
&= (\mathbb{E}_\sigma \|(\boldsymbol{x}_i - \boldsymbol{x}^*) + \eta \mathbf{s}_i\|^2)^{1/2} \\
&\le d_i + \eta (\mathbb{E}_\sigma \|\mathbf{s}_i\|^2)^{1/2} \\
&= d_i + \eta \sqrt{\frac{i(n-i)}{n-1}} G_* \le d_i + \eta \sqrt{n} G_*.
\end{aligned}
$$

Here the last line uses Equation (33).

For fixed $R = r \le i$, let $D_{i,r}$ denote the inner-product difference in Equation (38). Adding and subtracting $\langle \mathbf{h}_J(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i' - \boldsymbol{x}^* \rangle$ gives

$$
D_{i,r} = \langle \mathbf{h}_J(\boldsymbol{x}_i') - \mathbf{h}_J(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i' - \boldsymbol{x}^* \rangle + \langle \mathbf{h}_J(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i' - \widetilde{\boldsymbol{x}}_i \rangle .
$$

Cauchy–Schwarz, Lipschitzness, and $\|\mathbf{h}_J(\boldsymbol{x}_i)\| \le L \|\boldsymbol{x}_i - \boldsymbol{x}^*\|$ imply

$$
\begin{aligned}
\mathbb{E}_\sigma |D_{i,r}| &\le L (\mathbb{E}_\sigma \|\boldsymbol{x}_i' - \boldsymbol{x}_i\|^2)^{1/2} (d_i + \eta \sqrt{n} G_*) + L d_i q_i \\
&\le \frac{16}{3} \eta L^2 d_{r-1} d_i + \frac{10}{3} \eta L G_* d_i \\
&\quad + \frac{8}{3} \eta^2 L^2 \sqrt{n} G_* d_{r-1} + \frac{8}{3} \eta^2 L \sqrt{n} G_*^2,
\end{aligned}
$$

where the second inequality substitutes both bounds from Equation (39). Because R is uniform and the difference vanishes for $R > i$, averaging over R gives

$$
\sum_{i=0}^{n-1} |B_i| \le \frac{1}{n} \sum_{i=0}^{n-1} \sum_{r=1}^{i} \mathbb{E}_\sigma |D_{i,r}|.
$$

To sum the four terms above, use

$$
\sum_i d_i \le n \sqrt{W}, \qquad \sum_{0 \le j < i < n} d_j d_i = \frac{1}{2} \left[\left(\sum_i d_i \right)^2 - \sum_i d_i^2 \right] \le \frac{n^2 W}{2},
$$

$$
\frac{1}{n} \sum_{i=0}^{n-1} i d_i \le n \sqrt{W}, \qquad \frac{1}{n} \sum_{i=0}^{n-1} \sum_{r=1}^{i} d_{r-1} \le n \sqrt{W}, \qquad \frac{1}{n} \sum_{i=0}^{n-1} i = \frac{n-1}{2} \le \frac{n}{2},
$$

Combining these estimates with $\eta L \sqrt{n} \le 1/4$ gives

$$
\left| \sum_{i=0}^{n-1} B_i \right| \le \sum_{i=0}^{n-1} |B_i| \le \frac{8}{3} \eta n L^2 W + 4 \eta n L G_* \sqrt{W} + \frac{4}{3} \eta^2 n^{3/2} L G_*^2. \tag{40}
$$

Here the two mixed terms have total coefficient at most $10/3 + (8/3)\eta L \sqrt{n} \le 4$.

The epoch recursion. Strong convexity of F gives

$$
\langle \nabla F(\boldsymbol{x}_i), \boldsymbol{x}_i - \boldsymbol{x}^* \rangle \ge F(\boldsymbol{x}_i) - F^* + \frac{\mu}{2} \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 . \tag{41}
$$

Young's inequality and L-smoothness gives

$$
\begin{aligned}
2\eta^2 \left| \langle \nabla F(\boldsymbol{x}_i), \mathbf{s}_i \rangle \right| &\le \frac{\eta}{2L} \|\nabla F(\boldsymbol{x}_i)\|^2 + 2\eta^3 L \|\mathbf{s}_i\|^2 \\
&\le \eta (F(\boldsymbol{x}_i) - F^*) + 2\eta^3 L \|\mathbf{s}_i\|^2 .
\end{aligned} \tag{42}
$$

