Proceedings of Machine Learning Research vol 125:1–35, 2020 33rd Annual Conference on Learning Theory

How Good is SGD with Random Shuffling?

Itay Safran ITAY.SAFRAN@WEIZMANN.AC.IL
Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study the performance of stochastic gradient descent (SGD) on smooth and strongly-convex finite-sum optimization problems. In contrast to the majority of existing theoretical works, which assume that individual functions are sampled with replacement, we focus here on popular but poorly-understood heuristics, which involve going over random permutations of the individual functions. This setting has been investigated in several recent works, but the optimal error rates remain unclear. In this paper, we provide lower bounds on the expected optimization error with these heuristics (using SGD with any constant step size), which elucidate their advantages and disadvantages. In particular, we prove that after k passes over n individual functions, if the functions are re-shuffled after every pass, the best possible optimization error for SGD is at least $\Omega\left(1/(nk)^2 + 1/nk^3\right)$, which partially corresponds to recently derived upper bounds. Moreover, if the functions are only shuffled once, then the lower bound increases to $\Omega(1/nk^2)$. Since there are strictly smaller upper bounds for repeated reshuffling, this proves an inherent performance gap between SGD with single shuffling and repeated shuffling. As a more minor contribution, we also provide a non-asymptotic $\Omega(1/k^2)$ lower bound (independent of n) for the incremental gradient method, when no random shuffling takes place. Finally, we provide an indication that our lower bounds are tight, by proving matching upper bounds for univariate quadratic functions.

1. Introduction

We consider variants of stochastic gradient descent (SGD) for solving unconstrained finite-sum problems of the form

$$\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \;, \tag{1}$$

where $\mathcal{X}$ is some Euclidean space $\mathbb{R}^d$ (or more generally some real Hilbert space), F is a strongly convex function, and each individual function f_i is smooth (with Lipschitz gradients) and Lipschitz on a bounded domain. Such problems are extremely common in machine learning applications, which often boil down to minimizing the average loss over n data points with respect to a class of predictors parameterized by a vector $\mathbf{x}$. When n is large, perhaps the most common approach to solve such problems is via stochastic gradient descent, which initializes at some point in $\mathcal{X}$ and involves iterations of the form $\mathbf{x}' := \mathbf{x}-\eta\nabla f_i(\mathbf{x})$, where η is a step size parameter and $i \in \{1,\ldots,n\}$. The majority of existing theoretical works assume that each i is sampled independently across iterations (also known as with-replacement sampling). For example, if it is chosen independently and uniformly at random from $\{1,\ldots,n\}$, then $\mathbb{E}_i[\nabla f_i(\mathbf{x})|\mathbf{x}] = \nabla F(\mathbf{x})$, so the algorithm can be seen as a noisy version of exact gradient descent on F (with iterations of the form $\mathbf{x}' := \mathbf{x} - \eta\nabla F(\mathbf{x})$), which greatly facilitates its analysis.

However, this straightforward sampling approach suffers from practical drawbacks, such as requiring truly random data access and hence longer runtime. In practice, it is quite common to use *without-replacement* sampling heuristics, which utilize the individual functions in some random or even deterministic order (see for example Bottou (2009, 2012); Nedić and Bertsekas (2001); Recht and Ré (2012); Shalev-Shwartz and Zhang (2013); Bertsekas and Scientific (2015); Feng et al. (2012)). Moreover, to get sufficiently high accuracy, it is common to perform several passes over the data, where each pass either uses the same order as the previous one, or some new random order. The different algorithmic variants we study in this paper are presented as Algorithms 1 to 4 below. We assume that all algorithms take as input the functions $f_1, \ldots, f_n$, a step size parameter $\eta > 0$ (which remains constant throughout the iterations), and an initialization point $\mathbf{x}_0$. The algorithms then perform k passes (which we will also refer to as epochs) over the individual functions, but differ in their sampling strategies:

- Algorithm 1 (SGD with random reshuffling) chooses a new permutation of the functions at the beginning of every epoch, and processes the individual functions in that order.
- Algorithm 2 (SGD with single shuffling) uses the same random permutation for all k epochs.
- Algorithm 3 (usually referred to as the incremental gradient method, see Bertsekas and Scientific (2015)) performs k passes over the individual functions, each in the same fixed order (which we will assume without loss of generality to be the canonical order $f_1, \ldots, f_n$)

In contrast, Algorithm 4 presents SGD using with-replacement sampling, where each iteration an individual function is chosen uniformly and independently. To facilitate our analysis, we let $\mathbf{x}_t$ in the pseudocode denote the iterate at the end of epoch t.

Algorithm 1 SGD with Random Reshuffling

```
x := x_0
for t = 1, ..., k do
    Sample a permutation σ(1), ..., σ(n) of {1, ..., n} uniformly at random
    for j = 1, ..., n do
        x := x − η∇f_σ(j)(x)
    end for
    x_t := x
end for
```

Algorithm 2 SGD with Single Shuffling

```
x := x_0
Sample a permutation σ(1), ..., σ(n) of {1, ..., n} uniformly at random
for t = 1, ..., k do
    for j = 1, ..., n do
        x := x − η∇f_σ(j)(x)
    end for
    x_t := x
end for
```

Algorithm 3 Incremental Gradient Method

$\mathbf{x} := \mathbf{x}_0$
for $t = 1, \ldots, k$ **do**
  for $j = 1, \ldots, n$ **do**
    $\mathbf{x} := \mathbf{x} - \eta \nabla f_j(\mathbf{x})$
  end for
  $\mathbf{x}_t := \mathbf{x}$
end for

Algorithm 4 SGD with Replacement

$\mathbf{x} := \mathbf{x}_0$
for $t = 1, \ldots, k$ **do**
  for $j = 1, \ldots, n$ **do**
    Sample $i \in \{1, \ldots, n\}$ uniformly
    $\mathbf{x} := \mathbf{x} - \eta \nabla f_i(\mathbf{x})$
  end for
  $\mathbf{x}_t := \mathbf{x}$
end for

	Random Reshuffling	Single Shuffling	Incremental	With Replacement
Upper Bound	$1/k^2$ [1] $1/n$ (for $k = 1$) [3] $1/(nk)^2 + 1/k^3$ [4] $1/nk^2$ [5] $\mathbf{1/(nk)^2 + 1/nk^3}$ (for $1d$ quad.)	$1/k^2$ [2] $1/n$ (for $k = 1$) [3] $\mathbf{1/nk^2}$ (for $1d$ quad.)	$1/k^2$ [2]	$1/nk$
Lower Bound	$1/n$ (for $k = 1$) [4] $\mathbf{1/(nk)^2 + 1/nk^3}$	$\mathbf{1/nk^2}$	$1/k^2$ ([2], asymptotic) $\mathbf{1/k^2}$ (non-asymptotic)	$1/nk$

[1] Gürbüzbalaban et al. (2015b).
[2] Gürbüzbalaban et al. (2015a).
[3] Shamir (2016).
[4] HaoChen and Sra (2018).
[5] Jain et al. (2019).

Table 1: Upper and lower bounds on the expected optimization error $\mathbb{E}[F(\mathbf{x}_k) - \inf_{\mathbf{x}} F(\mathbf{x})]$ for constant-step-size SGD with various sampling strategies, after k passes over n individual functions, in terms of n, k. Boldface letters refer to new results in this paper. We note that the upper bound of [4] additionally requires that the Hessian of each f_i is Lipschitz, and the upper bounds of [4] and [5] require k to be larger than a problem-dependent parameter (depending for example on the condition number). Also, the upper bound of [3] requires functions which are generalized linear functions. Our lower bounds apply under all such assumptions. As to our upper bounds, note that they apply only to univariate quadratic functions. Finally, we note that the upper bound of [5] is actually not on the optimization error for $\mathbf{x}_k$, but rather on a certain averaging of several iterates – see Remark 4 for a further discussion.

These without-replacement sampling heuristics are often easier and faster to implement in practice. In addition, when using random permutations, they often exhibit faster error decay than with-replacement SGD Bottou (2009). A common intuitive explanation for this phenomenon is that random permutations force the algorithm to touch each individual function exactly once during each epoch, whereas with-replacement makes the algorithm touch each function once only in expectation. However, theoretically analyzing these sampling heuristics has proven to be very challenging, since the individual iterations are no longer statistically independent.

In the past few years, some progress has been made in this front, and we summarize the known results on the expected optimization error (or at least what these results imply[1]), as well as our new results, in Table 1. First, we note that for SGD with replacement, classical results imply an optimization error of $\mathcal{O}(1/nk)$ after nk stochastic iterations, and this is known to be tight (see for example Nemirovski et al. (2009)). For SGD with random reshuffling, better bounds have been shown in recent years, generally implying that when the number of epochs k is sufficiently large, such sampling schemes are better than with-replacement sampling, with optimization error decaying as $1/k^2$ rather than $1/k$. However, the optimal dependencies on n, k and other problem-dependent parameters remain unclear (HaoChen and Sra (2018) show that for $k = 1$, one cannot hope to achieve worst-case error smaller than $\Omega(1/n)$, but for $k > 1$ not much is known). Some other recent theoretical works on SGD with random reshuffling (but under somewhat different settings) include Recht and Ré (2012); Ying et al. (2018). For the incremental gradient method, an $\mathcal{O}(1/k^2)$ upper bound was shown in Gürbüzbalaban et al. (2015a), as well as a matching asymptotic lower bound in terms of k. For SGD with single shuffling, we are actually not aware of a rigorous theoretical analysis. Thus, we only have the $\mathcal{O}(1/k^2)$ upper bound trivially implied by the analysis for the incremental gradient method, and for $k = 1$, the $\mathcal{O}(1/n)$ upper bound implied by the analysis for random reshuffling (since in that case there is no distinction between single shuffling and random reshuffling). Indeed, for single shuffling, even different epochs are not statistically independent, which makes the analysis particularly challenging.

In this paper, we focus on providing bounds on the expected optimization error of SGD with these sampling heuristics, which complement the existing upper bounds and provide further insights on the advantages and disadvantages of each. We focus on constant-step size SGD, as it simplifies our analysis, and existing upper bounds in the literature are derived in the same setting. Our contributions are as follows:

- For SGD with random reshuffling, we provide in Sec. 3 a lower bound of $\Omega(1/(nk)^2 + 1/nk^3)$. Interestingly, it seems to combine the "best" behaviors of previous upper bounds: It behaves as $1/n$ for a small constant number k of passes (which is optimal as discussed above), interpolating to $\mathcal{O}(1/(nk)^2)$ when k is large enough, and contains a term decaying cubically with k. Moreover, the proof construction applies already for univariate quadratics.

- For SGD with a single shuffling, we provide in Sec. 4 a lower bound of $\Omega(1/nk^2)$. Although we are not aware of a previous upper bound to compare to, this lower bound already proves an inherent performance gap compared to random reshuffling: Indeed, in the latter case there is an upper bound of $\mathcal{O}(1/(nk)^2 + 1/k^3)$, which is smaller than the $\Omega(1/nk^2)$ lower bound

1. For example, some of these papers focus on bounding $\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2]$ where $\mathbf{x}^*$ is the minimum of $F(\cdot)$, rather than the expected optimization error $\mathbb{E}[F(\mathbf{x}_k) - F(\mathbf{x}^*)]$. However, for strongly convex and smooth functions, $\|\mathbf{x}_k - \mathbf{x}^*\|^2$ and $F(\mathbf{x}_k) - F(\mathbf{x}^*)$ are the same up to the strong convexity and smoothness parameters, see for example Nesterov (2018).

for single shuffling when k is sufficiently large. This implies that the added computational effort of repeatedly reshuffling the functions can provably pay off in terms of the optimization error.

- For the incremental gradient method, we provide in Sec. 5 an $\Omega(1/k^2)$ lower bound. We note that a similar bound (at least asymptotically and for a certain n) is already implied by (Gürbüzbalaban et al., 2015a, Theorem 3.4). Our contribution here is to present a more explicit and non-asymptotic lower bound.
- In Sec. 6, we provide an indication that our lower bounds are tight, by proving matching upper bounds in the specific setting of univariate quadratic functions. We conjecture that these bounds also hold for multivariate quadratics, and perhaps even to general smooth and strongly convex functions. This is based on our matching lower bounds, as well as the fact that the bounds for with-replacement SGD are known to be tight already for univariate quadratics.

We note that in a very recent work (appearing after the initial publication of our work), Rajput et al. (2020) show an upper bound of $\mathcal{O}(1/nk^3 + 1/n^2k^2)$ for SGD with random reshuffling for multivariate quadratics, as well as a $\Omega(1/nk^2)$ lower bound for general convex functions. This validates that our lower bounds are tight for quadratics in the random reshuffling case.

2. Preliminaries

We let bold-face letters denote vectors. A twice-differentiable function f on $\mathbb{R}^d$ is λ-strongly convex, if its Hessian satisfies $\nabla^2 F(\mathbf{x}) \succeq \lambda I$ for all $\mathbf{x}$. f is quadratic if it is of the form $f(\mathbf{x}) = \mathbf{x}'^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$ for some matrix A, vector $\mathbf{b}$ and scalar c.

We consider finite-sum optimization problems as in Eq. (1), and our lower bound constructions hold under the following conditions (for some positive parameters G, λ):

Assumption 1 *$F(\mathbf{x})$ is a quadratic finite-sum function of the form $\frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ for some $n > 1$, which is λ-strongly convex. Each f_i is convex and quadratic and of the form $f_i(x) = ax^2 + bx$, has λ-Lipschitz gradients, and moreover, is G-Lipschitz for any $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}^*\| \leq 1$ where $\mathbf{x}^* = \arg\min F(\mathbf{x})$. Also, the algorithm is initialized at some $\mathbf{x}_0$ for which $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq 1$.*

Before continuing, we make a few remarks about the setting and our results:

Remark 1 (Constant Condition Number) *In the above assumption, λ plays a double role as both the gradient Lipschitz and strong convexity parameter. This entails that the condition number (defined as the quotient of the two) is constant, hence our lower bounds stem from inherent limitations of each sampling method and not from the constructions being ill-conditioned. We leave the problem of deriving lower bounds for general condition numbers to future work.*

Remark 2 (Unconstrained Optimization) *For simplicity, in this paper we consider unconstrained SGD, where the iterates are not explicitly constrained to lie in some subset of the domain. However, we note that existing upper bounds for SGD on strongly convex functions often assume an explicit projection on such a subset, in order to ensure that the gradients remain bounded. That being said, it is not difficult to verify that all our constructions – which have a very simple structure – are such that the iterates remain in a region with bounded gradients (with probability 1, at least for reasonably small step sizes), in which case projections will not significantly affect the results.*

Remark 3 (Distance from Optimum) *In Assumption 1, we fix the initial distance from the optimum to be at most* 1*, rather than keeping it as a variable parameter. Besides simplifying the constructions, we note that existing SGD upper bounds for strongly convex functions often do not explicitly depend on the initial distance (both for with-replacement SGD and with random reshuffling, see for example Nemirovski et al. (2009); Rakhlin et al. (2012); Jain et al. (2019)). Thus, it makes sense to study lower bounds in which the initial distance is fixed to be some constant.*

Remark 4 (Applicability of the Lower Bounds) *We emphasize that in our lower bounds, we focus on (a) SGD with constant step size, and (b) the expected performance of the iterate* $\mathbf{x}_k$ *after exactly* k *epochs. Thus, they do not formally cover step sizes which change across iterations, the performance of other iterates, or the performance of some average of the iterates. However, it is not clear that these are truly necessary to achieve optimal error bounds in our setting (indeed, many existing analyses do not require them), and we conjecture that our lower bounds cannot be substantially improved even with non-constant step sizes and iterate averaging schemes.*

3. SGD with Random Reshuffling

We begin by discussing SGD with random reshuffling, where at the beginning of every epoch we choose a new random order for processing the individual functions (Algorithm 1). Our main result is the following:

Theorem 5 *For any* $k \geq 1, n > 1$*, and positive* G, λ *such that* $G \geq 6\lambda$*, there exists a function* F *on* $\mathbb{R}$ *and an initialization point* x_0 *satisfying Assumption 1, such that for any step size* $\eta > 0$*,*

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \;\geq\; c \cdot \min\left\{\lambda\,,\; \frac{G^2}{\lambda}\left(\frac{1}{(nk)^2} + \frac{1}{nk^3}\right)\right\}\,,$$

where $c > 0$ *is a universal constant.*

We remark that the λ term seems unavoidable (at least in the univariate setting), as it is a trivial lower bound that holds by Assumption 1 for most points in the domain[2]. However, for nk large, this lower bound is

$$\Omega\left(\frac{G^2}{\lambda}\left(\frac{1}{(nk)^2} + \frac{1}{nk^3}\right)\right)\,.$$

It is useful to compare this bound to the existing optimal bound for SGD with replacement, which is

$$\Theta\left(\frac{G^2}{\lambda nk}\right)$$

(see for example Rakhlin et al. (2012)). First, we note that the G^2/λ factor is the same in both of them. The dependence on n, k though is different: For $k = 1$ or constant k, our lower bound is $\Omega(1/n)$, similar to the with-replacement case, but as k increases, it decreases cubically (rather than linearly) with k. This indicates that even for small k, random reshuffling is superior to with-replacement sampling, which agrees with empirical observations. For k very large ($k > n$), a phase transition occurs and the bound becomes $1/(nk)^2$ – that is, scaling down quadratically with the total

2. e.g. for small enough c and when considering a uniform distribution over all points in the domain.

number of individual stochastic iterations. That being said, it should be emphasized that $k > n$ is often an unrealistic regime, especially in large-scale problems where n is a huge number.

