SGD without Replacement: Sharper Rates for General Smooth Convex Functions

Dheeraj Nagaraj [1] **Praneeth Netrapalli** [2] **Prateek Jain** [2]

Abstract

We study stochastic gradient descent *without replacement* (SGDo) for smooth convex functions. SGDo is widely observed to converge faster than true SGD where each sample is drawn independently *with replacement* (Bottou, 2009) and hence, is more popular in practice. But it's convergence properties are not well understood as sampling without replacement leads to coupling between iterates and gradients. By using method of exchangeable pairs to bound Wasserstein distance, we provide the first non-asymptotic results for SGDo when applied to *general smooth, strongly-convex* functions. In particular, we show that SGDo converges at a rate of $O(1/K^2)$ while SGD is known to converge at $O(1/K)$ rate, where K denotes the number of passes over data and is required to be *large enough*. Existing results for SGDo in this setting require additional *Hessian Lipschitz assumption* (Gürbüzbalaban et al., 2015; HaoChen & Sra, 2018). For *small* K, we show SGDo can achieve same convergence rate as SGD for *general smooth strongly-convex* functions. Existing results in this setting require $K = 1$ and hold only for generalized linear models (Shamir, 2016). In addition, by careful analysis of the coupling, for both large and small K, we obtain better dependence on problem dependent parameters like condition number.

1. Introduction

In this paper, we study the standard finite sum optimization problem that arises in most machine learning based optimization problems: $F(x) := \frac{1}{n}\sum_{i=1}^{n} f(x;i)$, where $f(x;i) : \mathbb{R}^d \to \mathbb{R}$ is the i-th component function. For example, in standard ERM and deep learning training $f(x;i)$ denotes the loss function w.r.t. the i-th data point. Stochastic Gradient Descent (SGD), originally proposed by Robbins & Monro (1985), has emerged as one of the most popular techniques to solve this problem.

At t-th step, SGD updates the iterate by $x_{t+1} = x_t - \eta\nabla f(x_t;i_t)$ where $\nabla f(x_t;i_t)$ is the gradient of $f(x_t;i_t)$ and i_t is selected uniformly at random *with replacement* yielding $\mathbb{E}_{i_t}[\nabla f(x_t;i_t)] = \nabla F(x_t)$. SGD has been extensively studied in literature and a vast number of results are known in many different settings, most prominent being that of convex optimization (Benaïm, 1999; Borkar, 2009; Kushner & Yin, 2003; Bubeck, 2015; Bottou et al., 2018; Schmidt & Roux, 2013; Gower et al., 2019; Nguyen et al., 2018; Vaswani et al., 2018). While SGD holds the rare distinction of being both theoretically well understood and practically widely used, there are still significant differences between the versions of SGD that are studied in theory vs those used in practice. Resolving this discrepancy is an important open question. One of the major differences is that SGD is widely used in practice *with out replacement* (SGDo). SGDo uses the standard SGD update but in each epoch/pass over data, every $i \in [n]$ is sampled *exactly* once but in a uniformly random position i.e., *without replacement*. This implies, that $\mathbb{E}_{i_t}[\nabla f(x_t;i_t)] = \nabla F(x_t)$ *does not* hold anymore, making the analysis of SGDo significantly more challenging.

Studies however, have shown *empirically* that SGDo converges significantly faster than SGD (Bottou, 2009). Gürbüzbalaban et al. (2015) provided first formal guarantee for this observation and proved that the suboptimality of SGDo after K epochs behaves as $O\left(1/K^2\right)$, where as the suboptimality of SGD is known to be $O\left(1/nK\right)$ (and this bound is tight). Under the same assumptions, (HaoChen & Sra, 2018) improve upon the result of (Gürbüzbalaban et al., 2015) and show a suboptimality bound of $O\left(1/n^2K^2 + 1/K^3\right)$ where n is the number of samples and K is the number of epochs. However, both the above given guarantees require Hessian Lipschitz, gradient Lipschitz (also known as smoothness) and strong convexity assumptions on F. In contrast, SGD's rate of $O\left(1/nK\right)$ requires only strong convexity. It is also known that this

[1] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA [2] Microsoft Research, Bengaluru, Karnataka, India. Correspondence to: Dheeraj Nagaraj <dheeraj@mit.edu>, Praneeth Netrapalli <praneeth@microsoft.com>, Prateek Jain <prajain@microsoft.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Paper	Guarantee	Assumptions	Step Sizes
(Gürbüzbalaban et al., 2015)	$O\left(\frac{C(n,d)}{K^2}\right)$	Lipschitz, Strong convexity Smoothness, **Hessian Lipschitz** $K > \kappa^{1.5}\sqrt{n}$	$\frac{1}{K}$
(HaoChen & Sra, 2018)	$\tilde{O}\left(\frac{1}{n^2K^2}+\frac{1}{K^3}\right)$		$\frac{\log nK}{\mu nK}$
This paper, Theorem 1	$\tilde{O}\left(\frac{1}{nK^2}\right)$	Lipschitz, Strong convexity Smoothness, $K > \kappa^2$	$\frac{\log nK}{\mu nK}$
(Shamir, 2016)	$O\left(\frac{1}{nK}\right)$	Lipschitz, Strong convexity, Smoothness **Generalized Linear Function**, $K = 1$	$\frac{1}{\mu nK}$
This Paper, Theorem 2	$O\left(\frac{1}{nK}\right)$	Lipschitz, Strong convexity, Smoothness	$\min\left(\frac{2}{L}, \frac{\log nK}{\mu nK}\right)$
(Shamir, 2016)	$O\left(\frac{1}{\sqrt{nK}}\right)$	Lipschitz **Generalized Linear Function**, $K = 1$	$\frac{1}{\sqrt{nK}}$
This Paper, Theorem 3	$O\left(\frac{1}{\sqrt{nK}}\right)$	Lipschitz, **Smoothness**	$\min\left(\frac{2}{L}, \frac{1}{\sqrt{nK}}\right)$

Table 1. Comparison of our results with previously known results in terms of number of functions n and number of epochs K. For simplicity, we suppress the dependence on other problem dependent parameters such as Lipschitz constant, strong convexity, smoothness etc. These dependencies are clearly stated in Theorems 1, 2 and 3.

rate cannot be improved with out smoothness (gradient Lipschitz). So, in this work, we ask the following fundamental question: *Does SGDo converge at a faster rate than SGD for general smooth, strongly-convex functions (with out Hessian Lipschitz assumption)?*

We answer the above question in affirmative and show that SGDo can achieve convergence rate of $\widetilde{O}\left(1/nK^2\right)$ for general smooth, strongly-convex functions. Moreover, for $K \lesssim n$, our result improves upon the best known rates (HaoChen & Sra, 2018). Our results also improve upon the $O(1/nK)$ rate of SGD once $K \geq O(\kappa^2)$ where κ is the condition number of the problem (2). In contrast, (HaoChen & Sra, 2018) requires $K \geq O\left(\kappa^{1.5} \cdot \sqrt{n}\right)$ to improve upon the rates of SGD. Note that in practice one takes only a few passes over the data and hence a practical method needs to demonstrate faster rate for a small number of epochs. Finally, our analysis yields improved dependence on problem dependent parameters like κ.

