Closing the convergence gap of SGD without replacement

Shashank Rajput [1] **Anant Gupta** [1] **Dimitris Papailiopoulos** [1]

Abstract

Stochastic gradient descent without replacement sampling is widely used in practice for model training. However, the vast majority of SGD analyses assumes data is sampled with replacement, and when the function minimized is strongly convex, an $\mathcal{O}\left(\frac{1}{T}\right)$ rate can be established when SGD is run for T iterations. A recent line of breakthrough works on SGD without replacement (SGDo) established an $\mathcal{O}\left(\frac{n}{T^2}\right)$ convergence rate when the function minimized is strongly convex and is a sum of n smooth functions, and an $\mathcal{O}\left(\frac{1}{T^2}+\frac{n^3}{T^3}\right)$ rate for sums of quadratics. On the other hand, the tightest known lower bound postulates an $\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$ rate, leaving open the possibility of better SGDo convergence rates in the general case. In this paper, we close this gap and show that SGD without replacement achieves a rate of $\mathcal{O}\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$ when the sum of the functions is a quadratic, and offer a new lower bound of $\Omega\left(\frac{n}{T^2}\right)$ for strongly convex functions that are sums of smooth functions.

1. Introduction

Stochastic gradient descent (SGD) is a widely used first order optimization technique used to approximately minimize a sum of functions

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

In its most general form, SGD produces a series of iterates

$$x_{i+1}=x_i-\alpha\cdot g(x,\xi_i)$$

where x_i is the i-th iterate, $g(x,\xi_i)$ is a stochastic gradient defined below, ξ_i is a random variable that determines the choice of a single or a subset of sampled functions f_i, and α represents the step size. With- and without replacement sampling of the individual component functions are regarded as some of the most popular variants of SGD. During SGD with replacement sampling, the stochastic gradient is equal to $g(x,\xi_i)=\nabla f_{\xi_i}(x)$ and ξ_i is a uniform number in $\{1,\dots,n\}$, *i.e.*, a with replacement sample from the set of gradients $\nabla f_1,\dots,\nabla f_n$. In the case of without replacement sapling, the stochastic gradient is equal to $g(x,\xi_i)=\nabla f_{\xi_i}(x)$ and ξ_i is the i-th ordered element in a random permutation of the numbers in $\{1,\dots,n\}$, *i.e.*, a without-replacement sample.

In practice, SGD without replacement is much more widely used compared to its with replacement counterpart, as it can empirically converge significantly faster (Bottou, 2009; Recht & Ré, 2013; 2012). However, in the land of theoretical guarantees, with replacement SGD has been the focal point of convergence analyses. This is because analyzing stochastic gradients sampled with replacement are significantly more tractable. The reason is simple: in expectation, the stochastic gradient is equal to the "true" gradient of F, *i.e.*, $\mathbb{E}_{\xi_i}\nabla f_{\xi_i}(x)=\nabla F(x)$. This makes SGD amenable to analyses very similar to that of vanilla gradient descent (GD), which has been extensively studied under a large variety of function classes and geometric assumptions, *e.g.*, see Bubeck et al. (2015).

Unfortunately, the same cannot be said for SGD without replacement, which has long resisted non-vacuous convergence guarantees. For example, although we have long known that SGD with replacement can achieve a $\mathcal{O}\left(\frac{1}{T}\right)$ rate for strongly convex functions F, for many years the best known bounds for SGD without replacement did not even match that rate, in contrast to empirical evidence. However, a recent series of breakthrough results on SGD without replacement has established similar or better convergence rates than SGD with replacement.

Gürbüzbalaban et al. (2015) established for the first time that for sums of quadratics or smooth functions, there exist parameter regimes under which SGDo achieves an $\mathcal{O}(n^2/T^2)$ rate compared to the $\mathcal{O}(1/T)$ rate of SGD with replacement sampling. In this case, if n is considered a constant, then SGDo becomes T times faster than SGD with replacement. Shamir (2016) showed that for one epoch, *i.e.,* one pass

[1]University of Wisconsin-Madison. Correspondence to: Shashank Rajput <rajput3@wisc.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

F is strongly convex and a sum of n quadratics		F is strongly convex and a sum of n smooth functions	
Lower bound, Safran & Shamir (2019)	$\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$	Lower bound, Safran & Shamir (2019)	$\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$
Upper bound, HaoChen & Sra (2018)	$\tilde{\mathcal{O}}\left(\frac{1}{T^2}+\frac{n^3}{T^3}\right)$	Upper bound, Nagaraj et al. (2019)	$\tilde{\mathcal{O}}\left(\frac{n}{T^2}\right)$
Our upper bound, Theorem 1	$\tilde{\mathcal{O}}\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$	Our lower bound, Theorem 2	$\Omega\left(\frac{n}{T^2}\right)$

Table 1. Comparison of our lower and upper bounds to current state-of-the-art results. Our matching bounds establish information theoretically optimal rates for SGD. We note that the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic factors.

over the n functions, SGDo achieves a convergence rate of $\mathcal{O}(1/T)$. More recently, HaoChen & Sra (2018) showed that for functions that are sums of quadratics, or smooth functions under a Hessian smoothness assumption, one could obtain an even faster rate of $\mathcal{O}\left(\frac{1}{T^2}+\frac{n^3}{T^3}\right)$. Nagaraj et al. (2019) show that for Lipschitz convex functions, SGDo is at least as fast as SGD with replacement, and for functions that are strongly convex and sum of n smooth components one can achieve a rate of $\mathcal{O}\left(\frac{n}{T^2}\right)$. This latter result was the first convergence rate that provably establishes the superiority of SGD without replacement even for the regime that n is not a constant, as long as the number of iterations T grows faster than the number n of function components.

