Minibatch Random Reshuffling for Smooth Convex Optimization: Batch-Size Dependence, Finite-Population Variance, and Quadratic Growth

Rajesh Kumar*

November 2026

Abstract

Random Reshuffling (RR) is the default data-ordering strategy in practice, and it is almost always run with minibatches. The recent theory of Liu [6, 7] shows that single-sample RR dominates SGD in smooth convex optimization for every stepsize below $2/\hat{L}$, but the role of the batch size has not been worked out. In this note we study *minibatch RR* with a fixed partition of the n components into $m = n/b$ batches and a fresh random batch order in each epoch. We show that this method is exactly Shuffling SGD on m "batch components", which lets us transfer the dominance theory of Liu while tracking how the batch size enters every constant. We obtain (i) an epoch-wise convergence rate for minibatch RR valid for all stepsizes $\eta < 2/\hat{L}_P$, where $\hat{L}_P \le \min\{\hat{L}, m\bar{L}\}$ is the maximal smoothness of a batch; (ii) a proof that minibatch RR dominates minibatch SGD on the same batches; (iii) an exact finite-population variance identity showing that, for a uniformly random partition, the effective gradient variance at the optimum is $\sigma_\star^2 \frac{n-b}{b(n-1)}$ per batch; (iv) a precise accounting of what batching does and does not change per epoch: the optimization term is multiplied by $b\hat{L}_P/\hat{L}$, the variance-driven terms change only by a factor $\sqrt{(n-b)/(n-1)}$ (resp. its cube root form), and the stepsize regime in which RR strictly beats SGD is enlarged by a factor b; and (v) a sample-complexity bound for functions with quadratic growth obtained by restarting minibatch RR. We also identify the setting that is *not* covered by our reduction, namely a fresh partition in every epoch, and state it as an open problem.

1 Introduction

A central task in machine learning is to minimize a finite sum $f(x) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x)$. In practice, the gradient method of choice processes the data in epochs, draws a new random permutation at the start of each epoch (Random Reshuffling, RR), and uses minibatches rather than single samples. The theory of RR has long lagged behind this practice [4, 5, 8, 10]. In smooth convex optimization, the known rates required a stepsize of order $1/n$ until the recent work of Liu [6], who proved that RR converges for every $\eta < 2/\hat{L}$ and dominates SGD after any finite number of epochs. That work, however, analyses single-sample RR. Minibatching changes three things at once: the number of updates per epoch drops from n to $m = n/b$, the smoothness constant of a step changes from L_i to a batch average, and the variance of the stochastic gradient at the optimum changes from $\sigma_\star^2$ to a batch-level quantity. A theory that explains the practice of minibatch RR must say how these three effects combine.

*Independent Researcher, Bangalore, India. ORCID: 0009-0009-7333-5411.

1.1 Our Contributions

- **Exact reduction.** We observe that minibatch RR with a fixed partition is, as an algorithm, *identical* to Shuffling SGD under RR applied to the m batch objectives $g_j = \frac{1}{b}\sum_{i\in B_j} f_i$. Combined with the rate of Liu [6], this yields Theorem 1: for every stepsize $\eta < 2/\hat{L}_P$,

$$\mathbb{E}[f(\bar{x}_K) - f_\star] \lesssim \frac{D^2}{(2-\eta\hat{L}_P)\,\eta\,mK} + \frac{\min\{1, \eta m\bar{L}\}\,\eta\,\sigma_P^2}{(2-\eta\hat{L}_P)^2},$$

  where $\hat{L}_P$ is the maximum batch smoothness and σ_P^2 is the batch-level gradient variance at the optimum. We prove $\bar{L} \le \hat{L}_P \le \min\{\hat{L}, m\bar{L}\}$ and $\sigma_P^2 \le \sigma_\star^2$ (Lemma 1).

- **Dominance over minibatch SGD.** Corollary 1 shows that, on the same batches and for any $\eta < 2/\hat{L}_P$, our bound is never worse than the corresponding bound for minibatch SGD and is strictly better when $\eta \lesssim 1/(m\bar{L})$.

