Understanding the Impact of Model Incoherence on Convergence of Incremental SGD with Random Reshuffle

Shaocong Ma [1]   **Yi Zhou** [1]

Abstract

Although SGD with random reshuffle has been widely-used in machine learning applications, there is a limited understanding of how model characteristics affect the convergence of the algorithm. In this work, we introduce model incoherence to characterize the diversity of model characteristics and study its impact on convergence of SGD with random reshuffle under weak strong convexity. Specifically, *minimizer incoherence* measures the discrepancy between the global minimizers of a sample loss and those of the total loss and affects the convergence error of SGD with random reshuffle. In particular, we show that the variable sequence generated by SGD with random reshuffle converges to a certain global minimizer of the total loss under full minimizer coherence. The other *curvature incoherence* measures the quality of condition numbers of the sample losses and determines the convergence rate of SGD. With model incoherence, our results show that SGD has a faster convergence rate and smaller convergence error under random reshuffle than those under random sampling, and hence provide justifications to the superior practical performance of SGD with random reshuffle.

1. Introduction

We study the following finite-sum optimization problem that covers many important machine learning applications.

$$\min_{\theta \in \mathbb{R}^d} f(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta), \qquad \text{(P)}$$

where $\theta \in \mathbb{R}^d$ corresponds to the model parameters, $f : \mathbb{R}^d \to \mathbb{R}$ denotes the total loss and each ℓ_i corresponds to

---

[1]Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, United States of America. Correspondence to: Yi Zhou <yi.zhou@utah.edu>.

the sample loss of the i-th data sample. Such a problem formulation covers a variety of machine learning problems including support vector machine, logistic regression, matrix completion and neural network training, etc. The common approach is minimizing the error of a predictive model over all data samples in a dataset, an i.i.d. assumption on the data typically decomposes the error into a sum of sample errors.

A standard and widely-applied algorithm that solves the problem (P) is the stochastic gradient descent (SGD) algorithm, which has been well studied in both convex optimization (Bottou et al., 2018; Robbins & Monro, 1951; Nemirovski et al., 2009; Lan, 2012) and nonconvex optimization (Bottou et al., 2018; Ghadimi et al., 2016; Ghadimi & Lan, 2016). In these works, the SGD adopts a random sampling with replacement scheme (referred to as *random sampling*) and its analysis is based on a bounded stochastic variance assumption. Although such an SGD framework yields theoretically optimal convergence rate (Rakhlin et al., 2012), it cannot fully explain the superior practical performance of SGD in modern machine learning applications where the models are typically over-parameterized and SGD usually adopts the incremental sampling with random reshuffle scheme (referred to as *random reshuffle*). Hence, it is desired to develop novel SGD frameworks that provide better understanding of the superior practical performance of SGD with random reshuffle.

Toward this goal, some existing works have proposed various novel analysis frameworks that lead to improved convergence rates of SGD with random sampling. In specific, (Tseng, 1998; Solodov, 1998) introduced a strong growth condition that bounds the maximum sample loss gradient norm in terms of the total loss gradient norm. Under such a condition, (Schmidt & Roux, 2013) established a sublinear convergence rate and a linear convergence rate for SGD with random sampling in convex and strongly convex optimization, respectively. More recently, (Vaswani et al., 2018) proposed a more relaxed weak growth condition and established similar convergence rate results for SGD with random sampling. In another recent work, (Ma et al., 2017) considered an interpolation setting where the model overfits all the data points so that all the sample losses share a unique global minimizer, and they showed that SGD with random sampling achieves a linear convergence rate un-

der strong convexity. On the other hand, another line of works studied SGD with random reshuffle and established sublinear convergence rates under strong convexity, e.g., (HaoChen & Sra, 2018; Nagaraj et al., 2019; Shamir, 2016). However, the analysis in these works are based on traditional assumptions (e.g., Lipshcitzness, boundedness) that do not emphasize model characteristics, and a sufficiently small step size (typically $\mathcal{O}(n^{-1})$) is required to justify the advantage of random reshuffle over random sampling. In particular, these technical settings are not practical in modern machine learning training scenarios where the models are typically *over-parameterized* and a *constant-level* step size is adopted. Therefore, it is of great importance and interest to develop a novel theoretical framework for SGD with random reshuffle that characterizes the impact of model characteristics on its convergence under a practical constant step size. In specific, we are interested in studying SGD with random reshuffle in the following aspects.

- The convergence results of SGD with random reshuffle studied in the existing works are *in-expectation* with regard to the randomness of reshuffle under a sufficiently small step size. Can we prove stronger type of convergence of SGD with random reshuffle under over-parameterized models and a constant step size?

- It has been observed in many practical scenarios that SGD with random reshuffle converges faster than SGD with random sampling under a constant step size. Therefore, the framework that we develop for analyzing SGD with random reshuffle is expected to provide theoretical justifications for this phenomenon.

- The existing theoretical frameworks for analyzing SGD either assume the loss is strongly convex or assume the sample losses share a single global minimizer, both of which rule out many practical machine learning problems that are nonconvex and have multiple global minimizers. Therefore, our framework for analyzing SGD with random reshuffle must cover nonconvex scenarios and allow the existence of multiple global minimizers.