Using $\widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^* = (\boldsymbol{x}_i - \boldsymbol{x}^*) + \eta \mathbf{s}_i$, we obtain

$$
\begin{aligned}
-2\eta \langle \nabla F(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i - \boldsymbol{x}^* \rangle &= -2\eta \langle \nabla F(\boldsymbol{x}_i), \boldsymbol{x}_i - \boldsymbol{x}^* \rangle - 2\eta^2 \langle \nabla F(\boldsymbol{x}_i), \mathbf{s}_i \rangle \\
&\le -2\eta (F(\boldsymbol{x}_i) - F^*) - \eta\mu \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 + 2\eta^2 \left| \langle \nabla F(\boldsymbol{x}_i), \mathbf{s}_i \rangle \right| \\
&\le -\eta (F(\boldsymbol{x}_i) - F^*) - \eta\mu \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 + 2\eta^3 L \|\mathbf{s}_i\|^2 \\
&\le -\frac{3}{2} \eta\mu \|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2 + 2\eta^3 L \|\mathbf{s}_i\|^2 .
\end{aligned} \tag{43}
$$

The first inequality is by strong convexity. The second uses Equation (42), and the last again uses strong convexity. Expanding the squared norm in Equation (32), taking expectations, and using the definition of B_i gives

$$\begin{aligned}\mathbb{E}\left\|\widetilde{\boldsymbol{x}}_{i+1}-\boldsymbol{x}^*\right\|^2 &= \mathbb{E}\left\|\widetilde{\boldsymbol{x}}_i-\boldsymbol{x}^*\right\|^2 - 2\eta\mathbb{E}\left\langle \mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i-\boldsymbol{x}^*\right\rangle \\ &\quad + \eta^2\mathbb{E}\left\|\mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i)\right\|^2 \\ &= \mathbb{E}\left\|\widetilde{\boldsymbol{x}}_i-\boldsymbol{x}^*\right\|^2 - 2\eta\mathbb{E}\left\langle \nabla F(\boldsymbol{x}_i), \widetilde{\boldsymbol{x}}_i-\boldsymbol{x}^*\right\rangle \\ &\quad - 2\eta B_i + \eta^2\mathbb{E}\left\|\mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i)\right\|^2.\end{aligned}$$

Summing over $i = 0, \dots, n-1$ telescopes the shifted squared distances. Since $\widetilde{\boldsymbol{x}}_0 = \boldsymbol{x}_0$ and $\widetilde{\boldsymbol{x}}_n = \boldsymbol{x}_n$, the endpoint terms are unshifted. We also use $\left\|\mathbf{h}_{\sigma(i+1)}(\boldsymbol{x}_i)\right\|^2 \le L^2\left\|\boldsymbol{x}_i-\boldsymbol{x}^*\right\|^2$ and $-2\eta\sum_i B_i \le 2\eta\left|\sum_i B_i\right|$.

Thus Equations (33), (40) and (43) give

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{x}_n-\boldsymbol{x}^*\right\|^2 &\le d_0^2 - \frac{3}{2}\eta\mu n W + \eta^2 n L^2 W + 2\eta\left|\sum_i B_i\right| + 2\eta^3 L\sum_i \mathbb{E}\left\|\mathbf{s}_i\right\|^2 \\ &\le d_0^2 - \frac{3}{2}\eta\mu n W + \frac{19}{3}\eta^2 n L^2 W + 8\eta^2 n L G_*\sqrt{W} \\ &\quad + \eta^3 L G_*^2\left(\frac{n(n+1)}{3} + \frac{8}{3}n^{3/2}\right). \end{aligned} \tag{44}$$

By Equation (31) and Young's inequality,

$$\frac{19}{3}\eta^2 n L^2 W \le \frac{\eta\mu n}{4}W, \qquad 8\eta^2 n L G_*\sqrt{W} \le \frac{\eta\mu n}{4}W + \frac{64\eta^3 n L^2 G_*^2}{\mu}.$$

Since $n(n+1)/3 + 8n^{3/2}/3 \le 3n^2$ for $n \ge 2$, substituting into Equation (44) and then using $W \ge d_0^2/2 - \eta^2 n G_*^2$ from Equation (34) gives

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{x}_n-\boldsymbol{x}^*\right\|^2 &\le d_0^2 - \eta\mu n W + 3\eta^3 n^2 L G_*^2 + \frac{64\eta^3 n L^2 G_*^2}{\mu} \\ &\le \left(1-\frac{\eta\mu n}{2}\right)d_0^2 + 4\eta^3 n^2 L G_*^2 + \frac{64\eta^3 n L^2 G_*^2}{\mu}, \end{aligned} \tag{45}$$

where the last step uses $\mu \le L$.

