On the Convergence to a Global Solution of Shuffling-Type Gradient Algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic gradient descent (SGD) algorithm is the method of choice in many machine learning tasks thanks to its scalability and efficiency in dealing with large-scale problems. In this paper, we focus on the shuffling version of SGD which matches the mainstream practical heuristics. We show the convergence to a global solution of shuffling SGD for a class of non-convex functions under over-parameterized settings. Our analysis employs more relaxed non-convex assumptions than previous literature. Nevertheless, we maintain the desired computational complexity as shuffling SGD has achieved in the general convex setting.

1 Introduction

In the last decade, neural network-based models have shown great success in many machine learning applications such as natural language processing [Collobert and Weston, 2008, Goldberg et al., 2018], computer vision and pattern recognition [Goodfellow et al., 2014, He and Sun, 2015]. The training task of many learning models boils down to the following finite-sum minimization problem:

$$\min_{w\in\mathbb{R}^d}\left\{F(w) := \frac{1}{n}\sum_{i=1}^{n} f(w;i)\right\}, \tag{1}$$

where $f(\cdot;i):\mathbb{R}^d\to\mathbb{R}$ is smooth and possibly non-convex for $i\in[n]:=\{1,\cdots,n\}$. Solving the empirical risk minimization (1) had been a difficult task for a long time due to the non-convexity and the complicated learning models. Later progress with stochastic gradient descent (SGD) and its variants [Robbins and Monro, 1951, Duchi et al., 2011, Kingma and Ba, 2014] have shown great performance in training deep neural networks. These stochastic first-order methods are favorable thanks to its scalability and efficiency in dealing with large-scale problems. At each iteration SGD samples an index i uniformly from the set $\{1,\ldots,n\}$, and uses the individual gradient $\nabla f(\cdot;i)$ to update the weight.

While there has been much attention on the theoretical aspect of the traditional i.i.d. (independently identically distributed) version of SGD [Nemirovski et al., 2009, Ghadimi and Lan, 2013, Bottou et al., 2018], practical heuristics often use without-replacement data sampling schemes. Also known as shuffling sampling schemes, these methods generate some random or deterministic permutations of the index set $\{1,2,\ldots,n\}$ and apply gradient updates using these permutation orders. Intuitively, a collection of such n individual updates is a pass over all the data, or an epoch. One may choose to create a new random permutation at the beginning of each epoch (in Random Reshuffling scheme) or use a random permutation for every epoch (in Single Shuffling scheme). Alternatively, one may use a Incremental Gradient scheme with a fixed deterministic order of indices. In this paper, we use the term unified shuffling SGD for SGD method using *any* data permutations, which includes the three special schemes described above.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Although shuffling sampling schemes usually show a better empirical performance than SGD [Bottou, 2009], the theoretical guarantees for these schemes are often more limited than vanilla SGD version, due to the lack of statistical independence. Recent works have shown improvement in computational complexity for shuffling schemes over SGD in various settings [Gürbüzbalaban et al., 2019, Haochen and Sra, 2019, Safran and Shamir, 2020, Nagaraj et al., 2019, Nguyen et al., 2021, Mishchenko et al., 2020, Ahn et al., 2020]. In particular, in a general non-convex setting, shuffling sampling schemes improve the computational complexity in terms of $\hat{\varepsilon}$ for SGD from $\mathcal{O}\left(\sigma^2/\hat{\varepsilon}^2\right)$ to $\mathcal{O}\left(n\sigma/\hat{\varepsilon}^{3/2}\right)$, where σ is the bounded variance constant [Ghadimi and Lan, 2013, Nguyen et al., 2021, Mishchenko et al., 2020][1]. We summarize the detailed literature for multiple settings later in Table 1.