1.1. Our Contributions

We analyze the convergence of SGD with random reshuffle under a constant step size by exploiting two notions of model incoherence. In specific, we introduce a minimizer incoherence that measures the discrepancy between the global minimizer of a sample loss and that of the total loss. In particular, full minimizer coherence implies that all the sample losses share a global minimum and hence the model is over-parameterized. We also introduce a curvature incoherence that measures the quality of the condition numbers of the sample losses. Our theoretical results are in two-fold:

- We first consider the case of full minimizer coherence where all the sample losses share a set of global minimiz-

ers. In such a case, we show that the variable sequence generated by SGD with random reshuffle converges to a certain global minimizer under a constant step size, and therefore the algorithm converges deterministically. Then, under full minimizer coherence and restricted strong convexity, we show that SGD with random reshuffle achieves a linear convergence rate, in which the contraction parameter is determined by the curvature incoherence of the sample losses. Moreover, we establish a linear convergence rate for SGD with random sampling in our framework and prove that SGD achieves a faster linear convergence rate under random reshuffle than that under random sampling, which provides justification to the superior performance of SGD with random reshuffle in training over-parameterized models. We further verify these theoretical results via experiments on over-parameterized neural network training.

- Then, we analyze SGD with random reshuffle in the case of minimizer incoherence where the sample losses do not share any global minimizer. Under a constant step size and restricted strong convexity, we show that SGD with random reshuffle converges to a neighborhood of the global minimizer set at a linear convergence rate. In specific, the convergence rate depends on the curvature incoherence of the sample losses and the convergence error is determined by the minimizer incoherence of the sample losses. We show that the convergence rate of SGD with random reshuffle is faster than that of SGD with random sampling, and the convergence error of SGD is smaller under random reshuffle than that under random sampling. We verify our theoretical results via experiments on nonconvex phase retrieval.

Our analysis shows that the convergence rate and convergence error of SGD with random reshuffle are in the form of geometric mean, whereas those of SGD with random sampling are in the form of arithmetic mean. Therefore, random reshuffle leads to a better convergence statistics for SGD than random sampling.

1.2. Related Works

SGD with random sampling: Various theoretical frameworks have been developed for analyzing SGD with random sampling. In specific, (Schmidt & Roux, 2013) exploited the strong growth condition to show that SGD with random sampling achieves a sublinear convergence rate in the convex case and achieves a linear convergence rate in the strongly convex case. (Ma et al., 2017) introduced an interpolation setting, in which they showed that SGD with random sampling achieves a linear convergence rate in the strongly convex case. (Vaswani et al., 2018) proposed a relaxed weak growth condition and established a linear convergence rate for SGD with random sampling under strong convexity. (Bottou et al., 2018) studied SGD with random

sampling under a second moment condition. In (Gower et al., 2019), they introduced an expected smooth condition and established linear convergence of SGD with random sampling to a neighborhood of the global minimum.

SGD with random reshuffle: It has been noticed that incremental SGD can achieve a faster convergence rate compared to SGD with random sampling in (Bottou, 2009). The first theoretical analysis was given in (Gürbüzbalaban et al., 2015), where incremental SGD is shown to outperform SGD with random sampling under a diminishing stepsize. Random reshuffle has been shown to further improve the convergence rate of traditional SGD from $\mathcal{O}(\frac{1}{k})$ to $\mathcal{O}(\frac{1}{k^2})$ in the strongly convex case. Then, in more recent works (HaoChen & Sra, 2018), (Nagaraj et al., 2019), and (Ying et al., 2018), it was shown that SGD with random reshuffle outperforms SGD with random sampling after finite epochs under a sufficiently small constant step size and strong convexity.

2. Introduction to Model Incoherence

In this section, we introduce two notions of model incoherence. Recall the finite-sum optimization problem

$$\min_{x \in \mathbb{R}^d} f(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta). \tag{P}$$

We make the following standard assumption on the existence of solution set of the problem (P).

Assumption 1 (Existence of solution set). *Each sample loss $\ell_i, i = 1, ..., n$ has a solution set $\Theta_i^* \subset \mathbb{R}^d$, on which its global minimum $\ell_i^* > -\infty$ is attained. The total loss f has a solution set $\Theta^* \subset \mathbb{R}^d$, on which its global minimum $f^* > -\infty$ is attained.*

In general, the solution sets $\{\Theta_i^*\}_{i=1}^n$ of the sample losses can be different from the solution set Θ^* of the total loss.

We also make the following standard assumptions on the sample losses, where we denote $\text{proj}_A(x)$ as the Euclidean projection of x onto set A.

Assumption 2. *The problem (P) satisfies:*

1. *The sample losses are L-smooth, i.e., $\forall i$ and $x, y \in \mathbb{R}^d$,*

$$\ell_i(x) \leq \ell_i(y) + \langle x - y, \nabla \ell_i(y) \rangle + \frac{L}{2} \|x - y\|^2;$$

2. *Every sample loss ℓ_i is μ_i-weakly strong convex on $\Theta_i^* \cup \Theta^*$, i.e.,*

$$\ell_i(\omega) \geq \ell_i(x) + \langle \omega - x, \nabla \ell_i(x) \rangle + \frac{\mu_i}{2} \|x - \omega\|^2$$

holds for all $x \in \mathbb{R}^d, \omega \in \text{proj}_{\Theta_i^*}(x), \text{proj}_{\Theta^*}(x)$.

