Random Reshuffling: Simple Analysis with Vast Improvements

Konstantin Mishchenko
KAUST
Thuwal, Saudi Arabia

Ahmed Khaled
Cairo University
Giza, Egypt

Peter Richtárik
KAUST
Thuwal, Saudi Arabia

Abstract

Random Reshuffling (RR) is an algorithm for minimizing finite-sum functions that utilizes iterative gradient descent steps in conjunction with data reshuffling. Often contrasted with its sibling Stochastic Gradient Descent (SGD), RR is usually faster in practice and enjoys significant popularity in convex and non-convex optimization. The convergence rate of RR has attracted substantial attention recently and, for strongly convex and smooth functions, it was shown to converge faster than SGD if 1) the stepsize is small, 2) the gradients are bounded, and 3) the number of epochs is large. We remove these 3 assumptions, improve the dependence on the condition number from κ^2 to κ (resp. from κ to $\sqrt{\kappa}$) and, in addition, show that RR has a different type of variance. We argue through theory and experiments that the new variance type gives an additional justification of the superior performance of RR. To go beyond strong convexity, we present several results for non-strongly convex and non-convex objectives. We show that in all cases, our theory improves upon existing literature. Finally, we prove fast convergence of the Shuffle-Once (SO) algorithm, which shuffles the data only once, at the beginning of the optimization process. Our theory for strongly convex objectives tightly matches the known lower bounds for both RR and SO and substantiates the common practical heuristic of shuffling once or only a few times. As a byproduct of our analysis, we also get new results for the Incremental Gradient algorithm (IG), which does not shuffle the data at all.

1 Introduction

We study the finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x)\right], \qquad (1)$$

where each $f_i:\mathbb{R}^d\to\mathbb{R}$ is differentiable and smooth, and are particularly interested in the big data machine learning setting where the number of functions n is large. Thanks to their scalability and low memory requirements, first-order methods are especially popular in this setting (Bottou et al., 2018). *Stochastic* first-order algorithms in particular have attracted a lot of attention in the machine learning community and are often used in combination with various practical heuristics. Explaining these heuristics may lead to further development of stable and efficient training algorithms. In this work, we aim at better and sharper theoretical explanation of one intriguingly simple but notoriously elusive heuristic: *data permutation/shuffling*.

1.1 Data permutation

In particular, the goal of our paper is to obtain deeper theoretical understanding of methods for solving (1) which rely on random or deterministic *permutation/shuffling* of the data $\{1,2,\ldots,n\}$ and

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

perform incremental gradient updates following the permuted order. We study three methods which belong to this class, described next.

An immensely popular but theoretically elusive method belonging to the class of data permutation methods is the **Random Reshuffling (RR)** algorithm (see Algorithm 1). This is the method we pay most attention to in this work, as reflected in the title. In each epoch t of RR, we sample indices $\pi_0, \pi_1, \ldots, \pi_{n-1}$ *without replacement* from $\{1, 2, \ldots, n\}$, i.e., $\{\pi_0, \pi_1, \ldots, \pi_{n-1}\}$ is a random permutation of the set $\{1, 2, \ldots, n\}$, and proceed with n iterates of the form

$$x_t^{i+1} = x_t^i - \gamma \nabla f_{\pi_i}(x_t^i),$$

where $\gamma > 0$ is a stepsize. We then set $x_{t+1} = x_t^n$, and repeat the process for a total of T epochs. Notice that in RR, a *new* permutation/shuffling is generated at the beginning of each epoch, which is why the term *re*shuffling is used.

Furthermore, we consider the **Shuffle-Once (SO)** algorithm, which is identical to RR with the exception that it shuffles the dataset only once—at the very beginning—and then reuses this random permutation in all subsequent epochs (see Algorithm 2). Our results for SO follow as corollaries of the tools we developed in order to conduct a sharp analysis of RR.

Finally, we also consider the **Incremental Gradient (IG)** algorithm, which is identical to SO, with the exception that the initial permutation is not random but deterministic. Hence, IG performs incremental gradient steps through the data in a *cycling* fashion. The ordering could be *arbitrary*, e.g., it could be selected *implicitly* by the ordering the data comes in, or chosen *adversarially*. Again, our results for IG follow as a byproduct of our efforts to understand RR.

Algorithm 1 Random Reshuffling (RR)

Input: Stepsize $\gamma > 0$, initial vector $x_0 = x_0^0 \in \mathbb{R}^d$, number of epochs T

```
for epochs t = 0, 1, ..., T − 1 do
    Sample a permutation π_0, π_1, ..., π_{n−1} of {1, 2, ..., n}
    for i = 0, 1, ..., n − 1 do
        x_t^{i+1} = x_t^i − γ∇f_{π_i}(x_t^i)
    x_{t+1} = x_t^n
```

Algorithm 2 Shuffle Once (SO)

Input: Stepsize $\gamma > 0$, initial vector $x_0 = x_0^0 \in \mathbb{R}^d$, number of epochs T

```
Sample a permutation π_0, π_1, ..., π_{n−1} of {1, 2, ..., n}
for epochs t = 0, 1, ..., T − 1 do
    for i = 0, 1, ..., n − 1 do
        x_t^{i+1} = x_t^i − γ∇f_{π_i}(x_t^i)
    x_{t+1} = x_t^n
```

1.2 Brief literature review

RR is usually contrasted with its better-studied sibling Stochastic Gradient Descent (SGD), in which each π_i is sampled uniformly *with replacement* from $\{1, 2, \ldots, n\}$. RR often converges faster than SGD on many practical problems (Bottou, 2009; Recht and Ré, 2013), is more friendly to cache locality (Bengio, 2012), and is in fact standard in deep learning (Sun, 2020).

The convergence properties of SGD are well-understood, with tightly matching lower and upper bounds in many settings (Rakhlin et al., 2012; Drori and Shamir, 2019; Nguyen et al., 2019). Sampling without replacement allows RR to leverage the finite-sum structure of (1) by ensuring that *each* function contributes to the solution once per epoch. On the other hand, it also introduces a significant complication: the steps are now *biased*. Indeed, in any iteration $i > 0$ within an epoch, we face the challenge of not having (conditionally) unbiased gradients since

$$\mathbb{E}\left[\nabla f_{\pi_i}(x_t^i) \mid x_t^i\right] \neq \nabla f(x_t^i).$$

This bias implies that individual iterations do not necessarily approximate a full gradient descent step. Hence, in order to obtain meaningful convergence rates for RR, it is necessary to resort to more involved proof techniques. In recent work, various convergence rates have been established for RR. However, a satisfactory, let alone complete, understanding of the algorithm's convergence remains elusive. For instance, the early line of attack pioneered by Recht and Ré (2012) seems to have hit the wall as their noncommutative arithmetic-geometric mean conjecture is not true (Lai and Lim, 2020). The situation is even more pronounced with the SO method, as Safran and Shamir (2020) point out that there are no convergence results specific for the method, and the only convergence rates for SO follow by applying the worst-case bounds of IG. Rajput et al. (2020) state that a common practical heuristic is to use methods like SO that do not reshuffle the data every epoch. Indeed, they add that *"current theoretical bounds are insufficient to explain this phenomenon, and a new theoretical breakthrough may be required to tackle it"*.

IG has a long history owing to its success in training neural networks (Luo, 1991; Grippo, 1994), and its asymptotic convergence has been established early (Mangasarian and Solodov, 1994; Bertsekas and Tsitsiklis, 2000). Several rates for non-smooth & smooth cases were established by Nedić and Bertsekas (2001); Li et al. (2019); Gürbüzbalaban et al. (2019a); Ying et al. (2019) and Nguyen et al. (2020). Using IG poses the challenge of choosing a specific permutation for cycling through the iterates, which Nedić and Bertsekas (2001) note to be difficult. Bertsekas (2011) gives an example that highlights the susceptibility of IG to bad orderings compared to RR. Yet, thanks to Gürbüzbalaban et al. (2019b) and Haochen and Sra (2019), RR is known to improve upon both SGD and IG for *twice*-smooth objectives. Nagaraj et al. (2019) also study convergence of RR for smooth objectives, and Safran and Shamir (2020); Rajput et al. (2020) give lower bounds for RR and related methods.

2 Contributions

In this work, we study the convergence behavior of the data-permutation methods RR, SO and IG. While existing proof techniques succeed in obtaining insightful bounds for RR and IG, they fail to fully capitalize on the intrinsic power reshuffling and shuffling offers, and are not applicable to SO at all[1]. Our proof techniques are dramatically novel, simple, more insightful, and lead to improved convergence results, all under weaker assumptions on the objectives than prior work.

2.1 New and improved convergence rates for RR, SO and IG

In Section 3, we analyze the RR and SO methods and present novel convergence rates for strongly convex, convex, and non-convex smooth objectives. Our results for RR are summarized in Table 1.

- **Strongly convex case.** If each f_i is strongly convex, we introduce a *new proof technique* for studying the convergence of RR/SO that allows us to obtain a *better dependence on problem constants*, such as the number of functions n and the condition number κ, compared to prior work (see Table 1). Key to our results is a *new notion of variance specific to RR/SO* (see Definition 2), which we argue explains the superior convergence of RR/SO compared to SGD in many practical scenarios. Our result for SO tightly *matches the lower bound* of Safran and Shamir (2020). We prove similar results in the more general setting when each f_i is convex and f is strongly convex (see Theorem 2), but in this case we are forced to use smaller stepsizes.
- **Convex case.** For convex but not necessarily strongly convex objectives f_i, we give the first result showing that *RR/SO can provably achieve better convergence than SGD* for a large enough number of iterations. This holds even when comparing against results that assume second-order smoothness, like the result of Haochen and Sra (2019).
- **Non-convex case.** For non-convex objectives f_i, we obtain for RR a *much better dependence on the number of functions* n compared to the prior work of Nguyen et al. (2020).

Furthermore, in the appendix we formulate and prove convergence results for IG for strongly convex objectives, convex, and non-convex objectives as well. The bounds are worse than RR by a factor of n in the noise/variance term, as IG does not benefit from randomization. Our result for strongly convex objectives *tightly matches the lower bound of Safran and Shamir (2020) up to an extra iteration and logarithmic factors, and is the first result to tightly match this lower bound.*

2.2 More general assumptions on the function class

Previous non-asymptotic convergence analyses of RR either obtain worse bounds that apply to IG, e.g., (Ying et al., 2019; Nguyen et al., 2020), or depend crucially on the assumption that each f_i is Lipschitz (Nagaraj et al., 2019; Haochen and Sra, 2019; Ahn and Sra, 2020). Unfortunately, requiring each f_i to be Lipschitz contradicts strong convexity (Nguyen et al., 2018) and is furthermore not satisfied in least square regression, matrix factorization, or for neural networks with smooth activations. In

[1]As we have mentioned before, the best known bounds for SO are those which apply to IG also, which means that the randomness inherent in SO is wholly ignored.

Table 1: Number of individual gradient evaluations needed by RR to reach an ε-accurate solution (defined in Section 3). Logarithmic factors and constants that are not related to the assumptions are ignored. For non-convex objectives, A and B are the constants given by Assumption 2.

Assumptions		μ-Strongly	Non-Strongly	Non-Convex	Citation
N.L.[(1)]	U.V.[(2)]	Convex	Convex		
✓	✓	$\kappa^2 n + \frac{\kappa n \sigma_*}{\mu\sqrt{\varepsilon}}$	–	–	Ying et al. (2019)
✗	✗	$\kappa^2 n + \frac{\kappa\sqrt{n}G}{\mu\sqrt{\varepsilon}}$	$\frac{LD^2}{\varepsilon} + \frac{G^2D^2}{\varepsilon^2}$ (3)	–	Nagaraj et al. (2019)
✗	✗	–	–	$\frac{Ln}{\varepsilon^2} + \frac{L^2 n G}{\varepsilon^3}$	Nguyen et al. (2020)
✓	✓	$\frac{\kappa^2 n}{\sqrt{\mu\varepsilon}} + \frac{\kappa^2 n \sigma_*}{\mu\sqrt{\varepsilon}}$ (4)	–	–	Nguyen et al. (2020)
✗	✗	$\frac{\kappa\alpha}{\varepsilon^{1/\alpha}} + \frac{\kappa\sqrt{n}G\alpha^{3/2}}{\mu\sqrt{\varepsilon}}$ (5)	–	–	Ahn and Sra (2020)
✓	✓	$\kappa n + \frac{\sqrt{n}}{\sqrt{\mu\varepsilon}} + \frac{\kappa\sqrt{n}G_0}{\mu\sqrt{\varepsilon}}$ (6)	–	–	Ahn et al. (2020)
✓	✓	$\kappa + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}$ (7) $\kappa n + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}$	$\frac{Ln}{\varepsilon} + \frac{\sqrt{Ln}\sigma_*}{\varepsilon^{3/2}}$	$\frac{Ln}{\varepsilon^2} + \frac{L\sqrt{n}(B+\sqrt{A})}{\varepsilon^3}$	This work

(1) Support for non-Lipschitz functions (N.L.): proofs without assuming that $\max_{i=1,\ldots,n} \|\nabla f_i(x)\| \le G$ for all $x \in \mathbb{R}^d$ and some $G > 0$. Note that $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_*)\|^2 \overset{\text{def}}{=} \sigma_*^2 \le G^2$ and $B^2 \le G^2$.

(2) Unbounded variance (U.V.): there may be no constant σ such that Assumption 2 holds with $A = 0$ and $B = \sigma$. Note that when the individual gradients are bounded, the variance is automatically bounded too.

(3) Nagaraj et al. (2019) require, for non-strongly convex functions, projecting at each iteration onto a bounded convex set of diameter D. We study the unconstrained problem.

(4) For strongly convex, Nguyen et al. (2020) bound $f(x) - f(x_*)$ rather than squared distances, hence we use strong convexity to translate their bound into a bound on $\|x - x_*\|^2$.

(5) The constant $\alpha > 2$ is a parameter to be specified in the stepsize used by (Ahn and Sra, 2020). Their full bound has several extra terms but we include only the most relevant ones.

(6) The result of Ahn et al. (2020) holds when f satisfies the Polyak-Łojasiewicz inequality, a generalization of strong convexity. We nevertheless specialize it to strong convexity for our comparison. The constant G_0 is defined as $G_0 \overset{\text{def}}{=} \sup_{x: f(x) \le f(x_0)} \max_{i \in [n]} \|\nabla f_i(x)\|$. Note that $\sigma_* \le G_0$.

(7) This result is the first to show that RR and SO work with any $\gamma \le \frac{1}{L}$, but it asks for each f_i to be strongly convex. The second result assumes that only f is strongly convex.

contrast, our work is the first to show how to leverage randomization to obtain better rates for RR without assuming each f_i to be Lipschitz. In concurrent work, Ahn et al. (2020) also obtain a result for non-convex objectives satisfying the Polyak-Łojasiewicz inequality, a generalization of strong convexity. Their result holds without assuming bounded gradients or bounded variance, but unfortunately with a worse dependence on κ and n when specialized to μ-strongly convex functions.

- **Strongly convex and convex case.** For strongly convex and convex objectives *we do not require any assumptions on the functions used beyond smoothness and convexity.*
- **Non-convex case.** For non-convex objectives we obtain our results under a significantly more general assumption than the bounded gradients assumptions employed in prior work. Our assumption is also provably satisfied when each function f_i is lower bounded, and hence is *not only more general but also a more realistic assumption to use.*

3 Convergence theory

We will derive results for strongly convex, convex as well as non-convex objectives. To compare between the performance of first-order methods, we define an ε-accurate solution as a point $\tilde{x} \in \mathbb{R}^d$ that satisfies (in expectation if $\tilde{x}$ is random)

$$\|\nabla f(\tilde{x})\| \le \varepsilon, \qquad \text{or} \qquad \|\tilde{x} - x_*\|^2 \le \varepsilon, \qquad \text{or} \qquad f(\tilde{x}) - f(x_*) \le \varepsilon$$

for non-convex, strongly convex, and non-strongly convex objectives, respectively, and where x_* is assumed to be a minimizer of f if f is convex. We then measure the performance of first-order methods by the number of individual gradients $\nabla f_i(\cdot)$ they access to reach an ε-accurate solution.