Unrolling the recursion and choosing the stepsize. Restore the epoch indices and take total expectations in Equation (45). Fresh independent permutations make this conditional bound valid at every epoch. Let $D_0 := \left\|\boldsymbol{x}_0-\boldsymbol{x}^*\right\|$ and $\rho := 1-\eta\mu n/2 \in (0,1)$. Using $\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_n^k$ and $\sum_{k=0}^{K-1}\rho^k \le 2/(\eta\mu n)$, we obtain

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{x}_n^K-\boldsymbol{x}^*\right\|^2 &\le \rho^K D_0^2 + \left(4\eta^3 n^2 L + \frac{64\eta^3 n L^2}{\mu}\right)G_*^2\sum_{k=0}^{K-1}\rho^k \\ &\le \left(1-\frac{\eta\mu n}{2}\right)^K D_0^2 + 8\eta^2 G_*^2(n\kappa + 16\kappa^2). \end{aligned} \tag{46}$$

This bound also holds when $G_* = 0$.

For $G_* > 0$, take the stated stepsize $\eta = \Lambda_{nK}/(\mu n K)$. The sufficient horizon condition

$$K \ge \Lambda_{nK}\max\left\{4\kappa, \frac{32\kappa^2}{n}\right\}$$

ensures both restrictions in Equation (31) and makes $K \gtrsim \max\{\kappa, \kappa^2/n\}$ precise. By the definition of Λ_{nK} in Section 2,

$$\left(1-\frac{\eta\mu n}{2}\right)^K D_0^2 \le \exp(-\Lambda_{nK}/2)D_0^2 \le \frac{G_*^2}{\mu^2(nK)^2}.$$

Substituting into Equation (46) gives

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \le \frac{G_*^2}{\mu^2 nK^2}\left[\frac{1}{n} + 8\Lambda_{nK}^2\left(\kappa + \frac{16\kappa^2}{n}\right)\right]$$
$$\le (1 + 136\Lambda_{nK}^2)\frac{LG_*^2}{\mu^3 nK^2}\max\left\{1, \frac{\kappa}{n}\right\}.$$

The last inequality uses $\kappa \ge 1$ and $\kappa + 16\kappa^2/n \le 17\kappa\max\{1, \kappa/n\}$. Suppressing the logarithmic factor proves the claim. □

D.3 Lower Bound

Theorem A.8 (RR Lower Bound without component convexity)**.** *For any even $n \ge 16$, $\kappa \ge 4096$, $K \ge \max\{\kappa, \kappa^2/n\}$, $G_* \ge 0$, and $D \ge 0$, there exists a 5-dimensional finite-sum objective F satisfying Assumption 2.3, along with an initialization point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = D$, such that for any constant stepsize $\eta > 0$, the final iterate $\boldsymbol{x}_n^K$ obtained by RR satisfies*

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \gtrsim \min\left\{D^2, \frac{LG_*^2}{\mu^3 nK^2}\max\left\{1, \frac{\kappa}{n}\right\}\right\}.$$

Proof. We construct three one-dimensional finite sums and one two-dimensional finite sum and combine them across coordinates so that a single objective gives the desired lower bound for every step size. The two constructions for moderate step sizes provide the two terms in the maximum. Write $\eta_0 := 1/(2\mu nK)$ and $g := G_*/\sqrt{2}$. Our constructions F_1, F_2, F_3, and F_4, with component functions $f_{j,i}$ for $j \in \{1, 2, 3, 4\}$ and $i \in [n]$, will have the following properties:

- **Small step size regime.** Starting from $u_0 = D/\sqrt{2}$, for every $0 \le \eta < \eta_0$, the final iterate on F_1 satisfies
$$|u_n^K|^2 \ge \frac{D^2}{8}.$$
- **Moderate step size regime I.** Starting from $x_0 = 0$, for every $\eta_0 \le \eta < 2/L$, the final iterate on F_2 satisfies
$$\mathbb{E}|x_n^K|^2 \ge 2^{-21}\frac{L^2G_*^2}{\mu^4 n^2K^2}.$$
- **Moderate step size regime II.** Starting from $\boldsymbol{y}_0 = \mathbf{0} \in \mathbb{R}^2$, for every $\eta_0 \le \eta < 2/L$, the final iterate on F_3 satisfies
$$\mathbb{E}\left\|\boldsymbol{y}_n^K\right\|^2 \ge 2^{-33}\frac{LG_*^2}{\mu^3 nK^2}.$$
(We obtain this bound from the two-dimensional construction in the proof of Theorem A.4.)
- **Large step size regime.** Starting from $v_0 = D/\sqrt{2}$, for every $\eta \ge 2/L$, the final iterate on F_4 satisfies
$$|v_n^K|^2 \ge \frac{D^2}{2}.$$

Each F_j is μ-strongly convex and minimized at zero, and all component functions are L-smooth. The component gradients of F_1 and F_4 vanish at zero, whereas those of F_2 and F_3 each have empirical second moment $g^2 = G_*^2/2$. We verify these properties below.

We aggregate the functions by defining

$$F(u, x, \boldsymbol{y}, v) := F_1(u) + F_2(x) + F_3(\boldsymbol{y}) + F_4(v),$$
$$f_i(u, x, \boldsymbol{y}, v) := f_{1,i}(u) + f_{2,i}(x) + f_{3,i}(\boldsymbol{y}) + f_{4,i}(v), \qquad i \in [n],$$
$$\boldsymbol{x}_0 := \left(\frac{D}{\sqrt{2}}, 0, \mathbf{0}, \frac{D}{\sqrt{2}}\right) \in \mathbb{R}^5.$$

Block diagonality preserves component L-smoothness and average μ-strong convexity. The minimizer is $\boldsymbol{x}^* = (0, 0, \mathbf{0}, 0)$, and the initialization has distance exactly D. Moreover,

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{x}^*)\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(|f'_{2,i}(0)|^2 + \|\nabla f_{3,i}(\mathbf{0})\|^2\right)$$
$$= g^2 + g^2 = G_*^2.$$

Under any realized sequence of permutations, the four blocks follow their respective updates. Hence,

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 = |u_n^K|^2 + \mathbb{E}|x_n^K|^2 + \mathbb{E}\left\|\boldsymbol{y}_n^K\right\|^2 + |v_n^K|^2,$$

and the stated bounds imply

$$\mathbb{E}\left\|\boldsymbol{x}_n^K - \boldsymbol{x}^*\right\|^2 \geq \min\left\{\frac{D^2}{8}, \max\left\{2^{-21}\frac{L^2G_*^2}{\mu^4n^2K^2}, 2^{-33}\frac{LG_*^2}{\mu^3nK^2}\right\}, \frac{D^2}{2}\right\}$$
$$\geq 2^{-33}\min\left\{D^2, \frac{LG_*^2}{\mu^3nK^2}\max\left\{1, \frac{\kappa}{n}\right\}\right\}.$$

It remains to give the four constructions and establish their claimed bounds.

Construction of F_1. Let

$$F_1(u) = \frac{\mu}{2}u^2, \qquad f_{1,i}(u) = F_1(u), \quad i \in [n].$$

Every component is L-smooth, the average is μ-strongly convex, and all component gradients vanish at the minimizer $u^* = 0$. Since the components are identical, the iterate is independent of the permutations. With $u_0 = D/\sqrt{2}$ and $0 \leq \eta < \eta_0$, Bernoulli's inequality gives

$$u_n^K = (1-\eta\mu)^{nK}\frac{D}{\sqrt{2}} \geq (1-\eta\mu nK)\frac{D}{\sqrt{2}} \geq \frac{D}{2\sqrt{2}}.$$

Consequently, $|u_n^K|^2 \geq D^2/8$ holds.