The proof of Thm. 5 appears in Appendix A.1. It is based on a set of very simple constructions, where $F(x) = \frac{\lambda}{2}x^2$, and the individual functions are all of the form $f_i(x) = a_i x^2 + b_i x$ for appropriate a_i, b_i. This allows us to write down the iterates $x_1, x_2, \ldots$ at the end of each epoch in closed form. The analysis then carefully tracks the decay of $\mathbb{E}[x_t^2]$ after each epoch, showing that it cannot decay to 0 too rapidly, hence implying a lower bound on $\mathbb{E}[F(x_k)]$ after k epochs. The main challenge is that unlike SGD with replacement, here the stochastic iterations in each epoch are not independent, so computing these expectations is not easy. To make it tractable, we identify two distinct sources contributing to the error in each epoch: A "bias" term, which captures the fact that the stochastic gradients at each epoch are statistically correlated, hence for a given iterate $\mathbf{x}$ during the algorithm's run, $\mathbb{E}[\nabla f_{\sigma(j)}(\mathbf{x})|\mathbf{x}] \neq \nabla F(\mathbf{x})$ (unlike the with-replacement case where equality holds), and a "variance" term, which captures the inherent noise in the stochastic sampling process. For different parameter regimes, we use different constructions and focus on either the bias or the variance component (which when studied in isolation are more tractable), and then combine the various bounds into the final lower bound appearing in Thm. 5.

We finish with the following remark about a possible extension of the lower bound:

Remark 6 (Convex Functions) *By allowing λ to decay to 0 at a rate governed by k (as well as the remaining problem parameters), we may consider the setting of convex functions which are not necessarily strongly convex (since that for large enough k, there exists no $c > 0$ such that $\lambda \geq c$). In such a regime, Thm. 5 seems to suggest a lower bound (in terms of n, k) of*

$$\Omega\left(G\sqrt{\frac{1}{(nk)^2} + \frac{1}{nk^3}}\right) \;=\; \Omega\left(G\left(\frac{1}{\sqrt{nk^3}} + \frac{1}{nk}\right)\right) \;,$$

since in this scenario we can set λ arbitrarily small, and in particular as $G\sqrt{1/(nk)^2 + 1/nk^3}$ so as to maximize the lower bound in Thm. 5. In contrast, Jain et al. (2019) shows a $\mathcal{O}(1/\sqrt{nk})$ upper bound in this setting for SGD with random reshuffling, and a similar upper bound hold for SGD with replacement. A similar argument can also be applied to the other lower bounds in our paper, extending them from the strongly convex to the convex case. However, we emphasize that some caution is needed, since our lower bounds do not quantify a dependence on the radius of the domain, which is usually explicit in bounds for this setting. We leave the task of proving a lower bound in the general convex case to future work.

4. SGD with a Single Shuffling

We now turn to the case of SGD where a single random order over the individual functions is chosen at the beginning, and the algorithm then cycles over the individual functions using that order (Algorithm 2). Our main result here is the following:

Theorem 7 *For any $k \geq 1, n > 1$, and positive G, λ such that $G \geq 6\lambda$, there exists a function F on $\mathbb{R}$ and an initialization point x_0 satisfying Assumption 1, such that for any step size $\eta > 0$,*

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \;\geq\; c \cdot \min\left\{\lambda\,,\; \frac{G^2}{\lambda n k^2}\right\}\;,$$

where $c > 0$ is a universal constant.

The proof appears in Appendix A.2. In the single shuffling case, we are not aware of a previously known upper bound to compare to (except the $\mathcal{O}(1/k^2)$ bound for the incremental gradient method below, which trivially applies also to SGD with single shuffling). However, the lower bound already implies an interesting separation between single shuffling and random reshuffling: In the former case, $\Omega(1/nk^2)$ is the best we can hope to achieve, whereas in the latter case, we have seen upper bounds which are strictly better when k is sufficiently large (i.e., $\mathcal{O}(1/(nk)^2)$). To the best of our knowledge, this is the first formal separation between these two shuffling schemes for SGD: It implies that the added computational effort of repeatedly reshuffling the functions can provably pay off in terms of the optimization error. It would be quite interesting to understand whether this separation might also occur for smaller values of k as well, which is definitely true if our $\Omega(1/(nk)^2 + 1/nk^3)$ lower bound for random reshuffling is tight. It would also be interesting to derive a good upper bound for SGD with single shuffling, which is a common heuristic (indeed, we prove such a bound in Sec. 6, but only for univariate quadratics).

5. Incremental Gradient Method

Next, we turn to discuss the incremental gradient method, where the individual functions are cycled over in a fixed deterministic order. We note that for this algorithm, an $\Omega(1/k^2)$ lower bound was already proven in Gürbüzbalaban et al. (2015a), but in an asymptotic form, and only for $n = 2$. Our contribution here is to provide an explicit, non-asymptotic bound:

Theorem 8 *For any $k \geq 1, n > 1$, and positive G, λ such that $G \geq 6\lambda$, there exists a function F on $\mathbb{R}$ and an initialization point x_0 satisfying Assumption 1, such that if we run the incremental gradient method for k epochs with any step size $\eta > 0$, then*

$$F(x_k) - \inf_x F(x) \;\geq\; c \cdot \min\left\{\lambda\,,\; \frac{G^2}{\lambda k^2}\right\}$$

where $c > 0$ is a universal constant.

The proof (which follows a strategy broadly similar to Thm. 5) appears in Appendix A.3. Comparing this theorem with our other lower bounds and the associated upper bounds, it is clear that there is a high price to pay (in a worst-case sense) for using a fixed, non-random order, as the bound does not improve at all with more individual functions n. Indeed, recalling that the bound for with-replacement SGD is $\mathcal{O}(G^2/\lambda nk)$, it follows that incremental gradient method can beat with-replacement SGD only when $\frac{G^2}{\lambda k^2} \leq \frac{G^2}{\lambda nk}$, or $k \geq n$. For large-scale problems where n is big, this is often an unrealistically large value of k.

6. Tight Upper Bounds for One-Dimensional Quadratics

As discussed in the introduction, for SGD with random reshuffling and single shuffling, there is a gap between the lower bounds we present here, and known upper bounds in the literature. In this section, we provide an indication that our lower bounds are tight, by proving matching upper bounds (up to log factors) for the setting of univariate quadratic functions[3]. Although this is a

3. I.e., $x \mapsto ax^2 + bx$. Note that for simplicity, we assume no constant term c as in $ax^2 + bx + c$, as it plays no role in the optimization process.

special case, we note that the standard $\Theta(1/nk)$ bounds for SGD with replacement on strongly convex functions are known to be tight already for univariate quadratics. This leads us to conjecture that even for without-replacement sampling schemes, the optimal rates for univariate quadratics are also the optimal rates for general strongly convex functions.

Before stating our upper bounds, we make the following assumption on the target functions f_i:

Assumption 2 *$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ is λ-strongly convex. Moreover, each $f_i(x) = \frac{a_i}{2}x^2 - b_i x$ is convex, has L-Lipschitz gradients, and satisfies $|f_i'(x^*)| \le G$ where $x^* = \arg\min_x F(x)$.*

For the single shuffling case we have the following theorem:

Theorem 9 *Let $F(x) := \frac{\lambda}{2}x^2 - bx = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, where $f_i(x) = \frac{1}{2}a_i x^2 - b_i x$ satisfy Assumption 2, and assume that $\frac{L}{\lambda} \le \frac{nk}{\log(n^{0.5}k)}$. Then single shuffling SGD with a fixed step size of $\eta = \frac{\log\left(n^{0.5}k\right)}{\lambda nk}$ satisfies*[4]

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \le \tilde{\mathcal{O}}\left(\frac{\lambda}{nk^2}(x_0 - x^*)^2 + \frac{G^2L^2}{\lambda^3 nk^2}\right),$$

where the expectation is taken over drawing a permutation $\sigma : [n] \to [n]$ uniformly at random, and the big O tilde notation hides a universal constant and factors poly-logarithmic in n and k.

For SGD with random reshuffling, we present the following theorem:

Theorem 10 *Let $F(x) := \frac{\lambda}{2}x^2 - bx = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, where $f_i(x) = \frac{1}{2}a_i x^2 - b_i x$ satisfy Assumption 2, and assume that $\frac{L}{\lambda} \le \frac{k}{2\log(nk)}$. Then random shuffling SGD with a fixed step size of $\eta = \frac{\log(nk)}{\lambda nk}$ satisfies*[5]

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \le \tilde{\mathcal{O}}\left(\frac{\lambda}{n^2k^2}(x_0 - x^*)^2 + \frac{G^2L^2}{\lambda^3}\left(\frac{1}{n^2k^2} + \frac{1}{nk^3}\right)\right),$$

where the expectation is taken over drawing k permutations $\sigma_i : [n] \to [n]$ uniformly at random, and the big O tilde notation hides a universal constant and factors poly-logarithmic in n and k.

The formal proofs appear in Appendix A.

It is easy to verify that these upper bounds match our lower bounds in Theorems 5 and 7 in terms of the dependence on n, k. Moreover, our requirement of $k \ge \tilde{\Omega}(\kappa)$ (recall that $\kappa := L/\lambda$) for random reshuffling is also made in HaoChen and Sra (2018). As to the other parameters, it is important to note that our lower bound constructions (which also utilize univariate quadratics) are in a regime where both L/λ and $x_0 - x^*$ are constants, and they match the upper bounds in this case. In particular, Thm. 9 then reduces to $\tilde{\mathcal{O}}\left(\frac{\lambda}{nk^2} + \frac{G^2}{\lambda nk^2}\right)$, which is $\tilde{\mathcal{O}}(\frac{G^2}{\lambda nk^2})$ under the assumption $G \ge 6\lambda$ which we make in the lower bound. Similarly, Thm. 10 reduces to

$$\tilde{\mathcal{O}}\left(\frac{\lambda}{n^2k^2} + \frac{G^2}{\lambda}\left(\frac{1}{n^2k^2} + \frac{1}{nk^3}\right)\right) = \tilde{\mathcal{O}}\left(\frac{G^2}{\lambda}\left(\frac{1}{n^2k^2} + \frac{1}{nk^3}\right)\right)$$

4. Letting $\kappa := L/\lambda$ denote the condition number, the second term in the right hand side can equivalently be written as $\frac{G^2\kappa^2}{\lambda nk^3}$.

5. Similarly to the above footnote, the second term in the right hand side can equivalently be written as $\frac{G^2\kappa^2}{\lambda}\left(\frac{1}{n^2k^2} + \frac{1}{nk^3}\right)$.

if $G \geq 6\lambda$. We leave the problem of getting matching upper and lower bounds in all parameter regimes of G, L, λ to future work.

While the assumption of univariate quadratics is restrictive, our main purpose here is to indicate the potential tightness of our lower bounds, and elucidate how without-replacement sampling can lead to faster convergence in a simple setting. Our proof is based on evaluating a closed-form expression for the iterate at the k-th epoch, splitting deteministic and stochastic terms, and then carefully bounding the stochastic terms using a Hoeffding-Serfling type inequality and the deterministic term using the AM-GM inequality.

We conjecture that our upper bounds can be generalized to general quadratic functions, and perhaps even to general smooth and strongly convex functions. The main technical barrier is that our proof crucially uses the commutativity of the scalar-valued a_i's. Once we deal with matrices, we essentially require (a special case of) a matrix-valued arithmetic-geometric mean inequality studied in Recht and Ré (2012) (See Eq. (20) for the part of the proof where we require this inequality). Unfortunately, as of today this conjectured inequality is not known to hold except in extremely special cases.

References

Rémi Bardenet, Odalric-Ambrym Maillard, et al. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

Dimitri P Bertsekas and Athena Scientific. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

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

L. Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*. Springer, 2012.

X. Feng, A. Kumar, B. Recht, and C. Ré. Towards a unified architecture for in-rdbms analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 325–336. ACM, 2012.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Convergence rate of incremental gradient and Newton methods. *arXiv preprint arXiv:1510.08562*, 2015a.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Why random reshuffling beats stochastic gradient descent. *arXiv preprint arXiv:1510.08560*, 2015b.

Jeffery Z HaoChen and Suvrit Sra. Random shuffling beats sgd after finite epochs. *arXiv preprint arXiv:1806.10077*, 2018.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. *arXiv preprint arXiv:1903.01463*, 2019.

A. Nedić and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic optimization: algorithms and applications*, pages 223–264. Springer, 2001.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

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

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

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*, pages 1571–1578. Omnipress, 2012.

B. Recht and C. Ré. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *COLT*, 2012.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in neural information processing systems*, pages 46–54, 2016.

Bicheng Ying, Kun Yuan, Stefan Vlaski, and Ali H Sayed. Stochastic learning under random reshuffling with constant step-sizes. *IEEE Transactions on Signal Processing*, 67(2):474–489, 2018.