This new wave of upper bounds has also been followed by new lower bounds. Safran & Shamir (2019) establish that there exist sums of quadratics on which SGDo cannot converge faster than $\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$. This lower bound gave rise to a gap between achievable rates and information theoretic impossibility. On one hand, SGDo on n quadratics has a rate of at least $\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$ and at most $\mathcal{O}\left(\frac{1}{T^2}+\frac{n^3}{T^3}\right)$. On the other hand, for the more general class of strongly convex functions that are sums of smooth functions the best rate is $\mathcal{O}\left(\frac{n}{T^2}\right)$. This leaves open the question of whether the upper or lower bounds are loose. This is precisely the gap we close in this work.

Our Contributions: In this work, we establish tight bounds for SGDo. We close the gap between lower and upper bounds on two of the function classes that prior works have focused on: strongly convex functions that are *i)* sums of quadratics and *ii)* sums of smooth functions. Specifically, for *i)*, we offer tighter convergence rates, *i.e.*, an upper bound that matches the lower bound given by Safran & Shamir (2019); as a matter of fact our convergence rates apply to general quadratic functions that are strongly convex, which is a little more general of a function class. For *ii)*, we provide a new lower bound that matches the upper bound by Nagaraj et al. (2019). A detailed comparison of current and proposed bounds can be found in Table 1.

A few words on the techniques used are in order. For our convergence rate on quadratic functions, we heavily rely on and combine the approaches used by Nagaraj et al. (2019) and HaoChen & Sra (2018). The convergence rate analyses proposed by HaoChen & Sra (2018) can be tightened by a more careful analysis that employs iterate coupling similar to the one used by Nagaraj et al. (2019), combined with new bounds on the deviation of the stochastic, without-replacement gradient from the true gradient of F.

For our lower bound, we use a similar construction to the one used by Safran & Shamir (2019), with the difference that each of the individual function components is not a quadratic function, but rather a piece-wise quadratic. This particular function has the property we need: it is smooth, but not quadratic. By appropriately scaling the sharpness of the individual quadratics we construct a function that behaves in a way that SGD without replacement cannot converge faster than a rate of n/T^2, no matter what step size one chooses.

We note that although our methods have an optimal dependence on n and T, we believe that the dependence on function parameters, *e.g.*, strong convexity, Lipschitz, and smoothness, can potentially be improved.

2. Related Work

The recent flurry of work on without replacement sampling in stochastic optimization extends to several variants of stochastic algorithms beyond SGD. In (Lee & Wright, 2019; Wright & Lee, 2017), the authors provide convergence rates for random cyclic coordinate descent, establishing for the first time that it can provably converge faster than stochastic coordinate descent with replacement sampling. This work is complemented by a lower bound on the gap between the random and non-random permutation variant of coordinate descent (Sun & Ye, 2019). Several other works have focused on the random permutation variant of coordinate descent, *e.g.*, see (Gurbuzbalaban et al., 2019b; Sun et al., 2019). In (Gurbuzbalaban et al., 2019a), novel bounds are given for incremental Newton based methods. In (Meng et al., 2019), Meng et al. present convergence bounds for with replacement sampling and distributed SGD. Finally, Ying et al.

(2018) present asymptotic bounds for SGDo for strongly convex functions, and show that with a constant step size it approaches the global optimizer to within smaller error radius compared to SGD with replacement. In (Shamir, 2016), linear convergence is established for a without replacement variant of SVRG.

3. Preliminaries and Notation

We focus on using SGDo to approximately find x^*, the global minimizer of the following unconstrained minimization problem

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

In our convergence bounds, we denote by T the total number of iterations of SGDo, and by K the number of epochs, *i.e.*, passes over the data. Hence,

$$T = nK.$$

In our derivations, we denote by x_i^j the i-th iterate of the j-th epoch. Consequentially, we have that $x_0^{j+1} \equiv x_n^j$.

Our results in the following sections rely on the following assumptions.

Assumption 1. *(Convexity of Components) f_i is convex for all $i \in [n]$.*

Assumption 2. *(Strong Convexity) F is strongly convex with strong convexity parameter μ, that is*

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

Assumption 3. *(Bounded Domain)*

$$\forall x : \|x - x^*\| \leq D.$$

Assumption 4. *(Bounded Gradients)*

$$\forall i, x : \|\nabla f_i(x)\| \leq G.$$

Assumption 5. *(Lipschitz Gradients) The functions f_i are L-smooth, that is*

$$\forall i, x, y : \|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|.$$

4. Optimal SGDo Rates for Quadratics

In this section, we will focus on strongly convex functions that are quadratic. We will provide a tight convergence rate that improves upon the the existing rates and matches the $\Omega\left(\frac{1}{T^2} + \frac{n^2}{T^3}\right)$ lower bound by Safran & Shamir (2019) up to logarithmic factors.

For strongly convex functions that are a sum of smooth functions, Nagaraj et al. (2019) offer a rate of $\mathcal{O}\left(\frac{n}{T^2}\right)$, whereas for strongly convex quadratics HaoChen & Sra (2018) give a convergence rate of $\mathcal{O}\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right)$. A closer comparison of these two rates reveals that neither of them can be tight due to the following observation. Assume that $n \ll K$. Then, that implies

$$\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right) < \frac{n}{T^2}.$$

At the same time, if we assume that the number of data points is significantly larger than the number of epochs that we run SGDo for, *i.e.*, $n \gg K$ we have that

$$\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right) > \frac{n}{T^2}.$$

In comparison, the known lower bound for quadratics given by Safran & Shamir (2019) is $\Omega\left(\frac{1}{T^2} + \frac{n^2}{T^3}\right)$. This makes one wonder what is the true convergence rate of SGDo in this case. We settle the optimal rates for quadratics here by providing an upper bound which, up to logarithmic factors, matches the best known lower bound.