We note that the weak strong convexity is a weaker condition than the usual strong convexity and covers a wide range of non-convex problems including phase retrieval (Zhou et al., 2016; Zhang et al., 2017), neural networks (Zhong et al., 2017; Zhou & Liang, 2017), low-rank matrix factorization (Tu et al., 2016), blind deconvolution (Li et al., 2018), etc. Also, the weak strong convexity implies the restricted strong convexity under an additional convexity condition. We refer to (Karimi et al., 2016) for further discussions.

2.1. Minimizer Incoherence

In this subsection, we introduce minimizer incoherence to measure the discrepancy between the sample loss solution sets $\{\Theta_i^*\}_{i=1}^n$ and the total loss solution set Θ^*. In Appendix B, we provide a discussion that outlines the connections between the minimizer incoherence and other loss conditions that have been studied in the existing literature.

Definition 1 (Minimizer incoherence). *The minimizer incoherence $\epsilon_i > 0$ of every sample loss ℓ_i is defined as*

$$\epsilon_i := \sup_{\theta \in \Theta^*} \ell_i(\theta) - \ell_i^*, \quad i = 1, ..., n.$$

To elaborate, the minimizer incoherence corresponds to the gap between the highest sample loss that is achievable on the total loss solution set and the global minimum of the sample loss. Intuitively, it measures the incoherence between the sample loss solution set Θ_i^* and the total loss solution set Θ^*. In particular, when minimizer incoherence vanishes, the following inclusion properties of the solution sets hold.

Proposition 1 (Minimizer coherence). *The definition of minimizer incoherence implies that*

1. *If $\epsilon_i = 0$ for some i, then $\Theta^* \subset \Theta_i^*$;*
2. *If $\epsilon_i = 0$ for all i, then $\Theta^* = \cap_{i=1}^n \Theta_i^*$.*

In particular, the second item corresponds to the case where we have full minimizer coherence, i.e., all the sample losses share the set of global minimizers Θ^*. This is common in deep learning applications where the models are over-parameterized to overfit all the data samples (hence have full minimizer coherence) and have multiple global minimizers. Moreover, our minimizer incoherence generalizes the interpolation condition proposed in (Ma et al., 2017), which requires all the sample losses to share a unique global minimizer under strong convexity.

Figure 1 illustrates the cases of both minimizer incoherence and full minimizer coherence via quadratic sample losses. In fact, many nonconvex machine learning problems have been shown to have either vanishing or small minimizer incoherence, and we provide two illustrative examples below.

- Phase retrieval (Zhang et al., 2017): In this problem, we take linear measurements of an underlying complex signal
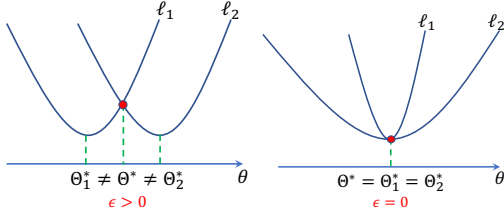
Figure 1. Left: Illustration of minimizer incoherence. Right: Illustration of full minimizer coherence.

$x_0 \in \mathbb{C}^d$ with multiple Gaussian vectors $\{a_i\}_{i=1}^n$ and make phaseless observations $y_i = |a_i^\mathsf{T} x_0|$. The goal is to recover the complex signal up to a global phase shift by solving the problem

$$\min_{x \in \mathbb{C}^d} f(x) := \frac{1}{2n} \sum_{i=1}^n \left(y_i - |a_i^\mathsf{T} x| \right)^2.$$

It is clear that all the sample losses share the set of minimizers $\{x_0 e^{j\phi} \mid \phi \in (0, 2\pi]\}$ and hence have full minimizer coherence.

- Over-parameterized neural networks: In deep learning, the neural network model θ is typically over-parameterized so that the predictor h_θ can be trained to overfit all the training samples, i.e., $h_\theta(x_i) \approx y_i$ for all $i = 1, ..., n$. Such overfitting usually achieves a small total loss as well as small sample losses. To justify this, we train a Resnet18 network to overfit the MNIST dataset with the cross-entropy loss. We do not apply any regularization. Figure 2 shows the distribution of the sample losses after 50 training epochs. One can see that most of the sample losses are below 3×10^{-3}, implying that deep models have very small minimizer incoherence.
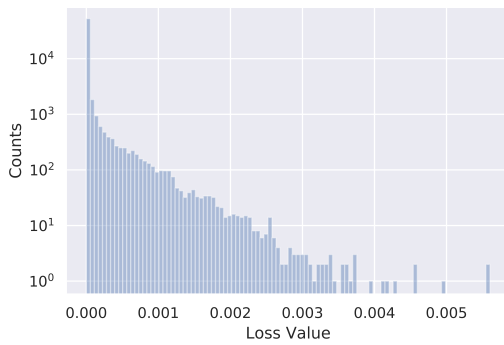

Figure 2. Distribution of sample losses after training ResNet18 for 50 epochs on MNIST dataset.