Appendix A. Proofs

A.1. Proof of Thm. 5

For simplicity, we will prove the theorem assuming the number of components n in our function is an even number. This is without loss of generality, since if $n > 1$ is odd, let $F_{n-1}(\mathbf{x}) = \frac{1}{n-1}\sum_{i=1}^{n-1} f_i(\mathbf{x})$ be the function achieving the lower bound using an even number $n-1$ of components, and define $F(\mathbf{x}) = \frac{1}{n}\left(\sum_{i=1}^{n-1} f_i(\mathbf{x}) + f_n(\mathbf{x})\right)$ where $f_n(\mathbf{x}) := 0$. $F()$ has the same Lipschitz parameter G as $F_{n-1}()$, and a strong convexity parameter λ smaller than that of $F_{n-1}()$ by a $\frac{n}{n+1}$ factor which is always in $[\frac{3}{4}, 1]$. Moreover, it is easy to see that for a fixed step size, the distribution of the iterates after k epochs is the same over $F()$ and $F_{n-1}()$, since SGD does not move on any iteration where f_n is chosen. Therefore, the lower bound on F_{n-1} translates to a lower bound on $F()$ up to a small factor which can be absorbed into the numerical constants. Thus, in what follows, we will assume that n is even and that $G \geq 4\lambda$, whereas in the theorem statement we make the slightly stronger assumption $G \geq 6\lambda$ so that the reduction described above will be valid.

The proof of the theorem is based on the following three propositions, each using a somewhat different construction and analysis:

Proposition 11 *For any even n and any positive G, λ such that $G \geq 2\lambda$, there exists a function F on $\mathbb{R}$ satisfying Assumption 1, such that for any step size $\eta > 0$,*

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \;\geq\; c \cdot \min\left\{\lambda\,,\; \frac{G^2}{\lambda n k^3}\right\}$$

where $c > 0$ is a universal constant.

Proposition 12 *Suppose that $k \geq n$ and that n is even. For any positive G, λ such that $G \geq 2\lambda$, there exists a function F on $\mathbb{R}$ satisfying Assumption 1, such that for any step size $\eta \geq \frac{1}{100\lambda n^2}$,*

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \;\geq\; c \cdot \frac{G^2}{\lambda(nk)^2}$$

where $c > 0$ is a numerical constant.

Proposition 13 *Suppose $k > 1$ and that n is even. For any positive G, λ such that $G \geq 4\lambda$, there exists a function F on $\mathbb{R}$ satisfying Assumption 1, such that for any step size $\eta \leq \frac{1}{100\lambda n^2}$,*

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \;\geq\; c \cdot \min\left\{\lambda \,,\, \frac{G^2}{\lambda(nk)^2}\right\}$$

where $c > 0$ is a numerical constant.

The proof of each proposition appears below, but let us first show how combining these implies our theorem. We consider two cases:

- If $k \leq n$, then $\frac{1}{nk^3} \geq \frac{1}{(nk)^2}$, so by Proposition 11,

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \;\geq\; c \cdot \min\left\{\lambda \,,\, \frac{G^2}{\lambda nk^3}\right\} \;\geq\; c \cdot \min\left\{\lambda \,,\, \frac{G^2}{2\lambda}\left(\frac{1}{(nk)^2} + \frac{1}{nk^3}\right)\right\} \,.$$

- If $k \geq n$ (which implies $k > 1$ since n is even), we have $\frac{1}{nk^3} \leq \frac{1}{(nk)^2}$, and by combining Proposition 12 and Proposition 13 (which together cover any positive step size),

$$\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \;\geq\; c \cdot \min\left\{\lambda \,,\, \frac{G^2}{\lambda(nk)^2}\right\} \;\geq\; c \cdot \min\left\{\lambda \,,\, \frac{G^2}{2\lambda}\left(\frac{1}{(nk)^2} + \frac{1}{nk^3}\right)\right\}$$

Thus, in any case we get $\mathbb{E}\left[F(x_k) - \inf_x F(x)\right] \geq c \cdot \min\left\{\lambda \,,\, \frac{G^2}{2\lambda}\left(\frac{1}{(nk)^2} + \frac{1}{nk^3}\right)\right\}$, from which the result follows.

A.1.1. PROOF OF PROPOSITION 11

We will need the following key technical lemma, whose proof (which is rather long and technical) appears in Appendix B:

Lemma 14 *Let $\sigma_0, \ldots, \sigma_{n-1}$ (for even n) be a random permutation of $(1, 1, \ldots, 1, -1, -1. \ldots, -1)$ (where both 1 and -1 appear exactly $n/2$ times). Then there is a numerical constant $c > 0$, such that for any $\alpha > 0$,*

$$\mathbb{E}\left[\left(\sum_{i=0}^{n-1} \sigma_i (1-\alpha)^i\right)^2\right] \;\geq\; c \cdot \min\left\{1 + \frac{1}{\alpha} \,,\, n^3\alpha^2\right\}$$

Let G, λ, n be fixed (assuming $G \geq 2\lambda$ and n is even). We will use the following function:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \;=\; \frac{\lambda}{2}x^2 \;,$$

where $\inf_x F(x) = 0$, and

$$f_i(x) = \begin{cases} \frac{\lambda}{2}x^2 + \frac{G}{2}x & i \leq \frac{n}{2} \\ \frac{\lambda}{2}x^2 - \frac{G}{2}x & i > \frac{n}{2} \end{cases} \;. \tag{2}$$

Also, we assume that the algorithm is initialized at $x_0 = 1$. On this function, we have that during any single epoch, we perform n iterations of the form

$$x_{new} = (1-\eta\lambda)x_{old} + \frac{\eta G}{2}\sigma_i,$$

where $\sigma_0, \ldots, \sigma_{n-1}$ are a random permutation of $\frac{n}{2}$ 1's and $\frac{n}{2}$ -1's. Repeatedly applying this inequality, we get that after n iterations, the relationship between the first and last iterates in the epoch satisfy

$$\begin{aligned} x_{t+1} \;&=\; (1-\eta\lambda)^n x_t + \frac{\eta G}{2}\sum_{i=0}^{n-1}\sigma_i(1-\eta\lambda)^{n-i-1} \\ &=\; (1-\eta\lambda)^n x_t + \frac{\eta G}{2}\sum_{i=0}^{n-1}\sigma_i(1-\eta\lambda)^{i} \;. \end{aligned} \tag{3}$$

(in the last equality, we used the fact that $\sigma_1, \ldots, \sigma_n$ are exchangeable). Using this and the fact that $\mathbb{E}[\sigma_i] = 0$, we get that

$$\mathbb{E}[x_{t+1}^2] \;=\; (1-\eta\lambda)^{2n}\mathbb{E}[x_t^2] + \left(\frac{\eta G}{2}\right)^2 \cdot \beta_{n,\eta,\lambda} \;, \tag{4}$$

where

$$\beta_{n,\eta,\lambda} \;:=\; \mathbb{E}\left[\left(\sum_{i=0}^{n-1}\sigma_i(1-\lambda\eta)^{n-i-1}\right)^2\right] \;=\; \mathbb{E}\left[\left(\sum_{i=0}^{n-1}\sigma_i(1-\lambda\eta)^{i}\right)^2\right]. \tag{5}$$

Note that if $\lambda\eta \geq 1$, then by Lemma 14, $\beta_{n,\eta,\lambda} \geq c$ for some positive constant c, and we get that

$$\mathbb{E}[x_{t+1}^2] \;\geq\; \left(\frac{\eta G}{2}\right)^2 \cdot c \;\geq\; \left(\frac{G}{2\lambda}\right)^2 \cdot c$$

for all t, and therefore $\mathbb{E}[F(x_k)] = \frac{\lambda}{2}\mathbb{E}[x_k^2] \geq c\frac{G^2}{8\lambda} \geq c\frac{G^2}{8\lambda n k^3}$, so the proposition we wish to prove holds. Thus, we will assume from now on that $\lambda\eta < 1$.

With this assumption, repeatedly applying Eq. (4) and recalling that $x_0 = 1$, we have

$$\begin{aligned} \mathbb{E}[x_k^2] \;&\geq\; (1-\eta\lambda)^{2nk} + \left(\frac{\eta G}{2}\right)^2 \cdot \beta_{n,\eta,\lambda}\sum_{t=0}^{k-1}(1-\eta\lambda)^{2nt} \\ &=\; (1-\eta\lambda)^{2nk} + \left(\frac{\eta G}{2}\right)^2 \cdot \beta_{n,\eta,\lambda} \cdot \frac{1-(1-\eta\lambda)^{2nk}}{1-(1-\eta\lambda)^{2n}} \;. \end{aligned} \tag{6}$$

We now consider a few cases (recalling that the case $\eta\lambda \geq 1$ was already treated earlier):

- If $\eta\lambda \leq \frac{1}{2nk}$, then we have

$$\mathbb{E}[x_k^2] \;\geq\; (1-\eta\lambda)^{2nk} \;\geq\; \left(1-\frac{1}{2nk}\right)^{2nk} \geq \frac{1}{4}$$

  for all n, k.

- If $\eta\lambda \in \left(\frac{1}{2nk}, \frac{1}{2n}\right)$ then by Bernoulli's inequality, we have $1 \geq (1-\eta\lambda)^{2n} \geq 1 - 2n\eta\lambda > 0$, and therefore, by Eq. (6)

$$\mathbb{E}[x_k^2] \;\geq\; \frac{\eta^2 G^2 \beta_{n,\eta,\lambda}(1-(1-1/2nk)^{2nk})}{4(1-(1-2n\eta\lambda))} \;\geq\; \frac{\eta G^2 \beta_{n,\eta,\lambda}(1-\exp(-1))}{8\lambda n}\,.$$

  Plugging in Lemma 14 and simplifying a bit, this is at least

$$\frac{c\eta G^2}{\lambda n}\cdot \min\left\{\frac{1}{\eta\lambda}, n^3(\eta\lambda)^2\right\} \;=\; \frac{c\eta G^2}{\lambda n}\cdot n^3(\eta\lambda)^2 \;=\; c\eta^3\lambda n^2 G^2$$

  for some numerical constant $c > 0$. Using the assumption that $\eta\lambda \geq \frac{1}{2nk}$ (which implies $\eta \geq \frac{1}{2\lambda nk}$), this is at least

$$\frac{c}{8}\cdot\frac{G^2}{\lambda^2 nk^3}\,.$$

- If $\eta\lambda \in \left[\frac{1}{2n}, 1\right)$, then $\frac{1-(1-\eta\lambda)^{2nk}}{1-(1-\eta\lambda)^{2n}}$ is at least some numerical constant $c > 0$, so Eq. (6) implies

$$\mathbb{E}[x_k^2] \;\geq\; c\left(\frac{\eta G}{2}\right)^2\cdot\beta_{n,\eta,\lambda}\,.$$

  By Lemma 14, this is at least

$$c'\left(\frac{\eta G}{2}\right)^2\cdot\min\left\{1+\frac{1}{\eta\lambda}\;,\; n^3(\eta\lambda)^2\right\} \;=\; c'\left(\frac{\eta G}{2}\right)^2\left(1+\frac{1}{\eta\lambda}\right) \;\geq\; \frac{c'\eta G^2}{4\lambda}$$

  Since $\eta \geq \frac{1}{2\lambda n}$, this is at least

$$\frac{c'G^2}{8\lambda^2 n} \;\geq\; \frac{c'G^2}{8\lambda^2 nk^3}\,.$$

Combining all the cases, we get overall that

$$\mathbb{E}[x_k^2] \;\geq\; c\cdot\min\left\{1, \frac{G^2}{\lambda^2 nk^3}\right\}$$

for some numerical constant $c > 0$. Noting that $\mathbb{E}[F(x_k)] = \mathbb{E}\left[\frac{\lambda}{2}x_k^2\right] = \frac{\lambda}{2}\mathbb{E}\left[x_k^2\right]$ and combining with the above, the result follows.

A.1.2. PROOF OF PROPOSITION 12

We use the same construction as in the proof of Proposition 11, where $F(x) = \frac{\lambda}{2}x^2$, and leading to Eq. (6), namely

$$\mathbb{E}[x_k^2] \;\geq\; (1-\eta\lambda)^{2nk} + \left(\frac{\eta G}{2}\right)^2 \cdot \beta_{n,\eta,\lambda} \cdot \frac{1-(1-\eta\lambda)^{2nk}}{1-(1-\eta\lambda)^{2n}} \;, \tag{7}$$

where $\beta_{n,\eta,\lambda} = \mathbb{E}\left[\left(\sum_{i=0}^{n-1}\sigma_i(1-\lambda\eta)^i\right)^2\right]$, $\sigma_0,\ldots,\sigma_n$ are a random permutation of $\frac{n}{2}$ 1's and $\frac{n}{2}$ -1's.

As in the proof of Proposition 11, we consider several regimes of $\eta\lambda$. In the same manner as in that proof, it is easy to verify that when $\eta\lambda > 1$ or $\eta\lambda \leq \frac{1}{2nk}$, then $\mathbb{E}[x_k^2]$ is at least a positive constant (hence $\mathbb{E}[F(x_k)] \geq \Omega(\lambda)$) , and when $\eta\lambda \in \left[\frac{1}{2n}, 1\right)$, $\mathbb{E}[x_k^2] \geq \frac{c' G^2}{2\lambda^2 n}$ for a numerical constant $c' > 0$ (hence $\mathbb{E}[F(x_k)] \geq \Omega(G^2/\lambda n)$). In both these cases, the statement in our proposition follows, so it is enough to consider the regime $\eta\lambda \in \left(\frac{1}{2nk}, \frac{1}{2n}\right)$.

In this regime, by Bernoulli's inequality, we have $0 < 1-(1-\eta\lambda)^{2n} \leq 1-(1-2n\eta\lambda) = 2n\eta\lambda$, so we can lower bound Eq. (7) by

$$\left(\frac{\eta G}{2}\right)^2 \cdot \beta_{n,\eta,\lambda}\frac{1-(1-\eta\lambda)^{2nk}}{2n\eta\lambda} \;=\; \frac{\eta G^2 \beta_{n,\eta,\lambda}(1-(1-\eta\lambda)^{2nk})}{8\lambda n} \;.$$

Since we assume $\eta\lambda \geq \frac{1}{2nk}$, it follows that $1-(1-\eta\lambda)^{2nk} \geq 1-(1-1/2nk)^{2nk} \geq c$ for some positive $c > 0$. Plugging this and the bound for $\beta_{n,\eta,\lambda}$ from Lemma 14, the displayed equation above is at least

$$\frac{c\eta G^2}{8\lambda n} \cdot \min\left\{\frac{1}{\eta\lambda}, n^3(\eta\lambda)^2\right\} \;=\; \frac{c\eta G^2}{8\lambda n} \cdot n^3(\eta\lambda)^2 \;=\; \frac{c}{8}G^2\lambda\eta^3 n^2 \;.$$

Since we assume $\eta \geq \frac{1}{100\lambda n^2}$, this is at least

$$c' \cdot \frac{G^2}{\lambda^2 n^4}$$

for some numerical $c' > 0$. Since we assume that $k \geq n$, this is at least $c' \cdot \frac{G^2}{\lambda^2 (nk)^2}$. Noting that $\mathbb{E}[F(x_k)] = \mathbb{E}\left[\frac{\lambda}{2}x_k^2\right] = \frac{\lambda}{2}\mathbb{E}\left[x_k^2\right]$ and combining with the above, the result follows.