Our first assumption is that the objective is bounded from below and smooth. This assumption is used in all of our results and is widely used in the literature.

Assumption 1. The objective f and the individual losses $f_1, \ldots, f_n$ are all L-smooth, i.e., their gradients are L-Lipschitz. Further, f is lower bounded by some $f_* \in \mathbb{R}$. If f is convex, we also assume the existence of a minimizer $x_* \in \mathbb{R}^d$.

Assumption 1 is necessary in order to obtain better convergence rates for RR compared to SGD, since without smoothness the SGD rate is optimal and cannot be improved (Nagaraj et al., 2019). The following quantity is key to our analysis and serves as an asymmetric distance between two points measured in terms of functions.

Definition 1. For any i, the quantity $D_{f_i}(x, y) \overset{\text{def}}{=} f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle$ is the *Bregman divergence* between x and y associated with f_i.

It is well-known that if f_i is L-smooth and μ-strongly convex, then for all $x, y \in \mathbb{R}^d$

$$\frac{\mu}{2}\|x - y\|^2 \leq D_{f_i}(x, y) \leq \frac{L}{2}\|x - y\|^2, \tag{2}$$

so each Bregman divergence is closely related to the Euclidian distance. Moreover, the difference between the gradients of a convex and L-smooth f_i is related to its Bregman divergence by

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L \cdot D_{f_i}(x, y). \tag{3}$$

3.1 Main result: strongly convex objectives

Before we proceed to the formal statement of our main result, we need to present the central finding of our work. The analysis of many stochastic methods, including SGD, rely on the fact that the iterates converge to x_* up to some noise. This is exactly where we part ways with the standard analysis techniques, since, it turns out, the intermediate iterates of shuffling algorithms converge to some other points. Given a permutation π, the real limit points are defined below,

$$x_*^i \overset{\text{def}}{=} x_* - \gamma \sum_{j=0}^{i-1} \nabla f_{\pi_j}(x_*), \qquad i = 1, \ldots, n-1. \tag{4}$$

In fact, it is predicted by our theory and later validated by our experiments that within an epoch the iterates *go away* from x_*, and closer to the end of the epoch they make a sudden comeback to x_*.

The second reason the vectors introduced in Equation (4) are so pivotal is that they allow us to define a new notion of variance. Without it, there seems to be no explanation for why RR sometimes overtakes SGD from the very beginning of optimization process. We define it below.

Definition 2 (Shuffling variance)**.** Given a stepsize $\gamma > 0$ and a random permutation π of $\{1, 2, \ldots, n\}$, define x_*^i as in (4). Then, the shuffling variance is given by

$$\sigma^2_{\text{Shuffle}} \overset{\text{def}}{=} \max_{i=1,\ldots,n-1} \left[\frac{1}{\gamma} \mathbb{E}\left[D_{f_{\pi_i}}(x_*^i, x_*) \right] \right], \tag{5}$$

where the expectation is taken with respect to the randomness in the permutation π.

Naturally, $\sigma^2_{\text{Shuffle}}$ depends on the functions $f_1, \ldots, f_n$, but, unlike SGD, it also depends in a non-trivial manner on the stepsize γ. The easiest way to understand the new notation is to compare it to the standard definition of variance used in the analysis of SGD. We argue that $\sigma^2_{\text{Shuffle}}$ is the natural counter-part for the standard variance used in SGD. We relate both of them by the following upper and lower bounds:

Proposition 1. Suppose that each of $f_1, f_2, \ldots, f_n$ is μ-strongly convex and L-smooth. Then $\frac{\gamma \mu n}{8} \sigma_*^2 \leq \sigma^2_{\text{Shuffle}} \leq \frac{\gamma L n}{4} \sigma_*^2$, where $\sigma_*^2 \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x_*)\|^2$.

In practice, $\sigma^2_{\text{Shuffle}}$ may be much closer to the lower bound than the upper bound; see Section 4. This leads to a dramatic difference in performance and provides additional evidence of the superiority of RR over SGD. The next theorem states how exactly convergence of RR depends on the introduced variance.

Theorem 1. Suppose that the functions $f_1, \ldots, f_n$ are μ-strongly convex and that Assumption 1 holds. Then for Algorithms 1 or 2 run with a constant stepsize $\gamma \le \frac{1}{L}$, the iterates generated by either of the algorithms satisfy

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \le (1-\gamma\mu)^{nT} \|x_0 - x_*\|^2 + \frac{2\gamma\sigma^2_{\text{Shuffle}}}{\mu}.$$

Proof. The key insight of our proof is that the intermediate iterates $x_t^1, x_t^2, \ldots$ do not converge to x_*, but rather converge to the sequence $x_*^1, x_*^2, \ldots$ defined by (4). Keeping this intuition in mind, it makes sense to study the following recursion:

$$\begin{aligned}&\mathbb{E}\left[\|x_t^{i+1} - x_*^{i+1}\|^2\right] \\ &= \mathbb{E}\left[\|x_t^i - x_*^i\|^2 - 2\gamma\langle \nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*), x_t^i - x_*^i\rangle + \gamma^2\|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)\|^2\right]. \end{aligned} \tag{6}$$

Once we have this recursion, it is useful to notice that the scalar product can be decomposed as

$$\begin{aligned}\langle \nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*), x_t^i - x_*^i\rangle &= [f_{\pi_i}(x_*^i) - f_{\pi_i}(x_t^i) - \langle \nabla f_{\pi_i}(x_t^i), x_*^i - x_t^i\rangle] \\ &\quad + [f_{\pi_i}(x_t^i) - f_{\pi_i}(x_*) - \langle \nabla f_{\pi_i}(x_*), x_t^i - x_*\rangle] \\ &\quad - [f_{\pi_i}(x_*^i) - f_{\pi_i}(x_*) - \langle \nabla f_{\pi_i}(x_*), x_*^i - x_*\rangle] \\ &= D_{f_{\pi_i}}(x_*^i, x_t^i) + D_{f_{\pi_i}}(x_t^i, x_*) - D_{f_{\pi_i}}(x_*^i, x_*). \end{aligned} \tag{7}$$

This decomposition is, in fact, very standard and is a special case of the so-called *three-point identity* (Chen and Teboulle, 1993). So, it should not be surprising that we use it.

The rest of the proof relies on obtaining appropriate bounds for the terms in the recursion. First, we need to lower bound the inner product, which we do by separately bounding each of the three Bregman divergence terms appearing in (7). Since we assume each function f_i to be μ-strongly convex, the first term in (7) can be bounded as

$$\frac{\mu}{2}\|x_t^i - x_*^i\|^2 \overset{(2)}{\le} D_{f_{\pi_i}}(x_*^i, x_t^i),$$

which will get combined with the first term in the expansion (6) to obtain contraction. The second term in (7) can be bounded via

$$\frac{1}{2L}\|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)\|^2 \overset{(3)}{\le} D_{f_{\pi_i}}(x_t^i, x_*),$$

which gets absorbed in the last term in the expansion of $\|x_t^{i+1} - x_*^{i+1}\|^2$. The expectation of the third divergence term in (7) is trivially bounded as follows:

$$\mathbb{E}\left[D_{f_{\pi_i}}(x_*^i, x_*)\right] \le \max_{i=1,\ldots,n-1}\left[\mathbb{E}\left[D_{f_{\pi_i}}(x_*^i, x_*)\right]\right] = \gamma\sigma^2_{\text{Shuffle}}.$$

Plugging these three bounds back into (7), and the resulting inequality into (6), we obtain

$$\begin{aligned}\mathbb{E}\left[\|x_t^{i+1} - x_*^{i+1}\|^2\right] &\le \mathbb{E}\left[(1-\gamma\mu)\|x_t^i - x_*^i\|^2 - 2\gamma(1-\gamma L)D_{f_{\pi_i}}(x_t^i, x_*)\right] + 2\gamma^2\sigma^2_{\text{Shuffle}} \\ &\le (1-\gamma\mu)\mathbb{E}\left[\|x_t^i - x_*^i\|^2\right] + 2\gamma^2\sigma^2_{\text{Shuffle}}. \end{aligned} \tag{8}$$

The rest of the proof is just solving this recursion, and is relegated to Section 8.2 in the appendix. ∎

We show (Corollary 1 in the appendix) that by carefully controlling the stepsize, the final iterate of RR after T epochs satisfies

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu n T}{L}\right)\|x_0 - x_*\|^2 + \frac{\kappa\sigma_*^2}{\mu^2 n T^2}\right), \tag{9}$$

where the $\tilde{\mathcal{O}}(\cdot)$ notation suppresses absolute constants and polylogarithmic factors. Note that Theorem 1 covers both RR and SO, and for SO, Safran and Shamir (2020) give almost the same *lower* bound. Stated in terms of the squared distance from the optimum, their lower bound is

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \Omega\left(\min\left\{1, \frac{\sigma_*^2}{\mu^2 n T^2}\right\}\right),$$

where we note that in their problem $\kappa = 1$. This translates to sample complexity $\mathcal{O}\left(\sqrt{n}\sigma_*/(\mu\sqrt{\varepsilon})\right)$ for $\varepsilon \le 1$[2]. Specializing $\kappa = 1$ in Equation (9) gives the sample complexity of $\tilde{\mathcal{O}}\left(1 + \sqrt{n}\sigma_*/(\mu\sqrt{\varepsilon})\right)$, matching the optimal rate up to an extra iteration. More recently, Rajput et al. (2020) also proved a similar lower bound for RR. We emphasize that Theorem 1 is not only tight, but it is also the first convergence bound that applies to SO. Moreover, it also immediately works if one permutes once every few epochs, which interpolates between RR and SO mentioned by Rajput et al. (2020).

Comparison with SGD To understand when RR is better than SGD, let us borrow a convergence bound for the latter. Several works have shown (e.g., see (Needell et al., 2014; Stich, 2019)) that for any $\gamma \le \frac{1}{2L}$ the iterates of SGD satisfy

$$\mathbb{E}\left[\left\|x_{nT}^{\mathrm{SGD}} - x_*\right\|^2\right] \le (1-\gamma\mu)^{nT}\left\|x_0 - x_*\right\|^2 + \frac{2\gamma\sigma_*^2}{\mu}.$$

Thus, the question as to which method will be faster boils down to which variance is smaller: $\sigma_{\mathrm{Shuffle}}^2$ or σ_*^2. According to Proposition 1, it depends on both n and the stepsize. Once the stepsize is sufficiently small, $\sigma_{\mathrm{Shuffle}}^2$ becomes smaller than σ_*^2, but this might not be true in general. Similarly, if we partition n functions into $\frac{n}{\tau}$ groups, i.e., use minibatches of size τ, then σ_*^2 decreases as $\mathcal{O}\left(\frac{1}{\tau}\right)$ and $\sigma_{\mathrm{Shuffle}}^2$ as $\mathcal{O}\left(\frac{1}{\tau^2}\right)$, so RR can become faster even without decreasing the stepsize. We illustrate this later with numerical experiments.

While Theorem 1 requires each f_i to be strongly convex, we can also obtain results in the case where the individual strong convexity assumption is replaced by convexity. However, in such a case, we need to use a smaller stepsize, as the next theorem shows.

Theorem 2. Suppose that each f_i is convex, f is μ-strongly convex, and Assumption 1 holds. Then provided the stepsize satisfies $\gamma \le \frac{1}{\sqrt{2}Ln}$ the final iterate generated by Algorithms 1 or 2 satisfies

$$\mathbb{E}\left[\left\|x_T - x_*\right\|^2\right] \le \left(1 - \frac{\gamma\mu n}{2}\right)^T \left\|x_0 - x_*\right\|^2 + \gamma^2\kappa n\sigma_*^2.$$

It is not difficult to show that by properly choosing the stepsize γ, the guarantee given by Theorem 2 translates to a sample complexity of $\tilde{\mathcal{O}}\left(\kappa n + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}\right)$, which matches the dependence on the accuracy ε in Theorem 1 but with $\kappa(n-1)$ additional iterations in the beginning. For $\kappa = 1$, this translates to a sample complexity of $\tilde{\mathcal{O}}\left(n + \frac{\sqrt{n}\sigma_*}{\mu\sqrt{\varepsilon}}\right)$ which is worse than the lower bound of Safran and Shamir (2020) when ε is large. In concurrent work, Ahn et al. (2020) obtain in the same setting a complexity of $\tilde{\mathcal{O}}\left(1/\varepsilon^{1/\alpha} + \frac{\sqrt{n}G}{\mu\sqrt{\varepsilon}}\right)$ (for a constant $\alpha > 2$), which requires that each f_i is Lipschitz and matches the lower bound only when the accuracy ε is large enough that $1/\varepsilon^{1/\alpha} \le 1$. Obtaining an optimal convergence guarantee for all accuracies ε in the setting of Theorem 2 remains open.

3.2 Non-strongly convex objectives

We also make a step towards better bounds for RR/SO without any strong convexity at all and provide the following convergence statement.

Theorem 3. Let functions $f_1, f_2, \ldots, f_n$ be convex. Suppose that Assumption 1 holds. Then for Algorithm 1 or Algorithm 2 run with a stepsize $\gamma \le \frac{1}{\sqrt{2}Ln}$, the average iterate $\hat{x}_T \stackrel{\text{def}}{=} \frac{1}{T}\sum_{j=1}^T x_j$ satisfies

$$\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] \le \frac{\|x_0 - x_*\|^2}{2\gamma nT} + \frac{\gamma^2 Ln\sigma_*^2}{4}.$$

Unfortunately, the theorem above relies on small stepsizes, but we still deem it as a valuable contribution, since it is based on a novel analysis. Indeed, the prior works showed that RR approximates a full gradient step, but we show that it is even closer to the implicit gradient step, see the appendix.

To translate the recursion in Theorem 3 to a complexity, one can choose a small stepsize and obtain (Corollary 2 in the appendix) the following bound for RR/SO:

$$\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] = \mathcal{O}\left(\frac{L\|x_0 - x_*\|^2}{T} + \frac{L^{1/3}\|x_0 - x_*\|^{4/3}\sigma_*^{2/3}}{n^{1/3}T^{2/3}}\right).$$

[2] In their problem, the initialization point x_0 satisfies $\|x_0 - x_*\|^2 \le 1$ and hence asking for accuracy $\varepsilon > 1$ does not make sense.

Stich (2019) gives a convergence upper bound of $\mathcal{O}\left(\frac{L\|x_0-x_*\|^2}{nT}+\frac{\sigma_*\|x_0-x_*\|}{\sqrt{nT}}\right)$ for SGD. Comparing upper bounds, we see that RR/SO beats SGD when the number of epochs satisfies $T \geq \frac{L^2\|x_0-x_*\|^2 n}{\sigma_*^2}$. To the best of our knowledge, there are no strict lower bounds in this setting. Safran and Shamir (2020) suggest a lower bound of $\Omega\left(\sigma_*\left(\frac{1}{\sqrt{nT^3}}+\frac{1}{nT}\right)\right)$ by setting μ to be small in their lower bound for μ-strongly convex functions, however this means the functions used are not strictly weakly convex, and hence we think this lower bound may be too optimistic.

3.3 Non-convex objectives

For non-convex objectives, we formulate the following assumption on the gradients variance.