2.2. Curvature Incoherence

In this subsection, we introduce the curvature incoherence. Recall that the condition number of each sample loss ℓ_i is $\frac{L}{\mu_i}$. Then, we define the following curvature incoherence.

Definition 2 (Curvature incoherence). *The curvature incoherence α of the sample losses is defined as*

$$\alpha := \prod_{i=1}^n \left(1 - \frac{\mu_i}{L} \right).$$

In particular, α belongs to the range $[0, 1)$.

As an intuitive understanding, if all the sample losses have a good condition number (i.e., $\frac{L}{\mu_i} \to 1$), then α vanishes and the curvatures of all sample losses are highly coherent.

3. SGD with Random Reshuffle

In this section, we introduce the SGD with random reshuffle algorithm and provide some preliminary results on it.

The SGD algorithm starts with an initialization $\theta_0 \in \mathbb{R}^d$ and applies the following update rule iteratively.

$$\text{(SGD): } \theta_{k+1} = \theta_k - \eta \nabla \ell_{\xi(k)}(\theta_k), \ k = 0, 1, ..., \quad (1)$$

where $\eta > 0$ is the step size and $\xi(k)$ corresponds to the index of data sample drawn from $\{1, ..., n\}$ randomly in the k-th iteration. In this work, we focus on the widely-used incremental sampling with random reshuffle scheme, which is formally defined as follows and is referred to as *random reshuffle* for simplicity throughout the paper, .

(Random reshuffle): *In each epoch, we apply a random permutation to the sample indexes, i.e., $\{1, 2, ..., n\} \xrightarrow{permute} \{\xi(0), \xi(1), ..., \xi(n-1)\}$. Then, the sample $\xi(k)$ is used in the k-th iteration of this epoch.*

We obtain the following preliminary result for SGD with random reshuffle.

Lemma 1. *Let Assumptions 1 and 2 hold. Apply SGD with random reshuffle to solve the problem (P) with step size $\eta \leq \frac{1}{L}$. Then, the variable sequence $\{\theta_k\}_k$ generated by the algorithm satisfies: for all $k \in \mathbb{N}$ and all $\omega \in \Theta_{\xi(k)}^*$,*

$$\|\theta_{k+1} - \omega\|^2 \leq \|\theta_k - \omega\|^2 - \eta \left(\ell_{\xi(k)}(\theta_{k+1}) - \ell_{\xi(k)}^* \right).$$

We note that the proof of Lemma 1 only requires the sample losses to be restricted convex (i.e., μ_i can be zero in Assumption 2.3). The above lemma characterizes the per-iteration progress of SGD with random reshuffle towards any global minimizer of the sample loss used in the k-th iteration. In particular, it implies that $\|\theta_{k+1} - \omega\| \leq \|\theta_k - \omega\|$, i.e., SGD makes monotonic progress towards the minimizer of the sample loss used in the k-th iteration.

4. Analysis under Full Minimizer Coherence

In this section, we study the convergence properties of SGD with random reshuffle under full minimizer coherence, i.e., $\epsilon_i = 0$ for all $i = 1, ..., n$ (see Definition 1) and hence all sample losses share a set of global minimizers Θ^*.

4.1. Convergence of SGD Trajectory

We note that all the results in this subsection only require the sample losses to be restricted convex. We first characterize the boundedness of the optimization trajectory of SGD with random reshuffle.

Lemma 2 (Bounded trajectory). *Let Assumptions 1 and 2 hold and assume that the problem (P) has full minimizer coherence. Apply SGD with random reshuffle with step size $\eta \leq \frac{1}{L}$ to solve the problem. Then, the SGD trajectory $\{\theta_k\}_k$ is bounded.*

As the trajectory of SGD with random reshuffle $\{\theta_k\}_k$ is bounded, it has a compact set of limit points and we denote it as $\mathfrak{X}$. Also, note that the iteration index sequence $\{k\}_{k\in\mathbb{N}}$ can be decomposed into n subsequences $\{i(T)\}_T, i = 1, ..., n$, each of which tracks the SGD iterations that sample the i-th data point in the epochs $T = 1, 2, ...$ In particular, we denote $\mathfrak{X}_i$ as the set of limit points of $\{\theta_{i(T)}\}_T$ and it holds that $\mathfrak{X} = \bigcup_{i=1}^{n} \mathfrak{X}_i$. Moreover, we obtain the following properties regarding the limit point sets of the trajectory of SGD with random reshuffle.

Proposition 2 (Limit points). *Under the same conditions as those of Lemma 2, the trajectory of SGD with random reshuffle satisfies the following properties.*

1. *$\mathfrak{X}_i \subset \Theta_i^*$ for all $i = 1, ..., n$;*
2. *$\mathfrak{X}_i = \mathfrak{X} \subset \Theta^*$ for all $i = 1, ..., n$.*

To elaborate, item 1 shows that each sub-trajectory $\{\theta_{i(T)}\}_T$ generated by SGD with random reshuffle is a minimizing sequence for the corresponding sample loss ℓ_i. Item 2 further strengthens item 1 by showing that all the sub-trajectories $\{i(T)\}_T, i = 1, ..., n$ share the same set of limit points, which is a subset of the global minimizer set of the total loss. Intuitively, this is due to the fact that all the sample losses share a set of global minimizers under full minimizer coherence, which guarantees the sub-trajectories of SGD with random reshuffle to have consistent asymptotic properties.