A.1.3. PROOF OF PROPOSITION 13

To simplify some of the notation, we will prove the result for a function which is $\lambda/2$-strongly convex (rather than λ-strongly convex), assuming $G \geq 2\lambda$, and notice that this only affects the universal constant c in the bound. Specifically, we use the following function:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \;=\; \frac{\lambda}{4}x^2 \;,$$

where $\inf_x F(x) = 0$, and

$$f_i(x) = \begin{cases} \frac{\lambda}{2}x^2 + \frac{G}{2}x & i \leq \frac{n}{2} \\ -\frac{G}{2}x & i > \frac{n}{2} \end{cases} \;.$$

Also, we assume that the algorithm is initialized at $x_0 = -1$. On this function, we have that during any single epoch, we perform n iterations of the form

$$x_{new} = (1-\eta\lambda\sigma_i)x_{old} + \frac{\eta G}{2}(1-2\sigma_i),$$

where $\sigma_0, \ldots, \sigma_{n-1}$ are a random permutation of $\frac{n}{2}$ 1's and $\frac{n}{2}$ 0's. Repeatedly applying this equation, we get that after n iterations, the relationship between the iterates x_t and x_{t+1} is

$$x_{t+1} \;=\; x_t \cdot \prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i) + \frac{\eta G}{2}\sum_{i=0}^{n-1}(1-2\sigma_i)\prod_{j=i+1}^{n-1}(1-\eta\lambda\sigma_j) \tag{8}$$

As a result, and using the fact that $\sigma_1, \ldots, \sigma_n$ are independent of x_t and in $\{0,1\}$, we have

$$\begin{aligned}\mathbb{E}[x_{t+1}^2] \;&\geq\; \mathbb{E}\left[x_t^2 \cdot \prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i)^2\right] + \eta G \cdot \mathbb{E}\left[x_t\left(\prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i)\right)\left(\sum_{i=0}^{n-1}(1-2\sigma_i)\prod_{j=i+1}^{n-1}(1-\eta\lambda\sigma_j)\right)\right]\\ &\geq\; (1-\eta\lambda)^{2n}\cdot\mathbb{E}[x_t^2] + \eta G\cdot\mathbb{E}[x_t]\cdot\mathbb{E}\left[\left(\prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i)\right)\left(\sum_{i=0}^{n-1}(1-2\sigma_i)\prod_{j=i+1}^{n-1}(1-\eta\lambda\sigma_j)\right)\right]\end{aligned} \tag{9}$$

We now wish to use Lemma 19 from Appendix C, in order to replace the products in the expression above by sums. To that end, and in order to simplify the notation, define

$$A := \prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i)\;,\;\; B_i := \prod_{j=i+1}^{n-1}(1-\eta\lambda\sigma_j)\;,\;\; \tilde{A} := 1-\eta\lambda\sum_{i=1}^{n}\sigma_i = 1-\frac{\eta\lambda n}{2}\;,\;\; \tilde{B}_i := 1-\eta\lambda\sum_{j=i+1}^{n}\sigma_i\;, \tag{10}$$

and note that by Lemma 19,

$$A\sum_{i=0}^{n-1}(1-2\sigma_i)B_i \;\leq\; \left(\tilde{A}\pm 2\left(\eta\lambda\sum_{i=0}^{n-1}\sigma_i\right)^2\right)\left(\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i \pm 2\sum_{i=0}^{n-1}\left(\eta\lambda\sum_{j=i+1}^{n-1}\sigma_j\right)^2\right)\,, \tag{11}$$

where $\pm$ is taken to be either plus or minus depending on the sign of $\tilde{A}$ and $\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i$, to make the inequality valid (we note that eventually we will show that these terms are relatively negligible). Opening the product, and using the deterministic upper bounds

$$|\tilde{A}| \leq 1\;,\quad \left(\eta\lambda\sum_{i=0}^{n-1}\sigma_i\right)^2 \leq (\eta\lambda n)^2 \tag{12}$$

and

$$\left|\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i\right| \leq n\;,\quad \sum_{i=0}^{n-1}\left(\eta\lambda\sum_{j=i+1}^{n-1}\sigma_j\right)^2 \;\leq\; n(\eta\lambda n)^2 \;\leq\; \frac{1}{10^4 n}\,, \tag{13}$$

(which follow from the assumption that $\eta \leq \frac{1}{100\lambda n^2}$), we can upper bound Eq. (11) by

$$\tilde{A}\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i + 2(\eta\lambda n)^2 \cdot \left(n + \frac{2}{100n}\right) + n(\eta\lambda n)^2 \overset{(*)}{\leq} \tilde{A}\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i + \frac{301}{100}(\eta\lambda)^2 n^3 ,$$

where in $(*)$ we used the fact that $n \geq 2$ and therefore $n + \frac{2}{100n} \leq n + \frac{1}{100} \leq (1+\frac{1}{200})n$. Substituting back the definitions of $\tilde{A}, \tilde{B}$ and plugging back into Eq. (11), we get that

$$\begin{aligned}
&\mathbb{E}\left[\left(\prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i)\right)\left(\sum_{i=0}^{n-1}(1-2\sigma_i)\prod_{j=i+1}^{n-1}(1-\eta\lambda\sigma_j)\right)\right] \\
&\leq \left(1-\frac{\eta\lambda n}{2}\right)\cdot \mathbb{E}\left[\left(\sum_{i=0}^{n-1}(1-2\sigma_i)(1-\eta\lambda\sum_{j=i+1}^{n}\sigma_j)\right)\right] + \frac{301}{100}(\eta\lambda)^2 n^3 \\
&\overset{(*)}{\leq} \eta\lambda n\left(-\left(1-\frac{\eta\lambda n}{2}\right)\frac{n+1}{4(n-1)} + \frac{301}{100}\eta\lambda n^2\right) ,
\end{aligned}$$

where $(*)$ is by Lemma 17. Using the assumptions that $\eta \leq \frac{1}{100\lambda n^2}$ (hence $\eta\lambda n \leq \eta\lambda n^2 \leq \frac{1}{100}$) and $n \geq 2$, this is at most $-c\eta\lambda n$ for a numerical constant $c > 0.2$. Summarizing this part of the proof, we have shown that

$$\mathbb{E}\left[\left(\prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i)\right)\left(\sum_{i=0}^{n-1}(1-2\sigma_i)\prod_{j=i+1}^{n-1}(1-\eta\lambda\sigma_j)\right)\right] \leq -c\eta\lambda n . \tag{14}$$

Next, we turn to analyze the $\mathbb{E}[x_t]$ term in Eq. (9). By Eq. (8), and the fact that σ_i is independent of x_t, we have

$$\mathbb{E}[x_{t+1}] = \mathbb{E}[x_t]\cdot\mathbb{E}\left[\prod_{i=0}^{n-1}(1-\eta\lambda\sigma_i)\right] + \frac{\eta G}{2}\mathbb{E}\left[\sum_{i=0}^{n-1}(1-2\sigma_i)\prod_{j=i+1}^{n-1}(1-\eta\lambda\sigma_j)\right] .$$

Again using the notation from Eq. (10), Lemma 19, and the deterministic upper bounds in Eq. (12) and Eq. (13), this can be written as

$$\begin{aligned}
\mathbb{E}[x_{t+1}] &= \mathbb{E}[x_t]\cdot\mathbb{E}[A] + \frac{\eta G}{2}\mathbb{E}\left[\sum_{i=0}^{n-1}(1-2\sigma_i)B_i\right] \\
&\leq \mathbb{E}[x_t]\cdot\left(\mathbb{E}[\tilde{A}] \pm 2\left(\eta\lambda\sum_{i=0}^{n-1}\sigma_i\right)^2\right) + \frac{\eta G}{2}\mathbb{E}\left[\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i \pm 2\sum_{i=0}^{n-1}\left(\eta\lambda\sum_{j=i+1}^{n-1}\sigma_j\right)^2\right] \\
&\leq \mathbb{E}[x_t]\cdot\left(\left(1-\frac{\eta\lambda n}{2}\right) \pm 2(\eta\lambda n)^2\right) + \frac{\eta G}{2}\mathbb{E}\left[\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i \pm 2n(\eta\lambda n)^2\right] .
\end{aligned}$$

Recalling that $\mathbb{E}\left[\sum_{i=0}^{n-1}(1-2\sigma_i)\tilde{B}_i\right] = \mathbb{E}\left[\sum_{i=0}^{n-1}(1-2\sigma_i)(1-\eta\lambda\sum_{j=i+1}^{n}\sigma_j)\right]$ and using Lemma 17, the above is at most

$$\mathbb{E}[x_t]\cdot\left(1-\eta\lambda n\left(\frac{1}{2}\pm 2\eta\lambda n\right)\right) - \frac{\eta^2\lambda n G}{2}\left(\frac{n+1}{4(n-1)} \pm 2n^2\eta\lambda\right) .$$

Using the assumption $\eta \leq \frac{1}{100\lambda n^2}$ and that $n \geq 2$, it follows that

$$\begin{aligned}
\mathbb{E}[x_{t+1}] \;&\leq\; \mathbb{E}\left[x_t\right] \cdot \left(1 - \eta\lambda n\left(\frac{1}{2} \pm \frac{2}{100}\right)\right) - \frac{\eta^2 \lambda n G}{2}\left(\frac{3}{4} \pm \frac{2}{100}\right) \\
&\leq\; \mathbb{E}[x_t] \cdot \left(1 - \eta\lambda n\left(\frac{1}{2} \pm \frac{2}{100}\right)\right) - \frac{\eta^2 \lambda n G}{2}\;.
\end{aligned}$$

This inequality implies that if $\mathbb{E}[x_t] \leq 0$, then $\mathbb{E}[x_{t+1}] \leq 0$. Since the algorithm is initialized at $x_0 = -1$, it follows by induction that $\mathbb{E}[x_t] \leq 0$ for all t, so the inequality above implies that

$$\mathbb{E}[x_{t+1}] \;\leq\; \mathbb{E}[x_t] \cdot \left(1 - \frac{\eta\lambda n}{3}\right) - \frac{\eta^2 \lambda n G}{2}\;.$$

Opening the recursion, and using the fact that $x_0 = -1$, it follows that

$$\begin{aligned}
\mathbb{E}[x_t] \;&\leq\; -\left(1 - \frac{\eta\lambda n}{3}\right)^t - \frac{\eta^2 \lambda n G}{2}\sum_{i=0}^{t-1}\left(1 - \frac{\eta\lambda n}{3}\right)^i \\
&=\; -\left(1 - \frac{\eta\lambda n}{3}\right)^t - \frac{\eta^2 \lambda n G}{2(\eta\lambda n/3)}\left(1 - \left(1 - \frac{\eta\lambda n}{3}\right)^t\right) \\
&=\; -\left(1 - \frac{\eta\lambda n}{3}\right)^t - \frac{3\eta G}{2}\left(1 - \left(1 - \frac{\eta\lambda n}{3}\right)^t\right)\;.
\end{aligned}$$

Plugging this and Eq. (14) into Eq. (9), we get that

$$\begin{aligned}
\mathbb{E}[x_{t+1}^2] \;&\geq\; (1-\eta\lambda)^{2n} \cdot \mathbb{E}[x_t^2] + \eta G \cdot \left(\left(1 - \frac{\eta\lambda n}{3}\right)^t + \frac{3\eta G}{2}\left(1 - \left(1 - \frac{\eta\lambda n}{3}\right)^t\right)\right) \cdot c\eta\lambda n \\
&\geq\; (1 - 2\eta\lambda n) \cdot \mathbb{E}[x_t^2] + c\eta^2 G\lambda n \cdot \left(\left(1 - \frac{\eta\lambda n}{3}\right)^t + \frac{3\eta G}{2}\left(1 - \left(1 - \frac{\eta\lambda n}{3}\right)^t\right)\right)\;,
\end{aligned}$$

where in the last step we used Bernoulli's inequality. Applying this inequality recursively and recalling that $x_0 = -1$, it follows that

$$\mathbb{E}[x_k^2] \;\geq\; (1-2\eta\lambda n)^k + c\eta^2 G\lambda n \sum_{t=0}^{k-1}\left(\left(1 - \frac{\eta\lambda n}{3}\right)^t + \frac{3\eta G}{2}\left(1 - \left(1 - \frac{\eta\lambda n}{3}\right)^t\right)\right) \cdot (1-2\eta\lambda n)^{k-1-t} \tag{15}$$

We now consider two cases:

- If $2\eta\lambda n \leq \frac{1}{2k}$, then Eq. (15) implies

$$\mathbb{E}[x_k^2] \;\geq\; (1-2\eta\lambda n)^k \;\geq\; \left(1 - \frac{1}{2k}\right)^k \;\geq\; \frac{1}{2}$$

  for all k.

- If $2\eta\lambda n \geq \frac{1}{2k}$, then Eq. (15) implies

$$\begin{aligned}
\mathbb{E}[x_k^2] \;&\geq\; c\eta^2 G\lambda n \sum_{t=0}^{k-1}\left(\frac{3\eta G}{2}\left(1-\left(1-\frac{\eta\lambda n}{3}\right)^t\right)\right)\cdot(1-2\eta\lambda n)^{k-1-t}\\
&=\; \frac{3c\eta^3 G^2\lambda n}{2}\sum_{t=0}^{k-1}\left(1-\left(1-\frac{\eta\lambda n}{3}\right)^t\right)\cdot(1-2\eta\lambda n)^{k-1-t}\\
&\geq\; \frac{3c\eta^3 G^2\lambda n}{2}\sum_{t=\lfloor k/2\rfloor}^{k-1}\left(1-\left(1-\frac{\eta\lambda n}{3}\right)^t\right)\cdot(1-2\eta\lambda n)^{k-1-t}\\
&\geq\; \frac{3c\eta^3 G^2\lambda n}{2}\sum_{t=\lfloor k/2\rfloor}^{k-1}\left(1-\left(1-\frac{\eta\lambda n}{3}\right)^{\lfloor k/2\rfloor}\right)\cdot(1-2\eta\lambda n)^{k-1-t}\;.
\end{aligned}$$

Since we assume $2\eta\lambda n \geq \frac{1}{2k}$, this is at least

$$\frac{3c\eta^3 G^2\lambda n}{2}\sum_{t=\lfloor k/2\rfloor}^{k-1}\left(1-\left(1-\frac{1}{12k}\right)^{\lfloor k/2\rfloor}\right)\cdot(1-2\eta\lambda n)^{k-1-t}\;.$$

Since we assume in the proposition $k>1$, $\left(1-\left(1-\frac{1}{2k}\right)^{\lfloor k/2\rfloor}\right)$ can be verified to be at least some positive constant $c' > 0.16$. Thus, we can lower bound the above by

$$\frac{3cc'\eta^3 G^2\lambda n}{2}\sum_{t=\lfloor k/2\rfloor}^{k-1}(1-2\eta\lambda n)^{k-1-t} \;=\; \frac{3cc'\eta^3 G^2\lambda n}{2}\cdot\sum_{t=0}^{k-1-\lfloor k/2\rfloor}(1-2\eta\lambda n)^t\;.$$

Since $\sum_{i=0}^r a^i = \frac{1-a^{r+1}}{1-a}$ for any $a\in(0,1)$ (and moreover, $2\eta\lambda n\in(0,1)$ by the assumption that $\eta\leq\frac{1}{100\lambda n^2}$), the above equals

$$\frac{3cc'\eta^3 G^2\lambda n}{2}\cdot\frac{1-(1-2\eta\lambda n)^{k-\lfloor k/2\rfloor}}{2\eta\lambda n} \;\geq\; \frac{3cc'\eta^2 G^2}{4}\cdot\left(1-\left(1-\frac{1}{2k}\right)^{k-\lfloor k/2\rfloor}\right)\;,$$

where again we used the assumption $2\eta\lambda n\geq\frac{1}{2k}$. It is easily verified that $1-\left(1-\frac{1}{2k}\right)^{k-\lfloor k/2\rfloor}$ is lower bounded by a positive constant >0.2, so we can lower bound the above by $c''(\eta G)^2$ for some numerical constant $c''>0$. Recalling that this is a lower bound on $\mathbb{E}[x_k^2]$, and once again using the assumption $2\eta\lambda n\geq\frac{1}{2k}$, it follows that

$$\mathbb{E}[x_k^2] \;\geq\; c''(\eta G)^2 \;\geq\; c''\left(\frac{G}{4\lambda nk}\right)^2\;.$$

Combining the two cases above, we get that there exist some positive numerical constant c''' so that

$$\mathbb{E}[x_k^2] \;\geq\; c'''\cdot\min\left\{1\;,\;\frac{G^2}{\lambda^2(nk)^2}\right\}\;.$$

Noting that $\mathbb{E}[F(x_k)] = \mathbb{E}[\frac{\lambda}{4}x_k^2] = \frac{\lambda}{4}\mathbb{E}[x_k^2]$ and combining with the above, the result follows.