- **Finite-population variance.** For a uniformly random partition we prove the identity $\mathbb{E}[\sigma_P^2] = \sigma_\star^2 \frac{n-b}{b(n-1)}$ (Lemma 2), which makes the dependence on the batch size explicit and vanishes at full batch $b = n$, as it must.

- **What batching changes.** Theorem 2 and Corollary 2 give tuned rates in terms of the number nK of individual gradient evaluations. The conclusion (Section 3.4) is a candid one: for a fixed budget of gradient evaluations, worst-case upper bounds do *not* reward batching through the variance terms, which change only by the finite-population factor, while the optimization term is multiplied by $b\hat{L}_P/\hat{L}$. Batching therefore helps in the worst-case bound exactly when a few components dominate the smoothness ($\hat{L} \gtrsim b\hat{L}_P$), and it always widens by a factor b the stepsize range $\eta \lesssim b/(n\bar{L})$ in which RR is provably faster than SGD.

- **Quadratic growth.** Restarting minibatch RR (Theorem 3, Corollary 3) gives a linearly convergent method up to a noise floor, under quadratic growth, with the batch size appearing explicitly in the gradient-evaluation complexity.

We stress at the outset what is and is not new. The deep technical ingredient, the convergence rate of RR for every stepsize below $2/\hat{L}$, is due to Liu [6, 7] and is used here as a black box (Fact 1). Our contribution is the batch-level reduction, the variance identity, the resulting dependence on the batch size, and the corollaries that follow.

1.2 Related Work

The first provable advantage of RR over SGD in smooth strongly convex optimization is due to Gürbüzbalaban et al. [4]. For smooth convex problems, Mishchenko et al. [8] and Nguyen et al. [10] proved the rate $\frac{D^2}{\eta nK} + \eta^2 nL\sigma_\star^2$ for $\eta \lesssim 1/(nL)$; Nagaraj et al. [9] treat larger steps under an additional Lipschitz assumption; Cha et al. [2] give lower bounds. Haochen and Sra [5] compare RR with SGD in a setting where all components share a minimizer. The last restriction on the stepsize was removed by Liu [6], whose results are the starting point of this note. The rate of SGD that we compare against is due to Garrigos and Gower [3] and, for stepsizes up to $2/L$, Attia et al. [1]. To our knowledge, the dependence of these bounds on the batch size, and the relation between the batch variance and the finite-population correction, have not been recorded in this regime.

2 Preliminaries

Notation. $[n] \triangleq \{1, \dots, n\}$; $\|\cdot\|$ is the Euclidean norm; $A \lesssim B$ means $A \le CB$ for an absolute constant $C > 0$. For differentiable h, $\nabla h(x)$ is its gradient.

Objective and assumptions. We study $\inf_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f_i(x)$ under the same three standard assumptions as Liu [6].

Assumption 1 (Minimizer)**.** *There is $x_\star \in \mathbb{R}^d$ with $f_\star \triangleq f(x_\star) = \inf_x f(x) \in \mathbb{R}$.*

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

Assumption 3 (Smoothness)**.** *Each f_i is L_i-smooth: $\|\nabla f_i(x) - \nabla f_i(y)\| \le L_i \|x - y\|$.*

Let $\sigma_\star^2 \triangleq \frac{1}{n} \sum_i \|\nabla f_i(x_\star)\|^2$, $\bar{L} \triangleq \frac{1}{n} \sum_i L_i$, $\hat{L} \triangleq \max_i L_i$, and $D \triangleq \|x_1^1 - x_\star\|$. We also use quadratic growth.