Assumption 2. There exists nonnegative constants $A, B \geq 0$ such that for any $x \in \mathbb{R}^d$ we have,

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)-\nabla f(x)\|^2 \leq 2A\left(f(x)-f(x_*)\right)+B^2. \tag{10}$$

Assumption 2 is quite general: if there exists some $G > 0$ such that $\|\nabla f_i(x)\| \leq G$ for all $x \in \mathbb{R}^d$ and $i \in \{1, 2, \ldots, n\}$, then Assumption 2 is clearly satisfied by setting $A = 0$ and $B = G$. Assumption 2 also generalizes the uniformly bounded variance assumption commonly invoked in work on non-convex SGD, which requires the existence of $\sigma^2 \geq 0$ such that the inequality $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)-\nabla f(x)\|^2 \leq \sigma^2$ holds for all $x \in \mathbb{R}^d$. Assumption 2 is a special case of the Expected Smoothness assumption of Khaled and Richtárik (2020), and it holds whenever each f_i is smooth and lower-bounded, as the next proposition shows.

Proposition 2. (Khaled and Richtárik, 2020, Special case of Proposition 3) Suppose that $f_1, f_2, \ldots, f_n$ are lower bounded by $f_1^*, f_2^*, \ldots, f_n^*$ respectively and that Assumption 1 holds. Then there exists constants $A, B \geq 0$ such that Assumption 2 holds.

We now give our main convergence theorem for RR without assuming convexity.

Theorem 4. Suppose that Assumptions 1 and 2 hold. Then for Algorithm 1 run for T epochs with a stepsize $\gamma \leq \min\left(\frac{1}{2Ln}, \frac{1}{(AL^2n^2T)^{1/3}}\right)$ we have

$$\min_{t=0,\ldots,T-1}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \frac{12(f(x_0)-f_*)}{\gamma nT}+2\gamma^2L^2nB^2.$$

Comparison with SGD. From Theorem 4, one can recover the complexity that we provide in Table 1, see Corollary 3 in the appendix. Ignore some constants not related to our assumptions and specialize to uniformly bounded variance. Then, the sample complexity of RR, $K_{\mathrm{RR}} \geq \frac{L\sqrt{n}}{\varepsilon^2}(\sqrt{n}+\frac{\sigma}{\varepsilon})$, becomes better than that of SGD, $K_{\mathrm{SGD}} \geq \frac{L}{\varepsilon^2}(1+\frac{\sigma^2}{\varepsilon^2})$, whenever $\sqrt{n}\varepsilon \leq \sigma$.

4 Experiments

We run our experiments on the ℓ_2-regularized logistic regression problem given by

$$\frac{1}{N}\sum_{i=1}^{N}\left(-\left(b_i\log\left(h(a_i^\top x)\right)+(1-b_i)\log\left(1-h(a_i^\top x)\right)\right)\right)+\frac{\lambda}{2}\|x\|^2,$$

where $(a_i, b_i) \in \mathbb{R}^d \times \{0, 1\}$, $i = 1, \ldots, N$ are the data samples and $h\colon t \to 1/(1+e^{-t})$ is the sigmoid function. For better parallelism, we use minibatches of size 512 for all methods and datasets. We set $\lambda = \frac{L}{\sqrt{N}}$ and use stepsizes decreasing as $\mathcal{O}(\frac{1}{t})$. See the appendix for more details on the parameters used, implementation details, and reproducibility.

Reproducibility. Our code is provided at https://github.com/konstmish/random_reshuffling. All used datasets are publicly available and all additional implementation details are provided in the appendix.

Observations. One notable property of all shuffling methods is that they converge with oscillations, as can be seen in Figure 1. There is nothing surprising about this as the proof of our Theorem 1 shows that the intermediate iterates converge to x_*^i instead of x_*. It is, however, surprising how striking the difference between the intermediate iterates within one epoch can be.

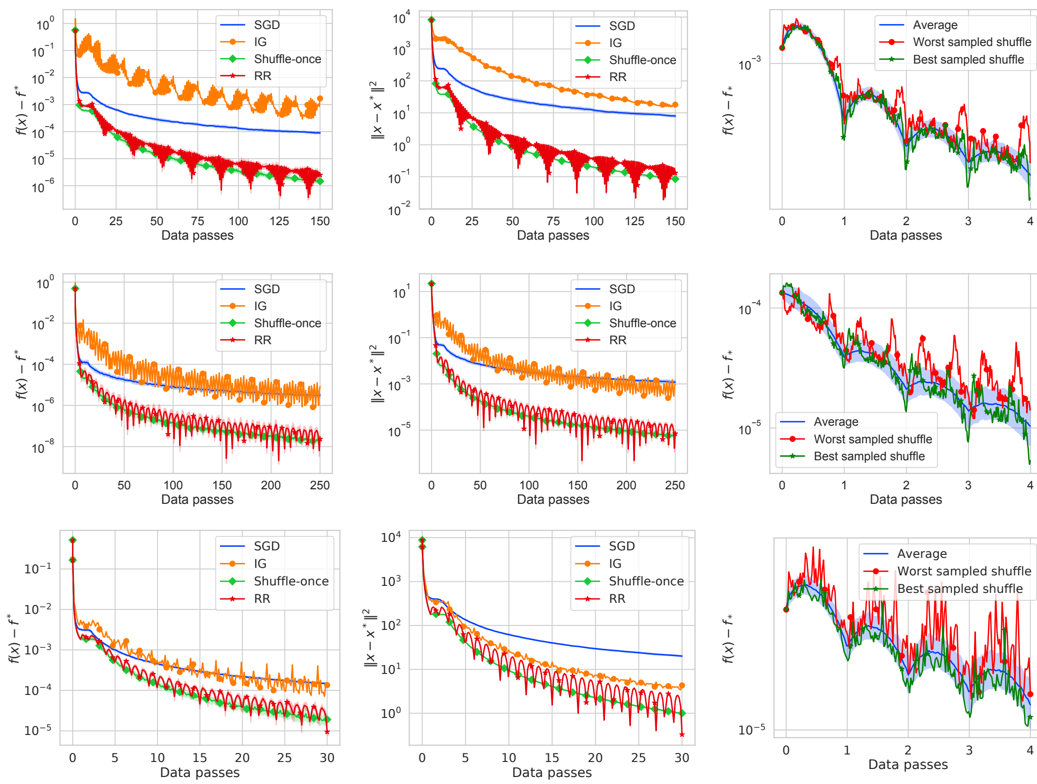

Figure 1: Top: `real-sim` dataset ($N = 72,309$; $d = 20,958$), middle row: `w8a` dataset ($N = 49,749$; $d = 300$), bottom: `RCV1` dataset ($N = 804,414$; $d = 47,236$). Left: convergence of $f(x_t^i)$, middle column: convergence of $\|x_t^i - x_*\|^2$, right: convergence of SO with different permutations .

Next, one can see that SO and RR converge almost exactly the same way, which is in line with Theorem 1. On the other hand, the contrast with IG is dramatic, even though we never managed to sample even a single permutation leading to similarly slow convergence. This suggests that perhaps the probability of getting such a permutation is negligible; see also the right plot in Figure 2.

Finally, we remark that the first two plots in Figure 2 demonstrate the importance of the new variance introduced in Definition 2. The upper and lower bounds from Proposition 1 are depicted in these two plots and one can observe that the lower bound is often closer to the actual value of $\sigma^2_{\text{Shuffle}}$ than the upper bound. And the fact that $\sigma^2_{\text{Shuffle}}$ very quickly becomes smaller than σ^2_* explains why RR often outperforms SGD starting from early iterations.

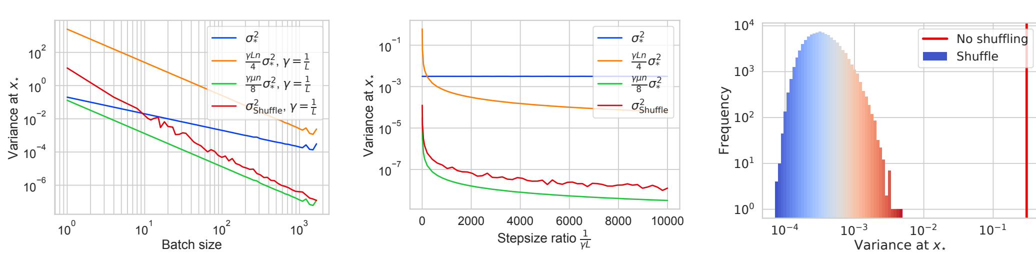

Figure 2: Estimated variance at the optimum, $\sigma^2_{\text{Shuffle}}$ and σ^2_*, for the `w8a` dataset. Left: the values of variance for different minibatch sizes with $\gamma = {}^1/L$. Middle: variance with fixed minibatch size 64 for different γ, starting with $\gamma = {}^1/L$ and ending with $\gamma = {}^{10^{-4}}/L$. Right: the empirical distribution of $\sigma^2_{\text{Shuffle}}$ for $500,000$ sampled permutations with $\gamma = {}^1/L$ and minibatch size 64.

Broader Impact

Our contribution is primarily theoretical. Moreover, we study methods that are already in use in practice, but are notoriously hard to analyze. We believe we have made a breakthrough in this area by developing new and remarkably simple proof techniques, leading to sharp bounds. This, we hope, will inspire other researchers to apply and further develop our techniques to other contexts and algorithms. These applications may one day push the state of the art in practice for existing or new supervised machine learning applications, which may then have broader impacts. Besides this, we do not expect any direct or short term societal consequences.

Acknowledgments and Disclosure of Funding

Ahmed Khaled acknowledges internship support from the Optimization and Machine Learning Lab led by Peter Richtárik at KAUST.

References

Kwangjun Ahn and Suvrit Sra. On Tight Convergence Rates of Without-replacement SGD. *arXiv preprint arXiv:2004.08657*, 2020. 3, 4

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: optimal rates without component convexity and large epoch requirements. *arXiv preprint arXiv:2006.06946. To appear in NeurIPS 2020*, 2020. 4, 7

Yoshua Bengio. Practical Recommendations for Gradient-Based Training of Deep Architectures. *Neural Networks: Tricks of the Trade*, page 437–478, 2012. ISSN 1611-3349. doi: 10.1007/978-3-642-35289-8_26. 2

Dimitri P. Bertsekas. Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey. In Suvrit Sra, Sebastan Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, chapter 4. The MIT Press, 2011. ISBN 9780262298773. 3

Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient Convergence in Gradient methods with Errors. *SIAM Journal on Optimization*, 10(3):627–642, January 2000. doi: 10.1137/s1052623497331063. 3

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. Unpublished open problem offered to the attendance of the SLDS 2009 conference, 2009. URL http://leon.bottou.org/papers/bottou-slds-open-problem-2009. 2, 15

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. 1

Gong Chen and Marc Teboulle. Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993. doi: 10.1137/0803026. 6

Yoel Drori and Ohad Shamir. The Complexity of Finding Stationary Points with Stochastic Gradient Descent. *arXiv preprint arXiv:1910.01845*, 2019. 2

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General Analysis and Improved Rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 15

Luigi Grippo. A class of unconstrained minimization methods for neural network training. *Optimization Methods and Software*, 4(2):135–150, January 1994. doi: 10.1080/10556789408805583. 3

Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. Convergence Rate of Incremental Gradient and Incremental Newton Methods. *SIAM Journal on Optimization*, 29(4):2542–2565, 2019a. doi: 10.1137/17M1147846. 3

Mert Gürbüzbalaban, Asuman Özdağlar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, Oct 2019b. ISSN 1436-4646. doi: 10.1007/s10107-019-01440-w. 3

Jeff Haochen and Suvrit Sra. Random Shuffling Beats SGD after Finite Epochs. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2624–2633, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 3

Ahmed Khaled and Peter Richtárik. Better Theory for SGD in the Nonconvex World. *arXiv preprint arXiv:2002.03329*, 2020. 8, 25, 27

Zehua Lai and Lek-Heng Lim. Recht-Ré Noncommutative Arithmetic-Geometric Mean Conjecture is False. *arXiv preprint arXiv:2006.01510*, 2020. 2

Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Jason D Lee. Incremental Methods for Weakly Convex Optimization. *arXiv preprint arXiv:1907.11687*, 2019. 3

Zhi-Quan Luo. On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks. *Neural Computation*, 3(2):226–245, June 1991. doi: 10.1162/neco.1991.3.2.226. 3

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019. 15

Olvi L. Mangasarian and Mikhail V. Solodov. Serial and parallel backpropagation convergence via nonmonotone perturbed minimization. *Optimization Methods and Software*, 4(2):103–116, 1994. doi: 10.1080/10556789408805581. 3

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without Replacement: Sharper Rates for General Smooth Convex Functions. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4703–4711, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 3, 4, 5

Angelia Nedić and Dimitri P. Bertsekas. Incremental Subgradient Methods for Nondifferentiable Optimization. *SIAM Journal on Optimization*, 12(1):109–138, January 2001. doi: 10.1137/s1052623499362111. 3

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1017–1025. Curran Associates, Inc., 2014. 7

Lam Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takáč. SGD and Hogwild! Convergence Without the Bounded Gradients Assumption. volume 80 of *Proceedings of Machine Learning Research*, pages 3750–3758, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 3

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. *arXiv preprint arXiv:2002.08246*, 2020. 3, 4, 25

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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3660–3669. Curran Associates, Inc., 2019. 2

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. *arXiv preprint arXiv:2002.10400*, 2020. 3, 7

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851. 2

Benjamin Recht and Christopher Ré. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. In S. Mannor, N. Srebro, and R. C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, page 11.1–11.24, 2012. Edinburgh, Scotland. 2

Benjamin Recht and Christopher Ré. Parallel Stochastic Gradient Algorithms for Large-Scale Matrix Completion. *Mathematical Programming Computation*, 5(2):201–226, April 2013. doi: 10.1007/s12532-013-0053-8. 2

Itay Safran and Ohad Shamir. How Good is SGD with Random Shuffling? In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of the 33rd Annual Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3250–3284. PMLR, 09–12 Jul 2020. 2, 3, 6, 7, 8, 30

Sebastian U. Stich. Unified Optimal Analysis of the (Stochastic) Gradient Method. *arXiv preprint arXiv:1907.04232*, 2019. 7, 8, 15

Ruo-Yu Sun. Optimization for Deep Learning: An Overview. *Journal of the Operations Research Society of China*, 8(2):249–294, Jun 2020. ISSN 2194-6698. doi: 10.1007/s40305-020-00309-6. 2

Bicheng Ying, Kun Yuan, Stefan Vlaski, and Ali H. Sayed. Stochastic Learning Under Random Reshuffling With Constant Step-Sizes. In *IEEE Transactions on Signal Processing*, volume 67, pages 474–489, 2019. doi: 10.1109/TSP.2018.2878551. 3, 4

Appendix

Contents

5 Additional experiment details

Objective properties. To better correspond to the theoretical setting of our main result, we use ℓ_2 regularization in every element of the finite-sum. To obtain minibatches for the RR, SO and IG we permute the dataset and then split it into $n = \lceil \frac{N}{\tau} \rceil$ groups of sizes $\tau, \ldots, \tau, N - \tau \left(\lceil \frac{N}{\tau} \rceil - 1 \right)$. In other words, the first $n-1$ groups are of size τ and the remaining samples go to the last group. For SO and IG, we split the data only once, and for RR, we do this at the beginning of each epoch. The permutation of samples used in IG is the one in which the datasets are stored online. The smoothness constant of the sum of logistic regression losses admits a closed form expression $L_f = \frac{1}{4N}\|A\|^2 + \lambda$. The individual losses are $L_{\max}$-smooth with $L_{\max} = \max_{i=1,\ldots,n} \|a_i\|^2 + \lambda$.