A.2. Proof of Thm. 7

We will assume without loss of generality that n is even (see the argument at the beginning of the proof of Thm. 5).

Using the same construction as in the proof of Proposition 11 (see Eq. (2)), we begin by observing that our analysis in the first epoch is identical to the random reshuffling case. Therefore, by recursively applying the relation in Eq. (3) (which in our case makes use of the same permutation in each epoch), we obtain the following relation between the initialization point x_0 and the k-th epoch x_k

$$\begin{aligned} x_k &= (1-\eta\lambda)^{nk}x_0 + \frac{\eta G}{2}\sum_{j=0}^{k-1}(1-\eta\lambda)^{nj}\sum_{i=0}^{n-1}\sigma_i(1-\eta\lambda)^i \\ &= (1-\eta\lambda)^{nk}x_0 + \frac{\eta G}{2}\cdot\frac{1-(1-\eta\lambda)^{nk}}{1-(1-\eta\lambda)^n}\sum_{i=0}^{n-1}\sigma_i(1-\eta\lambda)^i. \end{aligned}$$

From the above, the fact that $\mathbb{E}[\sigma_i]=0$, and the assumption $x_0=1$ we have

$$\mathbb{E}[x_k^2] = (1-\eta\lambda)^{2nk} + \left(\frac{\eta G}{2}\right)^2\beta_{n,\eta,\lambda}\left(\frac{1-(1-\eta\lambda)^{nk}}{1-(1-\eta\lambda)^n}\right)^2,$$

where $\beta_{n,\eta,\lambda}$ is as defined in Eq. (5).

The remainder of the proof now follows along a similar line as the proof of Proposition 11, where we consider different cases based on the value of $\eta\lambda$.

- If $\eta\lambda \geq 1$, then by Lemma 14, $\beta_{n,\eta,\lambda}$ is at least some positive constant $c>0$, and also $\left(\frac{1-(1-\eta\lambda)^{nk}}{1-(1-\eta\lambda)^n}\right)^2 \geq 1$ since it is the square of the geometric series $\sum_{j=0}^{k-1}(1-\eta\lambda)^{nj}$ with the first element being equal 1, and the other terms being positive (recall that n is even). Overall, we get for some constant $c>0$ that

$$\mathbb{E}[x_k^2] \geq c\left(\frac{\eta G}{2}\right)^2 \geq \frac{c}{4}\cdot\frac{G^2}{\lambda^2} \geq \frac{c}{4}\cdot\frac{G^2}{\lambda^2 nk^2}\,.$$

- If $\eta\lambda \leq \frac{1}{nk}$, then

$$\mathbb{E}[x_k^2] \geq (1-\eta\lambda)^{2nk} \geq \left(1-\frac{1}{nk}\right)^{2nk} \geq \left(\frac{1}{4}\right)^2 = \frac{1}{16}\,.$$

- If $\eta\lambda \in \left(\frac{1}{nk},\frac{1}{n}\right)$, then by Bernoulli's inequality we have $\exp(-1/k) \geq (1-\eta\lambda)^n \geq 1 - n\eta\lambda > 0$, implying that

$$\mathbb{E}[x_k^2] \geq \left(\frac{\eta G}{2}\right)^2\beta_{n,\eta,\lambda}\left(\frac{1-\exp(-1/k)^k}{1-(1-n\eta\lambda)}\right)^2 = \eta^2G^2\beta_{n,\eta,\lambda}\left(\frac{1-\exp(-1)}{2n\eta\lambda}\right)^2.$$

  Using Lemma 14 and recalling that $\eta\lambda \geq \frac{1}{nk}$, we have $\beta_{n,\eta,\lambda} \geq c\cdot\min\{1+1/\eta\lambda, n^3(\eta\lambda)^2\} \geq cn^3\eta^2\lambda^2$. Plugging this yields the above is at least

$$c'\frac{\eta^4G^2n^3\lambda^2}{n^2\eta^2\lambda^2} = c'\eta^2nG^2,$$

for some constant c'. Since $\eta\lambda \geq \frac{1}{nk} \iff \eta \geq \frac{1}{\lambda nk}$, this is lower bounded by

$$c'\frac{nG^2}{\lambda^2 n^2 k^2} = c'\frac{G^2}{\lambda^2 nk^2}.$$

- If $\eta\lambda \in \left[\frac{1}{n}, 1\right)$, then recalling $\left(\frac{1-(1-\eta\lambda)^{nk}}{1-(1-\eta\lambda)^n}\right)^2 \geq 1$ as the square of the sum of a geometric series with first element 1 and positive ratio, we have

$$\mathbb{E}[x_k^2] \;\geq\; \left(\frac{\eta G}{2}\right)^2 \cdot \beta_{n,\eta,\lambda}\ .$$

  By the assumption on $\eta\lambda$, we have that $n^3(\eta\lambda)^2 \geq 1/\eta\lambda$, therefore from Lemma 14 the above is at least

$$\begin{aligned} c\left(\frac{\eta G}{2}\right)^2 \cdot \min\left\{1+\frac{1}{\eta\lambda}\ ,\ n^3(\eta\lambda)^2\right\} \;&\geq\; c\left(\frac{\eta G}{2}\right)^2 \cdot \min\left\{\frac{1}{\eta\lambda}\ ,\ n^3(\eta\lambda)^2\right\} \\ &=\; c\left(\frac{\eta G}{2}\right)^2 \frac{1}{\eta\lambda} \;\geq\; \frac{c\eta G^2}{4\lambda}. \end{aligned}$$

  Since $\eta \geq \frac{1}{\lambda n}$, this is at least

$$\frac{cG^2}{4\lambda^2 n} \;\geq\; \frac{cG^2}{4\lambda^2 nk^2}\ .$$

Combining all previous cases, we have that

$$\mathbb{E}[x_k^2] \;\geq\; c\cdot\min\left\{1, \frac{G^2}{\lambda^2 nk^2}\right\}$$

for some numerical constant $c > 0$. Noting that $\mathbb{E}[F(x_k)] = \mathbb{E}\left[\frac{\lambda}{2}x_k^2\right] = \frac{\lambda}{2}\mathbb{E}\left[x_k^2\right]$ and combining with the above, the result follows.

A.3. Proof of Thm. 8

We will assume without loss of generality that n is even (see the argument at the beginning of the proof of Thm. 5).

First, we wish to argue that it is enough to consider the case where η is such that $\eta\lambda \in (0,1)$:

- If $\eta\lambda \geq 2$, it is easy to see that the algorithm may not converge. For example, consider the function $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ where $f_i(x) = \frac{\lambda}{2}x^2$ for all i. Then the algorithm performs iterations of the form $x_{new} = (1-\eta\lambda)x_{old}$, hence $|x_{new}| \geq |x_{old}|$. Assuming the initialization $x_0 = 1$, we have $F(x_k) = \frac{\lambda}{2}x_k^2 \geq \frac{\lambda}{2}x_0^2 = \frac{\lambda}{2}$, and the theorem statement holds.

- If $\eta\lambda \in [1,2)$, consider the function $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{\lambda}{2}x^2$ where $f_i(x) = \frac{\lambda}{2}x^2 - \frac{G}{2}x$ for odd i, and $f_i(x) = \frac{\lambda}{2}x^2 + \frac{G}{2}x$ for even i, initializing at $x_0 = 1$. Recalling that n is even, it is easy to verify that

$$x_{t+1} \;=\; (1-\eta\lambda)^n x_t + \frac{G\eta^2\lambda}{2}\sum_{i=0}^{n/2-1}(1-\eta\lambda)^{2i}\ .$$

Since $x_0 = 1$ and all terms above are non-negative, it follows that $x_k \geq 0$ for all $k \geq 1$. Moreover, since $\eta \geq 1/\lambda$, it follows that $x_k \geq \frac{G\eta^2\lambda}{2} \geq \frac{G}{2\lambda}$. Therefore, $F(x_k) = \frac{\lambda}{2}x_k^2 \geq \frac{G^2}{8\lambda} \geq \frac{G^2}{8\lambda k^2}$, and the theorem statement holds.

Assuming from now on that $\eta\lambda \in (0,1)$, we turn to our main construction. Consider the following function on $\mathbb{R}$:

$$F(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{\lambda}{2}x^2 ,$$

where

$$f_i(x) \;=\; \begin{cases} \frac{G}{2}x & i \leq \frac{n}{2} \\ \lambda x^2 - \frac{G}{2}x & i > \frac{n}{2} \end{cases} ,$$

Also, we assume that the initialization point x_0 is 1.

On this function, we have that during any single epoch, we perform $n/2$ iterations of the form

$$x_{new} \;=\; x_{old} - \frac{\eta G}{2} ,$$

followed by $n/2$ iterations of the form

$$x_{new} \;=\; (1-\eta\lambda)x_{old} + \frac{\eta G}{2} .$$

Thus, after n iterations, we get the following update for a single epoch:

$$\begin{aligned} x_{t+1} \;&=\; (1-\eta\lambda)^{n/2}\left(x_t - \frac{\eta G n}{4}\right) + \frac{\eta G}{2}\sum_{i=0}^{n/2-1}(1-\eta\lambda)^i \\ &=\; (1-\eta\lambda)^{n/2}x_t + \frac{\eta G}{2}\left(\sum_{i=0}^{n/2-1}(1-\eta\lambda)^i - \frac{n}{2}(1-\eta\lambda)^{n/2}\right) . \end{aligned} \tag{16}$$

Recalling that $\eta\lambda \in (0,1)$, we now consider two cases:

- If $\eta\lambda \in (1/n, 1)$, we have $\frac{1}{2\eta\lambda} < \frac{n}{2}$. Therefore,

$$\begin{aligned} \sum_{i=0}^{n/2-1}(1-\eta\lambda)^i - \frac{n}{2}(1-\eta\lambda)^{n/2} \;&=\; \sum_{i=0}^{n/2-1}\left((1-\eta\lambda)^i - (1-\eta\lambda)^{n/2}\right) \\ &\geq\; \sum_{i=0}^{\lceil 1/4\eta\lambda\rceil-1}\left((1-\eta\lambda)^i - (1-\eta\lambda)^{n/2}\right) \;=\; \sum_{i=0}^{\lceil 1/4\eta\lambda\rceil-1}(1-\eta\lambda)^i\left(1-(1-\eta\lambda)^{n/2-i}\right) \\ &\geq\; \sum_{i=0}^{\lceil 1/4\eta\lambda\rceil-1}(1-\eta\lambda)^i\left(1-(1-\eta\lambda)^{1/2\eta\lambda-i}\right) \;\geq\; \sum_{i=0}^{\lceil 1/4\eta\lambda\rceil-1}(1-\eta\lambda)^i\left(1-(1-\eta\lambda)^{1/4\eta\lambda}\right) . \end{aligned}$$

Since $1/\eta\lambda \geq 1$, and $(1-1/z)^{z/4} \leq \exp(-1/4)$ for any $z \geq 1$, the displayed equation above is at least

$$
\begin{aligned}
(1-\exp(-1/4)) \sum_{i=0}^{\lceil 1/4\eta\lambda \rceil - 1} (1-\eta\lambda)^i \;&=\; (1-\exp(-1/4)) \cdot \frac{1-(1-\eta\lambda)^{\lceil 1/4\eta\lambda \rceil}}{\eta\lambda} \\
&\geq\; (1-\exp(-1/4)) \cdot \frac{1-(1-\eta\lambda)^{1/4\eta\lambda}}{\eta\lambda} \\
&\geq\; \frac{(1-\exp(-1/4))^2}{\eta\lambda} \,.
\end{aligned}
$$

Denoting $c := (1-\exp(-1/4))^2 > 0.04$ and plugging this lower bound on $\sum_{i=0}^{n/2-1}(1-\eta\lambda)^i - \frac{n}{2}(1-\eta\lambda)^{n/2}$ into Eq. (16), we get that

$$
x_{t+1} \;\geq\; (1-\eta\lambda)^{n/2} x_t + \frac{\eta G}{2} \cdot \frac{c}{\eta\lambda},
$$

and hence $x_{t+1} \geq \frac{cG}{2\lambda}$. This holds for any t, and in particular $x_k \geq \frac{cG}{2\lambda}$, hence $F(x_k) = \frac{\lambda}{2}x_k^2 = \frac{c^2G^2}{8\lambda} \geq \frac{c^2G^2}{8\lambda k^2}$, which satisfies the theorem statement.