The proof of Proposition 2 consists of two major steps. We first exploit full minimizer coherence to prove item 1 and the stationary condition $\|\theta_{k+1} - \theta_k\| \overset{k}{\to} 0$. Then, the stationary condition further guarantees that all sub-trajectories share the same set of limit points and hence implies item 2.

Our main result below further strengthens the convergence properties of the SGD trajectory.

Theorem 1 (Trajectory convergence). *Under the same conditions as those of Lemma 2, every trajectory $\{\theta_k\}_k$ generated by SGD with random reshuffle converges to a certain global minimizer in Θ^*, i.e., it has a single limit point.*

The above result shows that the entire trajectory of SGD with random reshuffle converges to a certain global minimizer in the case of full minimizer coherence. This implies that full minimizer coherence helps suppress the randomness of the random reshuffle and leads to the point-wise convergence. Such a deterministic convergence result of SGD with random reshuffle is stronger than other *in-expectation* convergence results of SGD with random sampling that are established under various loss conditions (e.g., strong growth condition, interpolation) that imply full minimizer coherence.

4.2. Convergence Rate Analysis

In this subsection, we further study the convergence rate of SGD with random reshuffle under full minimizer coherence. For any point $\theta \in \mathbb{R}^d$, we denote its distance to an arbitrary set $A \subset \mathbb{R}^d$ as $\mathrm{dist}_A(\theta) := \inf_{u\in A} \|\theta - u\|$.

We obtain the following convergence rate result.

Theorem 2 (Random reshuffle). *Let Assumptions 1 and 2 hold and assume that the problem (P) has full minimizer coherence. Apply SGD with random reshuffle with step size $\eta = \frac{1}{L}$ to solve the problem. Then, for all epochs $B = 1, 2, ...,$ it holds that*

$$\mathrm{dist}_{\Theta^*}^2(\theta_{nB}) \leq \prod_{i=1}^{n} \left(1 - \frac{\mu_i}{L}\right)^B \mathrm{dist}_{\Theta^*}^2(\theta_0). \quad (2)$$

The above theorem establishes the linear convergence rate of SGD with random reshuffle under full minimizer coherence and the constant step size $\eta = \frac{1}{L}$. In particular, the convergence rate depends on the curvature incoherence parameter $\alpha = \prod_{i=1}^{n}(1 - \frac{\mu_i}{L})$ (see Definition 2) that characterizes the quality of the condition numbers of all the sample losses. We also note that in the special case that all the sample losses have the same condition number $\frac{\mu}{L}$, the above convergence rate of SGD with random reshuffle is of order $\mathcal{O}(1 - \frac{\mu}{L})^{nB}$, which meets the convergence rate of full gradient descent under strong convexity.

4.3. Comparison to Other Sampling Schemes

We further analyze the convergence rates of SGD with incremental sampling (i.e., cyclic sampling without random reshuffle) and random sampling under full minimizer coherence and compare them with that of SGD with random reshuffle.

In fact, under full minimizer coherence, our proof of Theorem 2 only rely on the fact that the random reshuffle scheme samples every data point once in each epoch, which is also satisfied by the incremental sampling scheme. Therefore, the convergence rate result in Theorem 2 also applies to SGD with incremental sampling and we obtain the following corollary.

Corollary 1 (Incremental sampling). *Under the same settings as those of Theorem 2 and apply SGD with incremental*

sampling and step size $\eta = \frac{1}{L}$ to solve the problem (P). Then, the convergence rate is also characterized by eq. (2).

On the other hand, we obtain the following result for SGD with random sampling.

Proposition 3 (Random sampling). *Under the same settings as those of Theorem 2 and apply SGD with random sampling and step size $\eta = \frac{1}{L}$ to solve the problem (P). Then, for all epochs $B = 1, 2, ...,$ it holds that*

$$\text{dist}_{\Theta^*}^2(\theta_{nB}) \leq \left(1 - \frac{\bar{\mu}}{L}\right)^{nB} \text{dist}_{\Theta^*}^2(\theta_0), \qquad (3)$$

where $\bar{\mu} := \frac{1}{n} \sum_{i=1}^n \mu_i$.

The above result establishes a linear convergence rate for SGD with random sampling. Note that the convergence rate depends on the average of the condition numbers of the sample losses. This is different from the convergence rate of SGD with random reshuffle, which depends on the product of the condition numbers of all the sample losses. In particular, by the arithmetic mean-geometric mean (AM-GM) inequality, it holds that

$$\prod_{i=1}^n \left(1 - \frac{\mu_i}{L}\right) \leq \left(1 - \frac{\bar{\mu}}{L}\right)^n. \qquad (4)$$

Therefore, under full minimizer coherence, SGD achieves a faster convergence rate under random reshuffle than that under random sampling. Such a result provides a theoretical justification for the superior performance of SGD with random reshuffle in training over-parameterized models.