Stepsizes. For all methods in Figure 1, we keep the stepsize equal to $\frac{1}{L}$ for the first $k_0 = \lfloor K/40 \rfloor$ iterations, where K is the total number of stochastic steps. This is important to ensure that there is an exponential convergence before the methods reach their convergence neighborhoods (Stich, 2019). After the initial k_0 iterations, the stepsizes used for RR, SO and IG were chosen as $\gamma_k = \min\left\{\frac{1}{L}, \frac{3}{\mu \max\{1, k-k_0\}}\right\}$ and for SGD as $\gamma_k = \min\left\{\frac{1}{L}, \frac{2}{\mu \max\{1, k-k_0\}}\right\}$. Although these stepsizes for RR are commonly used in practice (Bottou, 2009), we do not analyze them and leave decreasing-stepsize analysis for future work. We also note that although L is generally not available, it can be estimated using empirically observed gradients (Malitsky and Mishchenko, 2019). For our experiments, we estimate L of minibatches of size τ using the closed-form expressions from Proposition 3.8 in (Gower et al., 2019) as $L \le \frac{n(\tau-1)}{\tau(n-1)} L_f + \frac{n-\tau}{\tau(n-1)} L_{\max}$. The confidence intervals in Figure 1 are estimated using 20 random seeds.

For the experiments in Figure 2, we estimate the expectation from (5) with 20 permutations, which provides sufficiently stable estimates. In addition, we use $L = L_{\max}$ (instead of using the batch smoothness of (Gower et al., 2019)) as the plots in this figure use different minibatch sizes and we want to isolate the effect of reducing the variance by minibatching from the effect of changing L.

SGD implementation. For SGD, we used two approaches to minibatching. In the first, we sampled τ indices from $\{1, \ldots, N\}$ and used them to form the minibatch, where N is the total number of data samples. In the second approach, we permuted the data once and then at each iteration, we only sampled one index i and formed the minibatch from indices $i, (i+1) \mod N, \ldots, (i+\tau-1) \mod N$. The latter approach is much more cash-friendly and runs significantly faster, while the iteration convergence was the same in our experiments. Thus, we used the latter option to produce the final plots.

For all plots and methods, we use zero initialization, $x_0 = (0, \ldots, 0)^\top \in \mathbb{R}^d$. We obtain the optimum, x_*, by running Nesterov's accelerated gradient method until it reaches machine precision. The plots in the right column in Figure 2 were obtained by initializing the methods at an intermediate iterate of Nesterov's method, and we found the average, best and worst results by sampling 1,000 permutations.

6 Basic facts and notation

6.1 Basic identities and inequalities

For any two vectors $a, b \in \mathbb{R}^d$, we have

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2. \tag{11}$$

As a consequence of (11), we get

$$\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2. \tag{12}$$

Convexity, strong convexity and smoothness. A differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ is called μ-convex if for some $\mu \ge 0$ and for all $x, y \in \mathbb{R}^d$, we have

$$h(x) + \langle \nabla h(x), y - x\rangle + \frac{\mu}{2}\|y-x\|^2 \le h(y). \tag{13}$$

If h satisfies (13) with $\mu > 0$, then we say that h is μ-strongly convex, and if $\mu = 0$ then we say h is convex. A differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ is called L-smooth if for some $L \ge 0$ and for all $x, y \in \mathbb{R}^d$, we have

$$\|\nabla h(x) - \nabla h(y)\| \le L\|x-y\|. \tag{14}$$

A useful consequence of L-smoothness is the inequality

$$h(x) \le h(y) + \langle \nabla h(y), x-y\rangle + \frac{L}{2}\|x-y\|^2, \tag{15}$$

holding for all $x, y \in \mathbb{R}^d$. If h is L-smooth and lower bounded by h_*, then

$$\|\nabla h(x)\|^2 \le 2L\left(h(x) - h_*\right). \tag{16}$$

For any convex and L-smooth function h it holds

$$D_h(x,y) \ge \frac{1}{2L}\|\nabla h(x) - \nabla h(y)\|^2. \tag{17}$$

Jensen's inequality and consequences. For a convex function $h : \mathbb{R}^d \to \mathbb{R}$ and any vectors $y_1, \ldots, y_n \in \mathbb{R}^d$, Jensen's inequality states that

$$h\left(\frac{1}{n}\sum_{i=1}^n y_i\right) \le \frac{1}{n}\sum_{i=1}^n h(y_i).$$

Applying this to the squared norm, $h(y) = \|y\|^2$, we get

$$\left\|\frac{1}{n}\sum_{i=1}^n y_i\right\|^2 \le \frac{1}{n}\sum_{i=1}^n \|y_i\|^2. \tag{18}$$

After multiplying both sides of (18) by n^2, we get

$$\left\|\sum_{i=1}^n y_i\right\|^2 \le n\sum_{i=1}^n \|y_i\|^2. \tag{19}$$

Variance decomposition. We will use the following decomposition that holds for any random variable X with $\mathbb{E}\left[\|X\|^2\right] < +\infty$,

$$\mathbb{E}\left[\|X\|^2\right] = \|\mathbb{E}[X]\|^2 + \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right]. \tag{20}$$

We will make use of the particularization of (20) to the discrete case: let $y_1, \ldots, y_m \in \mathbb{R}^d$ be given vectors and let $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$ be their average. Then,

$$\frac{1}{m}\sum_{i=1}^m \|y_i\|^2 = \|\bar{y}\|^2 + \frac{1}{m}\sum_{i=1}^m \|y_i - \bar{y}\|^2. \tag{21}$$

Table 2: Summary of key notation used in the paper.

Symbol	Description
x_t	The iterate used at the start of epoch t.
π	A permutation $\pi = (\pi_0, \pi_1, \ldots, \pi_{n-1})$ of $\{1, 2, \ldots, n\}$. Fixed for Shuffle-Once and resampled every epoch for Random Reshuffling.
γ	The stepsize used when taking descent steps in an epoch.
x_t^i	The current iterate after i steps in epoch t, for $0 \leq i \leq n$.
g_t	The sum of gradients used over epoch t such that $x_{t+1} = x_t - \gamma g_t$.
σ_t^2	The variance of the individual loss gradients from the average loss at point x_t.
A, B	Assumption 2 constants.
L	The smoothness constant of f and $f_1, f_2, \ldots, f_n$.
μ	The strong convexity constant (for strongly convex objectives).
κ	The condition number $\kappa \stackrel{\text{def}}{=} L/\mu$ for strongly convex objectives.
δ_t	Functional suboptimality, $\delta_t = f(x_t) - f_*$, where $f_* = \inf_x f(x)$.
r_t	The squared iterate distance from an optimum for convex losses $r_t = \\|x_t - x_*\\|^2$.
$\mathbb{E}_t\left[\cdot\right]$	Expectation conditional on the history of the algorithm prior to timestep t, including x_t.

6.2 Notation

We define the epoch total gradient g_t as

$$g_t \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \nabla f_{\pi_i}(x_t^i).$$

We define the variance of the local gradients from their average at a point x_t as

$$\sigma_t^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} \left\| \nabla f_j(x_t) - \nabla f(x_t) \right\|^2.$$

By $\mathbb{E}_t\left[\cdot\right]$ we denote the expectation conditional on all information prior to iteration t, including x_t. To avoid issues with the special case $n = 1$, we use the convention $0/0 = 0$. A summary of the notation used in this work is given in Table 2.

7 A lemma for sampling without replacement

The following algorithm-independent lemma characterizes the variance of sampling a number of vectors from a finite set of vectors, without replacement. It is a key ingredient in our results on the convergence of the RR and SO methods.

Lemma 1. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be fixed vectors, $\overline{X} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n X_i$ be their average and $\sigma^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \left\|X_i - \overline{X}\right\|^2$ be the population variance. Fix any $k \in \{1, \ldots, n\}$, let $X_{\pi_1}, \ldots X_{\pi_k}$ be sampled uniformly without replacement from $\{X_1, \ldots, X_n\}$ and $\overline{X}_\pi$ be their average. Then, the sample average and variance are given by

$$\mathbb{E}\left[\overline{X}_\pi\right] = \overline{X}, \qquad \mathbb{E}\left[\left\|\overline{X}_\pi - \overline{X}\right\|^2\right] = \frac{n-k}{k(n-1)}\sigma^2. \tag{22}$$

Proof. The first claim follows by linearity of expectation and uniformity of sampling:

$$\mathbb{E}\left[\overline{X}_\pi\right] = \frac{1}{k}\sum_{i=1}^k \mathbb{E}\left[X_{\pi_i}\right] = \frac{1}{k}\sum_{i=1}^k \overline{X} = \overline{X}.$$

To prove the second claim, let us first establish that the identity $\mathrm{cov}(X_{\pi_i}, X_{\pi_j}) = -\frac{\sigma^2}{n-1}$ holds for any $i \neq j$. Indeed,

$$\begin{aligned}
\mathrm{cov}(X_{\pi_i}, X_{\pi_j}) = \mathbb{E}\left[\left\langle X_{\pi_i} - \overline{X}, X_{\pi_j} - \overline{X}\right\rangle\right] &= \frac{1}{n(n-1)}\sum_{l=1}^n \sum_{m=1, m\neq l}^n \left\langle X_l - \overline{X}, X_m - \overline{X}\right\rangle \\
&= \frac{1}{n(n-1)}\sum_{l=1}^n\sum_{m=1}^n \left\langle X_l - \overline{X}, X_m - \overline{X}\right\rangle - \frac{1}{n(n-1)}\sum_{l=1}^n \left\|X_l - \overline{X}\right\|^2 \\
&= \frac{1}{n(n-1)}\sum_{l=1}^n \left\langle X_l - \overline{X}, \sum_{m=1}^n (X_m - \overline{X})\right\rangle - \frac{\sigma^2}{n-1} \\
&= -\frac{\sigma^2}{n-1}.
\end{aligned}$$

This identity helps us to establish the formula for sample variance:

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{X}_\pi - \overline{X}\right\|^2\right] &= \frac{1}{k^2}\sum_{i=1}^k\sum_{j=1}^k \mathrm{cov}(X_{\pi_i}, X_{\pi_j}) \\
&= \frac{1}{k^2}\mathbb{E}\left[\sum_{i=1}^k \left\|X_{\pi_i} - \overline{X}\right\|^2\right] + \sum_{i=1}^k \sum_{j=1, j\neq i}^k \mathrm{cov}(X_{\pi_i}, X_{\pi_j}) \\
&= \frac{1}{k^2}\left(k\sigma^2 - k(k-1)\frac{\sigma^2}{n-1}\right) = \frac{n-k}{k(n-1)}\sigma^2.
\end{aligned}$$

∎

8 Proofs for convex objectives (Sections 3.1 and 3.2)

8.1 Proof of Proposition 1

Proof. Let us start with the upper bound. Fixing any i such that $1 \le i \le n-1$, we have $i(n-i) \le \frac{n^2}{4} \le \frac{n(n-1)}{2}$ and using smoothness and Lemma 1 leads to

$$\begin{aligned}\mathbb{E}\left[D_{f_{\pi_i}}(x_*^i, x_*)\right] \overset{(15)}{\le} \frac{L}{2}\mathbb{E}\left[\|x_*^i - x_*\|^2\right] &= \frac{L}{2}\mathbb{E}\left[\left\|\sum_{k=0}^{i-1}\gamma\nabla f_{\pi_k}(x_*)\right\|^2\right] \\ &\overset{(22)}{=} \frac{\gamma^2 L i(n-i)}{2(n-1)}\sigma_*^2 \\ &\le \frac{\gamma^2 L n}{4}\sigma_*^2.\end{aligned}$$

To obtain the upper bound, it remains to take maximum with respect to i on both sides and divide by γ. To prove the lower bound, we use strong convexity and the fact that $\max_i i(n-i) \ge \frac{n(n-1)}{4}$ holds for any integer n. Together, this leads to

$$\max_i \mathbb{E}\left[D_{f_{\pi_i}}(x_*^i, x_*)\right] \overset{(13)}{\ge} \max_i \frac{\mu}{2}\mathbb{E}\left[\|x_*^i - x_*\|^2\right] = \max_i \frac{\gamma^2 \mu i(n-i)}{2(n-1)}\sigma_*^2 \ge \frac{\gamma^2 \mu n}{8}\sigma_*^2,$$

as desired. ■

8.2 Proof Remainder for Theorem 1

Proof. We start from (8) proved in the main text:

$$\mathbb{E}\left[\|x_t^{i+1} - x_*^{i+1}\|^2\right] \le (1-\gamma\mu)\mathbb{E}\left[\|x_t^i - x_*^i\|^2\right] + 2\gamma^2\sigma_{\text{Shuffle}}^2.$$

Since $x_{t+1} - x_* = x_t^n - x_*^n$ and $x_t - x_* = x_t^0 - x_*^0$, we can unroll the recursion, obtaining the epoch level recursion

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] \le (1-\gamma\mu)^n\,\mathbb{E}\left[\|x_t - x_*\|^2\right] + 2\gamma^2\sigma_{\text{Shuffle}}^2\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right).$$

Unrolling this recursion across T epochs, we obtain

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \le (1-\gamma\mu)^{nT}\|x_0 - x_*\|^2 + 2\gamma^2\sigma_{\text{Shuffle}}^2\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right)\left(\sum_{j=0}^{T-1}(1-\gamma\mu)^{nj}\right). \tag{23}$$

The product of the two sums in (23) can be bounded by reparameterizing the summation as follows:

$$\begin{aligned}\left(\sum_{j=0}^{T-1}(1-\gamma\mu)^{nj}\right)\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right) &= \sum_{j=0}^{T-1}\sum_{i=0}^{n-1}(1-\gamma\mu)^{nj+i} \\ &= \sum_{k=0}^{nT-1}(1-\gamma\mu)^k \le \sum_{k=0}^{\infty}(1-\gamma\mu)^k = \frac{1}{\gamma\mu}.\end{aligned}$$

Plugging this bound back into (23), we finally obtain the bound

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \le (1-\gamma\mu)^{nT}\|x_0 - x_*\|^2 + 2\gamma\frac{\sigma_{\text{Shuffle}}^2}{\mu}.$$

■

8.3 Proof of complexity

In this subsection, we show how we get from Theorem 1 the complexity for strongly convex functions.

Corollary 1. Under the same conditions as those in Theorem 1, we choose stepsize

$$\gamma = \min\left\{\frac{1}{L}, \frac{2}{\mu n T}\log\left(\frac{\|x_0 - x_*\|\,\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_*}\right)\right\}.$$

The final iterate x_T then satisfies

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu n T}{L}\right)\|x_0 - x_*\|^2 + \frac{\kappa\sigma_*^2}{\mu^2 n T^2}\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ denotes ignoring absolute constants and polylogarithmic factors. Thus, in order to obtain error (in squared distance to the optimum) less than ε, we require that the total number of iterations nT satisfies

$$nT = \tilde{\Omega}\left(\kappa + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}\right).$$

Proof. Applying Theorem 1, the final iterate generated by Algorithms 1 or 2 after T epochs satisfies

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \le (1-\gamma\mu)^{nT}\|x_0 - x_*\|^2 + 2\gamma\frac{\sigma_{\text{Shuffle}}^2}{\mu}.$$

Using Proposition 1 to bound $\sigma_{\text{Shuffle}}^2$, we get

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \le (1-\gamma\mu)^{nT}\|x_0 - x_*\|^2 + \gamma^2\kappa n\sigma_*^2. \tag{24}$$

We now have two cases:

- **Case 1**: If $\frac{1}{L} \le \frac{2}{\mu n T}\log\left(\frac{\|x_0 - x_*\|\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_*}\right)$, then using $\gamma = \frac{1}{L}$ in (24) we have

$$\begin{aligned}\mathbb{E}\left[\|x_T - x_*\|^2\right] &\le \left(1 - \frac{\mu}{L}\right)^{nT}\|x_0 - x_*\|^2 + \frac{\kappa n\sigma_*^2}{L^2}\\ &\le \left(1 - \frac{\mu}{L}\right)^{nT}\|x_0 - x_*\|^2 + \frac{4\kappa\sigma_*^2}{\mu^2 n T^2}\log^2\left(\frac{\|x_0 - x_*\|\,\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_*}\right).\end{aligned}$$

  Using that $1 - x \le \exp(-x)$ in the previous inequality, we get

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu n T}{L}\right)\|x_0 - x_*\|^2 + \frac{\kappa\sigma_*^2}{\mu^2 n T^2}\right), \tag{25}$$

  where $\tilde{\mathcal{O}}(\cdot)$ denotes ignoring polylogarithmic factors and absolute (non-problem specific) constants.