For the special case of one dimensional quadratics, Safran & Shamir (2019) proved an upper bound matching the one we prove in this paper. Further, the paper conjectures that the proof can be extended to the generic multidimensional case. However, the authors say that the main technical barrier for this extension is that it requires a special case of a matrix-valued arithmetic-geometric mean inequality, which has only been conjectured to be true but not yet proven. The authors further conjecture that their proof can be extended to general smooth and strongly convex functions, which turns out to not be true, as we show in Corollary 1. On the other hand, we believe that our proof can be extended to the more general family of strongly convex functions, where the Hessian is Lipschitz, similar to the the way HaoChen & Sra (2018) extend their proof to that case.

In addition to Assumptions 1-5 above, here we also assume the following:

Assumption 6. *F is a quadratic function*

$$F(x) = \frac{1}{2}x^T Hx + b^T x + c,$$

where H is a positive semi-definite matrix.

Note that this assumption is a little more general than the assumption that F is a sum of quadratics. Also, note that this assumption, in combination with the assumptions on strong convexity and Lipschitz gradients implies bounds on the minimum and maximum eigenvalues of the Hessian of F, that is,

$$\mu I \preccurlyeq H \preccurlyeq LI,$$

where I is the identity matrix and $A \preccurlyeq B$ means that $x^T(A-B)x \leq 0$ for all x.

Theorem 1. *Under Assumptions 1-6, let the step size of SGDo be*

$$\alpha = \frac{8 \log T}{T\mu}$$

and the number of epochs be

$$K \geq 128 \frac{L^2}{\mu^2} \log T.$$

Then, after T iterations SGDo achieves the following rate

$$\mathbb{E}[\|x_T - x^*\|^2] = \tilde{\mathcal{O}}\left(\frac{1}{T^2} + \frac{n^2}{T^3}\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

The exact upper bound and the full proof of this Theorem are given in Appendix A, but we give a proof sketch in the next subsection.

At this point, we would like to remark that the bound on epochs $K \geq 128 \frac{L^2}{\mu^2} \log T$ may be a bit surprising as K and T are dependent. However, note that since $T = nK$, we can show that the bound on K above is satisfied if we set the number of epochs to be greater than $C \log n$ for some constant C. Furthermore, we note that the dependence of K on $\frac{L}{\mu}$ (*i.e.*, the condition number of F) is most probably not optimal. In particular both (Nagaraj et al., 2019) and (HaoChen & Sra, 2018) have a better dependence on the condition number.

The proof for Theorem 1 uses ideas from the works of HaoChen & Sra (2018) and Nagaraj et al. (2019). In particular, one of the central ideas in these two papers is that they aim to quantify the amount of progress made by SGDo over a single epoch. Both analyses decompose the progress of the iterates in an epoch as n steps of full gradient descent plus some noise term.

Similar to (HaoChen & Sra, 2018), we use the fact that the Hessian H of F is constant, which helps us better estimate the value of gradients around the minimizer. In contrast to that work, we do not require all individual components f_i to be quadratic, but rather the entire F to be a quadratic function.

An important result proved by (Nagaraj et al., 2019) is that during an epoch, the iterates do not steer off too far away from the starting point of the epoch. This allows one to obtain a reasonably good bound on the noise term, when one tries to approximate the stochastic gradient with the true gradient of F. In our analysis, we prove a slightly different version of the same result using an iterate coupling argument similar to the one in (Nagaraj et al., 2019).

The analysis of (Nagaraj et al., 2019) relies on computing the Wasserstein distance between the unconditional distribution of iterates and the distribution of iterates given a function sampled during an iteration. In our analysis, we use the same coupling, but we bypass the Wasserstein framework that (Nagaraj et al., 2019) suggests and directly obtain a bound on how far the coupled iterates move away from each other during the course of an epoch. This results, in our view, to a somewhat simpler and shorter proof.

4.1. Sketch of proof for Theorem 1

Now we give an overview of the proof. As mentioned before, similar to the previous works, the key idea is to perform a tight analysis of the progress made during an epoch. This is captured by the following Lemma.

Lemma 1. *Let the SGDo step size be $\alpha = \frac{4l \log T}{T\mu}$ and the total number of epochs be $K \geq 128 \frac{L^2}{\mu^2} \log T$, where $l \leq 2$. Then for any epoch,*

$$\mathbb{E}\left[\|x_0^j - x^*\|^2\right] \leq \left(1 - \frac{n\alpha\mu}{4}\right)\|x_0^{j-1} - x^*\|^2 + 16n\alpha^3 G^2 L^2 \mu^{-1} + 20n^3\alpha^4 G^2 L^2. \tag{1}$$

Given the result in Lemma 1, proving Theorem 1 is a simple exercise. To do so, we simply unroll the recursion (1) for K consecutive epochs. For ease of notation, define $C_1 := 16G^2L^2\mu^{-1}$ and $C_2 := 20G^2L^2$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\|x_n^K - x^*\|^2\right] \\
&\leq \left(1 - \frac{n\alpha\mu}{4}\right)\mathbb{E}\left[\|x_0^K - x^*\|^2\right] + C_1 n\alpha^3 + C_2 n^3\alpha^4 \\
&\leq \left(1 - \frac{n\alpha\mu}{4}\right)^2 \mathbb{E}\left[\|x_0^{K-1} - x^*\|^2\right] \\
&\quad + (C_1 n\alpha^3 + C_2 n^3\alpha^4)\left(1 + \left(1 - \frac{n\alpha\mu}{4}\right)\right) \\
&\qquad \vdots \\
&\leq \left(1 - \frac{n\alpha\mu}{4}\right)^{K+1} \mathbb{E}\left[\|x_0^0 - x^*\|^2\right] \\
&\quad + (C_1 n\alpha^3 + C_2 n^3\alpha^4)\sum_{j=1}^{K}\left(1 - \frac{n\alpha\mu}{4}\right)^{j-1} \\
&= \left(1 - \frac{n\alpha\mu}{4}\right)^{K+1} \|x_0^0 - x^*\|^2 \\
&\quad + (C_1 n\alpha^3 + C_2 n^3\alpha^4)\sum_{j=1}^{K}\left(1 - \frac{n\alpha\mu}{4}\right)^{j-1}.
\end{aligned}$$