While global convergence is a desirable property for neural network training, the non-convexity landscape of complex learning models leads to difficulties in finding the global minimizer. In addition, there is little to no work studying the convergence to a global solution of shuffling-type SGD algorithms for a general non-convex setting. The closest line of research investigates the Polyak-Lojasiewicz (PL) condition (a generalization of strong-convexity), which demonstrates similar convergence rates as the strongly convex rates for shuffling SGD methods [Haochen and Sra, 2019, Ahn et al., 2020, Nguyen et al., 2021]. In another direction, Gower et al. [2021] and Khaled and Richtárik [2020] investigates the global convergence for some class of non-convex functions, however for vanilla SGD method. Beznosikov and Takáč [2021] investigate a random shuffle version of variance reduction methods (e.g. SARAH algorithm Nguyen et al. [2017]), but this approach only can show convergence to stationary points. With a target on shuffling SGD methods and specific learning architectures, we come up with the central question of this paper:

How can we establish the convergence to global solution for a class of non-convex functions using shuffling-type SGD algorithms? Can we exploit the structure of neural networks to achieve this goal?

We answer this question affirmatively, and our contributions are summarized below. **Contributions**.

- We investigate a new framework for the convergence of a shuffling-type gradient algorithm to a global solution. We consider a relaxed set of assumptions and discuss their relations with previous settings. We show that our average-PL inequality (Assumption 3) holds for a wide range of neural networks equipped with squared loss function.
- Our analysis generalizes the class function called star-M-smooth-convex. This class contains non-convex functions and is more general than the class of star-convex smooth functions with respect to the minimizer (in the over-parameterized settings). In addition, our analysis does not use any bounded gradient or bounded weight assumptions.
- We show the total complexity of $\mathcal{O}(\frac{n}{\hat{\varepsilon}^{3/2}})$ for a class of non-convex functions to reach an $\hat{\varepsilon}$-accurate global solution. This result matches the same gradient complexity to a stationary point for unified shuffling methods in non-convex settings, however, we are able to show the convergence to a global minimizer.

1.1 Related Work

In recent years, there have been different approaches to investigate the global convergence for machine learning optimization. This includes a popular line of research that studies some specific neural networks and utilizes their architectures. The most early works show the global convergence of Gradient Descent (GD) for simple linear networks and two-layer networks [Brutzkus et al., 2018, Soudry et al., 2018, Arora et al., 2019, Du et al., 2019b]. These results are further extended to deep learning architectures [Allen-Zhu et al., 2019, Du et al., 2019a, Zou and Gu, 2019]. This line of research continues with Stochastic Gradient Descent (SGD) algorithm, which proves the global convergence of SGD for deep neural networks for some probability depending on the initialization process and the number of input data [Brutzkus et al., 2018, Allen-Zhu et al., 2019, Zou et al., 2018, Zou and Gu, 2019]. The common theme that appeared in most of these references is the over-parameterized setting, which means that the number of parameters in the network are excessively large [Brutzkus et al., 2018, Soudry et al., 2018, Allen-Zhu et al., 2019, Du et al., 2019a, Zou and Gu, 2019]. This fact is closely related to our setting, and we will discuss it throughout our paper.

[1]The computational complexity is the number of (individual) gradient computations needed to reach an $\hat{\varepsilon}$-accurate stationary point (i.e. a point $\hat{w} \in \mathbb{R}^d$ that satisfies $\|\nabla F(\hat{w})\|^2 \leq \hat{\varepsilon}$.)

Polyak-Lojasiewicz (PL) condition and related assumptions. An alternative approach is to investigate some conditions on the optimization problem that may guarantee global convergence. A popular assumption is the Polyak-Lojasiewicz (PL) inequality, a generalization of strong-convexity [Polyak, 1964, Karimi et al., 2016, Nesterov and Polyak, 2006]. Using this PL assumption, it can be shown that (stochastic) gradient descent achieves the same theoretical rate as in the strongly convex setting (i.e linear convergence for GD and sublinear convergence for SGD) [Karimi et al., 2016, De et al., 2017, Gower et al., 2021]. Recent works demonstrate similar results for shuffling type SGD [Haochen and Sra, 2019, Ahn et al., 2020, Nguyen et al., 2021], both for unified and randomized shuffling schemes. On the other hand, [Schmidt and Roux, 2013, Vaswani et al., 2019] propose to use a new assumption called the Strong Growth Condition (SGC) that controls the rates at which the stochastic gradients decay comparing to the full gradient. This condition implies that the stochastic gradients and their variances converge to zero at the optimum solution [Schmidt and Roux, 2013, Vaswani et al., 2019]. While the PL condition for F implies that every stationary point of F is also a global solution, the SGC implies that such a point is also a stationary point of every individual function. However, complicated models as deep feed-forward neural networks generally have non-optimal stationary points [Karimi et al., 2016]. Thus, these assumptions are somewhat strong for non-convex settings.