- **Case 2**: If $\frac{2}{\mu n T}\log\left(\frac{\|x_0 - x_*\|\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_*}\right) < \frac{1}{L}$, recall that by Theorem 1,

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \le (1-\gamma\mu)^{nT}\|x_0 - x_*\|^2 + \gamma^2\kappa n\sigma_*^2. \tag{26}$$

  Plugging in $\gamma = \frac{2}{\mu n T}\log\left(\frac{\|x_0 - x_*\|\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_*}\right)$, the first term in (26) satisfies

$$\begin{aligned}(1-\gamma\mu)^{nT}\|x_0 - x_*\|^2 &\le \exp\left(-\gamma\mu n T\right)\|x_0 - x_*\|^2\\ &= \exp\left(-2\log\left(\frac{\|x_0 - x_*\|\,\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_*}\right)\right)\|x_0 - x_*\|^2\\ &= \frac{\kappa\sigma_*^2}{\mu^2 n T^2}.\end{aligned} \tag{27}$$

  Furthermore, the second term in (26) satisfies

$$\gamma^2\kappa n\sigma_*^2 = \frac{4\kappa\sigma_*^2}{\mu^2 n T^2}\log^2\left(\frac{\|x_0 - x_*\|\,\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_*}\right). \tag{28}$$

  Substituting (27) and (28) into (26), we get

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \tilde{\mathcal{O}}\left(\frac{\kappa\sigma_*^2}{\mu^2 n T^2}\right). \tag{29}$$

  This concludes the second case.

It remains to take the maximum of (25) from the first case and (29) from the second case. ■

8.4 Two lemmas for Theorems 2 and 3

In order to prove Theorems 2 and 3, it will be useful to define the following quantity.

Definition 3. Let $x_t^0, x_t^1, \ldots, x_t^n$ be iterates generated by Algorithms 1 or 2. We define the forward per-epoch deviation over the t-th epoch $\mathcal{V}_t$ as

$$\mathcal{V}_t \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \left\| x_t^i - x_{t+1} \right\|^2. \tag{30}$$

We will now establish two lemmas. First, we will show that $\mathcal{V}_t$ can be efficiently upper bounded using Bregman divergences and the variance at the optimum. Subsequently use this bound to establish the convergence of RR/SO.

8.4.1 Bounding the forward per-epoch deviation

Lemma 2. Consider the iterates of Random Reshuffling (Algorithm 1) or Shuffle-Once (Algorithm 2). If the functions $f_1, \ldots, f_n$ are convex and Assumption 1 is satisfied, then

$$\mathbb{E}\left[\mathcal{V}_t\right] \leq 4\gamma^2 n^2 L \sum_{i=0}^{n-1} \mathbb{E}\left[D_{f_{\pi_i}}(x_*, x_t^i)\right] + \frac{1}{2}\gamma^2 n^2 \sigma_*^2, \tag{31}$$

where $\mathcal{V}_t$ is defined as in Definition 3, and σ_*^2 is the variance at the optimum given by $\sigma_*^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_*)\|^2$.

Proof. For any fixed $k \in \{0, \ldots, n-1\}$, by definition of x_t^k and x_{t+1} we get the decomposition

$$x_t^k - x_{t+1} = \gamma \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_t^i) = \gamma \sum_{i=k}^{n-1} (\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)) + \gamma \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*).$$

Applying Young's inequality to the sums above yields

$$\begin{aligned}
\|x_t^k - x_{t+1}\|^2 &\overset{(12)}{\leq} 2\gamma^2 \left\| \sum_{i=k}^{n-1} (\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)) \right\|^2 + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \\
&\overset{(19)}{\leq} 2\gamma^2 n \sum_{i=k}^{n-1} \left\| \nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*) \right\|^2 + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \\
&\overset{(17)}{\leq} 4\gamma^2 L n \sum_{i=k}^{n-1} D_{f_{\pi_i}}(x_*, x_t^i) + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \\
&\leq 4\gamma^2 L n \sum_{i=0}^{n-1} D_{f_{\pi_i}}(x_*, x_t^i) + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2.
\end{aligned}$$

Summing up and taking expectations leads to

$$\sum_{k=0}^{n-1} \mathbb{E}\left[\left\|x_t^k - x_{t+1}\right\|^2\right] \leq 4\gamma^2 L n^2 \sum_{i=0}^{n-1} \mathbb{E}\left[D_{f_{\pi_i}}(x_*, x_t^i)\right] + 2\gamma^2 \sum_{k=0}^{n-1} \mathbb{E}\left[\left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2\right]. \tag{32}$$

We now bound the second term in the right-hand side of (32). First, using Lemma 1, we get

$$\begin{aligned}
\mathbb{E}\left[\left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2\right] &= (n-k)^2 \mathbb{E}\left[\left\| \frac{1}{n-k} \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2\right] \\
&= (n-k)^2 \frac{k}{(n-k)(n-1)} \sigma_*^2 \\
&= \frac{k(n-k)}{n-1} \sigma_*^2.
\end{aligned}$$

Next, by summing this for k from 0 to $n-1$, we obtain

$$\sum_{k=0}^{n-1}\mathbb{E}\left[\left\|\sum_{i=k}^{n-1}\nabla f_{\pi_i}(x_*)\right\|^2\right]=\sum_{k=0}^{n-1}\frac{k(n-k)}{n-1}\sigma_*^2=\frac{1}{6}n(n+1)\sigma_*^2\le\frac{n^2\sigma_*^2}{4},$$

where in the last step we also used $n\ge 2$. The result follows. ■

8.4.2 Finding a per-epoch recursion

Lemma 3. Assume that functions $f_1,\dots,f_n$ are convex and that Assumption 1 is satisfied. If Random Reshuffling (Algorithm 1) or Shuffle-Once (Algorithm 2) is run with a stepsize satisfying $\gamma\le\frac{1}{\sqrt{2}Ln}$, then

$$\mathbb{E}\left[\|x_{t+1}-x_*\|^2\right]\le\mathbb{E}\left[\|x_t-x_*\|^2\right]-2\gamma n\mathbb{E}\left[f(x_{t+1})-f_*\right]+\frac{\gamma^3Ln^2\sigma_*^2}{2}.$$

Proof. Define the sum of gradients used in the t-th epoch as $g_t\overset{\text{def}}{=}\sum_{i=0}^{n-1}\nabla f_{\pi_i}(x_t^i)$. We will use g_t to relate the iterates x_t and x_{t+1}. By definition of x_{t+1}, we can write

$$x_{t+1}=x_t^n=x_t^{n-1}-\gamma\nabla f_{\pi_{n-1}}(x_t^{n-1})=\cdots=x_t^0-\gamma\sum_{i=0}^{n-1}\nabla f_{\pi_i}(x_t^i).$$

Further, since $x_t^0=x_t$, we see that $x_{t+1}=x_t-\gamma g_t$, which leads to

$$\begin{aligned}\|x_t-x_*\|^2=\|x_{t+1}+\gamma g_t-x_*\|^2&=\|x_{t+1}-x_*\|^2+2\gamma\langle g_t,x_{t+1}-x_*\rangle+\gamma^2\|g_t\|^2\\&\ge\|x_{t+1}-x_*\|^2+2\gamma\langle g_t,x_{t+1}-x_*\rangle\\&=\|x_{t+1}-x_*\|^2+2\gamma\sum_{i=0}^{n-1}\left\langle\nabla f_{\pi_i}(x_t^i),x_{t+1}-x_*\right\rangle.\end{aligned}$$

Observe that for any i, we have the following decomposition

$$\begin{aligned}\left\langle\nabla f_{\pi_i}(x_t^i),x_{t+1}-x_*\right\rangle&=[f_{\pi_i}(x_{t+1})-f_{\pi_i}(x_*)]+[f_{\pi_i}(x_*)-f_{\pi_i}(x_t^i)-\langle\nabla f_{\pi_i}(x_t^i),x_t^i-x_*\rangle]\\&\quad-[f_{\pi_i}(x_{t+1})-f_{\pi_i}(x_t^i)-\langle\nabla f_{\pi_i}(x_t^i),x_{t+1}-x_t^i\rangle]\\&=[f_{\pi_i}(x_{t+1})-f_{\pi_i}(x_*)]+D_{f_{\pi_i}}(x_*,x_t^i)-D_{f_{\pi_i}}(x_{t+1},x_t^i).\end{aligned}\tag{33}$$

Summing the first quantity in (33) over i from 0 to $n-1$ gives

$$\sum_{i=0}^{n-1}[f_{\pi_i}(x_{t+1})-f_{\pi_i}(x_*)]=n(f(x_{t+1})-f_*).$$

Now, we can bound the third term in the decomposition (33) using L-smoothness as follows:

$$D_{f_{\pi_i}}(x_{t+1},x_t^i)\le\frac{L}{2}\|x_{t+1}-x_t^i\|^2.$$

By summing the right-hand side over i from 0 to $n-1$ we get the forward deviation over an epoch $\mathcal{V}_t$, which we bound by Lemma 2 to get

$$\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}(x_{t+1},x_t^i)\right]\overset{(30)}{\le}\frac{L}{2}\mathbb{E}\left[\mathcal{V}_t\right]\overset{(31)}{\le}2\gamma^2L^2n^2\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}(x_*,x_t^i)\right]+\frac{\gamma^2Ln^2\sigma_*^2}{4}.$$

Therefore, we can lower-bound the sum of the second and the third term in (33) as

$$\begin{aligned}\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}(x_*,x_t^i)-D_{f_{\pi_i}}(x_{t+1},x_t^i)\right]&\ge\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}(x_*,x_t^i)\right]-2\gamma^2L^2n^2\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}(x_*,x_t^i)\right]\\&\quad-\frac{\gamma^2Ln^2\sigma_*^2}{4}\\&\ge(1-2\gamma^2L^2n^2)\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}(x_*,x_t^i)\right]-\frac{\gamma^2Ln^2\sigma_*^2}{4}\\&\ge-\frac{\gamma^2Ln^2\sigma_*^2}{4},\end{aligned}$$

where in the third inequality we used that $\gamma \leq \frac{1}{\sqrt{2}Ln}$ and that $D_{f_{\pi_i}}(x_*, x_t^i)$ is nonnegative. Plugging this back into the lower-bound on $\|x_t - x_*\|^2$ yields

$$\mathbb{E}\left[\|x_t - x_*\|^2\right] \geq \mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] + 2\gamma n \mathbb{E}\left[f(x_{t+1}) - f_*\right] - \frac{\gamma^3 L n^2 \sigma_*^2}{2}.$$

Rearranging the terms gives the result. ■

8.5 Proof of Theorem 2

Proof. We can use Lemma 3 and strong convexity to obtain

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] &\leq \mathbb{E}\left[\|x_t - x_*\|^2\right] - 2\gamma n \mathbb{E}\left[f(x_{t+1}) - f_*\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2} \\ &\overset{(13)}{\leq} \mathbb{E}\left[\|x_t - x_*\|^2\right] - \gamma n \mu \mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2},\end{aligned}$$

whence

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] &\leq \frac{1}{1+\gamma\mu n}\left(\mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}\right) \\ &= \frac{1}{1+\gamma\mu n}\mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{1}{1+\gamma\mu n}\frac{\gamma^3 L n^2 \sigma_*^2}{2} \\ &\leq \left(1 - \frac{\gamma\mu n}{2}\right)\mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}.\end{aligned}$$

Recursing for T iterations, we get that the final iterate satisfies

$$\begin{aligned}\mathbb{E}\left[\|x_T - x_*\|^2\right] &\leq \left(1 - \frac{\gamma\mu n}{2}\right)^T \|x_0 - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2}{2}\left(\sum_{j=0}^{T-1}\left(1 - \frac{\gamma\mu n}{2}\right)^j\right) \\ &\leq \left(1 - \frac{\gamma\mu n}{2}\right)^T \|x_0 - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2}{2}\left(\frac{2}{\gamma\mu n}\right) \\ &= \left(1 - \frac{\gamma\mu n}{2}\right)^T \|x_0 - x_*\|^2 + \gamma^2 \kappa n \sigma_*^2.\end{aligned}$$

■

8.6 Proof of Theorem 3

Proof. We start with Lemma 3, which states that the following inequality holds:

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] \leq \mathbb{E}\left[\|x_t - x_*\|^2\right] - 2\gamma n \mathbb{E}\left[f(x_{t+1}) - f(x_*)\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}.$$

Rearranging the result leads to

$$2\gamma n \mathbb{E}\left[f(x_{t+1}) - f(x_*)\right] \leq \mathbb{E}\left[\|x_t - x_*\|^2\right] - \mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}.$$

Summing these inequalities for $t = 0, 1, \ldots, T-1$ gives

$$\begin{aligned}2\gamma n \sum_{t=0}^{T-1} \mathbb{E}\left[f(x_{t+1}) - f(x_*)\right] &\leq \sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|x_t - x_*\|^2\right] - \mathbb{E}\left[\|x_{t+1} - x_*\|^2\right]\right) + \frac{\gamma^3 L n^2 \sigma_*^2 T}{2} \\ &= \|x_0 - x_*\|^2 - \mathbb{E}\left[\|x_T - x_*\|^2\right] + \frac{\gamma^3 L n^2 \sigma_*^2 T}{2} \\ &\leq \|x_0 - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2 T}{2},\end{aligned}$$

and dividing both sides by $2\gamma n T$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[f(x_{t+1}) - f(x_*)\right] \leq \frac{\|x_0 - x_*\|^2}{2\gamma n T} + \frac{\gamma^2 L n \sigma_*^2}{4}.$$

Finally, using the convexity of f, the average iterate $\hat{x}_T \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} x_t$ satisfies

$$\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[f(x_t) - f(x_*)\right] \leq \frac{\|x_0 - x_*\|^2}{2\gamma n T} + \frac{\gamma^2 L n \sigma_*^2}{4}.$$

■

8.7 Proof of complexity

Corollary 2. Under the same conditions as Theorem 3, choose the stepsize

$$\gamma = \min\left\{ \frac{1}{\sqrt{2}Ln}, \left(\frac{\|x_0 - x_*\|^2}{Ln^2 T \sigma_*^2} \right)^{1/3} \right\}.$$

Then

$$\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] \leq \frac{L\|x_0 - x_*\|^2}{\sqrt{2}T} + \frac{3L^{1/3}\|x_0 - x_*\|^{4/3}\sigma_*^{2/3}}{4n^{1/3}T^{2/3}}.$$

We can guarantee $\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] \leq \varepsilon^2$ provided that the total number of iterations satisfies

$$Tn \geq \frac{2\|x_0 - x_*\|^2 \sqrt{Ln}}{\varepsilon^2} \max\left\{ \sqrt{2Ln}, \frac{\sigma_*}{\varepsilon} \right\}.$$

Proof. We start with the guarantee of Theorem 3:

$$\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] \leq \frac{\|x_0 - x_*\|^2}{2\gamma n T} + \frac{\gamma^2 Ln\sigma_*^2}{4}. \tag{34}$$

We now have two cases depending on the stepsize:

- **Case 1**: If $\gamma = \frac{1}{\sqrt{2}Ln} \leq \left(\frac{\|x_0 - x_*\|^2}{Ln^2 T \sigma_*^2} \right)^{1/3}$, then plugging this γ into (34) gives

$$\begin{aligned}
\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] &\leq \frac{L\|x_0 - x_*\|^2}{\sqrt{2}T} + \frac{\gamma^2 Ln\sigma_*^2}{4} \\
&\leq \frac{L\|x_0 - x_*\|^2}{\sqrt{2}T} + \left(\frac{\|x_0 - x_*\|^2}{Ln^2 T \sigma_*^2} \right)^{2/3} \frac{Ln\sigma_*^2}{4} \\
&= \frac{L\|x_0 - x_*\|^2}{\sqrt{2}T} + \frac{L^{1/3}\sigma_*^{2/3}\|x_0 - x_*\|^{4/3}}{4n^{1/3}T^{2/3}}.
\end{aligned} \tag{35}$$

- **Case 2**: If $\gamma = \left(\frac{\|x_0 - x_*\|^2}{Ln^2 T \sigma_*^2} \right)^{1/3} \leq \frac{1}{\sqrt{2}Ln}$, then plugging this γ into (34) gives

$$\begin{aligned}
\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] &\leq \frac{L^{1/3}\|x_0 - x_*\|^{4/3}\sigma_*^{2/3}}{2n^{1/3}T^{2/3}} + \frac{L^{1/3}\sigma_*^{2/3}\|x_0 - x_*\|^{4/3}}{4n^{1/3}T^{2/3}} \\
&= \frac{3L^{1/3}\|x_0 - x_*\|^{4/3}\sigma_*^{2/3}}{4n^{1/3}T^{2/3}}.
\end{aligned} \tag{36}$$

Combining (35) and (36), we see that in both cases we have

$$\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] \leq \frac{L\|x_0 - x_*\|^2}{\sqrt{2}T} + \frac{3L^{1/3}\|x_0 - x_*\|^{4/3}\sigma_*^{2/3}}{4n^{1/3}T^{2/3}}.$$

Translating this to sample complexity, we can guarantee that $\mathbb{E}\left[f(\hat{x}_T) - f(x_*)\right] \leq \varepsilon^2$ provided

$$nT \geq \frac{2\|x_0 - x_*\|^2 \sqrt{Ln}}{\varepsilon^2} \max\left\{ \sqrt{Ln}, \frac{\sigma_*}{\varepsilon} \right\}.$$

■

9 Proofs for non-convex objectives (Section 3.3)

9.1 Proof of Proposition 2

Proof. This proposition is a special case of Lemma 3 in (Khaled and Richtárik, 2020) and we prove it here for completeness. Let $x \in \mathbb{R}^d$. We start with (16) (which does not require convexity) applied to each f_i:

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

Averaging we have,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)\|^2 &\leq 2L\left(f(x) - \frac{1}{n}\sum_{i=1}^{n} f_i^*\right) \\ &= 2L\left(f(x) - f_*\right) + 2L\left(f_* - \frac{1}{n}\sum_{i=1}^{n} f_i^*\right).\end{aligned}$$

Note that because f_* is the infimum of $f(\cdot)$ and $\frac{1}{n}\sum_{i=1}^{n} f_i^*$ is a lower bound on f then $f_* - \frac{1}{n}\sum_{i=1}^{n} f_i^* \geq 0$. We may now use the variance decomposition

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2 &\overset{(21)}{=} \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)\|^2 - \|\nabla f(x)\|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)\|^2 \\ &\leq 2L\left(f(x) - f_*\right) + 2L\left(f_* - \frac{1}{n}\sum_{i=1}^{n} f_i^*\right).\end{aligned}$$

It follows that Assumption 2 holds with $A = L$ and $B^2 = 2L\left(f_* - \frac{1}{n}\sum_{i=1}^{n} f_i^*\right)$. ■

9.2 Finding a per-epoch recursion

For this subsection and the rest of this section, we need to define the following quantity:

Definition 4. For Algorithm 1 we define the *backward per-epoch deviation* at timestep t by

$$V_t \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\left\|x_t^i - x_t\right\|^2.$$

We will study the convergence of Algorithm 1 for non-convex objectives as follows: we first derive a per-epoch recursion that involves V_t in Lemma 4, then we show that V_t can be bounded using smoothness and probability theory in Lemma 5, and finally combine these two to prove Theorem 4.

Lemma 4. Suppose that Assumption 1 holds. Then for iterates x_t generated by Algorithm 1 with stepsize $\gamma \leq \frac{1}{Ln}$, we have

$$f(x_{t+1}) \leq f(x_t) - \frac{\gamma n}{2}\|\nabla f(x_t)\|^2 + \frac{\gamma L^2}{2} V_t, \tag{37}$$

where V_t is defined as in Definition 4.

Proof. Our approach for establishing this lemma is similar to that of (Nguyen et al., 2020, Theorem 1), which we became aware of in the course of preparing this manuscript. Recall that $x_{t+1} = x_t - \gamma g_t$, where $g_t = \sum_{i=0}^{n-1}\nabla f_{\pi_i}(x_t^i)$. Using L-smoothness of f, we get

$$\begin{aligned}f(x_{t+1}) &\overset{(15)}{\leq} f(x_t) + \langle\nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \gamma n\left\langle\nabla f(x_t), \frac{g_t}{n}\right\rangle + \frac{\gamma^2 L n^2}{2}\left\|\frac{g_t}{n}\right\|^2 \\ &\overset{(11)}{=} f(x_t) - \frac{\gamma n}{2}\left(\|\nabla f(x_t)\|^2 + \left\|\frac{g_t}{n}\right\|^2 - \left\|\nabla f(x_t) - \frac{g_t}{n}\right\|^2\right) + \frac{\gamma^2 L n^2}{2}\left\|\frac{g_t}{n}\right\|^2 \\ &= f(x_t) - \frac{\gamma n}{2}\|\nabla f(x_t)\|^2 - \frac{\gamma n}{2}\left(1 - L\gamma n\right)\left\|\frac{g_t}{n}\right\|^2 + \frac{\gamma n}{2}\left\|\nabla f(x_t) - \frac{g_t}{n}\right\|^2. \end{aligned} \tag{38}$$

By assumption, we have $\gamma \le \frac{1}{Ln}$, and hence $1 - L\gamma n \ge 0$. Using this in (38), we get

$$f(x_{t+1}) \le f(x_t) - \frac{\gamma n}{2}\|\nabla f(x_t)\|^2 + \frac{\gamma n}{2}\left\|\nabla f(x_t) - \frac{g_t}{n}\right\|^2. \tag{39}$$

For the last term in (39), we note

$$\begin{aligned}\left\|\nabla f(x_t) - \frac{g_t}{n}\right\|^2 &= \left\|\frac{1}{n}\sum_{i=0}^{n-1}\left[\nabla f_{\pi_i}(x_t) - \nabla f_{\pi_i}(x_t^i)\right]\right\|^2 \\ &\overset{(18)}{\le} \frac{1}{n}\sum_{i=0}^{n-1}\left\|\nabla f_{\pi_i}(x_t) - \nabla f_{\pi_i}(x_t^i)\right\|^2 \\ &\overset{(14)}{\le} \frac{1}{n}\sum_{i=0}^{n-1} L^2\left\|x_t - x_t^i\right\|^2 = \frac{L^2}{n}V_t. \end{aligned} \tag{40}$$

Plugging in (40) into (39) yields the lemma's claim. ■

9.3 Bounding the backward per-epoch deviation

Lemma 5. Suppose that Assumption 1 holds (with each f_i possibly non-convex) and that Algorithm 1 is used with a stepsize $\gamma \le \frac{1}{2Ln}$. Then

$$\mathbb{E}_t\left[V_t\right] \le \gamma^2 n^3 \|\nabla f(x_t)\|^2 + \gamma^2 n^2 \sigma_t^2, \tag{41}$$

where V_t is defined as in Definition 4 and $\sigma_t^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n}\|\nabla f_j(x_t) - \nabla f(x_t)\|^2$.

Proof. Let us fix any $k \in [1, n-1]$ and find an upper bound for $\mathbb{E}_t\left[\left\|x_t^k - x_t\right\|^2\right]$. First, note that

$$x_t^k = x_t - \gamma\sum_{i=0}^{k-1}\nabla f_{\pi_i}(x_t^i).$$

Therefore, by Young's inequality, Jensen's inequality and gradient Lipschitzness

$$\begin{aligned}\mathbb{E}_t\left[\|x_t^k - x_t\|^2\right] &= \gamma^2\mathbb{E}_t\left[\left\|\sum_{i=0}^{k-1}\nabla f_{\pi_i}(x_t^i)\right\|^2\right] \\ &\overset{(12)}{\le} 2\gamma^2\mathbb{E}_t\left[\left\|\sum_{i=0}^{k-1}\left(\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_t)\right)\right\|^2\right] + 2\gamma^2\mathbb{E}_t\left[\left\|\sum_{i=0}^{k-1}\nabla f_{\pi_i}(x_t)\right\|^2\right] \\ &\overset{(19)}{\le} 2\gamma^2 k\sum_{i=0}^{k-1}\mathbb{E}_t\left[\left\|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_t)\right\|^2\right] + 2\gamma^2\mathbb{E}_t\left[\left\|\sum_{i=0}^{k-1}\nabla f_{\pi_i}(x_t)\right\|^2\right] \\ &\overset{(14)}{\le} 2\gamma^2 L^2 k\sum_{i=0}^{k-1}\mathbb{E}_t\left[\|x_t^i - x_t\|^2\right] + 2\gamma^2\mathbb{E}_t\left[\left\|\sum_{i=0}^{k-1}\nabla f_{\pi_i}(x_t)\right\|^2\right].\end{aligned}$$

Let us bound the second term. For any i we have $\mathbb{E}_t\left[\nabla f_{\pi_i}(x_t)\right] = \nabla f(x_t)$, so using Lemma 1 (with vectors $\nabla f_{\pi_0}(x_t), \nabla f_{\pi_1}(x_t), \ldots, \nabla f_{\pi_{k-1}}(x_t)$) we obtain

$$\begin{aligned}\mathbb{E}_t\left[\left\|\sum_{i=0}^{k-1}\nabla f_{\pi_i}(x_t)\right\|^2\right] &\overset{(20)}{=} k^2\|\nabla f(x_t)\|^2 + k^2\mathbb{E}_t\left[\left\|\frac{1}{k}\sum_{i=0}^{k-1}(\nabla f_{\pi_i}(x_t) - \nabla f(x_t))\right\|^2\right] \\ &\overset{(22)}{\le} k^2\|\nabla f(x_t)\|^2 + \frac{k(n-k)}{n-1}\sigma_t^2.\end{aligned}$$

where $\sigma_t^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n \|\nabla f_j(x_t) - \nabla f(x_t)\|^2$. Combining the produced bounds yields

$$\begin{aligned}\mathbb{E}_t\left[\left\|x_t^k - x_t\right\|^2\right] &\le 2\gamma^2 L^2 k \sum_{i=0}^{k-1} \mathbb{E}_t\left[\left\|x_t^i - x_t\right\|^2\right] + 2\gamma^2 k^2 \|\nabla f(x_t)\|^2 + 2\gamma^2 \frac{k(n-k)}{n-1}\sigma_t^2 \\ &\le 2\gamma^2 L^2 k \mathbb{E}\left[V_t\right] + 2\gamma^2 k^2 \|\nabla f(x_t)\|^2 + 2\gamma^2 \frac{k(n-k)}{n-1}\sigma_t^2,\end{aligned}$$

whence

$$\begin{aligned}\mathbb{E}\left[V_t\right] &= \sum_{k=0}^{n-1} \mathbb{E}_t\left[\|x_t^k - x_t\|^2\right] \\ &\le \gamma^2 L^2 n(n-1)\mathbb{E}\left[V_t\right] + \frac{1}{3}\gamma^2(n-1)n(2n-1)\|\nabla f(x_t)\|^2 + \frac{1}{3}\gamma^2 n(n+1)\sigma_t^2.\end{aligned}$$

Since $\mathbb{E}\left[V_t\right]$ appears in both sides of the equation, we rearrange and use that $\gamma \le \frac{1}{2Ln}$ by assumption, which leads to

$$\begin{aligned}\mathbb{E}\left[V_t\right] &\le \frac{4}{3}(1 - \gamma^2 L^2 n(n-1))\mathbb{E}\left[V_t\right] \\ &\le \frac{4}{9}\gamma^2(n-1)n(2n-1)\|\nabla f(x_t)\|^2 + \frac{4}{9}\gamma^2 n(n+1)\sigma_t^2 \\ &\le \gamma^2 n^3 \|\nabla f(x_t)\|^2 + \gamma^2 n^2 \sigma_t^2.\end{aligned}$$

■

9.4 A lemma for solving the non-convex recursion

Lemma 6. Suppose that there exist constants $a, b, c \ge 0$ and nonnegative sequences $(s_t)_{t=0}^T, (q_t)_{t=0}^T$ such that for any t satisfying $0 \le t \le T$ we have the recursion

$$s_{t+1} \le (1+a)\, s_t - b q_t + c. \tag{42}$$

Then, the following holds:

$$\min_{t=0,\ldots,T-1} q_t \le \frac{(1+a)^T}{bT} s_0 + \frac{c}{b}. \tag{43}$$

Proof. The first part of the proof (for $a > 0$) is a distillation of the recursion solution in Lemma 2 of Khaled and Richtárik (2020) and we closely follow their proof. Define

$$w_t \stackrel{\text{def}}{=} \frac{1}{(1+a)^{t+1}}.$$

Note that $w_t\,(1+a) = w_{t-1}$ for all t. Multiplying both sides of (42) by w_t,

$$w_t s_{t+1} \le (1+a)\, w_t s_t - b w_t q_t + c w_t = w_{t-1} s_t - b w_t q_t + c w_t.$$

Rearranging, we get $b w_t q_t \le w_{t-1} s_t - w_t s_{t+1} + c w_t$. Summing up as t varies from 0 to $T-1$ and noting that the sum telescopes leads to

$$\begin{aligned}\sum_{t=0}^{T-1} b w_t q_t &\le \sum_{t=0}^{T-1} (w_{t-1} s_t - w_t s_{t+1}) + c \sum_{t=0}^{T-1} w_t \\ &= w_0 s_0 - w_{T-1} s_T + c \sum_{t=0}^{T-1} w_t \\ &\le w_0 s_0 + c \sum_{t=0}^{T-1} w_t.\end{aligned}$$

Let $W_T = \sum_{t=0}^{T-1} w_t$. Dividing both sides by W_T, we have

$$\frac{1}{W_T} \sum_{t=0}^{T-1} b w_t q_t \le \frac{w_0 s_0}{W_T} + c. \tag{44}$$

Note that the left-hand side of (44) satisfies

$$b \min_{t=0,\ldots,T-1} q_t \leq \frac{1}{W_T} \sum_{t=0}^{T-1} b w_t q_t. \tag{45}$$

For the right-hand side of (44), we have

$$W_T = \sum_{t=0}^{T-1} w_t \geq T \min_{t=0,\ldots,T-1} w_t = T w_{T-1} = \frac{T}{(1+a)^T}. \tag{46}$$

Substituting with (46) in (45) and dividing both sides by b, we finally get

$$\min_{t=0,\ldots,T-1} q_t \leq \frac{(1+a)^T}{bT} s_0 + \frac{c}{b}.$$

■

9.5 Proof of Theorem 4

Proof. Taking expectation in Lemma 4 and then using Lemma 5, we have that for any $t \in \{0, 1, \ldots, T-1\}$,

$$\begin{aligned}
\mathbb{E}_t \left[f(x_{t+1}) \right] &\overset{(37)}{\leq} f(x_t) - \frac{\gamma n}{2} \|\nabla f(x_t)\|^2 + \frac{\gamma L^2}{2} \mathbb{E}_t \left[V_t \right] \\
&\overset{(41)}{\leq} f(x_t) - \frac{\gamma n}{2} \|\nabla f(x_t)\|^2 + \frac{\gamma L^2}{2} \left(\gamma^2 n^3 \|\nabla f(x_t)\|^2 + \gamma^2 n^2 \sigma_t^2 \right) \\
&= f(x_t) - \frac{\gamma n}{2} \left(1 - \gamma^2 L^2 n^2\right) \|\nabla f(x_t)\|^2 + \frac{\gamma^3 L^2 n^2 \sigma_t^2}{2}.
\end{aligned}$$