- If $\eta\lambda \in (0, 1/n]$, we have

$$
\begin{aligned}
\frac{\eta G}{2}\left(\sum_{i=0}^{n/2-1}(1-\eta\lambda)^i - \frac{n}{2}(1-\eta\lambda)^{n/2}\right) \;&=\; \frac{\eta G}{2}\left(\frac{1-(1-\eta\lambda)^{n/2}}{\eta\lambda} - \frac{n}{2}(1-\eta\lambda)^{n/2}\right) \\
&=\; \frac{G}{2\lambda}\left(1-(1-\eta\lambda)^{n/2} - \frac{\eta\lambda n}{2}(1-\eta\lambda)^{n/2}\right) \\
&=\; \frac{G}{2\lambda}\left(1-\left(1+\frac{\eta\lambda n}{2}\right)(1-\eta\lambda)^{n/2}\right) \\
&\overset{(*)}{\geq}\; \frac{G}{2\lambda}\left(1-\left(1+\frac{\eta\lambda n}{2}\right)\left(1-\frac{\eta\lambda n}{2}+\frac{(\eta\lambda n/2)^2}{2}\right)\right) \\
&=\; \frac{G}{2\lambda}\left(1-\left(1-\left(\frac{\eta\lambda n}{2}\right)^2+\left(1+\frac{\eta\lambda n}{2}\right)\frac{(\eta\lambda n)^2}{8}\right)\right) \\
&=\; \frac{G(\eta\lambda n)^2}{2\lambda}\left(\frac{1}{4}-\left(1+\frac{\eta\lambda n}{2}\right)\frac{1}{8}\right) \;\geq\; \frac{G(\eta\lambda n)^2}{2\lambda}\left(\frac{1}{4}-\left(1+\frac{1}{2}\right)\cdot\frac{1}{8}\right) \\
&=\; \frac{G\lambda(\eta n)^2}{32}\,,
\end{aligned}
$$

where $(*)$ is by Lemma 18. Plugging this back into Eq. (16), we get

$$
x_{t+1} \;\geq\; (1-\eta\lambda)^{n/2} x_t + \frac{G\lambda(\eta n)^2}{32}\,.
$$

Recalling that $x_0 = 1$, this implies that x_t remains positive for all t. Also, by Bernoulli's inequality, $1 \geq (1-\eta\lambda)^{n/2}) \geq 1-\eta\lambda n/2 \geq 0$. Therefore, the above displayed equation implies that

$$
x_{t+1} \;\geq\; \left(1-\frac{\eta\lambda n}{2}\right) x_t + \frac{G\lambda(\eta n)^2}{32}\,.
$$

Recurseively applying this inequality, and recalling that $x_0 = 1$, it follows that

$$\begin{aligned} x_k &\geq \left(1 - \frac{\eta\lambda n}{2}\right)^k + \frac{G\lambda(\eta n)^2}{32} \sum_{t=0}^{k-1} \left(1 - \frac{\eta\lambda n}{2}\right)^t \\ &= \left(1 - \frac{\eta\lambda n}{2}\right)^k + \frac{G\lambda(\eta n)^2}{32} \cdot \frac{1 - (1 - \eta\lambda n/2)^k}{\eta\lambda n/2} \\ &= \left(1 - \frac{\eta\lambda n}{2}\right)^k + \frac{G\eta n}{16} \left(1 - \left(1 - \frac{\eta\lambda n}{2}\right)^k\right) . \end{aligned}$$

We now consider two sub-cases:

- If $\eta\lambda \in (0, 1/nk)$, the above is at least $\left(1 - \frac{\eta\lambda n}{2}\right)^k \geq \left(1 - \frac{1}{2k}\right)^k \geq \frac{1}{2}$ for all $k \geq 1$, so we have $F(x_k) = \frac{\lambda}{2} x_k^2 \geq \frac{\lambda}{8}$, satisfying the theorem statement.
- If $\eta\lambda \in [1/nk, 1/n]$, we have $\left(1 - \frac{\eta\lambda n}{2}\right)^k \leq \left(1 - \frac{1}{2k}\right)^k \leq \exp\left(-\frac{1}{2}\right)$, so the displayed equation above is at least $\frac{G\eta n}{16}\left(1 - \exp\left(-\frac{1}{2}\right)\right)$, which by the assumption $\eta\lambda \geq \frac{1}{nk}$, is at least $\frac{1-\exp(-1/2)}{16} \cdot \frac{G}{\lambda k}$. Therefore,

$$F(x_k) = \frac{\lambda}{2} x_k^2 \geq \frac{1}{2} \cdot \left(\frac{1 - \exp(-1/2)}{16}\right)^2 \cdot \frac{G^2}{\lambda k^2} ,$$

which satisfies the theorem statement.

A.4. Proof of Thm. 9

We begin by assuming w.l.o.g. that $b = 0$. This is justified as seen by the transformation $f_i(x) \mapsto f_i(x - b/\lambda)$ which shifts each f_i to the right by a distance of b/λ, and consequentially shifting the initialization point x_0 to the right by the same distance to $x_0 + \frac{b}{\lambda}$. The derivative in the initialization point after transforming remains the same, and a simple inductive argument shows this persists throughout all the iterations of SGD where all the iterates are also shifted by b/λ. Additionally, this also entails $|b_i| \leq G$ for all i since by the gradient boundedness assumption we have $|a_i x^* - b_i| \leq G$ for all i.

Next, we evaluate an expression for the iterate on the k-th epoch x_k. First, for a selected permutation $\sigma_i : [n] \to [n]$ we have that the gradient update at iteration j in epoch i is given by

$$x_{new} = \left(1 - \eta a_{\sigma_i(j)}\right) x_{old} + \eta b_{\sigma_i(j)}.$$

Repeatedly applying the above relation, we have that in the end of each epoch the relation between the iterates x_t and x_{t+1} is given by

$$x_{t+1} = \prod_{j=1}^{n} \left(1 - \eta a_{\sigma_{t+1}(j)}\right) x_t + \eta \sum_{j=1}^{n} \left(\prod_{i=j+1}^{n} \left(1 - \eta a_{\sigma_{t+1}(i)}\right) \right) b_{\sigma_{t+1}(j)}.$$

Letting $S \coloneqq \prod_{j=1}^{n} \left(1 - \eta a_{\sigma_i(j)}\right) = \prod_{j=1}^{n} (1 - \eta a_j)$ and $X_{\sigma_t} \coloneqq \sum_{j=1}^{n} \left(\prod_{i=j+1}^{n} \left(1 - \eta a_{\sigma_t(i)}\right) \right) b_{\sigma_t(j)}$, this can be rewritten equivalently as

$$x_{t+1} = S x_t + \eta X_{\sigma_{t+1}}. \tag{17}$$

Iteratively applying the above, we have after k epochs that

$$x_k = S^k x_0 + \eta \sum_{i=1}^{k} S^{i-1} X_{\sigma_i}. \tag{18}$$

Squaring and taking expectation on both sides yields

$$\begin{aligned}\mathbb{E}\left[x_k^2\right] &= \mathbb{E}\left[\left(S^k x_0 + \eta \sum_{i=1}^{k} S^{i-1} X_{\sigma_i}\right)^2\right] \leq 2\mathbb{E}\left[S^{2k} x_0^2 + \eta^2 \left|\sum_{i=1}^{k} S^{i-1} X_{\sigma_i}\right|^2\right] \\ &\leq 2S^{2k} x_0^2 + 2\eta^2 k \sum_{i=1}^{k} \mathbb{E}\left[X_{\sigma_i}^2\right] = 2S^{2k} x_0^2 + 2\eta^2 k^2 \mathbb{E}\left[X_{\sigma_1}^2\right],\end{aligned} \tag{19}$$

where the first and second inequalities are application of Jensen's inequality on the function $x \mapsto x^2$ and the last equality is due to the fact that in single shuffling we have $\sigma_i = \sigma_1$ for all i.

Since $\frac{L}{\lambda} \leq \frac{nk}{\log(n^{0.5}k)}$ implies that $\eta L \leq 1$, we have $1 - \eta a_i \in (0, 1]$ for any $i \in \{1, \ldots, n\}$. Using the AM-GM inequality on $1 - \eta a_1, \ldots, 1 - \eta a_n$ we have

$$\sqrt[n]{S} = \sqrt[n]{\prod_{i=1}^{n}(1 - \eta a_i)} \leq \frac{1}{n}\sum_{i=1}^{n}(1 - \eta a_i) = 1 - \frac{\eta \sum_{i=1}^{n} a_i}{n} = 1 - \eta\lambda, \tag{20}$$

implying

$$S \leq (1 - \eta\lambda)^n. \tag{21}$$

Recall that $\eta = \frac{\log\left(n^{0.5}k\right)}{\lambda nk}$, we combine the above with Lemma 20 which together with the inequality $(1 - x/y)^y \leq \exp(-x)$ for all $x, y > 0$ yields that Eq. (19) is upper bounded by

$$2(1 - \eta\lambda)^{2nk} x_0^2 + 2\eta^4 n^3 k^2 G^2 L^2 \leq \tilde{\mathcal{O}}\left(\frac{1}{nk^2} x_0^2 + \frac{G^2 L^2}{\lambda^4 n k^2}\right),$$

and since $\mathbb{E}\left[F(x_k) - F(x^*)\right] \leq \frac{\lambda}{2}\mathbb{E}[x_k^2]$, the theorem follows.

A.5. Proof of Thm. 10

Similarly to the single shuffling case, we assume w.l.o.g. that $b = 0$ and $|b_i| \leq G$ for all $i \in [n]$ (see the argument in the beginning of the proof of Thm. 9 for justification). Continuing from Eq. (17), we square and take expectation on both sides to obtain

$$\mathbb{E}\left[x_{t+1}^2\right] = \mathbb{E}\left[\left(S x_t + \eta X_{\sigma_{t+1}}\right)^2\right] = S^2 \mathbb{E}[x_t^2] + 2\eta S \mathbb{E}\left[x_t X_{\sigma_{t+1}}\right] + \eta^2 \mathbb{E}\left[X_{\sigma_{t+1}}^2\right].$$

Since in random reshuffling the random component at iteration $t+1$, $X_{\sigma_{t+1}}$, is independent of the iterate at iteration t, x_t, and by plugging Eq. (18), the above equals

$$\begin{aligned}
\mathbb{E}\left[x_{t+1}^2\right] &= S^2\mathbb{E}[x_t^2] + 2\eta S\mathbb{E}\left[x_t\right]\mathbb{E}\left[X_{\sigma_{t+1}}\right] + \eta^2\mathbb{E}\left[X_{\sigma_{t+1}}^2\right] \\
&= S^2\mathbb{E}[x_t^2] + 2\eta S\mathbb{E}\left[S^t x_0 + \eta\sum_{i=1}^{t} S^{i-1}X_{\sigma_i}\right]\mathbb{E}\left[X_{\sigma_{t+1}}\right] + \eta^2\mathbb{E}\left[X_{\sigma_{t+1}}^2\right] \\
&= S^2\mathbb{E}[x_t^2] + 2\eta S^{t+1}x_0\mathbb{E}\left[X_{\sigma_{t+1}}\right] + 2\eta^2\sum_{i=1}^{t} S^i\mathbb{E}\left[X_{\sigma_i}\right]\mathbb{E}\left[X_{\sigma_{t+1}}\right] + \eta^2\mathbb{E}\left[X_{\sigma_{t+1}}^2\right] \\
&= S^2\mathbb{E}[x_t^2] + 2\eta S^{t+1}x_0\mathbb{E}\left[X_{\sigma_1}\right] + 2\eta^2\sum_{i=1}^{t} S^i\mathbb{E}\left[X_{\sigma_1}\right]^2 + \eta^2\mathbb{E}\left[X_{\sigma_1}^2\right],
\end{aligned}$$

where the last equality is due to X_{σ_i} being i.i.d for all i. Recursively applying the above relation and taking absolute value, we obtain

$$\mathbb{E}\left[x_k^2\right] = S^{2k}x_0^2 + 2\eta x_0\mathbb{E}\left[X_{\sigma_1}\right]\sum_{j=1}^{k} S^{k+j} + 2\eta^2\mathbb{E}\left[X_{\sigma_1}\right]^2\sum_{j=1}^{k} S^{2j}\sum_{i=1}^{j} S^{k-i} + \eta^2\mathbb{E}\left[X_{\sigma_1}^2\right]\sum_{j=1}^{k} S^{2j},$$

which entails an upper bound of

$$\begin{aligned}
\mathbb{E}\left[x_k^2\right] &\le S^{2k}x_0^2 + 2\eta\left|x_0\mathbb{E}\left[X_{\sigma_1}\right]\right|\sum_{j=1}^{k} S^{k+j} + 2\eta^2\mathbb{E}\left[X_{\sigma_1}\right]^2\sum_{j=1}^{k} S^{2j}\sum_{i=1}^{j} S^{k-i} + \eta^2\mathbb{E}\left[X_{\sigma_1}^2\right]\sum_{j=1}^{k} S^{2j} \\
&\le S^{2k}x_0^2 + 2\eta k S^k\left|x_0\right|\cdot\left|\mathbb{E}\left[X_{\sigma_1}\right]\right| + 2\eta^2k^2\mathbb{E}\left[X_{\sigma_1}\right]^2 + \eta^2k\mathbb{E}\left[X_{\sigma_1}^2\right].
\end{aligned}$$

Since $2S^k\left|x_0\right|\cdot\eta k\left|\mathbb{E}\left[X_{\sigma_1}\right]\right| \le S^{2k}x_0^2 + \eta^2k^2\mathbb{E}\left[X_{\sigma_1}\right]^2$, the above is at most

$$2S^{2k}x_0^2 + 3\eta^2k^2\mathbb{E}\left[X_{\sigma_1}\right]^2 + \eta^2k\mathbb{E}\left[X_{\sigma_1}^2\right],$$

and by virtue of Eq. (21), the inequality $(1-x/y)^y \le \exp(-x)$ for all $x, y > 0$ and Lemmas 20 and 22, we conclude

$$\begin{aligned}
\mathbb{E}\left[x_k^2\right] &\le 2S^{2k}x_0^2 + 48\eta^4n^2k^2G^2L^2 + 5\eta^4n^3kG^2L^2\log(2n) \\
&\le \tilde{\mathcal{O}}\left(\frac{1}{n^2k^2}x_0^2 + \frac{G^2L^2}{\lambda^4n^2k^2} + \frac{G^2L^2}{\lambda^4nk^3}\right),
\end{aligned}$$

and since $\mathbb{E}\left[F(x_k) - F(x^*)\right] \le \frac{\lambda}{2}\mathbb{E}[x_k^2]$, the theorem follows.

Appendix B. Proof of Lemma 14

Using Lemma 15 from Appendix C, we have that

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=0}^{n-1}\sigma_i(1-\alpha)^i\right)^2\right] &= \mathbb{E}\left[\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sigma_i\sigma_j(1-\alpha)^{i+j}\right] \\
&= \sum_{i=0}^{n-1}\mathbb{E}[\sigma_i^2](1-\alpha)^{2i} + \sum_{i,j\in\{0,\ldots,n-1\},i\neq j}\mathbb{E}[\sigma_i\sigma_j](1-\alpha)^{i+j} \\
&= \sum_{i=0}^{n-1}(1-\alpha)^{2i} - \frac{1}{n-1}\left(\left(\sum_{i=0}^{n-1}(1-\alpha)^i\right)^2 - \sum_{i=0}^{n-1}(1-\alpha)^{2i}\right) \\
&= \left(1+\frac{1}{n-1}\right)\sum_{i=0}^{n-1}(1-\alpha)^{2i} - \frac{1}{n-1}\left(\sum_{i=0}^{n-1}(1-\alpha)^i\right)^2 . \qquad (22)
\end{aligned}$$

Using the fact that $\sum_{i=0}^{r-1} s^i = \frac{1-s^r}{1-s}$ for any $a \neq 1$, the above can also be written as

$$\begin{aligned}
&\left(1+\frac{1}{n-1}\right)\frac{1-(1-\alpha)^{2n}}{1-(1-\alpha)^2} - \frac{(1-(1-\alpha)^n)^2}{(n-1)(1-(1-\alpha))^2} \\
&= \frac{n}{n-1}\cdot\frac{1-(1-\alpha)^{2n}}{\alpha(2-\alpha)} - \frac{(1-(1-\alpha)^n)^2}{\alpha^2(n-1)} \\
&= \frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot\left(1+(1-\alpha)^n - \frac{2-\alpha}{n\alpha}(1-(1-\alpha)^n)\right) \\
&= \frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot\left(1-\frac{2-\alpha}{n\alpha}+\left(1+\frac{2-\alpha}{n\alpha}\right)(1-\alpha)^n\right) . \qquad (23)
\end{aligned}$$

We now lower bound either Eq. (22) or (equivalently) Eq. (23), on a case-by-case basis, depending on the size of α.