As mentioned above, in many settings, we are interested in the performance of SGDo, when the number of passes K is quite small. (Shamir, 2016) considers an extreme version of this setting, and obtains suboptimality bounds for SGDo for the *first* pass, for the special case of *generalized linear models*. These bounds are similar to the standard suboptimality bounds for SGD of $O\left(1/n\right)$ and $O\left(1/\sqrt{n}\right)$ for convex functions with and with out strong convexity respectively (here number of passes $K = 1$).

For the small K regime, we obtain similar convergence rates of $O\left(1/nK\right)$ and $O(1/\sqrt{nK})$ for smooth convex functions with and with out strong convexity respectively. This improves upon (Shamir, 2016) by showing the result for *general* convex functions, for any number of epochs and also in terms of dependence on problem dependent parameters. These results are summarized in Table 1. The first three rows of the table compare our result for *large* K against those of (Gürbüzbalaban et al., 2015) and (HaoChen & Sra, 2018). The next two rows compare our result for *small* K (i.e., constant K) against that of (Shamir, 2016) in the presence of strong convexity. The final two rows compare our result for *small* K against that of (Shamir, 2016) *without* strong convexity.

As noted earlier, the main challenge in analyzing SGDo is that in expectation, the update does not follow gradient descent (GD). That is, $\mathbb{E}_{i_t}[\nabla f(x_t; i_t)] \neq \nabla F(x_t)$. The main proof strategy is to bound the bias in SGDo update, i.e., $\|\mathbb{E}_{i_t}[\nabla f(x_t; i_t)] - \nabla F(x_t)\|$ as well as the variance associated with the update, i.e., $\mathbb{E}_{i_t}[\|\nabla f(x_t; i_t)\|^2] - \|\nabla F(x_t)\|^2$. To bound the bias term, we use a novel coupling technique for limiting Wasserstein distance between the paths of SGDo and SGD. For the variance term, we use smoothness of the function to show that compared to SGD, SGDo naturally leads to variance reduction. We put together these two terms and analyze them in different settings of K (constant vs condition number dependent K) to obtain our final results (Theorems 1, 2 and 3).

Organization: We introduce problem setup, notations, and a brief overview of related works in Section 2. In Section 3, we present our main results, compare it with existing work and give a rough outline of our proof strategy. In Section 4, we introduce coupling and Wasserstein distances and use these ideas to state and prove some important lemmas in our context. Section 5 presents the proofs of our main results. Finally, we conclude with Section 6. Due to space limitations, some of the proofs are presented in the appendix.

2. Problem Setup

Given convex functions $f(;1),\ldots,f(;n):\mathbb{R}^d\to\mathbb{R}$, we consider the following optimization problem:

$$\min_{x\in\mathcal{W}} F(x) := \frac{1}{n}\sum_{i=1}^{n} f(x;i)\,, \tag{1}$$

where $\mathcal{W}\subset\mathbb{R}^d$ is a closed convex set. We will refer to F as the objective function and $f(\cdot;i)$ as the component functions. Henceforth, we let x^* denote the minimizer of F over $\mathcal{W}$ and $\Pi_{\mathcal{W}}$ denote the projection operator onto the set $\mathcal{W}$. We study SGDo when applied to the above problem. The algorithm takes K passes (epochs) over the data. In each pass, it goes through the component functions in a random order $\sigma_k:[n]\to[n]$ and requires a step size sequence $\alpha_{k,i}\geq 0$ for $k\in[K]$, $0\leq i\leq n-1$ for computing stochastic gradient. See Algorithm 2 for pseudocode. For simplicity of analysis and exposition, we assume constant step-sizes $\alpha_{k,i}$. For our analysis, we assume that the component functions are twice differentiable, uniformly G lipschitz and L smooth over $\mathcal{W}$.

Assumption 1 (Lipschitz Continuity)**.** *There exists $G>0$ such that $\|\nabla f(x;i)\|\leq G\;\forall\, x\in\mathcal{W}$ and $i\in[n]$.*

Assumption 2 (Smoothness/Gradient Lipschitz)**.** *There exists $L>0$ such that, $\|\nabla f(x;i)-\nabla f(y;i)\|\leq L\|x-y\|\;\forall\, x,y\in\mathcal{W}$ and $i\in[n]$.*

In addition, we require strong-convexity of $F(\cdot)$ for Theorem 1 and Theorem 2 to hold.

Assumption 3 (Strongly-convex)**.** *There exists $\mu>0$ s. t. $F(y)\geq F(x)+\langle\nabla F(x),y-x\rangle+\frac{\mu}{2}\|y-x\|^2\;\forall\, x,y\in\mathcal{W}$.*

We define condition number κ of the problem (1) as:

$$\kappa = L/\mu, \tag{2}$$

where L and μ are smoothness and strong convexity parameters defined by Assumptions 2 and 3, respectively. Finally, we denote the distance of initial point x_i^0 from the optimum by D i.e., $D\overset{\text{def}}{=}\|x_i^0-x^*\|$.

2.1. Related Work

Gradient descent (GD) and it's variants are well-studied in literature (Bubeck, 2015). If Assumption 1 is satisfied, then suboptimality of GD (more precisely subgradient descent) with averaging is bounded by $O(G\cdot D/\sqrt{K})$ where K is the number of GD iterations. With Assumption 2, the convergence rate improves to $O(LD^2/K)$ and with additional Assumption 3, it further improves to $O(e^{-K/\kappa}LD^2)$ where κ is defined by (2). For smooth functions, accelerated gradient descent (AGD) further improves the rates to $O(LD^2/K^2)$ and $O(e^{-K/\sqrt{\kappa}}LD^2)$, in the above two settings respectively (Bubeck, 2015).

Each iteration of GD requires a full pass over data and hence requires prohibitively large $O(n\cdot T_f)$ computation where T_f is the computation cost of evaluating gradient of any $f(x;i)$ at any x. In contrast, SGD (Algorithm 1) requires only $O(T_f)$ computation per step. Moreover, SGD's suboptimality after K passes over the data is $O(G\cdot D/\sqrt{nK})$ with Assumption 1. Similarly, it is $O(G^2/\mu\cdot 1/(nK))$ if Assumption 3 also holds. Without any additional assumptions, these rates are known to be tight.