Assumption 4 (Quadratic growth)**.** *$f(x) - f_\star \ge \frac{\mu}{2} \operatorname{dist}^2(x, X_\star)$ for all x, with $X_\star = \arg\min f$ and $\mu > 0$.*

Minibatch partitions. Assume $n = mb$ with $m \ge 2$. A *partition* $P = \{B_1, \dots, B_m\}$ splits $[n]$ into m batches of size b. Define the batch objectives and their parameters

$$g_j \triangleq \frac{1}{b} \sum_{i \in B_j} f_i, \quad L_{B_j} \triangleq \frac{1}{b} \sum_{i \in B_j} L_i, \quad \hat{L}_P \triangleq \max_{j \in [m]} L_{B_j}, \quad \sigma_P^2 \triangleq \frac{1}{m} \sum_{j=1}^m \|\nabla g_j(x_\star)\|^2.$$

Algorithm.

> **Minibatch Random Reshuffling (MB-RR)**
> **Input:** partition P, initial point x_1^1, stepsize $\eta > 0$, epochs K.
> **for** $k = 1, \dots, K$: draw π_k uniformly from the permutations of $[m]$;
> **for** $i = 1, \dots, m$: $x_k^{i+1} = x_k^i - \eta \nabla g_{\pi_k^i}(x_k^i)$;
> $x_{k+1}^1 = x_k^{m+1}$.
> **Output:** $\bar{x}_K \triangleq \frac{1}{mK} \sum_{k=1}^K \sum_{i=1}^m x_k^i$.

Each epoch costs n individual gradient evaluations (b per step, m steps). Minibatch SGD on the same batches (MB-SGD) is the same loop with π_k^i drawn independently and uniformly from $[m]$.

Result used as a black box. The following is the constant-stepsize form of Theorem 1 of Liu [6], as displayed in the proof of its Corollary 1, together with the SGD bound of its Theorem 11 and the restart bound of its Theorem 4.

Fact 1 (Liu 6, 7)**.** *Let $h_1, \dots, h_N$ ($N \ge 2$) satisfy Assumptions 1–3 with smoothness constants $\ell_1, \dots, \ell_N$, $\bar{\ell} = \frac{1}{N} \sum \ell_i$, $\hat{\ell} = \max \ell_i$, and let $s^2 = \frac{1}{N} \sum \|\nabla h_i(x_\star)\|^2$. Run Shuffling SGD under RR with constant stepsize $\eta < 2/\hat{\ell}$ for K epochs, and let $\bar{x}_K$ be the uniform average of the NK iterates.*

Then

(a) $\mathbb{E}[h(\bar{x}_K) - h_\star] \le C\Big[\frac{D^2}{(2-\eta\hat{\ell})\eta NK} + \frac{\min\{1,\eta N\bar{\ell}\}\,\eta s^2}{(2-\eta\hat{\ell})^2}\Big]$;

(b) *for SGD with the same data:* $\mathbb{E}[h(\bar{x}_K) - h_\star] \le \frac{2D^2}{(2-\eta\hat{\ell})\eta NK} + \frac{8\,\eta s^2}{(2-\eta\hat{\ell})^2}$;

(c) *if in addition* h *has* μ*-quadratic growth, then Restarted Shuffling SGD with* $K \asymp \frac{1}{(2-\eta\hat{\ell})\eta N\mu}$ *satisfies* $\mathbb{E}[h(x_{T+1}) - h_\star] \lesssim \frac{\mu\,\mathrm{dist}^2(x_1, X_\star)}{2^T} + \frac{\min\{1,\eta N\bar{\ell}\}\eta s^2}{(2-\eta\hat{\ell})^2}$.

Here C *is an absolute constant and* $D = \|x_1^1 - x_\star\|$.

3 Main Results

3.1 Batch Structure

Lemma 1 (Batch structure)**.** *Under Assumptions 1–3, for any partition* P*:*

1. *each* g_j *is convex and* L_{B_j}*-smooth,* $\frac{1}{m}\sum_j g_j = f$*, and* $x_\star$ *minimizes* f*;*
2. $\frac{1}{m}\sum_j L_{B_j} = \bar{L}$*, and* $\bar{L} \le \hat{L}_P \le \min\{\hat{L},\, m\bar{L}\}$*;*
3. $\sigma_P^2 \le \sigma_\star^2$*, and* $\sigma_P^2 = 0$ *whenever* $b = n$.