B.1. The case $\alpha \geq 1$

We will show that in this case, our equations are lower bounded by a positive numerical constant, which satisfies the lemma statement. We split this case into a few sub-cases:

- If $\alpha = 1$, then Eq. (22) equals $1 + \frac{1}{n-1} - \frac{1}{n-1} = 1$.
- If $\alpha \in (1,2)$, then $\frac{2-\alpha}{n\alpha} = \frac{2}{n\alpha} - \frac{1}{n} \leq \frac{2}{n} - \frac{1}{n} = \frac{1}{n}$. Using this fact, Eq. (23) can be lower bounded as

$$\begin{aligned}
2\cdot\frac{1-(1-\alpha)^n}{2(2-\alpha)}\cdot\left(1-\frac{2-\alpha}{n\alpha}\right) &\geq \frac{1-(1-\alpha)^n}{2-\alpha}\cdot\left(1-\frac{1}{n}\right) \\
&\geq \frac{1-(1-\alpha)^n}{2(2-\alpha)} \\
&\overset{(*)}{=} \frac{1-|1-\alpha|^n}{2(1-|1-\alpha|)} \geq \frac{1-|1-\alpha|}{2(1-|1-\alpha|)} = \frac{1}{2} ,
\end{aligned}$$

where in $(*)$ we used the facts that n is even and that since $\alpha \in (1,2)$, we have $2-\alpha = 1+1-\alpha = 1-|1-\alpha|$.

- If $\alpha = 2$, then using the assumption that n is even, Eq. (22) reduces to

$$\left(1+\frac{1}{n-1}\right)\sum_{i=0}^{n-1}(-1)^{2i} - \frac{1}{n-1}\left(\sum_{i=0}^{n-1}(-1)^i\right)^2 \;=\; \left(1+\frac{1}{n-1}\right)n - \frac{1}{n-1}\cdot 0 \geq n\,.$$

- If $\alpha > 2$, then noting that $1+\frac{2-\alpha}{n\alpha} = 1-\frac{1}{n}+\frac{2}{n\alpha} > 0$, Eq. (23) is lower bounded as

$$\begin{aligned} 2\cdot\frac{(1-\alpha)^n-1}{\alpha(\alpha-2)}\cdot\left(1-\frac{2-\alpha}{n\alpha}\right) \;&\geq\; 2\cdot\frac{(1-\alpha)^2-1}{\alpha(\alpha-2)}\cdot\left(1-\frac{2}{n\alpha}+\frac{1}{n}\right) \\ &\geq\; 2\cdot\left(1-\frac{1}{n}+\frac{1}{n}\right) \;=\; 2\,. \end{aligned}$$

B.2. The case $\alpha \in [1/13n, 1)$

In this case, we will show a lower bound of c/α for some positive numerical constant c, which implies the lemma statement in this case. To show this, we first focus on the term

$$1-\frac{2-\alpha}{n\alpha}+\left(1+\frac{2-\alpha}{n\alpha}\right)(1-\alpha)^n\,, \tag{24}$$

in Eq. (23), and argue that it is monotonically increasing in α. For that, it is enough to show that its derivative with respect to α is non-negative. With some straightforward computations, the derivative equals

$$(1-\alpha)^{n-1}\left(1-\frac{2}{\alpha}-n-\frac{2}{\alpha^2 n}+\frac{2}{\alpha n}\right)+\frac{2}{\alpha^2 n}\,.$$

this can also be written as

$$\begin{aligned} &\frac{2}{\alpha^2 n}\left((1-\alpha)^{n-1}\left(\frac{\alpha^2 n}{2}-\alpha n-\frac{\alpha^2 n^2}{2}-1+\alpha\right)+1\right) \\ &= \frac{2}{\alpha^2 n}\left(1-(1-\alpha)^{n-1}\left(1+\alpha(n-1)+\alpha^2\frac{n(n-1)}{2}\right)\right)\,. \end{aligned} \tag{25}$$

It is easy to verify that $1+\alpha(n-1)+\alpha^2\frac{n(n-1)}{2}$ is the third-order Taylor expansion of the function $g(\alpha) := (1-\alpha)^{1-n}$ around $\alpha = 0$, and moreover, it is a lower bound on the function (for $\alpha \in [1/13n, 1)$) since the Taylor remainder term (in Lagrange form) equals $\frac{g^{(3)}(\xi)}{3!}\alpha^3 = \frac{(n-1)n(n+1)}{3!(1-\xi)^{n+2}}\alpha^3$ for some $\xi \in [0,\alpha]$, which is strictly positive for any α in our range. Overall, we can lower bound Eq. (25) by

$$\frac{2}{\alpha^2 n}\left(1-(1-\alpha)^{n-1}\cdot(1-\alpha)^{1-n}\right) \;=\; 0.$$

This implies that Eq. (24) is monotonically increasing.

Using this monotonicity property, we get that Eq. (24) is minimized over the interval $\alpha \in [1/13n, 1)$ when $\alpha = 1/13n$, in which case it takes the value

$$1-\left(26-\frac{1}{n}\right)+\left(1+26-\frac{1}{n}\right)\left(1-\frac{1}{13n}\right)^n \;=\; \left(27-\frac{1}{n}\right)\left(1-\frac{1}{13n}\right)^n+\frac{1}{n}-25$$
$$=\; 27\left(1-\frac{1}{13n}\right)^n+\frac{1}{n}\left(1-\left(1-\frac{1}{13n}\right)^n\right)-25\,.$$

A numerical computation reveals that this expression is strictly positive (lower bounded by $7\cdot 10^{-4}$) for all $2 \leq n < 78$. For $n \geq 78$, noting that $(1-1/13n)^n$ is monotonically increasing in n, this expression can be lower bounded by

$$27\left(1-\frac{1}{13n}\right)^n-25 \;\geq\; 27\left(1-\frac{1}{13\cdot 78}\right)^{78}-25 \;>\; 2\cdot 10^{-7}.$$

In any case, we get that Eq. (24) is lower bounded by some positive numerical constant c. Plugging back into Eq. (23), and using that fact that $(1-1/13n)^n$ is upper bounded by $\exp(-1/13)$, we can lower bound that equation by

$$\frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot c \;\geq\; c\cdot\frac{1-(1-1/13n)^n}{2\alpha} \;\geq\; c\cdot\frac{1-\exp(-1/13)}{2\alpha}\,,$$

which equals c'/α for some numerical constant $c' > 0$.

B.3. The case $\alpha \in (0, 1/13n)$

In this case, we have $n^3\alpha^2 \leq \frac{1}{\alpha}$, so it is enough to prove a lower bound of $c\cdot n^3\alpha^2$ in order to satisfy the lemma statement. We analyze seperately the cases $n = 2$ and $n > 2$. If $n = 2$, then Eq. (23) equals

$$2\cdot 1\cdot\left(1-\frac{2-\alpha}{2\alpha}+\left(1+\frac{2-\alpha}{2\alpha}\right)(1-\alpha)^2\right)$$
$$=\; 2\left(\frac{3}{2}-\frac{1}{\alpha}+\left(\frac{1}{2}+\frac{1}{\alpha}\right)(1-\alpha)^2\right)$$
$$=\; 2\left(2-2\alpha\left(\frac{1}{2}+\frac{1}{\alpha}\right)+\alpha^2\left(\frac{1}{2}+\frac{1}{\alpha}\right)\right) \;=\; \alpha^2 \;=\; \frac{1}{8}n^3\alpha^2\,,$$

which satisfies the lower bound in the lemma statement. If $n > 2$, by Lemma 18, Eq. (23) equals

$$\frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot\left(1-\frac{2-\alpha}{\alpha n}+\left(1+\frac{2-\alpha}{\alpha n}\right)\left(1-\alpha n+\binom{n}{2}\alpha^2-\binom{n}{3}\alpha^3+c_{\alpha,n}\right)\right)\,,$$

where $|c_{\alpha,n}| \leq (\alpha n)^4/24$. Simplifying a bit, this equals

$$\frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot\left(2+\left(1+\frac{2-\alpha}{\alpha n}\right)\left(-\alpha n+\binom{n}{2}\alpha^2-\binom{n}{3}\alpha^3+c_{\alpha,n}\right)\right)$$
$$=\; \frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot\left(2+\left(1-\frac{1}{n}+\frac{2}{\alpha n}\right)\left(-\alpha n+\binom{n}{2}\alpha^2-\binom{n}{3}\alpha^3+c_{\alpha,n}\right)\right)\,.$$

Opening the inner product and collecting terms according to powers of α, this equals

$$\frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot\left(\left(-n+1+\frac{2}{n}\binom{n}{2}\right)\alpha+\left(\left(1-\frac{1}{n}\right)\binom{n}{2}-\frac{2}{n}\binom{n}{3}\right)\alpha^2+\left(1-\frac{1}{n}\right)\binom{n}{3}\alpha^3+\left(1-\frac{1}{n}+\frac{2}{\alpha n}\right)c_{\alpha,n}\right). \tag{26}$$

It is easily verified that

$$-n+1+\frac{2}{n}\binom{n}{2}=0\quad,\quad\left(1-\frac{1}{n}\right)\binom{n}{2}-\frac{2}{n}\binom{n}{3}\geq\frac{(n-1)^2}{6}\quad,\quad\left(1-\frac{1}{n}\right)\binom{n}{3}\leq n^3.$$

Plugging this into Eq. (26), and recalling that $|c_{\alpha,n}|\leq(\alpha n)^4/24$, we can lower bound Eq. (26) by

$$\frac{n}{n-1}\cdot\frac{1-(1-\alpha)^n}{\alpha(2-\alpha)}\cdot\left(\frac{(n-1)^2}{6}\alpha^2-(\alpha n)^3-\left|1-\frac{1}{n}+\frac{2}{\alpha n}\right|\frac{(\alpha n)^4}{24}\right).$$

Invoking again Lemma 18, and noting that $n/(n-1)\geq 1$ and $\alpha n\in(0,1/13)$, we can lower bound the above by

$$\begin{aligned}
&1\cdot\frac{1-(1-\alpha n+(\alpha n)^2/2)}{\alpha(2-\alpha)}\left(\frac{(n-1)^2}{6}\alpha^2-(\alpha n)^3-\frac{3}{\alpha n}\cdot\frac{(\alpha n)^4}{24}\right)\\
&=\frac{\alpha n(1-\alpha n/2)}{\alpha(2-\alpha)}\left(\frac{(n-1)^2}{6}\alpha^2-\frac{9}{8}(\alpha n)^3\right)\\
&\geq\frac{n}{2(2-\alpha)}\left(\frac{(n-1)^2}{6}\alpha^2-\frac{9}{8}(\alpha n)^3\right)\\
&\geq\frac{n^3\alpha^2}{4}\left(\frac{(n-1)^2}{6n^2}-\frac{9}{8}\alpha n\right)=\frac{n^3\alpha^2}{4}\left(\frac{1}{6}\left(1-\frac{1}{n}\right)^2-\frac{9}{8}\alpha n\right).
\end{aligned}$$

Since we can assume $n\geq 4$ (as n is even and the case $n=2$ was treated earlier), and $\alpha n\leq 1/13$, it can be easily verified that this is at least $cn^3\alpha^2$ for some positive constant $c>10^{-3}$.

Appendix C. Technical Lemmas

Lemma 15 *Let $\sigma_0,\ldots,\sigma_{n-1}$ be a random permutation of $(1,...,1,-1,...,-1)$ (where there are $n/2$ 1's and $n/2$ -1's). Then for any indices i,j,*

$$\mathbb{E}[\sigma_i\sigma_j]=\begin{cases}1 & \text{if } i=j\\ -\frac{1}{n-1} & \text{if } i\neq j\end{cases}.$$

Proof Note that each σ_i is uniformly distributed on $\{-1,+1\}$. Therefore, $\mathbb{E}[\sigma_i^2]=1$, and for any $i\neq j$,

$$\begin{aligned}
\mathbb{E}[\sigma_i\sigma_j]&=\frac{1}{2}\mathbb{E}[\sigma_i|\sigma_j=1]-\frac{1}{2}\mathbb{E}[\sigma_i|\sigma_j=-1]\\
&=\frac{1}{2}\left(\Pr(\sigma_i=1|\sigma_j=1)-\Pr(\sigma_i=-1|\sigma_j=1)-\Pr(\sigma_i=1|\sigma_j=-1)+\Pr(\sigma_i=-1|\sigma_j=-1)\right)\\
&=\frac{1}{2}\left(\frac{n/2-1}{n-1}-\frac{n/2}{n-1}-\frac{n/2}{n-1}+\frac{n/2-1}{n-1}\right)=-\frac{1}{n-1}.
\end{aligned}$$

■

Lemma 16 *Let $\sigma_0,\ldots,\sigma_{n-1}$ be a random permutation of $(1,...,1,0,...,0)$ (where there are $n/2$ 1's and $n/2$ 0's). Then for any indices i,j,*

$$\mathbb{E}[\sigma_i\sigma_j] \;=\; \begin{cases} \frac{1}{2} & \textit{if } i=j \\ \frac{1}{4}\left(1-\frac{1}{n-1}\right) & \textit{if } i\neq j \end{cases} .$$

Proof This follows from applying Lemma 15 on the random variables $\mu_0,\ldots,\mu_{n-1}$, where $\mu_i := 1-2\sigma_i$ for all i, and noting that $\mathbb{E}[\mu_i\mu_j] = \mathbb{E}[(1-2\sigma_i)(1-2\sigma_j)] = 4\mathbb{E}[\sigma_i\sigma_j]-1$ (using the fact that each σ_i is uniform on $\{0,1\}$). ■

Lemma 17 *Under the conditions of Lemma 16, we have that*

$$\mathbb{E}\left[\left(\sum_{i=0}^{n-1}(1-2\sigma_i)(1-\eta\lambda\sum_{j=i+1}^{n}\sigma_j)\right)\right] \;=\; -\eta\lambda\frac{n(n+1)}{4(n-1)}$$

Proof Using Lemma 16, and the fact that each σ_i is uniform on $\{0,1\}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=0}^{n-1}(1-2\sigma_i)(1-\eta\lambda\sum_{j=i+1}^{n}\sigma_j)\right)\right]\\
&= \mathbb{E}\left[n-2\sum_{i=0}^{n-1}\sigma_i-\eta\lambda\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\sigma_j+2\eta\lambda\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\sigma_i\sigma_j\right]\\
&= n-n-\eta\lambda\cdot\frac{n(n+1)}{2}\cdot\frac{1}{2}+2\eta\lambda\cdot\frac{n(n+1)}{2}\cdot\frac{1}{4}\left(1-\frac{1}{n-1}\right)\\
&-\eta\lambda\cdot\frac{n(n+1)}{4}+\eta\lambda\cdot\frac{n(n+1)}{4}\left(1-\frac{1}{n-1}\right)\\
&= -\eta\lambda\frac{n(n+1)}{4(n-1)} .
\end{aligned}$$

■

Lemma 18 *Let r be a positive integer and $x\in[0,1]$. Then for any positive integer $j<r$,*

$$(1-x)^n \;=\; \sum_{i=0}^{j}(-1)^i\binom{r}{i}x^i+a_{j,x} ,$$

where $\binom{r}{1},\binom{r}{2}$ etc. refer to binomial coefficients, and $a_{j,x}$ has the same sign as $(-1)^{j+1}$ and satisfies

$$|a_{j,x}| \;\leq\; \frac{(rx)^{j+1}}{(j+1)!} .$$

Proof The proof follows by a Taylor expansion of the function $g(x) = (1-x)^r$ around $x = 0$: It is easily verified that the first j terms are $\sum_{i=0}^{j}(-1)^i\binom{r}{i}x^i$. Moreover, by Taylor's theorem, the remainder term $\alpha_{j,x}$ (in Lagrange form) is $\frac{g^{(j+1)}(\xi)}{(j+1)!}x^{j+1}$ for some $\xi \in [0,x]$. Moreover, $g^{(j+1)}(\xi) = (-1)^{j+1}\binom{r}{j+1}(1-\xi)^{r-j-1}$, whose sign is $(-1)^{j+1}$ and absolute value at most

$$\sup_{\xi\in[0,x]}\binom{r}{j+1}(1-\xi)^{r-j-1}x^{j+1} \leq \frac{r^{j+1}}{(j+1)!}\cdot 1\cdot x^{j+1}\,.$$