Although there are plenty of works investigating the PL condition for the objective function F [De et al., 2017, Vaswani et al., 2019, Gower et al., 2021], not many materials devoted to study the PL inequality for the individual functions $f(\cdot; i)$. A recent work [Sankararaman et al., 2020] analyzes SGD with the specific notion of gradient confusion for over-parameterized settings where the individual functions satisfy PL condition. They show that the neighborhood where SGD converges linearly depends on the level of gradient confusion (i.e. how much the individual gradients are negatively correlated). Taking a different approach, we investigate the PL property for individual functions and further show that our condition holds for a general class of neural networks with quadratic loss.

Over-paramaterized settings for neural networks. Most of the modern learning architectures contain deep and large networks, where the number of parameters are often far more than the number of input data. This leads to the fact that the objective loss function is trained closer and closer to zero. Understandably, in such settings all the individual functions $f(\cdot; i)$ are minimized simultaneously at 0 and they share a common minimizer. This condition is called the interpolation property (see e.g. [Schmidt and Roux, 2013, Ma et al., 2018, Meng et al., 2020, Loizou et al., 2021]) and is studied well in the literature (see e.g. [Zhou et al., 2019, Gower et al., 2021]). For a comparison, functions satisfying the strong growth condition necessarily satisfy the interpolation property. This property implies zero variance of individual gradients at the global minimizer, which allows good behavior for SGD near the solution. In this work, we slightly change this assumption which requires a small variance up to some level of the threshold ε. Note that when letting $\varepsilon \to 0$, our assumption exactly recovers the interpolation property.

Star-convexity and related conditions. There have been many attentions to a class of structured non-convex functions called star-convex [Nesterov and Polyak, 2006, Lee and Valiant, 2016, Bjorck et al., 2021]. Star-convexity can be understood as convexity between an arbitrary point w and the global minimizer w_*. The name star-convex comes from the fact that each sublevel set is star-shaped [Nesterov and Polyak, 2006, Lee and Valiant, 2016]. Zhou et al. [2019] shows that if SGD follows a star-convex path and there exists a common global minimizer for all component functions, then SGD converges to a global minimum.

In recent progress, Hinder et al. [2020] considers the class of quasar-convex functions, which further generalizes star-convexity. This property was introduced originally in [Hardt et al., 2018] under the name 'weakly quasi-convex', and investigated recently in literature [Hinder et al., 2020, Jin, 2020, Gower et al., 2021]. This class uses a parameter $\zeta \in (0, 1]$ to control the non-convexity of the function, where $\zeta = 1$ yeilds the star-convexity and ζ approaches 0 indicates more non-convexity [Hinder et al., 2020]. Intuitively, quasar-convex functions are unimodal on all lines that pass through a global minimizer. Gower et al. [2021] investigates the performance of SGD for smooth and quasar-convex functions using an expected residual assumption (which is comparable to the interpolation property). They show a convergence rate of $\mathcal{O}(1/\sqrt{K})$ for i.i.d. sampling SGD with the number of total iterations K, which translates to the computational complexity of $\mathcal{O}\left(1/\hat{\varepsilon}^2\right)$. To the best of our knowledge, this paper is the first work studying the relaxation of star-convexity and global convergence for SGD with shuffling sample schemes, not for the i.i.d. version.

2 Theoretical Setting

We first present the shuffling-type gradient algorithm below. Our convergence results hold for any permutation of the training data $\{1, 2, \ldots, n\}$, including deterministic and random ones. Thus, our theoretical framework is general and applicable for any shuffling strategy, including Incremental Gradient, Single Shuffling, and Random Reshuffling.