With additional Assumption 2, people have designed acceleration methods for SGD such as SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013), SDCA (Shalev-Shwartz & Zhang, 2013) and SAG (Schmidt et al., 2017) - these methods achieve variance reduction using previous iterates in the algorithm and obtain faster rates of convergence. Note that none of these results applies for SGDo as sampling without replacement introduces dependencies between iterates and gradients. But, at a high-level, our result shows that SGDo naturally achieves some amount of variance reduction giving better convergence rate than SGD. There have also been other works that study SGDo. (Recht & Re, 2012) relate the performance of SGDo to a noncommutative version of arithmetic-geometric mean inequality (Zhang, 2014; Israel et al., 2016). However, this conjecture has not yet been fully resolved. (Ying et al., 2018) shows that for a small enough fixed step size, the distibution of SGDo converges closer to the optimum than SGD.

3. Main Results

In this section, we present our main results for SGDo and the main ideas behind the proofs. Recall that x_i^k denotes the iterates of SGDo and let x^* be a minimizer of $F(\cdot)$ over $\mathcal{W}$. We define $d_{i,k}:=\|x_i^k-x^*\|$. We now present our first result that improves upon the convergence rate of SGD for large K.

Theorem 1. *Suppose $F(\cdot)$ satisfies Assumptions 1-3. Fix $l>0$ and let number of epochs $K>32l\kappa^2$. Let x_i^k be the iterates of SGDo (Algorithm 2) when applied to $F(\cdot)$ with constant learning rate $\alpha_{k,i}=\alpha\overset{\text{def}}{=}4l\frac{\log nK}{\mu nK}$. Then the following holds for the tail average $\hat{x}\overset{\text{def}}{=}\frac{1}{K-\lceil K/2\rceil+1}\sum_{k=\lceil K/2\rceil}^{K}x_0^k$ of the iterates:*

$$\mathbb{E}[F(\hat{x})]-F(x^*)\leq O\left(\mu\frac{d_{0,1}^2}{(nK)^l}\right)+O\left(\frac{\kappa^2G^2}{\mu}\frac{(\log nK)^2}{nK^2}\right).$$

Remarks:

- The error has two terms – first term depending on initial error $d_{0,1}$ and second term depending on problem parameters L,G and μ. The dependence on initial error can be made to decay very fast by choosing l to be

Algorithm 1 SGD: SGD with replacement

```
Input: Functions $f(x;i), i \in [n]$, convex set $\mathcal{W}$, maximum number of epochs $K$, step-size sequence $\alpha_{k,i},\ k \in [K],\ i \in [n]$
$x_n^0 \leftarrow 0$
for $k \in [K]$ do
    $x_0^k \leftarrow x_n^{k-1}$
    for $0 \leq i \leq n-1$ do
        $j_i^k \leftarrow Unif[n]$
        $x_{i+1}^k \leftarrow \Pi_{\mathcal{W}}\left(x_i^k - \alpha_{k,i}\nabla f\left(x_i^k; j_i^k\right)\right)$
    end for
end for
```

Algorithm 2 SGDo: SGD without replacement

```
Input: Functions $f(x;i), i \in [n]$, convex set $\mathcal{W}$, number of epochs $K$, step-size sequence $\alpha_{k,i},\ k \in [K],\ i \in [n]$
$x_n^0 \leftarrow 0$
for $k \in [K]$ do
    $x_0^k \leftarrow x_n^{k-1}$
    $\sigma_k \leftarrow$ uniformly random permutation of $[n]$
    for $0 \leq i \leq n-1$ do
        $x_{i+1}^k \leftarrow \Pi_{\mathcal{W}}\left(x_i^k - \alpha_{k,i}\nabla f\left(x_i^k; \sigma_k(i+1)\right)\right)$
    end for
end for
```

a large enough constant, i.e., $K = \Omega(\kappa^2)$. In this case, the leading order term is the second term which decays as $O\left(\frac{1}{nK^2}\right)$. Our result improves upon the $O\left(\frac{G^2}{\mu nK}\right)$ rate of SGD once $K > O\left(\kappa^2\right)$.

- Our result improves upon the state of the art result for SGDo by (HaoChen & Sra, 2018) as long as $K \leq \kappa n$, which captures the most interesting setting in practice. Furthermore, we do not require the additional Hessian Lipschitz assumption. For the sake of clarity, (HaoChen & Sra, 2018) keeps all parameters other than μ constant and takes $\kappa = \Theta(1/\mu)$ to get suboptimality of $\tilde{O}\left(\frac{\kappa^4}{n^2K^2} + \frac{\kappa^4}{K^3} + \frac{\kappa^6}{K^4}\right)$. By the same token, our suboptimality is $\tilde{O}(\frac{\kappa^3}{nK^2})$.

Note that Theorem 1 requires the number of passes $K > \kappa^2$. We now present results that apply even for small number of passes. In this setting, we match the rates of SGD. The problem setting is the same as Theorem 1.

Theorem 2. *Suppose $F(\cdot)$ satisfies Assumptions 1-3. Let x_i^k be the iterates of SGDo (Algorithm 2) when applied to $F(\cdot)$ with constant learning rate $\alpha_{k,i} = \alpha \overset{\text{def}}{=} \min\left(\frac{2}{L}, 4l\frac{\log nK}{\mu nK}\right)$ for a fixed $l > 0$. Then the following holds for the tail average $\hat{x} \overset{\text{def}}{=} \frac{1}{K-\lceil\frac{K}{2}\rceil+1}\sum_{k=\lceil\frac{K}{2}\rceil}^{K} x_0^k$ of the iterates:*

$$\mathbb{E}[F(\hat{x})] - F(x^*) = O\left(\mu\frac{\|x_0^1-x^*\|^2}{(nK)^l} + L\frac{\|x_0^1-x^*\|^2}{(nK)^{(l+1)}}\right) + O\left(\frac{G^2\log nK}{\mu nK} + \frac{L^2G^2\log nK}{\mu^3n^2K^2}\right).$$

Remarks:

- The dependence on initial error can be made to decay as fast as any polynomial by choosing l to be a large enough constant.
- Our result is the first such result for general smooth, strongly-convex functions and for arbitrary K; recall that the result of (Shamir, 2016) requires F to be a generalized linear function and requires $K = 1$. Furthermore, even in setting of (Shamir, 2016), our result improves upon best known bounds when $nK > \kappa^2$. In this case, our error rate is $O\left(\frac{G^2\log nK}{\mu nK}\right)$ that matches the rate of SGD upto log factors. The result of (Shamir, 2016) does not obtain this rate even when $n \to \infty$.

The above two theorems require $F(\cdot)$ to be strongly convex (Assumption 3). We now present our result for F that need not satisfy the strong convexity assumption.