Let $\delta_t \overset{\text{def}}{=} f(x_t) - f_*$. Adding $-f_*$ to both sides and using Assumption 2,

$$\begin{aligned}
\mathbb{E}_t \left[\delta_{t+1} \right] &\leq \delta_t - \frac{\gamma n}{2} \left(1 - \gamma^2 L^2 n^2\right) \|\nabla f(x_t)\|^2 + \frac{\gamma^3 L^2 n^2 \sigma_t^2}{2} \\
&\leq \left(1 + \gamma^3 A L^2 n^2\right) \delta_t - \frac{\gamma n}{2} \left(1 - \gamma^2 L^2 n^2\right) \|\nabla f(x_t)\|^2 + \frac{\gamma^3 L^2 n^2 B^2}{2}.
\end{aligned}$$

Taking unconditional expectations in the last inequality and using that by assumption on γ we have $1 - \gamma^2 L^2 n^2 \geq \frac{1}{2}$, we get the estimate

$$\mathbb{E} \left[\delta_{t+1} \right] \leq \left(1 + \gamma^3 A L^2 n^2\right) \mathbb{E} \left[\delta_t \right] - \frac{\gamma n}{4} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \frac{\gamma^3 L^2 n^2 B^2}{2}. \tag{47}$$

Comparing (42) with (47) verifies that the conditions of Lemma 6 are readily satisfied. Applying the lemma, we get

$$\min_{t=0,\ldots,T-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq \frac{4 \left(1 + \gamma^3 A L^2 n^2\right)^T}{\gamma n T} \left(f(x_0) - f_* \right) + 2\gamma^2 L^2 n B^2.$$

Using that $1 + x \leq \exp(x)$ and that the stepsize γ satisfies $\gamma \leq \left(A L^2 n^2 T\right)^{-1/3}$, we have

$$\left(1 + \gamma^3 A L^2 n^2\right)^T \leq \exp\left(\gamma^3 A L^2 n^2 T\right) \leq \exp(1) \leq 3.$$

Using this in the previous bound, we finally obtain

$$\min_{t=0,\ldots,T-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq \frac{12 \left(f(x_0) - f_* \right)}{\gamma n T} + 2\gamma^2 L^2 n B^2.$$

■

9.6 Proof of complexity

Corollary 3. Choose the stepsize γ as

$$\gamma = \min \left\{ \frac{1}{2Ln}, \frac{1}{A^{1/3} L^{2/3} n^{2/3} T^{1/3}}, \frac{\varepsilon}{2L\sqrt{n}B} \right\}.$$

Then the minimum gradient norm satisfies

$$\min_{t=0,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \le \varepsilon^2$$

provided the total number of iterations satisfies

$$Tn \ge \frac{48\delta_0 L\sqrt{n}}{\varepsilon^2} \max\left\{\sqrt{n}, \frac{\sqrt{6\delta_0 A}}{\varepsilon}, \frac{B}{\varepsilon}\right\}.$$

Proof. From Theorem 4

$$\min_{t=0,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \le \frac{12\left(f(x_0) - f_*\right)}{\gamma n T} + 2\gamma^2 L^2 n B^2.$$

Note that by condition on the stepsize $2L^2\gamma^2 nB^2 \le \varepsilon^2/2$, hence

$$\min_{t=0,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \le \frac{12\left(f(x_0) - f_*\right)}{\gamma n T} + \frac{\varepsilon^2}{2}.$$

Thus, to make the squared gradient norm smaller than ε^2 we require

$$\frac{12\left(f(x_0) - f_*\right)}{\gamma n T} \le \frac{\varepsilon^2}{2},$$

or equivalently

$$nT \ge \frac{24\left(f(x_0) - f_*\right)}{\varepsilon^2 \gamma} = \frac{24\delta_0}{\varepsilon^2} \max\left\{2Ln, \left(AL^2 n^2 T\right)^{1/3}, \frac{2L\sqrt{n}B}{\varepsilon}\right\}, \tag{48}$$

where $\delta_0 \overset{\text{def}}{=} f(x_0) - f_*$ and where we plugged in the value of the stepsize γ we use. Note that nT appears on both sides in the second term in the maximum in (48), hence we can cancel out and simplify:

$$nT \ge \frac{24\delta_0}{\varepsilon^2}(AL^2n^2T)^{1/3} \iff nT \ge \frac{(24\delta_0)^{3/2} L\sqrt{An}}{\varepsilon^3}.$$

Using this simplified bound in (48) we obtain that $\min_{t=0,\ldots,T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \le \varepsilon^2$ provided

$$nT \ge \frac{48\delta_0 L\sqrt{n}}{\varepsilon^2} \max\left\{\sqrt{n}, \frac{\sqrt{6\delta_0 A}}{\varepsilon}, \frac{B}{\varepsilon}\right\}.$$

∎

10 Convergence results for IG

In this section we present results that are extremely similar to the previously obtained bounds for RR and SO. For completeness, we also provide a full description of IG in Algorithm 3.

Algorithm 3 Incremental Gradient (IG)

Input: Stepsize $\gamma > 0$, initial vector $x_0 = x_0^0 \in \mathbb{R}^d$, number of epochs T
1: **for** epochs $t = 0, 1, \ldots, T-1$ **do**
2:     **for** $i = 0, 1, \ldots, n-1$ **do**
3:         $x_t^{i+1} = x_t^i - \gamma \nabla f_{i+1}(x_t^i)$
4:     $x_{t+1} = x_t^n$

Theorem 5. Suppose that Assumption 1 is satisfied. Then we have the following results for the Incremental Gradient method:

- **If each f_i is μ-strongly convex**: if $\gamma \le \frac{1}{L}$, then

$$\|x_T - x_*\|^2 \le (1-\gamma\mu)^{nT}\|x_0 - x_*\|^2 + \frac{\gamma^2 L n^2 \sigma_*^2}{\mu}.$$

By carefully choosing the stepsize as in Corollary 1, we see that this result implies that IG has sample complexity $\tilde{\mathcal{O}}\left(\kappa + \frac{\sqrt{\kappa} n \sigma_*}{\mu\sqrt{\varepsilon}}\right)$ in order to reach a point $\tilde{x}$ with $\|\tilde{x} - x_*\|^2 \le \varepsilon$.

- **If f is μ-strongly convex and each f_i is convex**: if $\gamma \le \frac{1}{\sqrt{2}nL}$, then

$$\|x_T - x_*\|^2 \le \left(1 - \frac{\gamma\mu n}{2}\right)^T \|x_0 - x_*\|^2 + 2\gamma^2\kappa n^2\sigma_*^2.$$

Using the same approach for choosing the stepsize as Corollary 1, we see that IG in this setting reaches an ε-accurate solution after $\tilde{\mathcal{O}}\left(n\kappa + \frac{\sqrt{\kappa}n\sigma_*}{\mu\sqrt{\varepsilon}}\right)$ individual gradient accesses.

- **If each f_i is convex**: if $\gamma \le \frac{1}{\sqrt{2}nL}$, then

$$f(\hat{x}_T) - f(x_*) \le \frac{\|x_0 - x_*\|^2}{2\gamma nT} + \frac{\gamma^2 Ln^2\sigma_*^2}{2},$$

where $\hat{x}_T \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^T x_t$. Choosing the stepsize $\gamma = \min\left\{\frac{1}{\sqrt{2}nL}, \frac{\sqrt{\varepsilon}}{\sqrt{L}n\sigma_*}\right\}$, then the average of iterate generated by IG is an ε-accurate solution (i.e., $f(\hat{x}_T) - f(x_*) \le \varepsilon$) provided that the total number of iterations satisfies

$$nT \ge \frac{\|x_0 - x_*\|^2}{\varepsilon}\max\left\{\sqrt{8}nL, \frac{\sqrt{L}\sigma_* n}{\sqrt{\varepsilon}}\right\}.$$

- **If each f_i is possibly non-convex**: if Assumption 2 holds with constants $A, B \ge 0$ and $\gamma \le \min\left\{\frac{1}{\sqrt{8}nL}, \frac{1}{(4L^2n^3AT)^{1/3}}\right\}$, then

$$\min_{t=0,\ldots,T-1}\|\nabla f(x_t)\|^2 \le \frac{12\left(f(x_0) - f_*\right)}{\gamma nT} + 8\gamma^2L^2n^2B^2.$$

Using an approach similar to Corollary 3, we can establish that IG reaches a point with gradient norm less than ε provided that the total number of iterations exceeds

$$nT \ge \frac{48\left(f(x_0) - f_*\right)Ln}{\varepsilon^2}\max\left\{\sqrt{2}, \frac{\sqrt{24\left(f(x_0) - f_*\right)A}}{\varepsilon}, \frac{2B}{\varepsilon}\right\}.$$

The proof of Thoerem 5 is given in the rest of the section, but first we briefly discuss the convergence rates and the relation of the result on strongly convex objectives to the lower bound of Safran and Shamir (2020).

Discussion of the convergence rates. A brief comparison between the sample complexities given for IG in Theorem 5 and those given for RR (in Table 1) reveals that IG has similar rates to RR but with a worse dependence on n in the variance term (the term associated with σ_* in the convex case and B in the non-convex case), in particular IG is worse by a factor of $\sqrt{n}$. This difference is significant in the large-scale machine learning regime, where the number of data points n can be on the order of thousands to millions.

Discussion of existing lower bounds. Safran and Shamir (2020) give the lower bound (in a problem with $\kappa = 1$)

$$\|x_T - x_*\|^2 = \Omega\left(\frac{\sigma_*^2}{\mu^2T^2}\right).$$

This implies a sample complexity of $\mathcal{O}\left(\frac{n\sigma_*}{\mu\sqrt{\varepsilon}}\right)$, which matches our upper bound (up to an extra iteration and log factors) in the case each f_i is strongly convex and $\kappa = 1$.

10.1 Preliminary Lemmas for Theorem 5

10.1.1 Two lemmas for convex objectives

Lemma 7. Consider the iterates of Incremental Gradient (Algorithm 3). Suppose that functions $f_1, \ldots, f_n$ are convex and that Assumption 1 is satisfied. Then it holds

$$\sum_{k=0}^{n-1}\left\|x_t^k - x_{t+1}\right\|^2 \le 4\gamma^2Ln^2\sum_{i=0}^{n-1}D_{f_{i+1}}(x_*, x_t^i) + 2\gamma^2n^3\sigma_*^2, \tag{49}$$

where σ_*^2 is the variance at the optimum given by $\sigma_*^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_*)\|^2$.

Proof. The proof of this Lemma is similar to that of Lemma 2 but with a worse dependence on the variance term, since there is no randomness in IG. Fix any $k \in \{0,\dots,n-1\}$. It holds by definition

$$x_t^k - x_{t+1} = \gamma \sum_{i=k}^{n-1} \nabla f_{i+1}(x_t^i) = \gamma \sum_{i=k}^{n-1}(\nabla f_{i+1}(x_t^i) - \nabla f_{i+1}(x_*)) + \gamma \sum_{i=k}^{n-1} \nabla f_{i+1}(x_*).$$

Applying Young's inequality to the sums above yields

$$\begin{aligned}
\|x_t^k - x_{t+1}\|^2 &\overset{(12)}{\leq} 2\gamma^2 \left\|\sum_{i=k}^{n-1}(\nabla f_{i+1}(x_t^i) - \nabla f_{i+1}(x_*))\right\|^2 + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{i+1}(x_*)\right\|^2 \\
&\overset{(19)}{\leq} 2\gamma^2 n \sum_{i=k}^{n-1} \left\|\nabla f_{i+1}(x_t^i) - \nabla f_{i+1}(x_*)\right\|^2 + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{i+1}(x_*)\right\|^2 \\
&\overset{(17)}{\leq} 4\gamma^2 Ln \sum_{i=k}^{n-1} D_{f_{i+1}}(x_*, x_t^i) + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{i+1}(x_*)\right\|^2 \\
&\leq 4\gamma^2 Ln \sum_{i=0}^{n-1} D_{f_{i+1}}(x_*, x_t^i) + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{i+1}(x_*)\right\|^2.
\end{aligned}$$

Summing up,

$$\sum_{k=0}^{n-1} \left\|x_t^k - x_{t+1}\right\|^2 \leq 4\gamma^2 Ln^2 \sum_{i=0}^{n-1} D_{f_{i+1}}(x_*, x_t^i) + 2\gamma^2 \sum_{k=0}^{n-1} \left\|\sum_{i=k}^{n-1} \nabla f_{i+1}(x_*)\right\|^2. \tag{50}$$

We now bound the second term in (50). We have

$$\begin{aligned}
\sum_{k=0}^{n-1} \left\|\sum_{i=k}^{n-1} \nabla f_{i+1}(x_*)\right\|^2 &\overset{(19)}{\leq} \sum_{k=0}^{n-1}(n-k) \sum_{i=k}^{n-1} \|\nabla f_{i+1}(x_*)\|^2 \\
&\leq \sum_{k=0}^{n-1}(n-k) \sum_{i=0}^{n-1} \|\nabla f_{i+1}(x_*)\|^2 \\
&= \sum_{k=0}^{n-1}(n-k)\, n\sigma_*^2 = \frac{n^2(n+1)}{2}\sigma_*^2 \leq n^3\sigma_*^2.
\end{aligned} \tag{51}$$

Using (51) in (50), we derive

$$\sum_{k=0}^{n-1} \left\|x_t^k - x_{t+1}\right\|^2 \leq 4\gamma^2 Ln^2 \sum_{i=0}^{n-1} D_{f_{i+1}}(x_*, x_t^i) + 2\gamma^2 n^3 \sigma_*^2.$$ ∎

Lemma 8. Assume the functions $f_1, \dots, f_n$ are convex and that Assumption 1 is satisfied. If Algorithm 3 is run with a stepsize $\gamma \leq \frac{1}{\sqrt{2}Ln}$, then

$$\|x_{t+1} - x_*\|^2 \leq \|x_t - x_*\|^2 - 2\gamma n\left(f(x_{t+1}) - f(x_*)\right) + \gamma^3 Ln^3 \sigma_*^2.$$

Proof. The proof for this lemma is identical to Lemma 3 but with the estimate of Lemma 7 used for $\sum_{i=0}^{n-1}\left\|x_t^i - x_{t+1}\right\|^2$ instead of Lemma 2. We only include it for completeness. Define the sum of gradients used in the t-th epoch as $g_t \overset{\text{def}}{=} \sum_{i=0}^{n-1} \nabla f_{i+1}(x_t^i)$. By definition of x_{t+1}, we have $x_{t+1} = x_t - \gamma g_t$. Using this,

$$\begin{aligned}
\|x_t - x_*\|^2 = \|x_{t+1} + \gamma g_t - x_*\|^2 &= \|x_{t+1} - x_*\|^2 + 2\gamma\left\langle g_t, x_{t+1} - x_*\right\rangle + \gamma^2\|g_t\|^2 \\
&\geq \|x_{t+1} - x_*\|^2 + 2\gamma\left\langle g_t, x_{t+1} - x_*\right\rangle \\
&= \|x_{t+1} - x_*\|^2 + 2\gamma \sum_{i=0}^{n-1}\left\langle \nabla f_{i+1}(x_t^i), x_{t+1} - x_*\right\rangle.
\end{aligned}$$