Algorithm 1 (Shuffling-Type Gradient Algorithm for Solving (1))

Initialization: Choose an initial point $\tilde{w}_0 \in \mathrm{dom}\,(F)$.
for $t = 1, 2, \ldots, T$ **do**
  Set $w_0^{(t)} := \tilde{w}_{t-1}$;
  Generate any permutation $\pi^{(t)}$ of $[n]$ (either deterministic or random);
  for $i = 1, \ldots, n$ **do**
    Update $w_i^{(t)} := w_{i-1}^{(t)} - \eta_i^{(t)} \nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))$;
  end for
  Set $\tilde{w}_t := w_n^{(t)}$;
end for

We further specify the choice of learning rate $\eta_i^{(t)}$ in the detailed analysis. Now we proceed to describe the set of assumptions used in our paper.

Assumption 1. *Suppose that* $f_i^* := \min_{w \in \mathbb{R}^d} f(w; i) > -\infty$, $i \in \{1, \ldots, n\}$.

Assumption 2. *Suppose that* $f(\cdot; i)$ *is* L*-smooth for all* $i \in \{1, \ldots, n\}$*, i.e. there exists a constant* $L \in (0, +\infty)$ *such that:*

$$\|\nabla f(w; i) - \nabla f(w'; i)\| \leq L\|w - w'\|, \quad \forall w, w' \in \mathbb{R}^d. \tag{2}$$

Assumption 1 is required in any algorithm to guarantee the well-definedness of (1). In most applications, the component losses are bounded from below. By Assumption 2, the objective function F is also L-smooth. This Lipschitz smoothness Assumption is widely used for gradient-type methods. In addition, we denote the minimum value of the objective function $F_* = \min_{w \in \mathbb{R}^d} F(w)$. It is worthy to note the following relationship between F_* and the component minimum values:

$$F_* = \min_{w \in \mathbb{R}^d} F(w) = \frac{1}{n} \min_{w \in \mathbb{R}^d} \left(\sum_{i=1}^n f(w; i) \right) \geq \frac{1}{n} \sum_{i=1}^n \min_{w \in \mathbb{R}^d} (f(w; i)) = \frac{1}{n} \sum_{i=1}^n f_i^*. \tag{3}$$

We are interested in the case where the set of minimizers of F is not empty. The equality $F_* = \frac{1}{n} \sum_{i=1}^n f_i^*$ attains if and only if a minimizer of F is also the common minimizer for all component functions. This condition implies that the variance of individual functions is 0 at the common minimizer.

2.1 PL Condition for Component Functions

Now we are ready to discuss the Polyak-Lojasiewicz condition as follows.

Definition 1 (Polyak-Lojasiewicz condition)**.** *We say that* f *satisfies Polyak-Lojasiewicz (PL) inequality for some constant* $\mu > 0$ *if*

$$\|\nabla f(w)\|^2 \geq 2\mu[f(w) - f^*], \quad \forall w \in \mathbb{R}^d, \tag{4}$$

where $f^* := \min_{w \in \mathbb{R}^d} f(w)$.

The PL condition for the objective function F is sufficient to show a global convergence for (stochastic) gradient descent [Karimi et al., 2016, Nesterov and Polyak, 2006, Polyak, 1964]. It is well known that a function satisfying the PL condition is not necessarily convex [Karimi et al., 2016]. However, this assumption on F is somewhat strong because it implies that every stationary point of F is also a global minimizer. Our goal is to consider a class of non-convex function which is more relaxed than the PL condition on F, while still having the good global convergence properties. In this paper, we formulated an assumption called "average PL inequality", specifically for the finite sum setting:

Assumption 3. *Suppose that $f(\cdot; i)$ satisfies average PL inequality for some constant $\mu > 0$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w;i)\|^2 \geq 2\mu\frac{1}{n}\sum_{i=1}^{n}[f(w;i) - f_i^*], \quad \forall w \in \mathbb{R}^d. \tag{5}$$

where $f_i^* := \min_{w\in\mathbb{R}^d} f(w; i)$.

Comparisons. Assumption 3 is weaker than assuming the PL inequality for every component function $f(\cdot; i)$. In general setting, Assumption 3 is not comparable to assuming the PL inequality for F. Formally, if F satisfies PL the condition for some parameter $\tau > 0$, then we have:

$$2\tau[F(w) - F_*] \leq \|\nabla F(w)\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w;i)\|^2. \tag{6}$$

However, by equation (3) we have that $[F(w) - F_*] \leq \frac{1}{n}\sum_{i=1}^{n}[f(w;i) - f_i^*]$. Therefore, the PL inequality for each function $f(\cdot; i)$, cannot directly imply the PL condition on F and vice versa.