We note that (HaoChen & Sra, 2018) also obtains a similar comparison of convergence rate between SGD with random sampling and SGD with random reshuffle. However, their analysis requires the loss to be uniformly strongly convex, whereas our result applies to the broader class of restricted strongly convex functions. Moreover, we established trajectory convergence of SGD with random reshuffle under the existence of multiple global minimizers, whereas their result establishes convergence in expectation under the existence of a unique global minimizer.

4.4. Empirical Verification

In this subsection, we verify our theoretical results via experiments. We first study the impact of curvature incoherence α on the convergence of SGD with random reshuffle. In specific, we train a Resnet 18 network using SGD with random reshuffle on a mini MNIST dataset that consists of 1000 images of digit "1" and 1000 images of digit "8". To model different distributions of curvature incoherence, we divide the data samples evenly into 50 fixed mini-batches and consider two different settings: 1) each mini-batch contains 50% images of digit "1" and 50% images of digit "8"; and

2) each mini-batch contains either images of digit "1" or images of digit "8". In both settings, the average condition numbers of the sample losses are different. Figure 3 (Left) shows the training loss curves of SGD with random reshuffle in these two settings starting from the same initialization point. It can be observed that SGD with random reshuffle converges faster in the second setting. This implies that the average condition number of the sample losses in the first setting is better than that in the second setting.

Next, we further compare the empirical convergence of SGD under random reshuffle with that under incremental sampling and random sampling. We train a Resnet18 on 4096 images sampled from CIFAR10 using SGD with the three sampling schemes. We use learn rate $\eta = 0.03$, batch-size 128 and a fixed initialization model that is trained by SGD with random reshuffle for one epoch (with learning rate 0.0025) using a pre-trained ImageNet model. Figure 3 (Right) shows the training loss curves of the three algorithms. It can be seen that SGD with random reshuffle and incremental sampling have a comparable convergence speed, both of which are faster than that of SGD with random sampling. This observation fully supports our theoretical comparison in eq. (4).
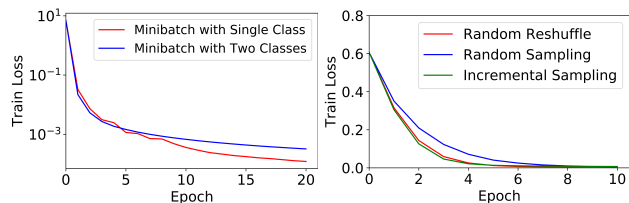

Figure 3. Left: Comparison of convergence of SGD with random reshuffle under different curvature incoherence distributions. Right: Comparison of convergence of SGD under random reshuffle, incremental sampling and random sampling.

5. Analysis under Minimizer Incoherence

In this section, we study the convergence of SGD with random reshuffle under minimizer incoherence where $\epsilon_i > 0$ for some $i \in \{1, ..., n\}$. In such a case, the minimizer sets of the sample losses are different from that of the total loss. For simplicity, we assume the minimizer incoherences of all the sample losses are bounded by $\epsilon := \max_{i \in \{1,...,n\}} \epsilon_i$.

5.1. Convergence Rate Analysis

We first show the boundedness of the trajectory of SGD with random reshuffle under minimizer incoherence and a constant step size.

Lemma 3 (Bounded trajectory). *Let Assumptions 1 and 2 hold and assume that the problem (P) has bounded minimizer incoherence ϵ. Apply SGD with random reshuffle*

with step size $\eta = \frac{1}{L}$ to solve the problem. Then, the SGD trajectory $\{\theta_k\}_k$ is bounded.

The above result generalizes the bounded trajectory result in Lemma 2, which is proved under full minimizer coherence. Next, we obtain the following result regarding the convergence rate of SGD with random reshuffle under minimizer incoherence and a constant step size.

Theorem 3 (Random reshuffle). *Let Assumptions 1 and 2 hold. Apply SGD with random reshuffle with step size $\eta = \frac{1}{L}$ to solve the problem (P). Then, for all epochs $B = 1, 2, ...,$*

$$\mathbb{E}\mathrm{dist}_{\Theta^*}^2(\theta_{nB}) \leq \alpha^B \left(\mathrm{dist}_{\Theta^*}^2(\theta_0) - \frac{2\epsilon\overline{M}}{L(1-\alpha)} \right) + \frac{2\epsilon\overline{M}}{L(1-\alpha)},$$

where $\overline{M} = \mathbb{E}_\xi \left[\sum_{k=0}^{n-1} \prod_{s=k+1}^{n-1} \left(1 - \frac{\mu_{\xi(s)}}{L} \right) \right]$ and α corresponds to the curvature incoherence.

The above result shows that SGD with random reshuffle converges linearly to a neighborhood of the global minimizer set under minimizer incoherence. Similar to the full minimizer coherence case, the convergence rate coefficient is determined by the curvature incoherence α. Moreover, the size of the neighborhood is characterized by the minimizer incoherence ϵ and the condition numbers of the sample losses. This explains why over-parametrized models such as neural networks can be trained to achieve a small loss by SGD with constant step size: they have very small minimizer incoherence, as demonstrated by the experiment in Figure 2. In general, a higher minimizer incoherence and worse condition numbers lead to a larger convergence error of SGD.