Proof. (1) A sum of convex functions is convex, and $\|\nabla g_j(x) - \nabla g_j(y)\| \le \frac{1}{b}\sum_{i\in B_j} L_i\|x - y\|$. Averaging the g_j over j returns f because the batches partition $[n]$. (2) The average of the L_{B_j} equals $\frac{1}{mb}\sum_i L_i = \bar{L}$, which is at most their maximum $\hat{L}_P$. Each L_{B_j} is an average of numbers at most $\hat{L}$, hence $\hat{L}_P \le \hat{L}$; and $L_{B_j} \le \frac{1}{b}\sum_{i=1}^n L_i = \frac{n}{b}\bar{L} = m\bar{L}$. (3) Write $a_i = \nabla f_i(x_\star)$. By convexity of $\|\cdot\|^2$, $\|\nabla g_j(x_\star)\|^2 = \|\frac{1}{b}\sum_{i\in B_j} a_i\|^2 \le \frac{1}{b}\sum_{i\in B_j}\|a_i\|^2$; averaging over j gives $\sigma_P^2 \le \frac{1}{n}\sum_i \|a_i\|^2 = \sigma_\star^2$. For $b = n$ there is one batch, $m = 1$, and $\nabla g_1(x_\star) = \nabla f(x_\star) = 0$ (the case $m = 1$ is excluded from our theorems, since it is plain gradient descent, but the identity holds). □

Lemma 2 (Finite-population variance)**.** *Let* P *be a uniformly random partition of* $[n]$ *into batches of size* b *(so each batch is a uniformly random* b*-subset). Then*

$$\mathbb{E}[\sigma_P^2] = \sigma_\star^2\,\frac{n-b}{b\,(n-1)}.$$

Proof. Let $a_i = \nabla f_i(x_\star)$, so $\sum_i a_i = \nabla(nf)(x_\star) = 0$. For a uniformly random b-subset S, $\Pr[i \in S] = \frac{b}{n}$ and $\Pr[i, j \in S] = \frac{b(b-1)}{n(n-1)}$ for $i \ne j$. Hence

$$\mathbb{E}\Big\|\sum_{i\in S} a_i\Big\|^2 = \frac{b}{n}\sum_i \|a_i\|^2 + \frac{b(b-1)}{n(n-1)}\sum_{i\ne j}\langle a_i, a_j\rangle.$$

Since $\sum_{i\ne j}\langle a_i, a_j\rangle = \|\sum_i a_i\|^2 - \sum_i\|a_i\|^2 = -\sum_i\|a_i\|^2$, this equals $\frac{b}{n}\big(1 - \frac{b-1}{n-1}\big)\sum_i\|a_i\|^2 = \frac{b(n-b)}{n(n-1)}\sum_i\|a_i\|^2$. Dividing by b^2 and using $\frac{1}{n}\sum\|a_i\|^2 = \sigma_\star^2$ gives $\mathbb{E}\|\nabla g_j(x_\star)\|^2 = \sigma_\star^2\frac{n-b}{b(n-1)}$ for each j; averaging over j proves the claim. □

3.2 Convergence Rate of Minibatch RR

Theorem 1 (MB-RR). *Under Assumptions 1–3, let P be any partition with $m \geq 2$ batches, and run MB-RR with constant $\eta < 2/\hat{L}_P$. Then*

$$\mathbb{E}[f(\bar{x}_K) - f_\star] \;\leq\; C\left[\frac{D^2}{(2-\eta\hat{L}_P)\,\eta\,mK} + \frac{\min\{1,\eta m\bar{L}\}\;\eta\,\sigma_P^2}{(2-\eta\hat{L}_P)^2}\right],$$

with the absolute constant C of Fact 1.