Theorem 3. *Suppose $F(\cdot)$ satisfies Assumptions 1-2 and that $\mathsf{diam}(\mathcal{W}) \leq D$. The average $\hat{x} \overset{\text{def}}{=} \frac{\sum_{i=0}^{n-1}\sum_{k=1}^{K} x_i^k}{Kn}$ of SGDo (Algorithm 2) with constant learning rate $\alpha_{k,i} = \alpha \overset{\text{def}}{=} \min\left(\frac{2}{L}, \frac{D}{G\sqrt{Kn}}\right)$ satisfies:*

$$\mathbb{E}[F(\hat{x})] - F(x^*) \leq \frac{D^2L}{4nK} + \frac{3GD}{\sqrt{nK}}.$$

Remarks:

- The second term of $O\left(\frac{GD}{\sqrt{nK}}\right)$ is the same as the rate of SGD in this setting. This becomes the leading order term once $nK \geq \frac{L^2D^2}{G^2}$.
- Our result is the first such result for general smooth, Lipschitz convex functions. The earlier result by (Shamir, 2016) applied only for generalized linear models but does not require smoothness assumption.

3.1. Necessity of Smoothness

In the classical analysis of SGD for $O\left(\frac{1}{nK}\right)$ rate, one only requires Assumptions 1 and 3. In this section, we outline an argument showing that obtaining a better rate than $O\left(\frac{1}{nK}\right)$ for SGDo as in Theorem 1, requires additional Assumption 2 (smoothness). In contrast, it is well known that the rate of $O\left(\frac{1}{nK}\right)$ is tight for SGD even with additional Assumption 2. Consider the example where all the component functions are same. i.e, $f(x;i) = g(x)$ for all $1 \leq i \leq n$ and $x \in \mathbb{R}^d$. Then, running SGDo for optimizing $F(x) := \frac{1}{n}\sum_{i=1}^{n} f(x;i) = g(x)$ for K epochs (over a closed convex set $\mathcal{W}$) is the same as running gradient descent over $F(x)$ for nK iterations.

Given any $T = nK$, (Bubeck, 2015, Theorem 3.13) shows the existence of a function satisfying Assumptions 1 and 3 and a closed convex set $\mathcal{W}$ such that the suboptimality of all iterates up to the T^{th} iteration of GD–hence, for all the iterates up to K^{th} epoch of SGDo–is lower bounded by $\frac{G^2}{8\mu n K}$. This establishes the necessity of Assumption 2 for obtaining improved rates over SGD as in Theorem 1.

3.2. Proof Strategy

As a general note, in the proofs, we assume that $\mathcal{W} = \mathbb{R}^d$, which avoids the projection operator $\Pi_{\mathcal{W}}$. All the steps go through in a straight forward fashion even with this projection operator. When we try to apply the classical proof of rate of convergence of SGD to SGDo, the major problem we encounter is that $\mathbb{E}[f(x_i^k;\sigma_k(i+1))] \neq \mathbb{E}[F(x_i^k)]$. In section 4, we propose a coupling sequence and use it to bound a certain Wasserstein distance to argue that $\mathbb{E}[f(x_i^k;\sigma_k(i+1))] \approx \mathbb{E}[F(x_i^k)]$. This along with standard analysis tools then yields Theorems 2 and 3.

However, this technique does not suffice to obtain faster rate as in Theorem 1. So, to prove Theorem 1, we show that in expectation, SGDo over one epoch approximates one step of GD applied to F. Therefore, K epochs of SGDo approximates GD iterates after K iterations. Recall

$$x_0^{k+1} = x_0^k - \alpha_k \sum_{i=0}^{n-1} \nabla f(x_i^k, \sigma_k(i+1)).$$

If $x_i^k \approx x_0^k$, then the equation above implies:

$$x_0^{k+1} \approx x_0^k - \alpha_k \sum_{i=0}^{n-1} \nabla f(x_0^k, \sigma_k(i+1)) = x_0^k - n\alpha_k \nabla F(x_0^k)\,.$$

We observe that the right hand side is one step of gradient descent. Lemma 5 in Section 4 makes this argument rigorous as it shows that $\mathbb{E}[\|x_i^k - x_0^k\|^2]$ becomes small as $F(x_0^k) \to F(x^*)$.

4. Coupling and Wasserstein distance

In this section, we develop the required machinery to show:

$$\mathbb{E}[f(x_i^k;\sigma_k(i+1))] \approx \mathbb{E}[F(x_i^k)]. \tag{3}$$

Define the following exchangeable pair: suppose we run the algorithm for $k-1$ epochs using permutations $\sigma_1, \dots, \sigma_{k-1}$ to obtain x_0^k. When $k = 1$, this means that we start with the same starting point x_0^1. We draw two independent uniform permutations: σ_k and σ_k'. If we run the k-th epoch with permutation σ_k, we denote the k-th epoch iterates by $(x_i(\sigma_k))_{i=1}^n$ to explicity show the dependence on σ_k. Similarly, the sequence obtained by using the permutation σ_k' for the k-th epoch is denoted by $(x_i(\sigma_k'))_{i=1}^n$). It is clear that $(x_i(\sigma_k'))_{i=1}^n$ is independent and indentically distributed as $(x_i(\sigma_k))_{i=1}^n$. We note:

$$\mathbb{E}[f(x_i(\sigma_k');\sigma_k(i+1))] = \mathbb{E}[f(x_i(\sigma_k'))] = \mathbb{E}[F(x_i^k)]\,. \tag{4}$$

Here the first equality follows from the fact that σ_k is independent of σ_k' (and applying Fubini's theorem which allows us to exchange the order of integration with respect to σ_k and σ_k'). The second equality follows from the fact that $x_i(\sigma_k')$ and $x_i(\sigma_k)$ are identically distributed. Therefore, to show (3), we need to show that: $\mathbb{E}[f(x_i(\sigma_k');\sigma_k(i+1))] - \mathbb{E}[f(x_i(\sigma_k);\sigma_k(i+1))] \approx 0$. Since $f(\cdot;j)$ is uniformly lipschitz, a bound on the Wasserstein distance between $x_i(\sigma_k)$ and $x_i(\sigma_k')$ would imply the above result. That is, Lemma 1 shows that $\mathbb{E}[f(x_i(\sigma_k');\sigma_k(i+1))] - \mathbb{E}[f(x_i(\sigma_k);\sigma_k(i+1))]$ is bounded by the Wasserstein distance between $x_i(\sigma_k)$ and $x_i(\sigma_k')$, and Lemma 4 then bounds the Wasserstein distance, to bound the above quantity.