5.2. Comparison to Other Sampling Schemes

We also obtain the convergence rates of SGD with incremental sampling and random sampling under minimizer incoherence and a constant step size.

In specific, for SGD with incremental sampling, we denote $\{\sigma(0), \sigma(1), ..., \sigma(n-1)\}$ as a specific permutation of the data sample indexes. Such a permutation is fixed throughout the entire training process under incremental sampling. We obtain the following result on SGD with incremental sampling.

Corollary 2 (Incremental sampling). *Let Assumptions 1 and 2 hold. Apply SGD with incremental sampling with step size $\eta = \frac{1}{L}$ to solve the problem (P). Then, for all epochs $B = 1, 2, ...,$*

$$\mathrm{dist}_{\Theta^*}^2(\theta_{nB}) \leq \alpha^B \left(\mathrm{dist}_{\Theta^*}^2(\theta_0) - \frac{2\epsilon\widetilde{M}}{L(1-\alpha)} \right) + \frac{2\epsilon\widetilde{M}}{L(1-\alpha)},$$

where $\widetilde{M} = \sum_{k=0}^{n-1} \prod_{s=k+1}^{n-1} \left(1 - \frac{\mu_{\sigma(s)}}{L} \right)$ and α corresponds to the curvature incoherence.

To elaborate, the permutation map σ used by the incremental sampling can be viewed as a particular realization of the random permutation of the random reshuffle scheme. In particular, the convergence error term $\overline{M}$ in Theorem 3 corresponds to the average of the convergence errors over all possible random permutations of the data indexes, whereas the convergence error term $\widetilde{M}$ in Corollary 2 is determined by the specific permutation map σ used. Therefore, depending on the quality of the permutation map, the convergence error of SGD under incremental sampling can be either larger or smaller than that of SGD under random reshuffle.

For SGD with random sampling, we obtain the following convergence rate under minimizer incoherence and a constant step size.

Proposition 4 (Random sampling). *Let Assumptions 1 and 2 hold. Apply SGD with random sampling with learning rate $\eta = \frac{1}{L}$ to solve the problem (P). Then, for all $k = nB, n = 1, 2, ...,$*

$$\mathbb{E}\mathrm{dist}_{\Theta^*}^2(\theta_{nB}) \leq \left(1 - \frac{\overline{\mu}}{L} \right)^{nB} \left(\mathrm{dist}_{\Theta^*}^2(\theta_0) - \frac{2\epsilon}{\overline{\mu}} \right) + \frac{2\epsilon}{\overline{\mu}}.$$

where $\bar{\mu} := \frac{1}{n} \sum_{i=1}^n \mu_i$.

Comparing the above result with that in Theorem 3, one can see that under minimizer incoherence, SGD with random reshuffle has a better convergence rate coefficient $\alpha = \prod_{i=1}^n (1 - \frac{\mu_i}{L})$ than that $(1 - \frac{\overline{\mu}}{L})^n$ of SGD with random sampling (due to the AM-GM inequality). Moreover, regarding the convergence error, one can show that the convergence error $\frac{2\epsilon\overline{M}}{L(1-\alpha)}$ of SGD with random reshuffle is smaller than that $\frac{2\epsilon}{\overline{\mu}}$ of SGD with random sampling, and we outline the proof below.

$$\frac{2\epsilon\overline{M}}{L(1-\alpha)} = \frac{2\epsilon \sum_{k=0}^{n-1} \mathbb{E}_\xi \left[\prod_{s=k+1}^{n-1} \left(1 - \frac{\mu_{\xi(s)}}{L} \right) \right]}{L(1-\alpha)}$$

$$\overset{(i)}{\leq} \frac{2\epsilon \sum_{k=0}^{n-1} \left(1 - \frac{\overline{\mu}}{L} \right)^{n-k-1}}{L(1-\alpha)}$$

$$= \frac{2\epsilon L}{\overline{\mu}} \frac{1 - \left(1 - \frac{\overline{\mu}}{L} \right)^n}{L(1-\alpha)}$$

$$\overset{(ii)}{\leq} \frac{2\epsilon}{\overline{\mu}}.$$

To elaborate, consider the quantity $\mathbb{E}_\xi \left[\prod_{s=k+1}^{n-1} \left(1 - \frac{\mu_{\xi(s)}}{L} \right) \right]$ in $\overline{M}$. Note that for each fixed k, the samples $\{\xi(s)\}_{s=k+1}^{n-1}$ are drawn from $\{1, ..., n\}$ uniformly at random *without replacement* due to the random reshuffle scheme, and hence the expectation over $\{\xi(s)\}_{s=k+1}^{n-1}$ consists of $\binom{n}{n-k-1}$ number of different combinations. Therefore, the inequality (i) follows from the Maclaurin's inequality. Moreover, the inequality (ii) follows from the AM-GM inequality. Such a comparison result reveals the statistical advantage of random

reshuffle over random sampling: random reshuffle visits all data permutations in expectation and leads to a convergence error in spirit of geometric series (i.e., the $\sum \prod$ term in $\overline{M}$), whereas random sampling samples each data uniformly with replacement and leads to a convergence error in spirit of arithmetic mean (i.e., the $\overline{\mu}$ term).