Construction of F_2. Set $a := L/2 - \mu$ and define

$$f_{2,i}(x) := \begin{cases} \dfrac{\mu+a}{2}x^2 + gx, & 1 \leq i \leq n/2, \\ \dfrac{\mu-a}{2}x^2 - gx, & n/2 < i \leq n. \end{cases}$$

Their average is $F_2(x) = \mu x^2/2$, with minimizer $x^* = 0$. Since $\kappa \geq 8$, we have

$$\frac{3L}{8} \leq a < \frac{L}{2}, \qquad \mu + a = \frac{L}{2}, \qquad \mu - a = 2\mu - \frac{L}{2} < 0.$$

Thus each component is L-smooth, although half of the components are nonconvex. Also, $|f'_{2,i}(0)| = g$ for every i. We use the initialization $x_0 = 0$.

Write $f_{2,i}(x) = h_i x^2/2 + b_i x$, where (h_i, b_i) equals $(\mu + a, g)$ for the first half of the components and $(\mu - a, -g)$ for the second half. A component update is

$$x^+ = (1 - \eta h_i)x - \eta b_i.$$

Composing the updates in epoch k yields

$$x_n^k = P(\eta)x_0^k + Q_{\sigma_k}(\eta),$$
$$Q_\sigma(\eta) := -\eta\sum_{r=1}^{n} b_{\sigma(r)}\prod_{s=r+1}^{n}(1 - \eta h_{\sigma(s)}),$$

where

$$H(t) := (1 - t(\mu+a))(1 - t(\mu-a)) = 1 - 2\mu t - (a^2 - \mu^2)t^2,$$
$$P(t) := H(t)^{n/2}.$$

Note that the multiplier $P(\eta)$ does not depend on the permutation because the scalar multipliers commute. On $0 \le t < 2/L$, $H(t)$ is positive and decreasing, with $H(0) = 1$. In particular, $0 < P(\eta) < 1$ for $0 < \eta < 2/L$.

Let $q(\eta) := \mathbb{E}_\sigma[Q_\sigma(\eta)]$ for a uniform permutation σ. Since $x_0 = 0$, unrolling the epoch recursion gives

$$-\mathbb{E}[x_n^K] = \frac{-q(\eta)}{1 - P(\eta)}\big(1 - P(\eta)^K\big). \tag{47}$$

To obtain a lower bound valid for every $\eta \in [\eta_0, 2/L)$, we will show that both factors $-q(\eta)/(1-P(\eta))$ and $1 - P(\eta)^K$ are nonnegative and nondecreasing in η on $(0, 2/L)$. Their product, the right-hand side of Equation (47), is then minimized over $[\eta_0, 2/L)$ at $\eta = \eta_0$, so it suffices to bound this product from below at η_0. The second factor already has these properties because $P(\eta)$ is decreasing and lies in $(0, 1)$.

To establish these properties for the first factor, we begin by computing $q'(\eta)$ directly from the permutation distribution. Fix a component label i and a subset $A \subseteq [n] \setminus \{i\}$. The probability that every element of A appears after i is $1/(|A| + 1)$, since this is exactly the event that i appears first among the labels in $A \cup \{i\}$. Expanding the products in Q_σ and taking expectations therefore gives

$$q(\eta) = -\sum_{i=1}^{n} b_i \sum_{A \subseteq [n]\setminus\{i\}} \frac{(-1)^{|A|}\eta^{|A|+1}}{|A| + 1} \prod_{j \in A} h_j.$$

This expression is a finite polynomial, and differentiation yields

$$\begin{aligned}
q'(\eta) &= -\sum_{i=1}^{n} b_i \sum_{A \subseteq [n]\setminus\{i\}} (-\eta)^{|A|} \prod_{j \in A} h_j \\
&= -\sum_{i=1}^{n} b_i \prod_{j \neq i} (1 - \eta h_j) \\
&= -\frac{ng}{2}(1 - \eta(\mu + a))^{n/2-1}(1 - \eta(\mu - a))^{n/2} \\
&\quad + \frac{ng}{2}(1 - \eta(\mu + a))^{n/2}(1 - \eta(\mu - a))^{n/2-1} \\
&= -\frac{ng}{2} H(\eta)^{n/2-1}\big((1 - \eta(\mu - a)) - (1 - \eta(\mu + a))\big).
\end{aligned}$$

Hence,

$$q'(\eta) = -nag\eta H(\eta)^{n/2-1}, \qquad q(0) = 0. \tag{48}$$

In particular, $q(\eta) < 0$ for $0 < \eta < 2/L$.