We first introduce some notation to prove the result. Let $\mathcal{D}_{i,k} := \mathcal{L}(x_i(\sigma_k))$ and $\mathcal{D}_{i,k}^{(r)} := \mathcal{L}\left(x_i(\sigma_k)|\sigma_k(i+1) = r\right)$. Here $\mathcal{L}(X)$ denotes the distribution of the random variable X. We let $\mathsf{Lip}_d(\beta)$ be the set of all β lipschitz functions from $\mathbb{R}^d \to \mathbb{R}$.

Definition 1. *Let P and Q be two probability measures over $\mathbb{R}^d$ such that $\mathbb{E}_{X\sim P}[\|X\|^2] < \infty$ and $\mathbb{E}_{Y\sim Q}[\|Y\|^2] < \infty$. Let $X \sim P$ and $Y \sim Q$ be random vectors defined on a common measure space (i.e, they are coupled). We define Wasserstein-1 and Wasserstein-2 distances between P and Q as:*

$$\mathsf{D}_{\mathsf{W}}^{(1)}(P,Q) \overset{\text{def}}{=} \inf_{\substack{(X,Y):\\ X\sim P\\ Y\sim Q}} \mathbb{E}[\|X-Y\|]\,, \textit{ and}$$

$$\mathsf{D}_{\mathsf{W}}^{(2)}(P,Q) \overset{\text{def}}{=} \inf_{\substack{(X,Y):\\ X\sim P\\ Y\sim Q}} \sqrt{\mathbb{E}[\|X-Y\|^2]},$$

respectively. Here the infimum is over all joint distributions over (X,Y) with prescribed marginals.

By Jensen's inequality, we have $\mathsf{D}_{\mathsf{W}}^{(2)}(P,Q) \geq \mathsf{D}_{\mathsf{W}}^{(1)}(P,Q)$. The following result gives a fundamental characterization of Wasserstein distance (Santambrogio, 2015).

Theorem 4 (Kantorovich Duality)**.** *Let P and Q satisfy the conditions in Definition 1. Let $X \sim P$ and $Y \sim Q$ then:*

$$\mathsf{D}_{\mathsf{W}}^{(1)}(P,Q) = \sup_{g\in\mathsf{Lip}_d(1)} \mathbb{E}[g(X)] - \mathbb{E}[g(Y)].$$

We can use Theorem 4 to bound the approximation error in (3) in terms of average Wasserstein-1 distance between $\mathcal{D}_{i,k}$ and $\mathcal{D}_{i,k}^{(r)}$.

Lemma 1.

$$\left|\mathbb{E}[F(x_i^k)]-\mathbb{E}[f(x_i^k;\sigma_k(i+1))]\right| \le \frac{G}{n}\sum_{r=1}^{n} \mathsf{D}_{\mathsf{W}}^{(1)}\left(\mathcal{D}_{i,k},\mathcal{D}_{i,k}^{(r)}\right)$$

Proof. Let $R_j := \sigma_k(j)$ for all $j \in [n]$. Using (4):

$$\begin{aligned}
&\left|\mathbb{E}[F(x_i^k)] - \mathbb{E}[f(x_i^k;R_{i+1})]\right| \\
&= \left|\mathbb{E}[f(x_i(\sigma'_k);R_{i+1})] - \mathbb{E}[f(x_i(\sigma_k);R_{i+1})]\right| \\
&\le \frac{1}{n}\sum_{r=1}^{n}\left|\mathbb{E}\left[f(x_i(\sigma'_k);r)\right] - \mathbb{E}\left[f(x_i(\sigma_k);r)\middle|R_{i+1}=r\right]\right| \\
&\le \frac{1}{n}\sum_{r=1}^{n}\sup_{g\in\mathsf{Lip}_d(G)}\left(\mathbb{E}\left[g(x_i(\sigma'_k))\right] - \mathbb{E}\left[g(x_i(\sigma_k))\middle|R_{i+1}=r\right]\right) \\
&= \frac{1}{n}\sum_{r=1}^{n} G\cdot\mathsf{D}_{\mathsf{W}}^{(1)}\left(\mathcal{D}_{i,k},\mathcal{D}_{i,k}^{(r)}\right),
\end{aligned}$$

where the second step follows from triangle inequality and the fact that $x_i(\sigma'_k)$ is independent of σ_k. Second to last inequality follows from the fact that $f \in Lip_d(G)$ and the last inequality follows from Theorem 4. We also used the fact that conditioned on $\sigma_k(i+1) = r$, $x_i^k(\sigma'_k) \sim \mathcal{D}_{i,k}$ □

From Lemma 1, we see that we only need to upper bound $\mathsf{D}_{\mathsf{W}}^{(1)}\left(\mathcal{D}_{i,k},\mathcal{D}_{i,k}^{(r)}\right)$. We hope to use the definition of Wasserstein-1 distance (Definition 1) by constructing a nice coupling between $\mathcal{D}_{i,k}$ and $\mathcal{D}_{i,k}^{(r)}$, that we present in the following lemma; see Appendix B for a proof of the lemma. We note that in essence, this lemma is similar to the stability analysis in (Hardt et al., 2015).

Lemma 2. *Given k, suppose $\alpha_{k,i}$ is a non-increasing function of i and $\alpha_{k,0} \le \frac{2}{L}$. Then almost surely, $\forall\ i\ \in\ [n]$,*

$$\|x_i(\sigma'_k) - x_i(\sigma_k)\| \le 2G\alpha_{k,0}\cdot\left|\{j \le i : \sigma_k(j) \ne \sigma'_k(j)\}\right|. \tag{5}$$

Here $|\{j \le i : \sigma_k(j) \ne \sigma'_k(j)\}|$ is the number of iterations till i where the two permutations σ_k and σ'_k choose different component functions.

A key ingredient in the proof of Lemma 2 is the following standard result which says that gradient step with small enough step size is contracting for smooth convex functions.