We can now use the fact that $(1-x) \leq e^{-x}$ and $\left(1 - \frac{n\alpha\mu}{4}\right) \leq 1$, to get the following bound:

$$\mathbb{E}\left[\|x_n^K - x^*\|^2\right] \leq e^{-\frac{n\alpha\mu}{4}K}\|x_0^0 - x^*\|^2 + (C_1 n\alpha^3 + C_2 n^3\alpha^4)K.$$

By setting the step size to be $\alpha = \frac{4l \log T}{T\mu}$ and noting that $T = nK$, we get that

$$\begin{aligned}\mathbb{E}\left[\|x_n^K - x^*\|^2\right] &\leq e^{-n\frac{4l\log T}{T\mu}\frac{\mu}{4}K}\|x_0^0 - x^*\|^2 \\ &\quad + (n\alpha^3 C_1 + \alpha^4 n^3 C_2)K \\ &= e^{-l\log T}\|x_0^0 - x^*\|^2 + \tilde{\mathcal{O}}\left(\frac{1}{T^2} + \frac{n^2}{T^3}\right) \\ &= \frac{\|x_0^0 - x^*\|^2}{T^l} + \tilde{\mathcal{O}}\left(\frac{1}{T^2} + \frac{n^2}{T^3}\right).\end{aligned}$$

Noting that $\|x_0^0 - x^*\| \leq D$ and choosing $l = 2$ gives us the result of Theorem 1.

4.2. With- and without-replacement stochastic gradients are close

One of the key lemmas in (Nagaraj et al., 2019) establishes that once SGDo iterates get close enough to the global minimizer x^*, then any iterate at any time during an epoch x_i^j stays close to the iterate at the beginning of that epoch. To be more precise, the lemma we refer to is the following.

Lemma 2. [Nagaraj et al. (2019, Lemma 5)] *Under the assumptions of Theorem 1,*

$$\mathbb{E}[\|x_i^j - x_0^j\|^2] \leq 5i\alpha^2 G^2 + 2i\alpha(F(x_0^j) - F(x^*)).$$

We would like to note that Lemma 2 is slightly different from the one in (Nagaraj et al., 2019), which instead uses $\mathbb{E}[F(x_i^j) - F(x^*)]$ rather than $(F(x_i^j) - F(x^*))$, but their proof can be adapted to obtain the version written above. For the formal version of Lemma 2, please see Lemma 6 in the Appendix.

Now, consider the case when the iterates are very close to the optimum and hence $F(x_i^j) - F(x^*) \approx 0$. Then, Lemma 2 implies that $\mathbb{E}[\|x_i^j - x_0^j\|^2]$ does not grow quadratically in i which would generically happen for i gradient steps, but it rather grows linearly in i. This is an important and useful fact for SGDo: it shows that all iterates within an epoch remain close to x_0^j.

Hence, since the iterates of SGDo do not move too much during an epoch, then the gradients computed throughout the epoch at points x_i^j should be well approximated by gradients computed on the x_0^j iterate. Roughly, this translates to the following observation: the n gradient steps taken through a single epoch are almost equal to n steps of full gradient descent computed at x_0^j. This is in essence what allows SGDo to achieve better convergence than SGD - an epoch can be approximated by n steps of gradient descent.

Now, let σ^j represent the random permutation of the n functions f_i during the j-th epoch. Thus, $\sigma^j(i)$ is the index of the function chosen at the i-th iteration of the j-th epoch.

Proving Lemma 2 requires proving that the function value of $f_{\sigma^j(i)}(x_i^j)$, in expectation, is almost equal to $F(x_i^j)$. In particular, we prove the following claim in our supplemental material.

Claim 1. [Nagaraj et al. (2019, Lemma 4)] *If $\alpha \leq \frac{2}{L}$, then for any epoch j and i-th ordered iterate during that epoch*

$$\left|\mathbb{E}\left[F(x_i^j) - f_{\sigma^j(i)}(x_i^j) \;\middle|\; x_0^j\right]\right| \leq 2\alpha G^2. \tag{2}$$

This claim establishes that SGDo behaves almost like SGD with replacement, for which the following is true: $\mathbb{E}[f_{\sigma^j(i)}(x_i^j)] = \mathbb{E}[F(x_i^j)]$. To prove this claim, (Nagaraj et al., 2019) consider the conditional distribution of iterates, given the current function index, that is $x_i^j|\sigma_i(j)$, and the unconditional distribution of the iterates x_i^j. Then, they prove that the absolute difference $|\mathbb{E}[F(x_i^j)] - \mathbb{E}[f_{\sigma^j(i)}(x_i^j)]|$ can be upper bounded by the Wasserstein distance between these two distributions. To further upper bound the Wasserstein distance, they propose a coupling between the two distributions. To prove our slightly different version of Lemma 2, we proved (2) without using this Wasserstein framework. Instead, we use the same coupling argument to directly get a bound on (2). Below we explain the coupling and provide a short intuition.

Consider the conditional distribution of $\sigma^j|\sigma^j(i) = s$. If we take the distribution of $\sigma|\sigma(i) = 1$, we can generate the support of $\sigma^j|\sigma^j(i) = s$ by taking all permutations $\sigma|\sigma(i) = 1$ and by swapping 1 and s among them. This is essentially a coupling between these two distributions, proposed in (Nagaraj et al., 2019). Now, if we use this coupling to convert a permutation in $\sigma|\sigma(i) = 1$ to a permutation $\sigma|\sigma(i) = s$, the corresponding $x_i|\sigma(i) = 1$ and $x_i|\sigma(i) = s$ would be within a distance of $2\alpha G$. This distance bound is Lemma 2 of (Nagaraj et al., 2019).