■

Lemma 19 *Let $a_1,\ldots,a_n$ be a sequence of elements in $\left[0,\frac{1}{10n}\right]$. Then*

$$\left|\prod_{i=1}^{n}(1-a_i)-\left(1-\sum_{i=1}^{n}a_i\right)\right| \;\leq\; 2\left(\sum_{i=1}^{n}a_i\right)^2\,.$$

Proof We have $\prod_{i=1}^{n}(1-a_i) = \exp\left(\sum_{i=1}^{n}\log(1-a_i)\right)$. By a standard Taylor expansion of $\log(1-x)$ around $x=0$, we have for any $a_i\in[0,1/10n]$

$$|\log(1-a_i)+a_i| \;\leq\; \frac{a_i^2}{2(1-a_i)^2} \;\leq\; \frac{1}{2(9/10)^2}a_i^2 \;\leq\; \frac{5}{8}a_i^2\,.$$

In particular, this implies that

$$\left|\sum_{i=1}^{n}\log(1-a_i)+\sum_{i=1}^{n}a_i\right| \leq \frac{5}{8}\sum_{i=1}^{n}a_i^2\,. \tag{27}$$

Since $a_i\in[0,1/10n]$, this means that

$$\left|\sum_{i=1}^{n}\log(1-a_i)\right| \;\leq\; \sum_{i=1}^{n}a_i+\frac{5}{8}\sum_{i=1}^{n}a_i^2 \;\leq\; \frac{1}{10}+\frac{5}{8\cdot 100n} \;<\; \frac{1}{9}\,.$$

Using the above two inequalities, and a Taylor expansion of $\exp(x)$ around $x=0$, we have

$$\begin{aligned}\left|\exp\left(\sum_{i=1}^{n}\log(1-a_i)\right)-\left(1+\sum_{i=1}^{n}\log(1-a_i)\right)\right| \;&\leq\; \max_{\xi\in[\sum_i\log(1-a_i),0]}\frac{\exp(\xi)}{2}\left(\sum_{i=1}^{n}\log(1-a_i)\right)^2\\ &\leq\; \frac{1}{2}\left(\sum_{i=1}^{n}a_i+\frac{5}{8}\sum_{i=1}^{n}a_i^2\right)^2\\ &\leq\; \frac{1}{2}\left(\frac{13}{8}\sum_{i=1}^{n}a_i\right)^2\,.\end{aligned}$$

Combining this with Eq. (27), and using the fact that $\exp(\sum_i\log(1-a_i)) = \prod_i(1-a_i)$, we get that

$$\left|\prod_{i=1}^{n}(1-a_i)-\left(1-\sum_{i=1}^{n}a_i\right)\right| \;\leq\; \frac{5}{8}\sum_{i=1}^{n}a_i^2+\frac{1}{2}\left(\frac{13}{8}\sum_{i=1}^{n}a_i\right)^2\,.$$

Simplifying, the result follows. ■

Lemma 20 *Let $X_\sigma := \sum_{j=1}^{n}\left(\prod_{i=j+1}^{n}\left(1-\eta a_{\sigma(i)}\right)\right) b_{\sigma(j)}$ where each $f_i(x) = \frac{a_i}{2}x^2 + b_i x$ satisfies Assumption 2, $\sum_{i=1}^{n} b_i = 0$ and $\eta L \leq 1$. Then*

$$\mathbb{E}_\sigma\left[X_\sigma^2\right] \leq 5\eta^2 n^3 L^2 G^2 \log(2n),$$

where the expectation is over sampling a permutation $\sigma : [n] \to [n]$ uniformly at random.

Proof Using summation by parts on $\alpha_j = \prod_{i=j+1}^{n}\left(1-\eta a_{\sigma(i)}\right)$ and $\beta_j = b_{\sigma(j)}$, we have

$$\begin{aligned}
X_\sigma^2 &= \left(\sum_{j=1}^{n}\left(\prod_{i=j+1}^{n}\left(1-\eta a_{\sigma(i)}\right)\right) b_{\sigma(j)}\right)^2 \\
&= \left(\sum_{j=1}^{n} b_{\sigma(j)} - \sum_{j=1}^{n-1}\left(\prod_{i=j+2}^{n}\left(1-\eta a_{\sigma(i)}\right) - \prod_{i=j+1}^{n}\left(1-\eta a_{\sigma(i)}\right)\right)\sum_{i=1}^{j} b_{\sigma(i)}\right)^2 \\
&= \left(\eta\sum_{j=1}^{n-1} a_{\sigma(j+1)}\prod_{i=j+2}^{n}\left(1-\eta a_{\sigma(i)}\right)\sum_{i=1}^{j} b_{\sigma(i)}\right)^2 \\
&\leq \left(\eta L\sum_{j=1}^{n-1}\left|\sum_{i=1}^{j} b_{\sigma(i)}\right|\right)^2 \leq \eta^2 n^2 L^2\left(\sum_{j=1}^{n-1}\left|\frac{1}{j}\sum_{i=1}^{j} b_{\sigma(i)}\right|\right)^2,
\end{aligned} \tag{28}$$

where the first inequality is due to $0 \leq a_i \leq L$ for all i and $\eta L \leq 1$ which implies $1-\eta a_{\sigma(i)} \in [0,1]$ for all i. Next, without any assumptions on σ we derive a worst-case bound. Since $|b_i| \leq G$ for all i, we have

$$X_\sigma^2 \leq \eta^2 n^4 G^2 L^2. \tag{29}$$

The above worst-case bound can be used to show a $\tilde{\mathcal{O}}(1/k^2)$ upper bound on the sub-optimality of the incremental gradient method which accords with known results (see Table 1). However, a more careful examination of the random sum reveals that when choosing σ uniformly at random, a concentration of measure phenomenon occurs which allows us to establish the stronger bound in the lemma (with linear dependence rather than quadratic in n), and improve the sub-optimality. We use the following version of the Hoeffding-Serfling inequality (Bardenet et al., 2015, Corollary 2.5), stated here for completeness.

Theorem 21 (Hoeffding-Serfling inequality) *Suppose $n \geq 2$, $x_1, \ldots, x_n \in [a,b]$ with mean $\bar{x}$ and $\sigma : [n] \to [n]$ is a permutation sampled uniformly at random. Then for all $j \leq n$, for all $\delta \in [0,1]$, w.p. at least $1-\delta$ it holds that*

$$\frac{1}{j}\sum_{i=1}^{j}\left(x_{\sigma(i)} - \bar{x}\right) \leq (b-a)\sqrt{\frac{\rho_j \log(1/\delta)}{2j}},$$

where

$$\rho_j = \min\left\{1 - \frac{j-1}{n}, \left(1-\frac{j}{n}\right)\left(1+\frac{1}{j}\right)\right\}.$$

Since $\rho_j \leq 1$ for all $j \in [n]$ and by applying the inequality on $-x_1, \dots, -x_n$ and using the union bound, we have w.p. at least $1-\delta$ that

$$\left| \frac{1}{j} \sum_{i=1}^{j} \left(x_{\sigma(i)} - \bar{x} \right) \right| \leq (b-a) \sqrt{\frac{\log(2/\delta)}{2j}}.$$

Using the union bound again for the n events where each of the n partial sums do not deviate, we have

$$\begin{aligned}
\sum_{j=1}^{n} \left| \frac{1}{j} \sum_{i=1}^{j} \left(x_{\sigma(i)} - \bar{x} \right) \right| &\leq (b-a) \sqrt{\frac{\log(2n/\delta)}{2}} \sum_{j=1}^{n} \frac{1}{\sqrt{j}} \leq (b-a) \sqrt{\frac{\log(2n/\delta)}{2}} \left(1 + \int_{2}^{n} \frac{1}{\sqrt{x-1}} dx \right) \\
&= (b-a) \sqrt{\frac{\log(2n/\delta)}{2}} (2\sqrt{n-1} - 1) \ \leq 2(b-a)\sqrt{n \log(2n/\delta)}.
\end{aligned}$$

Using the above to bound Eq. (28) w.h.p. we have that w.p. at least $1-\delta$

$$X_\sigma^2 \leq \eta^2 n^2 L^2 \cdot 2G^2 n \log(2n/\delta) = 2\eta^2 n^3 G^2 L^2 \log(2n/\delta).$$

Letting $\delta = \frac{1}{n}$, we denote the event where $X_\sigma^2 \leq 4\eta^2 n^3 G^2 L^2 \log(2n)$ as E, and we have that the complement of E satisfies $\Pr\left[\bar{E}\right] \leq \frac{1}{n}$ and

$$\mathbb{E}\left[X_\sigma^2 | E\right] \leq 4\eta^2 n^3 G^2 L^2 \log(2n).$$

Finally, from the above, the law of total expectation and Eq. (29) we have

$$\begin{aligned}
\mathbb{E}\left[X_\sigma^2\right] &= \mathbb{E}\left[X_\sigma^2 | E\right] \Pr\left[E\right] + \mathbb{E}\left[X_\sigma^2 | \bar{E}\right] \Pr\left[\bar{E}\right] \\
&\leq 4\eta^2 n^3 G^2 L^2 \log(2n) \cdot 1 + \eta^2 n^4 G^2 L^2 \cdot \frac{1}{n} \\
&\leq 5\eta^2 n^3 G^2 L^2 \log(2n).
\end{aligned}$$

■

Lemma 22 *Let $X_\sigma := \sum_{j=1}^{n} \left(\prod_{i=j+1}^{n} \left(1 - \eta a_{\sigma(i)}\right) \right) b_{\sigma(j)}$ where each $f_i(x) = \frac{a_i}{2}x^2 + b_i x$ satisfies Assumption 2, $\sum_{i=1}^{n} b_i = 0$ and $\eta n L \leq 0.5$. Then*

$$\left| \mathbb{E}_\sigma \left[X_\sigma \right] \right| \leq 4\eta n G L,$$

where the expectation is over sampling a permutation $\sigma : [n] \to [n]$ uniformly at random.

Proof Letting $Y_j := \left(\prod_{i=j+1}^{n} \left(1 - \eta a_{\sigma(i)}\right) \right) b_{\sigma(j)}$, we expand Y_j to obtain

$$\begin{aligned}
\mathbb{E}\left[Y_j\right] &= \mathbb{E}\left[b_{\sigma(j)}\right] + \sum_{m=1}^{n-j} (-\eta)^m \binom{n-j}{m} \mathbb{E}\left[\sum_{j+1 \leq i_1, \dots, i_m \leq n \text{ distinct}} \left(\prod_{l=1}^{m} a_{\sigma(i_l)} \right) b_{\sigma(j)} \right] \\
&= \sum_{m=1}^{n-j} (-\eta)^m \binom{n-j}{m} \mathbb{E}\left[\sum_{j+1 \leq i_1, \dots, i_m \leq n \text{ distinct}} \left(\prod_{l=1}^{m} a_{\sigma(i_l)} \right) b_{\sigma(j)} \right] \qquad (30)
\end{aligned}$$

Repeatedly using the law of total expectation, the expectation term in the right hand side above equals

$$
\begin{aligned}
&\sum_{t_1\in[n]} \mathbb{E}\left[\sum_{j+1\le i_1,\dots,i_m\le n \text{ distinct}} \left(\prod_{l=1}^{m} a_{\sigma(i_l)}\right) b_{\sigma(j)} \middle| \sigma(i_1)=t_1\right] \Pr\left[\sigma(i_1)=t_1\right] \\
=&\frac{1}{n}\sum_{t_1\in[n]} a_{t_1}\mathbb{E}\left[\sum_{j+1\le i_2,\dots,i_m\le n \text{ distinct}} \left(\prod_{l=2}^{m} a_{\sigma(i_l)}\right) b_{\sigma(j)} \middle| \sigma(i_1)=t_1\right] \\
=&\frac{1}{n(n-1)}\sum_{t_1\in[n]}\sum_{t_2\in[n]\setminus\{t_1\}} a_{t_1}a_{t_2}\mathbb{E}\left[\sum_{j+1\le i_3,\dots,i_m\le n \text{ distinct}} \left(\prod_{l=3}^{m} a_{\sigma(i_l)}\right) b_{\sigma(j)} \middle| \sigma(i_1)=t_1,\sigma(i_2)=t_2\right] \\
=&\dots \\
=&\frac{(n-m)!}{n!}\sum_{t_1\in[n]}\sum_{t_2\in[n]\setminus\{t_1\}}\dots\sum_{t_m\in[n]\setminus\{t_1,\dots,t_{m-1}\}} a_{t_1}a_{t_2}\dots a_{t_m}\mathbb{E}\left[b_{\sigma(j)}\middle|\sigma(i_1)=t_1,\dots,\sigma(i_m)=t_m\right] \\
=&\frac{(n-m)!}{n!}\sum_{t_1\in[n]}\sum_{t_2\in[n]\setminus\{t_1\}}\dots\sum_{t_m\in[n]\setminus\{t_1,\dots,t_{m-1}\}} a_{t_1}a_{t_2}\dots a_{t_m}\frac{1}{n-m}\sum_{t_{m+1}\in[n]\setminus\{t_1,\dots,t_m\}} b_{t_{m+1}} \\
=&-\frac{(n-m)!}{n!}\sum_{t_1\in[n]}\sum_{t_2\in[n]\setminus\{t_1\}}\dots\sum_{t_m\in[n]\setminus\{t_1,\dots,t_{m-1}\}} a_{t_1}a_{t_2}\dots a_{t_m}\frac{1}{n-m}\sum_{t_{m+1}\in\{t_1,\dots,t_m\}} b_{t_{m+1}}.
\end{aligned}
$$

Recalling that $|a_i|\le L$ and $|b_i|\le G$, the above is upper bounded in absolute value by.

$$
\frac{(n-m)!}{n!}\sum_{t_1\in[n]}\sum_{t_2\in[n]\setminus\{t_1\}}\dots\sum_{t_m\in[n]\setminus\{t_1,\dots,t_{m-1}\}} L^m\frac{1}{n-m}\sum_{t_{m+1}\in\{t_1,\dots,t_m\}} G \le \frac{m}{n-m}L^m G.
$$

Plugging this back in Eq. (30) we obtain

$$
\begin{aligned}
\left|\mathbb{E}\left[Y_j\right]\right| &\le \sum_{m=1}^{n-j}\left|(-\eta)^m\binom{n-j}{m}\frac{m}{n-m}L^m G\right| \le \sum_{m=1}^{n}\eta^m\binom{n}{m}\frac{m}{n-m}L^m G \\
&= \sum_{m=1}^{n}\eta^m\binom{n}{m-1}\frac{n-m+1}{n-m}L^m G \le 2G\sum_{m=1}^{\infty}\eta^m n^{m-1}L^m \\
&\le 2G\frac{\eta L}{1-\eta n L} \le 4\eta G L.
\end{aligned}
$$

Where the last two inequalities are by the assumption $\eta n L \le 0.5$ which guarantees that the sum converges. Finally, we conclude

$$
\left|\mathbb{E}\left[X_\sigma\right]\right| \le \sum_{j=1}^{n}\left|\mathbb{E}\left[Y_j\right]\right| \le 4\eta n G L.
$$

■