Lemma 3. *(Nesterov, 2013, Theorem 2.1.5) Let $\nabla^2 g$ denote the Hessian of g. If g is convex and $\|\nabla^2 g\| \le L$, then,*

$$\|\nabla g(x) - \nabla g(y)\|^2 \le L\langle\nabla g(x) - \nabla g(y), x - y\rangle.$$

4.1. Coupling σ_k and σ'_k

In this section, we construct a coupling between σ_k and σ'_k that minimizes the bound in Lemma 2. Let $\mathcal{S}_n$ be the set of all permutations over n letters. For $a, b \in [n]$, we define the exchange function $E_{a,b} : \mathcal{S}_n \to \mathcal{S}_n$: for any $\tau \in \mathcal{S}_n$, $E_{a,b}(\tau)$ gives a new permutation where a-th and b-th entries of τ are exchanged and it keeps everything else same. We construct the operator $\Lambda_{r,i} : \mathcal{S}_n \to \mathcal{S}_n$:

$$\Lambda_{r,i}(\tau) = \begin{cases}\tau & \text{if } \tau(i+1) = r\\ E_{i+1,j}(\tau) & \text{if } \tau(j) = r \text{ and } j \ne i+1\end{cases}$$

Basically, $\Lambda_{r,i}$ makes a single swap so that $i+1$-th position of the permutation is r. Clearly, if σ_k is a uniformly random permutation, then $\Lambda_{r,i}(\sigma_k)$ has the same distribution as $\sigma_k|\sigma_k(i+1) = r$. We use the defintion of $\mathsf{D}_{\mathsf{W}}^{(1)}$ to conclude:

Lemma 4. *Let k be fixed. When $\alpha_{k,0} \le \frac{2}{L}$ and $\alpha_{k,i}$ be a non-increasing function of i,*

$$\mathsf{D}_{\mathsf{W}}^{(1)}\left(\mathcal{D}_{i,k},\mathcal{D}_{i,k}^{(r)}\right) \le \mathsf{D}_{\mathsf{W}}^{(2)}\left(\mathcal{D}_{i,k},\mathcal{D}_{i,k}^{(r)}\right) \le 2\alpha_{k,0}G,$$

where $\mathcal{D}_{i,k}$, $\mathcal{D}_{i,k}^{(r)}$ are defined above. Consequently, from Lemma 1, we conclude:

$$\left|\mathbb{E}[F(x_i^k)] - \mathbb{E}[f(x_i^k;\sigma_k(i+1))]\right| \le 2\alpha_{k,0}G^2.$$

Proof. Let σ_k be a uniformly random permutation and $r \in [n]$. Therefore, $x_i(\sigma_k) \sim \mathcal{D}_{i,k}$ and $x_i\left(\Lambda_{r,i}(r)\right) \sim \mathcal{D}_{i,k}^{(r)}$. This gives a coupling between $\mathcal{D}_{i,k}$ and $\mathcal{D}_{i,k}^{(r)}$. By definition of Wasserstein distance:

$$\mathsf{D}_{\mathsf{W}}^{(2)}\left(\mathcal{D}_{i,k},\mathcal{D}_{i,k}^{(r)}\right) \le \sqrt{\mathbb{E}\|x_i(\sigma_k) - x_i(\Lambda_{r,i}\sigma_k)\|^2} \tag{6}$$

It is clear that $\left|\{j \le i : \sigma_k(j) \ne \left[\Lambda_{r,i}(\sigma_k)\right](j)\}\right| \le 1$ almost surely. Therefore, from Lemma 2, we conclude that $\|x_i(\sigma_k) - x_i(\Lambda_{r,i}\sigma_k)\| \le 2\alpha_{k,0}G$ almost surely. Together with Equation 6, and the fact that $\mathsf{D}_{\mathsf{W}}^{(2)} \ge \mathsf{D}_{\mathsf{W}}^{(1)}$ we conclude the result. □

The lemmas presented above tightly bound the difference in suboptimality between iterates of SGD and SGDo. These will be used in proving Theorems 2 and 3, matching the rates of SGD. For Theorem 1, we need to show that there is some amount of automatic variance reduction while running SGDo. In order to do this, we need to show that the iterates x_i^k do not move much when they are close to the optimum. The following lemma makes this precise.

Lemma 5. *Recall that $d_{i,k} \overset{\text{def}}{=} \|x_i^k - x^*\|$. Let $\alpha_{k,0} < \frac{2}{L}$ and $\alpha_{k,j}$ be a non-increasing sequence in j for a given k.*

For any $i \in [n]$, we have:

$$\mathbb{E}[\|x_i^k - x_0^k\|^2] \le 5i\alpha_{k,0}^2 G^2 + 2i\alpha_{k,0}\cdot \mathbb{E}\left[F(x_0^k) - F(x^*)\right],$$
$$\textit{and } \mathbb{E}[d_{i,k}^2] \le \mathbb{E}[d_{0,k}^2] + 5i\alpha_{k,0}^2 \cdot G^2,$$

See Appendix B for a detailed proof of the above lemma.

5. Proofs of Main Results

In this section, we will present proofs of Theorems 1, 2 and 3 using the results from the previous section.

5.1. Proof of Theorem 1

In this subsection, for the sake of clarity of notation, we take $R_j \overset{\text{def}}{=} \sigma_k(j)$ for every $j \in [n]$. Recall the definition $d_{i,k} := \|x_i^k - x^*\|$. From the definition of SGDo, and the choice of step sizes $\alpha_{k,i} = \alpha = 4l\frac{\log nK}{\mu n K}$, we have:

$$x_0^{k+1} = x_0^k - \alpha\sum_{i=0}^{n-1}\nabla f(x_i^k, R_{i+1}).$$

Using the hypothesis that $\alpha \le \frac{2}{L}$ (since $\frac{\mu}{L} \le 1$) and taking norm squared on both sides,

$$\begin{aligned}
d_{0,k+1}^2 &= d_{0,k}^2 - 2\alpha\sum_{i=0}^{n-1}\langle\nabla f(x_i^k, R_{i+1}), x_0^k - x^*\rangle \\
&\quad + \alpha^2\left\|\sum_{i=0}^{n-1}\nabla f(x_i^k, R_{i+1})\right\|^2 \\
&= d_{0,k}^2 - 2n\alpha\langle\nabla F(x_0^k), x_0^k - x^*\rangle \\
&\quad - 2\alpha\sum_{i=0}^{n-1}\langle\nabla f(x_i^k, R_{i+1}) - \nabla F(x_0), x_0^k - x^*\rangle \\
&\quad + \alpha^2\left\|\sum_{i=0}^{n-1}\nabla f(x_i^k, R_{i+1})\right\|^2 \\
&\le d_{0,k}^2(1 - n\alpha\mu) - 2n\alpha(F(x_0^k) - F(x^*)) \\
&\quad - 2\alpha\sum_{i=0}^{n-1}\langle\nabla f(x_i^k, R_{i+1}) - \nabla F(x_0), x_0^k - x^*\rangle \\
&\quad + \alpha^2\left\|\sum_{i=0}^{n-1}\nabla f(x_i^k, R_{i+1})\right\|^2, \qquad (7)
\end{aligned}$$

where we used strong convexity of F in the third step. We consider the term:

$$\begin{aligned}
T_1 &\overset{\text{def}}{=} -2\alpha\sum_{i=0}^{n-1}\langle\nabla f(x_i^k, R_{i+1}) - \nabla F(x_0), x_0^k - x^*\rangle \\
&= -2\alpha\sum_{i=0}^{n-1}\langle\nabla f(x_i^k; R_{i+1}) - \nabla f(x_0^k; R_{i+1}), x_0^k - x^*\rangle.
\end{aligned}$$