We can now use such distance bound, and let $v_{(1,s)}$ denote a (random) vector whose norm is less than $2\alpha G$. Then,

$$\begin{aligned}\mathbb{E}\left[f_{\sigma(i)}(x_i)\right] &= \frac{1}{n}\sum_{s=1}^{n}\mathbb{E}\left[f_{\sigma(i)}(x_i)\,|\sigma(i) = s\right] \\ &= \frac{1}{n}\sum_{s=1}^{n}\mathbb{E}\left[f_s(x_i)\,|\sigma(i) = s\right] \\ &= \frac{1}{n}\sum_{s=1}^{n}\mathbb{E}\left[f_s\left(x_i + v_{(1,s)}\right)|\sigma(i) = 1\right] \\ &\leq \frac{1}{n}\sum_{s=1}^{n}\mathbb{E}\left[f_s(x_i) + (2\alpha G^2)|\sigma(i) = 1\right] \\ &= \mathbb{E}\left[F(x_i)\,|\sigma(i) = 1\right] + 2\alpha G^2.\end{aligned}$$

Similarly, for any $s \in \{1, \ldots, n\}$:

$$\mathbb{E}\left[f_{\sigma(i)}(x_i)\right] \leq \mathbb{E}\left[F(x_i)\,|\sigma(i) = s\right] + 2\alpha G^2.$$

Therefore,

$$\mathbb{E}\left[f_{\sigma(i)}\left(x_i\right)\right] \leq \frac{1}{n}\sum_{s=1}^{n}\mathbb{E}\left[F\left(x_i\right)|\sigma(i)=s\right] + 2\alpha G^2$$
$$\leq \mathbb{E}\left[F\left(x_i\right)\right] + 2\alpha G^2.$$

Similarly, we can prove that

$$\mathbb{E}\left[f_{\sigma(i)}\left(x_i\right)\right] \geq \mathbb{E}\left[F\left(x_i\right)\right] - 2\alpha G^2.$$

Combining these two results we obtain (2). The detailed proof of Claim 1 is provided in the appendix.

The full proof of Theorem 1 requires some more nuanced bounding derivations, and the complete details can be found in Appendix A.

5. Lower Bound for General Case

In the previous section, we establish that for quadratic functions the $\Omega\left(\frac{1}{T^2}+\frac{n^2}{T^3}\right)$ lower-bound by Safran & Shamir (2019) is essentially tight. This still leaves open the possibility that a tighter lower bound may exist for strongly convex functions that are not quadratic. After all, the best convergence rate known for strongly convex functions that are sums of smooth functions is of the order of n/T^2.

Indeed, in this section, we show that the convergence rate of $\mathcal{O}\left(\frac{n}{T^2}\right)$ established by Nagaraj et al. (2019) is tight.

For a certain constant C (see Appendix B for the formal version of the theorem), we show the following theorem

Theorem 2. *There exists a strongly convex function F that is the sum of n smooth convex functions, such that for any step size*

$$\frac{1}{T} \leq \alpha \leq \frac{C}{n},$$

the error after T total iterations of SGDo satisfies

$$\mathbb{E}[\|x_T - x^*\|^2] = \Omega\left(\frac{n}{T^2}\right).$$

The full proof of this theorem is provided in Appendix B, but we give an intuitive explanation of the proof later in this section.

Note that the theorem above establishes the existence of a function for which SGDo converges at rate $\Omega(\frac{n}{T^2})$, but only for the step size range $\frac{1}{T} \leq \alpha \leq \frac{C}{n}$. This is the range of the most interest because most of the upper bounds and convergence guarantees of SGDo (and SGD) work in this step size range. However, it would still be desirable to get a function on which SGDo converges at rate $\Omega(n/T^2)$ for all step sizes. Such a function would be difficult to optimize, no matter how much we tune the step size. Indeed, we show that based on Theorem 2, we can create such a function. To do that, we use a function proposed by Safran & Shamir (2019, Proposition 1), which converges slowly outside of the step size range $\frac{1}{T} \leq \alpha \leq \frac{C}{n}$.

Safran & Shamir (2019) show that there exists a strongly convex function F_2, which is the sum of n quadratics, such that for step size $\alpha \leq \frac{1}{T}$, the expected error satisfies $\mathbb{E}[\|x_T - x^*\|^2] = \Omega(1)$ (see the proof of Proposition 1, pg. 10-12 in their paper). Further, for the same function F_2, the proof of that proposition can be adapted directly to get $\mathbb{E}[\|x_T - x^*\|^2] = \Omega\left(\frac{1}{n}\right)$ for any step size $\alpha \geq \frac{C}{n}$, for any constant C.

Using this function F_2 and the function F from Theorem 2, we can create a function on which SGDo converges at rate $\Omega(\frac{n}{T^2})$ for all step sizes.

Corollary 1. *There exists a 2-Dimensional strongly convex function that is the sum of n smooth convex functions, such that for any $\alpha > 0$*

$$\mathbb{E}[\|x_T - x^*\|^2] = \Omega\left(\frac{n}{T^2}\right).$$

The proof of this corollary is provided in Appendix C.

Thus overall, we get that for any fixed step size $\mathbb{E}[\|x_T - x^*\|^2] = \Omega\left(\frac{n}{T^2}\right)$. Next, we try to explain the function construction and proof technique behind Theorem 2. The construction of the lower bound is similar to the one used by Safran & Shamir (2019). The difference is that the prior work considers quadratic functions, while we consider a slightly modified piece-wise quadratic function.