Proof. MB-RR performs exactly the iteration of Shuffling SGD under RR on the $N = m$ functions $g_1, \ldots, g_m$: both draw a uniform permutation of $[m]$ per epoch and take gradient steps on the permuted functions in order. By Lemma 1 these functions satisfy Assumptions 1–3 with $\ell_j = L_{B_j}$, $\bar{\ell} = \bar{L}$, $\hat{\ell} = \hat{L}_P$, $s^2 = \sigma_P^2$, and the same minimizer $x_\star$ and the same average iterate. Fact 1(a) with $N = m$ is the claim. □

Remark 1 (Sanity checks). For $b = 1$ we recover the single-sample rate of Liu [6]. The rate is stated per epoch, so it already accounts for the fact that an epoch has m steps. The variance term uses σ_P^2, which can be much smaller than $\sigma_\star^2$ (Lemma 2); the factor $\min\{1, \eta m\bar{L}\}$ shows that RR's advantage over SGD is active as long as $\eta \lesssim \frac{1}{m\bar{L}} = \frac{b}{n\bar{L}}$.

Corollary 1 (Dominance over minibatch SGD). *Under the conditions of Theorem 1, MB-SGD on the same batches with the same $\eta < 2/\hat{L}_P$ satisfies the bound in Fact 1(b) with $N = m$, $\hat{\ell} = \hat{L}_P$, $s^2 = \sigma_P^2$. Consequently the upper bound of Theorem 1 is at most a constant times the upper bound for MB-SGD for every such η and K, and is smaller by the factor $\min\{1, \eta m\bar{L}\}$ in the variance term.*

Proof. Both statements are instances of Fact 1(a) and (b) on the same m functions. The two bounds have the same optimization term up to the constants C and 2, and variance terms that differ by the factor $\min\{1, \eta m\bar{L}\} \leq 1$. □

3.3 Random Partitions and Tuned Rates

In practice the partition is itself random. Let P be uniform and independent of the epoch permutations. To avoid tuning η to the realized partition, we choose a stepsize $\eta < 2/\hat{L}$ (valid for every P by Lemma 1(2)). Because the right side of Theorem 1 is nondecreasing in $\hat{L}_P$ and $\hat{L}_P \leq \hat{L}$, we may replace $\hat{L}_P$ by $\hat{L}$, and the only random quantity left is σ_P^2, which enters linearly.

Theorem 2 (MB-RR with a random partition). *Under Assumptions 1–3, with P uniformly random and independent of the permutations, and $\eta < 2/\hat{L}$,*

$$\mathbb{E}[f(\bar{x}_K) - f_\star] \leq C\left[\frac{D^2}{(2-\eta\hat{L})\,\eta\,mK} + \frac{\min\{1,\eta m\bar{L}\}\,\eta\,\rho_b\sigma_\star^2}{b\,(2-\eta\hat{L})^2}\right], \qquad \rho_b \triangleq \frac{n-b}{n-1}.$$

Proof. Condition on P and apply Theorem 1, then relax $\hat{L}_P \leq \hat{L}$ (monotonicity), then take expectation over P using Lemma 2; the factor $\min\{1, \eta m\bar{L}\}$ does not depend on P because $\frac{1}{m}\sum_j L_{B_j} = \bar{L}$. □

Theorem 2 is Fact 1(a) with $N = m$ and s^2 replaced by $\rho_b\sigma_\star^2/b$. Optimizing over η exactly as in the proof of Corollary 1 of Liu [6] gives the following.

Corollary 2 (Tuned rate)**.** *With the stepsize optimized over $0 < \eta < 2/\hat{L}$, and using $mb = n$,*

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

If the stepsize is instead tuned to the realized partition (so $\hat{L}_P$ replaces $\hat{L}$ in the first term), the same bound holds with $b\hat{L}$ replaced by $b\,\mathbb{E}[\hat{L}_P]$, by Jensen's inequality applied to the concave maps $s \mapsto \sqrt{s}$ and $s \mapsto s^{1/3}$ in the two branches of the minimum.