We now show that the right-hand side of Equation (47) is nondecreasing in η. Differentiating P and using Equation (48), we obtain

$$\begin{aligned}
P'(\eta) &= -n\big(\mu + (a^2 - \mu^2)\eta\big) H(\eta)^{n/2-1}, \\
\frac{q'(\eta)}{P'(\eta)} &= \frac{ag\eta}{\mu + (a^2 - \mu^2)\eta}.
\end{aligned}$$

This ratio is increasing on $(0, 2/L)$ because

$$\begin{aligned}
\frac{d}{d\eta}\left(\frac{q'(\eta)}{P'(\eta)}\right) &= \frac{d}{d\eta}\left(\frac{ag\eta}{\mu + (a^2 - \mu^2)\eta}\right) \\
&= \frac{ag\big(\mu + (a^2 - \mu^2)\eta\big) - ag\eta(a^2 - \mu^2)}{\big(\mu + (a^2 - \mu^2)\eta\big)^2} \\
&= \frac{ag\mu}{\big(\mu + (a^2 - \mu^2)\eta\big)^2} > 0.
\end{aligned}$$

For $0 < \eta < 2/L$, set $\phi(\eta) := -q(\eta)/(1 - P(\eta))$. By Cauchy's mean value theorem, there exists $\xi \in (0, \eta)$ such that

$$\phi(\eta) = \frac{q(\eta) - q(0)}{P(\eta) - P(0)} = \frac{q'(\xi)}{P'(\xi)} \leq \frac{q'(\eta)}{P'(\eta)}.$$

It follows that

$$\begin{aligned}\phi'(\eta) &= \frac{-q'(\eta)\big(1 - P(\eta)\big) - q(\eta)P'(\eta)}{\big(1 - P(\eta)\big)^2} \\ &= \frac{-q'(\eta) + P'(\eta)\phi(\eta)}{1 - P(\eta)} \\ &= \frac{-P'(\eta)}{1 - P(\eta)} \left(\frac{q'(\eta)}{P'(\eta)} - \phi(\eta)\right) \geq 0.\end{aligned}$$

Both $\phi(\eta)$ and $1 - P(\eta)^K$ are nonnegative and nondecreasing. Therefore, for every $\eta_0 \leq \eta < 2/L$,

$$-\mathbb{E}[x_n^K] \geq \phi(\eta_0)\big(1 - P(\eta_0)^K\big).$$

It remains to lower-bound $\phi(\eta_0)\big(1 - P(\eta_0)^K\big)$ at $\eta_0 = 1/(2\mu n K)$. The condition $K \geq \kappa^2/n$ implies $a^2/(\mu^2 n K) \leq 1$ and $\eta_0 < 2/L$. Recall that

$$\phi(\eta_0) = \frac{-q(\eta_0)}{1 - P(\eta_0)}, \qquad P(\eta) = H(\eta)^{n/2}.$$

We first lower-bound $\phi(\eta_0)$ by obtaining a lower bound on its numerator $-q(\eta_0)$ and an upper bound on its positive denominator $1 - P(\eta_0)$.

To estimate the numerator, the fundamental theorem of calculus, together with $q(0) = 0$ and Equation (48), gives

$$\begin{aligned}q(\eta) &= q(0) + \int_0^\eta q'(t)\, dt \\ &= -nag \int_0^\eta t H(t)^{n/2-1}\, dt.\end{aligned}$$

We therefore seek a uniform positive lower bound on $H(t)^{n/2-1}$ for $0 \leq t \leq \eta_0$. To obtain such a bound using Bernoulli's inequality, we first control $\left(\frac{n}{2} - 1\right)(1 - H(t))$. Since $1 - H(t) = 2\mu t + (a^2 - \mu^2)t^2$, for every $0 \leq t \leq \eta_0$ we have

$$\begin{aligned}\left(\frac{n}{2} - 1\right)(1 - H(t)) &\leq n\mu\eta_0 + \frac{n}{2}a^2\eta_0^2 \\ &= \frac{1}{2K} + \frac{a^2}{8\mu^2 n K^2} \leq \frac{5}{8}.\end{aligned}$$

Bernoulli's inequality consequently gives

$$H(t)^{n/2-1} \geq 1 - \left(\frac{n}{2} - 1\right)(1 - H(t)) \geq \frac{3}{8}.$$