$$\begin{aligned}
\implies \mathbb{E}[T_1] \\
&= -2\alpha\mathbb{E}\sum_{i=0}^{n-1}\langle\nabla f(x_i^k; R_{i+1}) - \nabla f(x_0^k; R_{i+1}), x_0^k - x^*\rangle \\
&\le 2\alpha L\sum_{i=0}^{n-1}\mathbb{E}\left[\|x_i^k - x_0^k\|.\|x_0^k - x^*\|\right] \\
&\le 2\alpha L\sum_{i=0}^{n-1}\sqrt{\mathbb{E}\|x_i^k - x_0^k\|^2}\sqrt{\mathbb{E}\|x_0^k - x^*\|^2} \\
&\le 2\alpha L n\sqrt{\mathbb{E}[d_{0,k}^2]}\sqrt{5n\alpha^2G^2 + 2n\alpha\mathbb{E}\left[F(x_0^k) - F(x^*)\right]}, \qquad (8)
\end{aligned}$$

where we used Cauchy-Schwarz and smoothness in the second step and Lemma 5 in the last step. Applying aritmetic mean - geometric mean inequality on (8), we have:

$$\begin{aligned}
\mathbb{E}[T_1] &\le \alpha L n\left[\tfrac{\mu\mathbb{E}[d_{0,k}^2]}{4L} + \tfrac{4L\left(5n\alpha^2G^2 + 2n\alpha\mathbb{E}\left[F(x_0^k) - F(x^*)\right]\right)}{\mu}\right] \\
&= \tfrac{\alpha\mu n}{4}\mathbb{E}[d_{0,k}^2] + \tfrac{20L^2\alpha^3n^2G^2}{\mu} \\
&\quad + \tfrac{8\alpha^2L^2n^2}{\mu}\mathbb{E}\left[F(x_0^k) - F(x^*)\right] \qquad (9)
\end{aligned}$$

We now consider

$$T_2 \overset{\text{def}}{=} \alpha^2\left\|\sum_{i=0}^{n-1}\nabla f(x_i^k; R_{i+1})\right\|^2.$$

We use the fact that: $\nabla F(x^*) = 0 = \sum_{i=0}^{n-1}\nabla f(x^*; R_{i+1})$ in the equation above to conclude:

$$\begin{aligned}
T_2 &= \alpha^2\left\|\sum_{i=0}^{n-1}\nabla f(x_i^k; R_{i+1}) - \nabla f(x^*; R_{i+1})\right\|^2 \\
&\le \alpha^2\left[\sum_{i=0}^{n-1}\left\|\nabla f(x_i^k; R_{i+1}) - \nabla f(x^*; R_{i+1})\right\|\right]^2 \\
&\le \alpha^2L^2\left[\sum_{i=0}^{n-1}\left\|x_i^k - x^*\right\|\right]^2 = \alpha^2L^2\sum_{i=1}^{n}\sum_{j=1}^{n}d_{i,k}d_{j,k}.
\end{aligned}$$

Taking expectation, we have

$$\begin{aligned}
\mathbb{E}[T_2] &\le \alpha^2L^2\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[d_{i,k}d_{j,k}] \\
&\le \alpha^2L^2\sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{\mathbb{E}[d_{i,k}^2]}\sqrt{\mathbb{E}[d_{j,k}^2]} \\
&\le \alpha^2L^2n^2\left[\mathbb{E}[d_{0,k}^2] + 5n\alpha^2G^2\right], \qquad (10)
\end{aligned}$$

where we again used Cauchy-Schwarz inequality and Lemma 5.

Plugging the inequalities (9) and (10) in (7), we conclude:

$$\begin{aligned}\mathbb{E}[d_{0,k+1}^2] &\leq \mathbb{E}[d_{0,k}^2](1-n\alpha\mu) - 2n\alpha\mathbb{E}(F(x_0^k)-F(x^*)) \\ &\quad + \tfrac{\alpha\mu n}{4}\mathbb{E}[d_{0,k}^2] + \tfrac{20L^2\alpha^3n^2G^2}{\mu} + \tfrac{8\alpha^2L^2n^2\mathbb{E}\left[F(x_0^k)-F(x^*)\right]}{\mu} \\ &\quad + \alpha^2L^2n^2\mathbb{E}[d_{0,k}^2] + 5\alpha^4L^2G^2n^3 \\ &\leq \mathbb{E}[d_{0,k}^2](1-\tfrac{3n\alpha\mu}{4}+\alpha^2n^2L^2) \\ &\quad - 2n\alpha\left(1-\tfrac{4\alpha nL^2}{\mu}\right)\mathbb{E}\left[F(x_0^k)-F(x^*)\right] \\ &\quad + \tfrac{20L^2\alpha^3n^2G^2}{\mu} + 5\alpha^4L^2G^2n^3. \qquad (11)\end{aligned}$$

It is clear that $1-\frac{3n\alpha\mu}{4}+\alpha^2n^2L^2 \leq 1-\frac{n\alpha\mu}{2}$. In (11), we use the fact that $F(x_0^k) \geq F(x^*)$ to conclude:

$$\mathbb{E}[d_{0,k}^2] \leq \left(1-\tfrac{n\alpha\mu}{2}\right)\mathbb{E}[d_{0,k-1}^2]+\tfrac{20L^2\alpha^3n^2G^2}{\mu}+5\alpha^4L^2G^2n^3$$

Unrolling the recursion above, we have:

$$\begin{aligned}\mathbb{E}[d_{0,k}^2] &\leq \left(1-\tfrac{n\alpha\mu}{2}\right)^k d_{0,1}^2 \\ &\quad + \sum_{t=0}^{\infty}\left(1-\tfrac{n\alpha\mu}{2}\right)^t\left[\tfrac{20L^2\alpha^3n^2G^2}{\mu}+5\alpha^4L^2G^2n^3\right] \\ &\leq \exp\left(-\tfrac{nk\alpha\mu}{2}\right)d_{0,1}^2 \\ &\quad + \tfrac{2}{n\alpha\mu}\left[\tfrac{20L^2\alpha^3n^2G^2}{\mu}+5\alpha^4L^2G^2n^3\right] \\ &\leq \exp\left(-\tfrac{nk\alpha\mu}{2}\right)d_{0,1}^2 + \tfrac{40L^2\alpha^2nG^2}{\mu^2} + \tfrac{10\alpha^3L^2G^2n^2}{\mu}\end{aligned}$$

Taking $\alpha = 4l\frac{\log nK}{\mu nK}$ and $k = \frac{K}{2}$, we have:

$$\mathbb{E}[d_{0,\frac{K}{2}}^2] \leq \tfrac{d_{0,1}^2}{(nK)^l} + \tfrac{40L^2\alpha^2nG^2}{\mu^2} + \tfrac{10\alpha^3L^2G^2n^2}{\mu}.$$