In the interpolated setting where there is a common minimizer for all component function $f(\cdot; i)$, it can be shown that the PL condition on F is stronger than our average PL assumption:

$$2\tau\frac{1}{n}\sum_{i=1}^{n}[f(w;i) - f_i^*] = 2\tau[F(w) - F_*] \leq \|\nabla F(w)\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w;i)\|^2.$$

On the other hand, our assumption cannot imply the PL inequality on F unless we impose a strong relationship that upper bound the sum of individual squared gradients $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w;i)\|^2$ in terms of the full squared gradient $\|\nabla F(w)\|^2$, for every $w \in \mathbb{R}^d$. For these reasons, the average PL Assumption 3 is arguably more reasonable than assuming the PL inequality for the objective function F. Moreover, we show that Assumption 3 holds for a general class of neural networks with a final bias layer and squared loss function. We have the following theorem.

Theorem 1. *Let $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is a training data set where $x^{(i)} \in \mathbb{R}^m$ is the input data and $y^{(i)} \in \mathbb{R}^c$ is the output data for $i = 1, \ldots, n$. We consider an architecture $h(w; i)$ with w be the vectorized weight and h consists of a final bias layer b:*

$$h(w;i) = W^T z(\theta; i) + b,$$

where $w = \textbf{vec}(\{\theta, W, b\})$ and $z(\theta; i)$ are some inner architectures, which can be chosen arbitrarily. Next, we consider the squared loss $f(w; i) = \frac{1}{2}\|h(w; i) - y^{(i)}\|^2$. Then

$$\|\nabla f(w;i)\|^2 \geq 2[f(w;i) - f_i^*], \quad \forall w \in \mathbb{R}^d, \tag{7}$$

where $f_i^* := \min_{w\in\mathbb{R}^d} f(w; i)$.

Therefore, for this application, Assumption 3 holds with $\mu = 1$.

2.2 Small Variance at Global Solutions

In this section, we change the interpolation property in previous literature [Ma et al., 2018, Meng et al., 2020, Loizou et al., 2021] by a small threshold. For any global solution w_* of F, let us define

$$\sigma_*^2 := \inf_{w_*\in\mathcal{W}_*}\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_*;i)\|^2\right). \tag{8}$$

We can show that when there is a common minimizer for all component functions (i.e. when the equality $F_* = \frac{1}{n}\sum_{i=1}^{n} f_i^*$ holds), the best variance σ_*^2 is 0. It is sufficient for our Theorem to impose a $\mathcal{O}(\varepsilon)$-level upper bound on the variance σ_*^2:

Assumption 4. *Suppose that the best variance at w_* is small, that is, for $\varepsilon > 0$*

$$\sigma_*^2 \leq P\varepsilon, \tag{9}$$

for some $P > 0$.

It is important to note that in current literature, Assumption 4 alone (or, assuming $\sigma_*^2 = 0$ alone) is not sufficient enough to guarantee a global convergence property for SGD. Typically, some other conditions on the good landscape of the loss function are needed to complement the over-parameterized setting. Thus, we have motivation to introduce our next assumption.

2.3 Generalized Star-Smooth-Convex Condition for Shuffling Type Algorithm

We introduce the definition of star-smooth-convex function as follows.

Definition 2. *The function g is star-M-smooth-convex with respect to a reference point $\hat{w} \in \mathbb{R}^d$ if*

$$\|\nabla g(w) - \nabla g(\hat{w})\|^2 \leq M\langle \nabla g(w) - \nabla g(\hat{w}), w - \hat{w}\rangle, \quad \forall w \in \mathbb{R}^d. \tag{10}$$

It is well known that when g is L-smooth and convex [Nesterov, 2004], we have the following general inequality for every $w, w' \in \mathbb{R}^d$:

$$\|\nabla g(w) - \nabla g(w')\|^2 \leq L\langle \nabla g(w) - \nabla g(w'), w - w'\rangle \tag{11}$$