Substituting this bound into the integral representation at $\eta = \eta_0$ yields the desired lower bound on the numerator:

$$\begin{aligned}-q(\eta_0) &= nag \int_0^{\eta_0} t H(t)^{n/2-1}\, dt \\ &\geq \frac{3}{8} nag \int_0^{\eta_0} t\, dt = \frac{3}{16} nag\eta_0^2.\end{aligned}$$

We next upper-bound the denominator $1 - P(\eta_0)$ of $\phi(\eta_0)$. Using $P(\eta_0) = H(\eta_0)^{n/2}$ and applying Bernoulli's inequality again, we obtain

$$1 - P(\eta_0) = 1 - H(\eta_0)^{n/2}$$

$$
\begin{aligned}
&\le \frac{n}{2}(1 - H(\eta_0)) \\
&\le n\mu\eta_0 + \frac{n}{2}a^2\eta_0^2 \\
&= n\mu\eta_0\left(1 + \frac{a^2\eta_0}{2\mu}\right) \\
&= n\mu\eta_0\left(1 + \frac{a^2}{4\mu^2 nK}\right) \le \frac{5}{4}n\mu\eta_0.
\end{aligned}
$$

Combining these numerator and denominator bounds gives

$$
\begin{aligned}
\phi(\eta_0) &= \frac{-q(\eta_0)}{1 - P(\eta_0)} \\
&\ge \frac{\frac{3}{16}nag\eta_0^2}{\frac{5}{4}n\mu\eta_0} = \frac{3ag\eta_0}{20\mu} = \frac{3ag}{40\mu^2 nK}.
\end{aligned}
$$

It remains to lower-bound the other factor $1 - P(\eta_0)^K$. Since $H(\eta_0) \le 1 - 2\mu\eta_0 \le e^{-2\mu\eta_0}$, we have

$$
P(\eta_0)^K = H(\eta_0)^{nK/2} \le e^{-nK\mu\eta_0} = e^{-1/2}.
$$

Using $1 - e^{-1/2} > 1/3$, we conclude that

$$
-\mathbb{E}[x_n^K] \ge \frac{ag}{40\mu^2 nK} \qquad \text{for every } \eta_0 \le \eta < 2/L.
$$

Finally, Jensen's inequality and $a \ge 3L/8$ imply

$$
\begin{aligned}
\mathbb{E}|x_n^K|^2 &\ge \left(\mathbb{E}[x_n^K]\right)^2 \\
&\ge \frac{a^2 g^2}{1600\mu^4 n^2 K^2} \\
&\ge \frac{9L^2 g^2}{102400\mu^4 n^2 K^2} \\
&\ge 2^{-20}\frac{L^2 g^2}{\mu^4 n^2 K^2} = 2^{-21}\frac{L^2 G_*^2}{\mu^4 n^2 K^2}.
\end{aligned}
$$

Construction of F_3. Let F_3, with component functions $f_{3,i}$, be the two-dimensional construction denoted by F_2 in Appendix C.3, with G_* replaced by g. Its component functions are convex and L-smooth, and its average is μ-strongly convex and minimized at $\mathbf{0}$. Moreover,

$$
\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_{3,i}(\mathbf{0})\|^2 = g^2 = \frac{G_*^2}{2}.
$$

Starting from $\boldsymbol{y}_0 = \mathbf{0}$, the bound established there gives, for every $\eta_0 \le \eta < 2/L$,

$$
\begin{aligned}
\mathbb{E}\left\|\boldsymbol{y}_n^K\right\|^2 &\ge 2^{-32}\frac{g^2}{\mu^2 nK}\min\left\{1, \frac{\kappa}{K}\right\} \\
&= 2^{-33}\frac{LG_*^2}{\mu^3 nK^2},
\end{aligned}
$$

where the equality uses $K \ge \kappa$.

Construction of F_4. Let

$$
F_4(v) = \frac{L}{2}v^2, \qquad f_{4,i}(v) = F_4(v), \quad i \in [n].
$$

Every component is L-smooth, the average is μ-strongly convex, and all component gradients vanish at $v^* = 0$. With $v_0 = D/\sqrt{2}$, for every $\eta \ge 2/L$ we have

$$
|v_n^K| = |1 - \eta L|^{nK}\frac{D}{\sqrt{2}} \ge \frac{D}{\sqrt{2}},
$$

since $|1 - \eta L| \ge 1$. Thus, $|v_n^K|^2 \ge D^2/2$. □