3.4 What Batching Does and Does Not Change

Comparing Corollary 2 with the single-sample tuned rate $\frac{\hat{L}D^2}{nK} + \min\{\frac{\sigma_\star D}{\sqrt{nK}}, (\frac{\bar{L}\sigma_\star^2 D^4}{nK^2})^{1/3}\}$ of Liu [6], at the same number nK of gradient evaluations:

- **Variance-driven terms:** multiplied by $\sqrt{\rho_b}$, respectively $\rho_b^{1/3}$, where $\rho_b = \frac{n-b}{n-1} \le 1$. The factor $1/b$ in $\mathbb{E}[\sigma_P^2]$ is exactly cancelled by the b-fold reduction in the number of steps, so the familiar "variance reduction by batching" is already paid for by taking fewer steps. The finite-population correction ρ_b is a genuine gain but matters only when b is a constant fraction of n.
- **Optimization term:** $\frac{\hat{L}D^2}{nK}$ becomes $\frac{b\hat{L}_P D^2}{nK}$. Since $\hat{L}_P \le \min\{\hat{L}, m\bar{L}\}$, this is at most $\frac{\bar{L}D^2}{K}$ and is a loss when $b\hat{L}_P > \hat{L}$.
- **Smoothness heterogeneity can reverse the loss.** If one component dominates, say $L_1 = \hat{L} \approx n\bar{L}$, then a partition that places it in a batch gives $\hat{L}_P \le m\bar{L} = \frac{n}{b}\bar{L} \approx \hat{L}/b$, so $b\hat{L}_P \lesssim \hat{L}$ and the optimization term does not deteriorate. (Heavy-tailed per-sample smoothness is common, for instance in linear models with a few large-norm feature vectors.)
- **Wider RR regime:** the advantage over SGD in Theorem 1 is active for $\eta \lesssim \frac{b}{n\bar{L}}$, a range b times larger than for $b = 1$.

Example 1 (Uniform smoothness). If $L_i = L$ for all i, then $\hat{L}_P = \hat{L} = \bar{L} = L$ for every partition, and Corollary 2 reads $\frac{bLD^2}{nK} + \min\{\cdots\}$: batching strictly worsens the optimization term by the factor b while the variance terms are unchanged up to ρ_b. In this regime our bound recommends the smallest batch consistent with hardware efficiency, in agreement with the folklore that small batches are sample-efficient.

3.5 Quadratic Growth

Theorem 3 (Restarted MB-RR)**.** *Under Assumptions 1–4, let P be a partition with $m \ge 2$ batches, $\eta < 2/\hat{L}_P$, and run Restarted Shuffling SGD (Algorithm 2 of Liu 6) with MB-RR as the inner solver, $K \asymp \frac{1}{(2-\eta\hat{L}_P)\eta m\mu}$ epochs per stage, and T stages. Then*

$$\mathbb{E}[f(x_{T+1}) - f_\star] \lesssim \frac{\mu\ \mathrm{dist}^2(x_1, X_\star)}{2^T} + \frac{\min\{1, \eta m\bar{L}\}\,\eta\,\sigma_P^2}{(2-\eta\hat{L}_P)^2}.$$

Proof. Apply Fact 1(c) to $g_1, \ldots, g_m$; Assumption 4 concerns $f = \frac{1}{m}\sum g_j$ only and is unaffected by batching. □

Corollary 3 (Gradient-evaluation complexity)**.** *Given* $0 < \varepsilon < \mu\,\mathrm{dist}^2(x_1, X_\star)$ *sufficiently small, take* $T = \lceil \log_2 \frac{\mu\,\mathrm{dist}^2(x_1,X_\star)}{\varepsilon} \rceil$ *and* $\eta = \min\{\frac{1}{\hat{L}_P}, \max\{\frac{\varepsilon}{\sigma_P^2}, \frac{1}{\sigma_P}\sqrt{\frac{\varepsilon}{m\bar{L}}}\}\}$*. Then* $\mathbb{E}[f(x_{T+1}) - f_\star] \lesssim \varepsilon$ *using*

$$b\left(\frac{\hat{L}_P}{\mu} + \min\left\{\frac{\sigma_P^2}{\mu\varepsilon},\ \frac{\sigma_P}{\mu}\sqrt{\frac{m\bar{L}}{\varepsilon}}\right\}\right)\log\frac{\mu\,\mathrm{dist}^2(x_1, X_\star)}{\varepsilon}$$

individual gradient evaluations.