Specifically, we construct the following function $F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ as

$$F(x) = \begin{cases} \frac{x^2}{2}, & \text{if } x \geq 0 \\ \frac{Lx^2}{2}, & \text{if } x < 0, \end{cases}$$

where n is an even number. Of the n component functions f_i, half of them are defined as follows:

$$\text{if } i \leq \frac{n}{2}, \text{ then } f_i(x) = \begin{cases} \frac{x^2}{2} + \frac{Gx}{2}, & \text{if } x \geq 0 \\ \frac{Lx^2}{2} + \frac{Gx}{2}, & \text{if } x < 0, \end{cases}$$

and the other half of the functions are defined as follows:

$$\text{if } i > \frac{n}{2}, \text{ then } f_i(x) = \begin{cases} \frac{x^2}{2} - \frac{Gx}{2}, & \text{if } x \geq 0 \\ \frac{Lx^2}{2} - \frac{Gx}{2}, & \text{if } x < 0. \end{cases}$$

For our construction, we set L to be a big enough positive constant. See for example, Fig. 1.

Next we ought to verify that this function abides to Assumptions 1-5. Note that Assumption 1 is satisfied, as it can

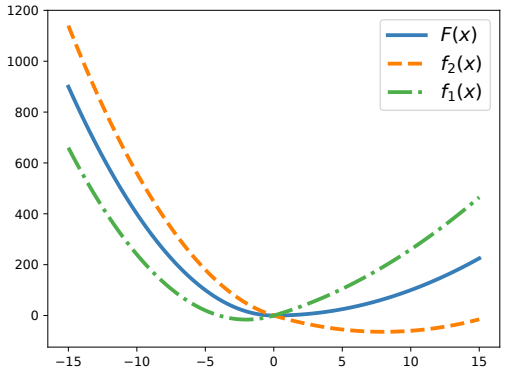

Figure 1. Lower bound construction. Note that $f_1(x)$ represents the component functions of the first kind, and $f_2(x)$ represents the component functions of the second kind, and $F(x)$ represents the overall function.

be seen that functions f_i's are all continuous and convex. Next, we need to show that Assumption 2 holds, that is F is strongly convex. We will show that this is true by proving the following equivalent definition of strong convexity: a function f is μ-strongly convex if $g(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex. We can see that this is true for F with $\mu = 1$.

In the proof of Theorem 2, we initialize at the origin. In that case, in the proof we also prove that Assumptions 3 and 4 hold. In particular, we show that the iterates do not go outside of a bounded domain, and inside this domain, the gradient is bounded by G. Finally, let us focus on Assumption 5. To prove that these functions have Lipschitz gradients, we need to show

$$\forall x, y : |\nabla f_i(x) - \nabla f_i(y)| \leq L|x - y|.$$

If $xy \geq 0$, that is x and y lie on the same side of the origin, then this is simple to see because they both lie on the same quadratic. Otherwise WLOG, assume $x < 0$ and $y > 0$. Also, assume WLOG that f_i is function of the first kind, that is $i \leq \frac{n}{2}$ and hence the linear term in $f_i(x)$ is $\frac{Gx}{2}$. Then,

$$\begin{aligned}|\nabla f_i(x) - \nabla f_i(y)| &= \left|Lx + \frac{G}{2} - y - \frac{G}{2}\right| \\ &= y - Lx \\ &\leq Ly - Lx \\ &\leq L|y - x|.\end{aligned}$$

Overall, the difficulty in the analysis comes from the fact that unlike the functions considered by Safran & Shamir (2019), our functions are piece-wise quadratics.

Let us initialize at $x_0^1 = 0$ (the minimizer). We will show that in expectation, at the end of K epochs, the iterate would be at a certain distance (in expectation). Note that the progress made over an epoch is just the sum of gradients (multiplied by $-\alpha$) over the epoch:

$$x_n^j - x_0^j = -\alpha \sum_{i=1}^{n} \nabla f_{\sigma^j(i)}(x_i)$$

where $\sigma^j(i)$ represents the index of the i-th function chosen in the j-th epoch. Next, note that the gradients from the linear components $\pm\frac{G}{2}x$ are equal to $\pm\frac{G}{2}$, that is they are constant. Thus, they will cancel out over an epoch.

However the gradients from the quadratic components do not cancel out, and in fact that part of the gradient will not even be unbiased, in the sense that if $x_t \geq 0$, the gradient at x_t from the quadratic component $\frac{x^2}{2}$ will be less in magnitude than the gradient from the quadratic component $\frac{Lx^2}{2}$ at $-x_t$.

The idea is to now ensure that if an epoch starts off near the minimizer, then the iterates spend a certain amount of time in the $x < 0$ region, so that they "accumulate" a lot of gradients of the form Lx, which makes the sum of the gradients at the end of the epoch biased away from the minimizer.

To ensure that the iterates spend some time in the $x < 0$ region, we analyze the contribution of the linear components during the epoch. This is because when the iterates are already near the minimizer $x \approx 0$, the gradient contribution of the quadratic terms would be small, and the dominating component during an epoch would come from the linear terms. What this means is that in the middle of an epoch, it is the linear terms which contribute the most towards the "iterate movement", even though at the end of that epoch their gradients get cancelled out and what remains is the contribution of the quadratic terms.

Then, to obtain a lower bound matching the upper bound given by Nagaraj et al. (2019), observe that it is indeed this contribution of the linear terms that we require to get a tight bound on. This is because, the upper bound from the aforementioned work was also in fact directly dependent on the movement of iterates away from the minimizer during an epoch, caused by the stochasticity in the gradients (cf. Lemma 5 of Nagaraj et al. (2019)). We give below the informal version of the main lemma for the proof:

Lemma 3. [Informal] *Let* $(\sigma_1, \ldots, \sigma_n)$ *be a random permutation of* $\{\underbrace{+1, \ldots, +1}_{\frac{n}{2} \text{ times}}, \underbrace{-1, \ldots, -1}_{\frac{n}{2} \text{ times}}\}$. *Then for* $i < n/2$,

$$\mathbb{E}\left[\left|\sum_{j=1}^{i} \sigma_j\right|\right] \geq C\sqrt{i},$$

where C *is a universal constant.*

For the formal version, please see Lemma 12 in Appendix B.