We now analyze the suffix averaging scheme given. Adding (11) from $k=\frac{K}{2}$ to $k=K$, we conclude:

$$\begin{aligned}& n\alpha\frac{\sum_{k=\lceil\frac{K}{2}\rceil}^{K}\mathbb{E}\left[F(x_0^k)-F(x^*)\right]}{K-\lceil\frac{K}{2}\rceil+1} \\ &\leq \frac{\mathbb{E}[d_{0,K/2}^2]}{K-\lceil\frac{K}{2}\rceil+1} + \tfrac{20L^2\alpha^3n^2G^2}{\mu} + 5\alpha^4L^2G^2n^3.\end{aligned}$$

Here we have used the fact that $2n\alpha\left(1-\frac{4\alpha nL^2}{\mu}\right) \geq n\alpha$. Since $F(\hat{x}) \leq \frac{\sum_{k=\lceil K/2\rceil}^{K}F(x_0^k)}{K-\lceil K/2\rceil+1}$ by convexity of F, we have:

$$\begin{aligned}&\mathbb{E}\left[F(\hat{x})-F(x^*)\right] \\ &\leq \tfrac{2}{nK\alpha}\left[\tfrac{d_{0,1}^2}{(nK)^l}\right] + \left[\tfrac{80L^2\alpha G^2}{\mu^2K} + \tfrac{20\alpha^2L^2G^2n}{\mu K}\right] \\ &\quad + \tfrac{20L^2\alpha^2nG^2}{\mu} + 5\alpha^3L^2G^2n^2 \\ &= O\left(\mu\tfrac{d_{0,1}^2}{(nK)^l}\right) + O\left(\tfrac{\kappa^2G^2}{\mu}\tfrac{(\log nK)^2}{nK^2}\right). \qquad \square\end{aligned}$$

5.2. Proof of Theorem 3

We note that we have taken $\alpha_{k,i} = \alpha < \frac{2}{L}$. By definition, $x_{i+1}^k = \Pi_{\mathcal{W}}\left(x_i^k - \alpha\nabla f(x_i^k;\sigma_k(i+1))\right)$. We take $r = \sigma_k(i+1)$ below. Taking norm squared and using Lemma 6

$$\begin{aligned}\|x_{i+1}^k - x^*\|^2 &\leq \|x_i^k - x^*\|^2 + \alpha^2\|\nabla f(x_i^k;r)\|^2 \\ &\quad - 2\alpha\langle\nabla f(x_i;r), x_i^k - x^*\rangle \\ &\leq \|x_i^k - x^*\|^2 + \alpha^2G^2 - 2\alpha(f(x_i^k;r) - f(x^*;r))\end{aligned}$$

In the last step we have used convexity of of $f(;j)$ and the fact that $\|\nabla f(;j)\| \leq G$. Taking expectation above, and noting that $\mathbb{E}f(x^*;\sigma_k(i+1)) = F(x^*)$ and using Lemma 4, we have:

$$\begin{aligned}\mathbb{E}\|x_{i+1}^k - x^*\|^2 &\leq \mathbb{E}\|x_i^k - x^*\|^2 - 2\alpha\mathbb{E}(F(x_i^k) - F(x^*)) \\ &\quad + 5\alpha^2G^2 \qquad (12)\end{aligned}$$

Summing from $i=0$ to $n-1$ and $k=1$ to K, we conclude:

$$\begin{aligned}&\tfrac{1}{nK}\sum_{k=1}^{K}\sum_{i=0}^{n-1}(F(x_i^k) - F(x^*)) \leq \tfrac{\mathbb{E}\|x_0^1 - x^*\|^2}{2\alpha nK} + \tfrac{5}{2}\alpha G^2 \\ &\leq \tfrac{D^2}{2nK}\max(\tfrac{L}{2},\tfrac{G\sqrt{nK}}{D}) + \tfrac{5G^2}{2}\min(\tfrac{2}{L},\tfrac{D}{G\sqrt{nK}}) \\ &\leq \tfrac{D^2}{2nK}\left(\tfrac{L}{2} + \tfrac{G\sqrt{nK}}{D}\right) + \tfrac{5G^2}{2}.\tfrac{D}{G\sqrt{nK}} = \tfrac{D^2L}{4nK} + \tfrac{3GD}{\sqrt{nK}}.\end{aligned}$$

By convexity of $F(\cdot)$, we conclude:

$$F(\hat{x}) \leq \tfrac{1}{nK}\sum_{k=1}^{K}\sum_{i=0}^{n-1}F(x_i^k)\,.$$

6. Conclusions

In this paper, we study stochastic gradient descent with out replacement (SGDo), which is widely used in practice. When the number of passes is large, we present the first convergence result for SGDo, that is faster than SGD, under standard smoothness, strong convexity and Lipschitz assumptions where as prior work uses additional Hessian Lipschitz assumption. Our convergence rates also improve upon existing results in practically interesting regimes. When the number of passes is small, we present convergence results for SGDo that match those of SGD for general smooth convex functions. These are the first such results for general smooth convex functions as previous work only showed such results for generalized linear models. In order to prove these results, we use techniques from optimal transport theory to couple variants of SGD and relate their performances. These ideas may be of independent interest in the analysis of SGD style algorithms with some dependencies.

Acknowledgements

This research was partially supported by ONR N00014-17-1-2147 and MIT-IBM Watson AI Lab.

References

Benaïm, M. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pp. 1–68. Springer, 1999.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

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

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

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

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtarik, P. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. Why random reshuffling beats stochastic gradient descent. *arXiv preprint arXiv:1510.08560*, 2015.

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

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Israel, A., Krahmer, F., and Ward, R. An arithmetic–geometric mean inequality for products of three matrices. *Linear Algebra and its Applications*, 488:1–12, 2016.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

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

Nguyen, L. M., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. Sgd and hogwild! convergence without the bounded gradients assumption. *arXiv preprint arXiv:1802.03801*, 2018.

Recht, B. and Re, C. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies. Technical report, and consequences. Technical report, University of Wisconsin-Madison, 2012.

Robbins, H. and Monro, S. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pp. 102–109. Springer, 1985.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, pp. 99–102, 2015.

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

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

Shamir, O. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pp. 46–54, 2016.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

Ying, B., Yuan, K., Vlaski, S., and Sayed, A. H. Stochastic learning under random reshuffling. *arXiv preprint arXiv:1803.07964*, 2018.

Zhang, T. A note on the non-commutative arithmetic-geometric mean inequality. *arXiv preprint arXiv:1411.5058*, 2014.