Proof. Identical to the proof of Corollary 2 of Liu [6] with $(n, \hat{L}, \sigma_\star)$ replaced by $(m, \hat{L}_P, \sigma_P)$, which counts evaluations of batch gradients. Each batch gradient costs b individual gradients, hence the leading factor b. □

Remark 2. For a random partition, $\mathbb{E}\sigma_P^2 = \rho_b\sigma_\star^2/b$, so $b\,\sigma_P^2/(\mu\varepsilon) \approx \rho_b\sigma_\star^2/(\mu\varepsilon)$ and $b\,\sigma_P\sqrt{m\bar{L}/\varepsilon}/\mu \approx \sqrt{\rho_b}\,\sigma_\star\sqrt{n\bar{L}/\varepsilon}/\mu$: the noise-dominated terms coincide with the single-sample ones up to ρ_b, and only the first term $b\hat{L}_P/\mu$ depends on how batching interacts with the smoothness profile, as in Section 3.4.

4 Conclusion and Open Problems

We have shown that, for a fixed partition, minibatch RR is Shuffling SGD on batch objectives, which transfers the stepsize-unrestricted dominance of RR over SGD to minibatches and exposes how the batch size enters. The exact finite-population variance identity shows that batching reduces the variance-driven terms only through the factor $\rho_b = \frac{n-b}{n-1}$ once the reduction in steps is accounted for; the optimization term pays a factor $b\hat{L}_P/\hat{L}$, which is favorable only under heterogeneous smoothness.

Several questions remain open.

1. **Fresh partitions.** Practical minibatch RR draws a new partition each epoch. Our reduction needs a fixed partition, because the batch objectives must be the same across epochs. Whether the same dominance holds for a fresh random partition per epoch, and whether it yields an improvement over a fixed partition, requires a new analysis.

2. **Lower bounds.** The lower bounds of Liu [6] apply to the class of all m-component objectives and therefore to the batch objectives, but a lower bound over objectives that are *induced* by a partition of n components, and that tracks b, is not available.

3. **Partition design.** Since σ_P^2 and $\hat{L}_P$ depend on the partition, one may ask how to choose a partition that minimizes the bound of Theorem 1, for instance by balancing batch smoothness.

4. **Last iterate and no restarts.** As in Liu [6], guarantees for the last iterate and for quadratic growth without restarts remain open.

AI Disclosure

The reduction, lemmas, and corollaries of this note were developed with the assistance of Claude (Anthropic) in drafting and checking derivations, based on the published results of Liu [6]. The author is responsible for the content, and all results that depend on external theorems are stated as such (Fact 1).

References

[1] A. Attia, M. Schliserman, U. Sherman, and T. Koren. Fast last-iterate convergence of SGD in the smooth interpolation regime. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[2] J. Cha, J. Lee, and C. Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pages 3855–3912, 2023.

[3] G. Garrigos and R. M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

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

[5] J. Haochen and S. Sra. Random shuffling beats SGD after finite epochs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 2624–2633, 2019.

[6] Z. Liu. Random reshuffling dominates stochastic gradient descent. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *PMLR*, pages 4859–4882, 2026.

[7] Z. Liu. Random reshuffling for smooth convex optimization: Dominates stochastic gradient descent. Extended version, 2026. (Add arXiv identifier.)

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

[9] D. Nagaraj, P. Jain, and P. Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 4703–4711, 2019.

[10] L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.