For the purpose of intuition, ignore the contribution of gradients from the quadratic terms. Then, the lemma above says that during an epoch, the gradients from the linear terms would move the iterates approximately $\Omega\left(\alpha\sqrt{n}\frac{G}{2}\right)$ away from the minimizer (after we multiply by the step size α).

This implies that in the middle of an epoch, with (almost) probability $1/2$ the iterates would be near $x \approx -\Omega\left(\alpha\sqrt{n}\frac{G}{2}\right)$ and with (almost) probability $1/2$ the iterates would be near $x \approx \Omega\left(\alpha\sqrt{n}\frac{G}{2}\right)$. Hence, over the epoch, the accumulated quadratic gradients multiplied by the step size would look like

$$\begin{aligned}&\sum_{i=1}^{n}\mathbb{E}\left[-\alpha(L\mathbb{1}_{x_i^j<0}+\mathbb{1}_{x_i^j\geq 0})x_i^j\right]\\&\approx -\alpha\sum_{i=1}^{n}\left(\frac{1}{2}L\Omega\left(-\alpha\sqrt{n}\frac{G}{2}\right)+\frac{1}{2}\Omega\left(\alpha\sqrt{n}\frac{G}{2}\right)\right)\\&=\Omega(L\alpha^2 n\sqrt{n}).\end{aligned}$$

If this happens for K epochs, we get that the accumulated error would be $\Omega(L\alpha^2 n\sqrt{n}K) = \Omega\left(\frac{1}{\sqrt{n}K}\right)$ for $\alpha \in [1/nK, 1/n]$. Since $E[|x_T|] \geq 1/\sqrt{n}K$, we know that $E[|x_T - 0|^2] \geq 1/nK^2 = n/T^2$. Since 0 is the minimizer of our function in this setting, we have constructed a case where SGDo achieves error

$$E[|x_T - x^*|^2] \geq n/T^2.$$

This completes the sketch of the proof and the complete proof of Theorem 2 is given in Appendix B.

5.1. Numerical verification

To verify our lower bound of Theorem 2, we ran SGDo on the function described in Eq. (5) with $L = 4$. The step size regimes that were considered were $\alpha = \frac{1}{T}, \frac{2\log T}{T}, \frac{4\log T}{T}, \frac{8\log T}{T}$, and $\frac{1}{n}$. The plot for $\alpha = \frac{4\log T}{T}$ is shown in Figure 2. The plots for the other step size regimes are provided in Appendix D.

The step size regimes considered cover the range specified in the statement of Theorem 2. Looking at Figure 2 (and the figures in Appendix D for the other step size regimes), the dependence of the convergence rate on K indeed looks exactly like $1/K^2$. However, looking at the figures for the dependence of the convergence rate on n, we see that they look like $\frac{(\log T)^2}{n}$. This suggests that the tightest possible lower bound for SGDo with constant step size on strongly convex smooth functions might have a logarithmic term in the numerator. Next, we explain the details of the experiment.

Consider any one of the step size regimes specified above, say $\alpha = \frac{4\log T}{T}$. For this regime, we ran two experiments:

1. We fix $n = 500$ and vary K from 30 to 200, and

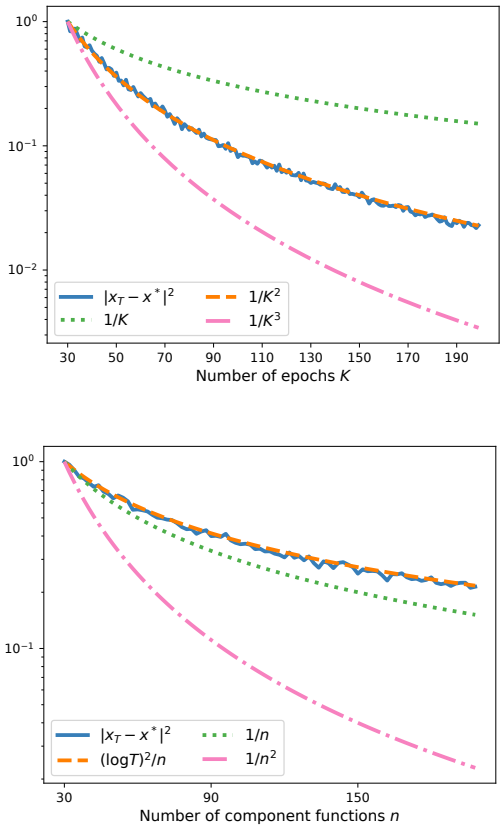

Figure 2. Running SGDo on the function F used in our lower bound (Theorem 2) confirms that the rate of convergence of SGDo on this function is indeed $\Omega(\frac{1}{nK^2}) = \Omega(\frac{n}{T^2})$. The curves are normalized so that they begin at the same point.

2. we fix $K = 500$ and vary n from 30 to 200.

Consider the first experiment, where $n = 500$ and K is varied. For each value of K, say $K = 50$, we set $\alpha = \frac{4\log T}{T} = \frac{4\log(nK)}{nK} = \frac{4\log(500*50)}{500*50}$ and ran SGDo with this constant step size α on the sum of $n = 500$ functions for $K = 50$ epochs, and the final error was recorded. This was repeated 1000 times to reduce variance. The final mean error after these 1000 runs gave us one point, which we plotted for $K = 50$ on the top subfigure of Figure 2. Repeating the same for all values of K from 30 to 200 gave us the top subfigure of Figure 2. The same procedure was followed for the second experiment where we fix K and vary n, and that gave us the bottom subfigure of Figure 2. The optimization was initialized at the origin, that is $x_0^1 = 0$. These pairs of experiments were performed for all values of step size regimes in the list $(\frac{1}{T}, \frac{2\log T}{T}, \frac{4\log T}{T}, \frac{8\log T}{T}$, and $\frac{1}{n})$.