For any i we have the following decomposition

$$\begin{aligned}\left\langle \nabla f_{i+1}(x_t^i), x_{t+1} - x_* \right\rangle &= [f_{i+1}(x_{t+1}) - f_{i+1}(x_*)] \\ &\quad + [f_{i+1}(x_*) - f_{i+1}(x_t^i) - \langle \nabla f_{i+1}(x_t^i), x_t^i - x_* \rangle] \\ &\quad - [f_{i+1}(x_{t+1}) - f_{i+1}(x_t^i) - \langle \nabla f_{i+1}(x_t^i), x_{t+1} - x_t^i \rangle] \\ &= [f_{i+1}(x_{t+1}) - f_{i+1}(x_*)] + D_{f_{i+1}}(x_*, x_t^i) - D_{f_{i+1}}(x_{t+1}, x_t^i). \end{aligned} \tag{52, 53}$$

Summing the first quantity in (53) over i from 0 to $n-1$ gives

$$\sum_{i=0}^{n-1} [f_{i+1}(x_{t+1}) - f_{i+1}(x_*)] = n(f(x_{t+1}) - f_*).$$

Now let us work out the third term in the decomposition (53) using L-smoothness,

$$D_{f_{i+1}}(x_{t+1}, x_t^i) \le \frac{L}{2} \|x_{t+1} - x_t^i\|^2.$$

We next sum the right-hand side over i from 0 to $n-1$ and use Lemma 7

$$\begin{aligned}\sum_{i=0}^{n-1} D_{f_{i+1}}(x_{t+1}, x_t^i) &\le \frac{L}{2} \sum_{i=0}^{n-1} \left\|x_{t+1} - x_t^i\right\|^2 \\ &\overset{(49)}{\le} 2\gamma^2 L^2 n^2 \sum_{i=0}^{n-1} D_{f_{i+1}}(x_*, x_t^i) + \gamma^2 L n^3 \sigma_*^2. \end{aligned}$$

Therefore, we can lower-bound the sum of the second and the third term in (53) as

$$\begin{aligned}\sum_{i=0}^{n-1} (D_{f_{i+1}}(x_*, x_t^i) - D_{f_{i+1}}(x_{t+1}, x_t^i)) &\ge \sum_{i=0}^{n-1} D_{f_{i+1}}(x_*, x_t^i) \\ &\quad - \left(2\gamma^2 L^2 n^2 \sum_{i=0}^{n-1} D_{f_{i+1}}(x_*, x_t^i) - \gamma^2 L n^3 \sigma_*^2 \right) \\ &= (1 - 2\gamma^2 L^2 n^2) \sum_{i=0}^{n-1} D_{f_{i+1}}(x_*, x_t^i) - \gamma^2 L n^3 \sigma_*^2 \\ &\ge -\gamma^2 L n^3 \sigma_*^2, \end{aligned}$$

where in the third inequality we used that $\gamma \le \frac{1}{\sqrt{2}Ln}$ and that $D_{f_{i+1}}(x_*, x_t^i)$ is nonnegative. Plugging this back into the lower-bound on $\|x_t - x_*\|^2$ yields

$$\|x_t - x_*\|^2 \ge \|x_{t+1} - x_*\|^2 + 2\gamma n \left(f(x_{t+1}) - f_*\right) - \gamma^3 L n^3 \sigma_*^2.$$

Rearranging the terms gives the result. ■

10.1.2 A lemma for non-convex objectives

Lemma 9. Suppose that Assumption 1 holds. Suppose that Algorithm 3 is used with a stepsize $\gamma > 0$ such that $\gamma \le \frac{1}{2Ln}$. Then we have,

$$\sum_{i=1}^{n} \left\|x_t^i - x_t\right\|^2 \le 4\gamma^2 n^3 \|\nabla f(x_t)\|^2 + 4\gamma^2 n^3 \sigma_t^2, \tag{54}$$

where $\sigma_t^2 \overset{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} \|\nabla f_j(x_t) - \nabla f(x_t)\|^2$.

Proof. Let $i \in \{1, 2, \dots, n\}$. Then we can bound the deviation of a single iterate as,

$$\begin{aligned}\left\|x_t^i - x_t\right\|^2 = \left\|x_t^0 - \gamma \sum_{j=0}^{i-1} \nabla f_{j+1}(x_t^j) - x_t\right\|^2 &= \gamma^2 \left\|\sum_{j=0}^{i-1} \nabla f_{j+1}(x_t^j)\right\|^2 \\ &\overset{(19)}{\le} \gamma^2 i \sum_{j=0}^{i-1} \left\|\nabla f_{j+1}(x_t^j)\right\|^2. \end{aligned}$$

Because $i \le n$, we have

$$\left\|x_t^i - x_t\right\|^2 \le \gamma^2 i \sum_{j=0}^{i-1} \left\|\nabla f_{i+1}(x_t^j)\right\|^2 \le \gamma^2 n \sum_{j=0}^{i-1} \left\|\nabla f_{i+1}(x_t^j)\right\|^2 \le \gamma^2 n \sum_{j=0}^{n-1} \left\|\nabla f_{i+1}(x_t^j)\right\|^2. \quad (55)$$

Summing up allows us to estimate V_t:

$$\begin{aligned} V_t &= \sum_{i=1}^{n} \left\|x_t^i - x_t\right\|^2 \\ &\overset{(55)}{\le} \sum_{i=1}^{n} \left(\gamma^2 n \sum_{j=0}^{n-1} \left\|\nabla f_{j+1}(x_t^j)\right\|^2\right) \\ &= \gamma^2 n^2 \sum_{j=0}^{n-1} \left\|\nabla f_{j+1}(x_t^j)\right\|^2 \\ &\overset{(12)}{\le} 2\gamma^2 n^2 \sum_{j=0}^{n-1} \left(\left\|\nabla f_{j+1}(x_t^j) - \nabla f_{i+1}(x_t)\right\|^2 + \|\nabla f_{j+1}(x_t)\|^2\right) \\ &= 2\gamma^2 n^2 \sum_{j=0}^{n-1} \left\|\nabla f_{j+1}(x_t^j) - \nabla f_{j+1}(x_t)\right\|^2 + 2\gamma^2 n^2 \sum_{j=0}^{n-1} \|\nabla f_{j+1}(x_t)\|^2. \end{aligned} \quad (56)$$

For the first term in (56) we can use the smoothness of individual losses and that $x_t^0 = x_t$:

$$\sum_{j=0}^{n-1} \left\|\nabla f_{j+1}(x_t^j) - \nabla f_{j+1}(x_t)\right\|^2 \overset{(14)}{\le} L^2 \sum_{j=0}^{n-1} \left\|x_t^j - x_t\right\|^2 = L^2 \sum_{j=1}^{n-1} \left\|x_t^j - x_t\right\|^2 = L^2 V_t. \quad (57)$$

The second term in (56) is a sum over all the individual gradient evaluated at the same point x_t. Hence, we can drop the permutation subscript and then use the variance decomposition:

$$\begin{aligned} \sum_{j=0}^{n-1} \|\nabla f_{i+1}(x_t)\|^2 &= \sum_{j=1}^{n} \|\nabla f_j(x_t)\|^2 \\ &\overset{(21)}{=} n\|\nabla f(x_t)\|^2 + \sum_{j=1}^{n} \|\nabla f_j(x_t) - \nabla f(x_t)\|^2 \\ &= n\|\nabla f(x_t)\|^2 + n\sigma_t^2. \end{aligned} \quad (58)$$

We can then use (57) and (58) in (56),

$$V_t \le 2\gamma^2 L^2 n^2 V_t + 2\gamma^2 n^3 \|\nabla f(x_t)\|^2 + 2\gamma^2 n^3 \sigma_t^2.$$

Since V_t shows up in both sides of the equation, we can rearrange to obtain

$$\left(1 - 2\gamma^2 L^2 n^2\right) V_t \le 2\gamma^2 n^3 \|\nabla f(x_t)\|^2 + 2\gamma^2 n^3 \sigma_t^2.$$

If $\gamma \le \frac{1}{2Ln}$, then $1 - 2\gamma^2 L^2 n^2 \ge \frac{1}{2}$ and hence

$$V_t \le 4\gamma^2 n^3 \|\nabla f(x_t)\|^2 + 4\gamma^2 n^3 \sigma_t^2. \qquad \blacksquare$$

10.2 Proof of Theorem 5

Proof. • **If each f_i is μ-strongly convex**: The proof follows that of Theorem 1. Define

$$x_*^i = x_* - \gamma \sum_{j=0}^{i-1} \nabla f_{j+1}(x_*).$$

First, we have

$$\begin{aligned} &\|x_t^{i+1} - x_*^{i+1}\|^2 \\ &= \|x_t^i - x_*^i\|^2 - 2\gamma\langle \nabla f_{i+1}(x_t^i) - \nabla f_{i+1}(x_*), x_t^i - x_*^i\rangle + \gamma^2 \|\nabla f_{i+1}(x_t^i) - \nabla f_{i+1}(x_*)\|^2. \end{aligned}$$

Using the same three-point decomposition as Theorem 1 and strong convexity, we have

$$\begin{aligned} -\left\langle \nabla f_{i+1}(x_t^i) - \nabla f_{i+1}(x_*), x_t^i - x_*^i \right\rangle &= -D_{f_{i+1}}(x_*^i, x_t^i) - D_{f_{i+1}}(x_t^i, x_*) + D_{f_{i+1}}(x_*^i, x_*) \\ &\le -\frac{\mu}{2}\left\|x_t^i - x_*^i\right\|^2 - D_{f_{i+1}}(x_t^i, x_*) + D_{f_{i+1}}(x_*^i, x_*). \end{aligned}$$

Using smoothness and convexity

$$\frac{1}{2L}\left\|\nabla f_{i+1}(x_t^i) - \nabla f_{i+1}(x_*)\right\|^2 \le D_{f_{i+1}}(x_t^i, x_*).$$

Plugging in the last two inequalities into the recursion, we get

$$\begin{aligned} \left\|x_t^{i+1} - x_*^{i+1}\right\|^2 &\le (1-\gamma\mu)\left\|x_t^i - x_*^i\right\|^2 - 2\gamma(1-\gamma L) D_{f_{i+1}}(x_t^i, x_*) + 2\gamma D_{f_{i+1}}(x_*^i, x_*). \\ &\le (1-\gamma\mu)\left\|x_t^i - x_*^i\right\|^2 + 2\gamma D_{f_{i+1}}(x_*^i, x_*). \end{aligned} \tag{59}$$

For the last Bregman divergence, we have

$$\begin{aligned} D_{f_{i+1}}(x_*^i, x_*) &\overset{(15)}{\le} \frac{L}{2}\left\|x_*^i - x_*\right\|^2 \\ &= \frac{\gamma^2 L}{2}\left\|\sum_{j=0}^{i-1} \nabla f_{j+1}(x_*)\right\|^2 \\ &\overset{(19)}{\le} \frac{\gamma^2 L i}{2}\sum_{j=0}^{i-1}\left\|\nabla f_{j+1}(x_*)\right\|^2 \\ &= \frac{\gamma^2 L i n}{2}\sigma_*^2 \le \frac{\gamma^2 L n^2}{2}\sigma_*^2. \end{aligned}$$

Plugging this into (59), we get

$$\left\|x_t^{i+1} - x_*^{i+1}\right\|^2 \le (1-\gamma\mu)\left\|x_t^i - x_*^i\right\|^2 + \gamma^3 L n^2 \sigma_*^2.$$

We recurse and then use that $x_*^n = x_*$, $x_{t+1} = x_t^n$, and that $x_*^0 = x_*$, obtaining

$$\begin{aligned} \left\|x_{t+1} - x_*\right\|^2 = \left\|x_t^n - x_*^n\right\|^2 &\le (1-\gamma\mu)^n\left\|x_t^0 - x_*^0\right\|^2 + \gamma^3 L n^2 \sigma_*^2 \sum_{j=0}^{n-1}(1-\gamma\mu)^j \\ &= (1-\gamma\mu)^n\left\|x_t - x_*\right\|^2 + \gamma^3 L n^2 \sigma_*^2 \sum_{j=0}^{n-1}(1-\gamma\mu)^j. \end{aligned}$$

Recursing again,

$$\begin{aligned} \left\|x_T - x_*\right\|^2 &\le (1-\gamma\mu)^{nT}\left\|x_0 - x_*\right\|^2 + \gamma^3 L n^2 \sigma_*^2 \sum_{j=0}^{n-1}(1-\gamma\mu)^j \sum_{t=0}^{T-1}(1-\gamma\mu)^{nt} \\ &= (1-\gamma\mu)^{nT}\left\|x_0 - x_*\right\|^2 + \gamma^3 L n^2 \sigma_*^2 \sum_{k=0}^{nT-1}(1-\gamma\mu)^k \\ &\le (1-\gamma\mu)^{nT}\left\|x_0 - x_*\right\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2}{\gamma\mu} \\ &= (1-\gamma\mu)^{nT}\left\|x_0 - x_*\right\|^2 + \gamma^2 \kappa n^2 \sigma_*^2. \end{aligned}$$

- **If f is μ-strongly convex and each f_i is convex**: the proof is identical to that of Theorem 2 but using Lemma 8 instead of Lemma 3, and we omit it for brevity.
- **If each f_i is convex**: the proof is identical to that of Theorem 3 but using Lemma 8 instead of Lemma 3, and we omit it for brevity.

- **If each f_i is possibly non-convex**: note that Lemma 4 also applies to IG without change, hence if $\gamma \leq \frac{1}{Ln}$ we have

$$f(x_{t+1}) \leq f(x_t) - \frac{\gamma n}{2}\|\nabla f(x_t)\|^2 + \frac{\gamma L^2}{2}\sum_{i=1}^{n}\left\|x_t - x_t^i\right\|^2.$$

We may then apply Lemma 9 to get for $\gamma \leq \frac{1}{2Ln}$

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{\gamma n}{2}\|\nabla f(x_t)\|^2 + \frac{\gamma L^2}{2}\left(4\gamma^2 n^3\|\nabla f(x_t)\|^2 + 4\gamma^2 n^3 \sigma_t^2\right) \\ &= f(x_t) - \frac{\gamma n}{2}\left(1 - 4\gamma^2 L^2 n^2\right)\|\nabla f(x_t)\|^2 + 2\gamma^3 L^2 n^3 \sigma_t^2. \end{aligned}$$

Using that $\gamma \leq \frac{1}{\sqrt{8}Ln}$ and subtracting f_* from both sides, we derive

$$f(x_{t+1}) - f_* \leq (f(x_t) - f_*) - \frac{\gamma n}{4}\|\nabla f(x_t)\|^2 + 2\gamma^3 L^2 n^3 \sigma_t^2.$$

Using Assumption 2, we get

$$f(x_{t+1}) - f_* \leq \left(1 + 4\gamma^3 L^2 A n^3\right)(f(x_t) - f_*) - \frac{\gamma n}{4}\|\nabla f(x_t)\|^2 + 2\gamma^3 L^2 n^3 B^2. \tag{60}$$

Applying Lemma 6 to (60), thus, gives

$$\min_{t=0,\ldots,T-1}\|\nabla f(x_t)\|^2 \leq \frac{4\left(1 + 4\gamma^3 L^2 A n^3\right)^T}{\gamma n T}(f(x_0) - f_*) + 8\gamma^2 L^2 n^2 B^2. \tag{61}$$

Note that by our assumption on the stepsize, $4\gamma^3 L^2 A n^3 T \leq 1$, hence,

$$\left(1 + 4\gamma^3 L^2 A n^3\right)^T \leq \exp\left(4\gamma^3 L^2 A n^3 T\right) \leq \exp(1) \leq 3.$$

It remains to use this in (61). ■