5.3. Empirical Verification

We verify our theoretical results obtained in this section via experiments on nonconvex phase retrieval. In specific, consider an underlying complex signal $x_0 \in \mathbb{C}^d$ with a set of Gaussian measurement vectors $\{a_i\}_{i=1}^m$. The nonconvex phase retrieval model is written as $y_i = |\langle a_i, x_0 \rangle| + n$, where $\{y_i\}_{i=1}^m$ are the phaseless observations and n denotes a Gaussian random noise. To retrieve the signal based on the phaseless observations and the Gaussian measurement vectors, we aim to solve the following nonconvex problem.

$$\min_{x \in \mathbb{C}^d} \frac{1}{2n} \sum_{i=1}^n \left(y_i - |a_i^\intercal x| \right)^2.$$

Due to noise corruption, the sample losses do not share a minimizer and hence have minimizer incoherence. In particular, the minimizer incoherence increases as the noise level increases. Specifically, we generate x_0, a_i, and y_i from normal distribution with $d = 128, n = 512$. We repeat each experiment for 300 times and use learning rate $\eta = 0.1$.
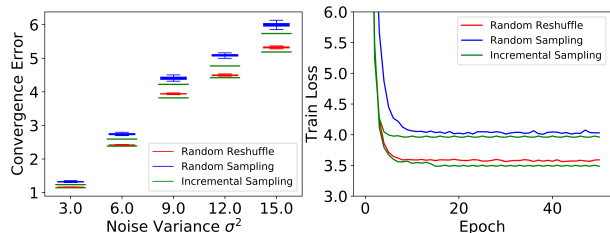

Figure 4. Left: Impact of minimizer incoherence on convergence error of SGD. Right: Convergence curves of SGD with different sampling schemes under minimizer incoherence.

We first explore how the level of noise (i.e., level of minimizer incoherence) in phase retrieval affects the convergence error of SGD with different sampling schemes. Figure 4 (Left) presents the box plot of convergence errors of SGD with random sampling and random reshuffle under different levels of Gaussian noise corruptions. For SGD with incremental sampling, we plot the smallest and largest errors achieved in the repeated experiments. It can be seen that as the noise increases (i.e., minimizer incoherence increases), the convergence errors of these SGDs increase accordingly, which matches our theoretical characterizations of the convergence error. In particular, it can be observed that SGD with random reshuffle consistently has smaller convergence error than SGD with random sampling. Moreover, SGD

with incremental sampling can sometimes outperforms SGD with random reshuffle when the permutation map happen to be good. Figure 4 (Right) shows the training loss curves of these algorithms under noise variance $\sigma^2 = 9$. Under minimizer incoherence, it can be seen that SGD with random reshuffle and SGD with incremental sampling have a comparable convergence speed (i.e., a comparable slope of the training curves), and both of them converge faster than SGD with random sampling. These empirical results validate our convergence rate results of SGD under minimizer incoherence obtained in this section.

6. Conclusion

In this paper, we propose a model incoherence framework to study the impact of model incoherence on convergence of SGD. When the model has full minimizer coherence, we prove that SGD with random reshuffle converges to a global minimum deterministically and achieves a faster convergence rate than that of SGD with random sampling. When the sample losses have incoherent minimizers, we further show that SGD with random reshuffle has a smaller convergence error than that of SGD with random sampling. Our results reveal the statistical difference between the two random sampling schemes and characterize the impact of model incoherence on the optimization convergence. In the future work, we will further explore the generalization ability of SGD under different sampling schemes and develop a proper analysis framework for it.

Acknowledgement

References

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.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, Mar 2016.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1): 267–305, Jan 2016.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5200–5209, Long Beach, California, USA, 09–15 Jun 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.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, Jun 2012.

Li, X., Ling, S., Strohmer, T., and Wei, K. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 2018.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *arXiv preprint arXiv:1712.06559*, 2017.

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.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proc. International Coference on Machine Learning (ICML)*, pp. 1571–1578, 2012.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, Sep 1951.

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

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

Solodov, M. V. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

Tseng, P. An incremental gradient (-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proc. 33rd International Conference on Machine Learning (ICML)*, pp. 964–973, 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 with constant stepsizes. *IEEE Transactions on Signal Processing*, 67(2): 474–489, 2018.

Zhang, H., Zhou, Y., Liang, Y., and Chi, Y. A nonconvex approach for phase retrieval: reshaped Wirtinger flow and

incremental algorithms. *Journal of Machine Learning Research (JMLR)*, 18(141):1–35, 2017.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. In *Proc. 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 4140–4149, Aug 2017.

Zhou, Y. and Liang, Y. Characterization of gradient dominance and regularity conditions for neural networks. *ArXiv:1710.06910v2*, Oct 2017.

Zhou, Y., Zhang, H., and Liang, Y. Geometrical properties and accelerated gradient solvers of non-convex phase retrieval. *In Proc. 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 331–335, 2016.