Now, we justify the ranges of n and K considered in our experiments. We wanted to verify that the lower bound on

the error of SGDo is indeed

$$\Omega\left(\frac{n}{T^2}\right) = \Omega\left(\frac{1}{nK^2}\right) \qquad \text{[Theorem 2]}$$

instead of the previously known best lower bound

$$\Omega\left(\frac{1}{T^2} + \frac{n^2}{T^3}\right) = \Omega\left(\frac{1}{nK^2}\left(\frac{1}{n} + \frac{1}{K}\right)\right). \qquad \text{[Safran \& Shamir (2019)]}$$

Looking at the RHS of the two equations above, we can see that the dependence of the two lower bounds on K differs only when $n \gg K$ and the dependence on n differs only when $K \gg n$. Thus for example, when we wanted to check dependence on K, we set $n = 500$ which was bigger than every K in the range 30 to 200.

The code for these experiments is available at `https://github.com/shashankrajput/SGDo`.

5.2. Discussion on possible improvements

Theorem 2 hints that for faster convergence rates in the epoch based random shuffling SGD, we would not just require smooth and strongly convex functions, but also potentially require that the Hessians of such functions to be Lipschitz.

We conjecture that Hessian Lipschitzness is sufficient to get the convergence rate of Theorem 1. We think that this is interesting, because the optimal rates for both SGD with replacement and vanilla gradient descent only require strong convexity and gradient smoothness. However, here we prove that an optimal rate for SGDo requires the function to be quadratic as well (or at the very least have a Lipschitz Hessian), and SGDo seems to converge slower if the Hessian is not Lipschitz.

6. Conclusions and Future Work

SGD without replacement has long puzzled researchers. From a practical point of view, it always seems to outperform SGD with replacement, and is the algorithm of choice for training modern machine learning models. From a theoretical point of view, SGDo has resisted tight convergence analysis that establish its performance benefits. A recent wave of work established that indeed SGDo can be faster than SGD with replacement sampling, however a gap still remained between the achievable rates and the best known lower bounds.

In this paper we settle the optimal performance of SGD without replacement for functions that are quadratics, and strongly convex functions that are sums of n smooth functions. Our results indicate that a possible improvement in convergence rates may require a fundamentally different step size rule and significantly different function assumptions.

As future directions, we believe that it would be interesting to establish rates for variants of SGDo that do not repermute the functions at every epoch. This is something that is common in practice, where a random permutation is only performed once every few epochs without a significant drop in performance. Current theoretical bounds are inadequate to explain this phenomenon, and a new theoretical breakthrough may be required to tackle it.

We however believe that one of the strongest new theoretical insights introduced by (Nagaraj et al., 2019) and used in our analyses can be of significance in a potential attempt to analyze other variants of SGDo as the one above. This insight is that of iterate coupling. That is the property that SGDo iterates are only mildly perturbed after swapping only two elements of a permutation. Such a property is reminiscent to that of algorithmic stability, and a deeper connection between that and iterate coupling is left as a meaningful intellectual endeavor for future work.

Acknowledgements

We would like to thank the ICML reviewers for their constructive feedback in improving the structure of the Appendix. The authors also attribute the motivation for Corollary 1 (see the paragraph after Theorem 2) to the comments of Reviewer #3.

This research is supported by an NSF CAREER Award #1844951, a Sony Faculty Innovation Award, an AFOSR & AFRL Center of Excellence Award FA9550-18-1-0166, and an NSF TRIPODS Award #1740707.

References

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

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, pp. 1–36, 2015.

Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P. Convergence rate of incremental gradient and incremental newton methods. *SIAM Journal on Optimization*, 29(4): 2542–2565, 2019a.

Gurbuzbalaban, M., Ozdaglar, A., Vanli, N. D., and Wright, S. J. Randomness and permutations in coordinate descent methods. *Mathematical Programming*, pp. 1–28, 2019b.

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

Lee, C.-P. and Wright, S. J. Random permutations fix a worst case for cyclic coordinate descent. *IMA Journal of Numerical Analysis*, 39(3):1246–1275, 2019.

Meng, Q., Chen, W., Wang, Y., Ma, Z.-M., and Liu, T.-Y. Convergence analysis of distributed stochastic gradient descent with shuffling. *Neurocomputing*, 337:46–57, 2019.

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

Nagaraj, D., Jain, P., and Netrapalli, P. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pp. 4703–4711, 2019.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Recht, B. and Ré, C. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. *arXiv preprint arXiv:1202.4184*, 2012.

Recht, B. and Ré, C. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

Safran, I. and Shamir, O. How good is sgd with random shuffling? *arXiv preprint arXiv:1908.00045*, 2019.

Shamir, O. Without-replacement sampling for stochastic gradient methods. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 46–54. Curran Associates, Inc., 2016.

Sun, R. and Ye, Y. Worst-case complexity of cyclic coordinate descent: $\mathcal{O}(n^2)$ gap with randomized version. *Mathematical Programming*, pp. 1–34, 2019.

Sun, R., Luo, Z.-Q., and Ye, Y. On the efficiency of random permutation for admm and coordinate descent. *Mathematics of Operations Research*, 2019.

Wright, S. J. and Lee, C.-p. Analyzing random permutations for cyclic coordinate descent. *arXiv preprint arXiv:1706.00908*, 2017.

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