Our class of star-smooth-convex function requires a similar inequality to hold only for the special point $w' = \hat{w}$. Interestingly, this is related to a class of star-convex functions, which satisfies the convex inequality for the minimizer $\hat{w}$:

$$\text{(star-convexity w.r.t } \hat{w}) \quad g(w) - g(\hat{w}) \leq \langle \nabla g(w), w - \hat{w}\rangle, \ \forall w \in \mathbb{R}^d, \tag{12}$$

This class of functions contains non-convex objective losses and is well studied in the literature (see e.g. [Zhou et al., 2019]). Our Lemma 2 in the Appendix shows that the class of star-smooth-convex function is broader than the class of L-smooth and star-convex functions. Therefore, our problem of interest is non-convex in general.

For the analysis of shuffling type algorithm in this paper, we consider the general assumption called the *generalized star-smooth-convex condition for shuffling algorithms*:

Assumption 5. *Using Algorithm 1, let us assume that there exist some constants $M > 0$ and $N > 0$ such that at each epoch $t = 1, \ldots, T$, we have for $i = 1, \ldots, n$:*

$$\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i)) - \nabla f(w_*; \pi^{(t)}(i))\|^2 \leq M\langle \nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i)) - \nabla f(w_*; \pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle + N\frac{1}{n}\sum_{i=1}^{n} \|w_i^{(t)} - w_0^{(t)}\|^2, \tag{13}$$

where w_* is a global solution of F.

We note that when the individual function $f(\cdot; i)$ is star-M-smooth-convex with respect to w_* for every $i = 1, \ldots, n$, Assumption 5 holds for the case $N = 0$. Assumption 5 is more flexible than the one in (10) because the right-hand side term $\langle \nabla f(w; i) - \nabla f(w_*; i), w - w_*\rangle$ could be negative for some $w \in \mathbb{R}^d$. An additional term $N\frac{1}{n}\sum_{i=1}^{n} \|w_i^{(t)} - w_0^{(t)}\|^2$ for some constant $N > 0$ will allow for extra flexibility in our setting. Note that we do not impose any assumptions on bounded weights or bounded gradients. Therefore, the term $\frac{1}{n}\sum_{i=1}^{n} \|w_i^{(t)} - w_0^{(t)}\|^2$ cannot be uniformly bounded by any universal constant.

3 New Framework for Convergence to a Global Solution

In this section, we present our theoretical results. Our Lemma 1 first provides a recursion to bound the squared distance term $\|\tilde{w}_t - w_*\|^2$:

Lemma 1. *Assume that Assumptions 1, 2, 3, and 5 hold. Let $\{\tilde{w}_t\}_{t=1}^T$ be the sequence generated by Algorithm 1 with $0 < \eta_t \leq \min\left\{\frac{n}{2M}, \frac{1}{2L}\right\}$. For every $\gamma > 0$ we have*

$$\|\tilde{w}_t - w_*\|^2 \leq \left(1 + C_1\eta_t^3\right)\|\tilde{w}_{t-1} - w_*\|^2 + C_2\eta_t\sigma_*^2 - C_3\eta_t[F(\tilde{w}_{t-1}) - F_*]. \tag{14}$$

where w_* is a global solution of F, $F_* = \min_{w\in\mathbb{R}^d} F(w)$, and

$$\begin{cases} C_1 = \frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4\gamma L^4}{6M}, \\ C_2 = \frac{2}{M} + 1 + \frac{5}{6L^2} + \frac{8N}{3ML^2} + \frac{5\gamma}{12M}, \\ C_3 = \frac{\gamma}{\gamma+1}\frac{\mu}{M}. \end{cases} \tag{15}$$

Rearranging the results of Lemma 1, we have

$$F(\tilde{w}_{t-1}) - F_* \leq \frac{1}{C_3}\left(\frac{1}{\eta_t} + C_1\eta_t^2\right)\|\tilde{w}_{t-1} - w_*\|^2 - \frac{1}{C_3\eta_t}\|\tilde{w}_t - w_*\|^2 + \frac{C_2}{C_3}\sigma_*^2. \tag{16}$$

Therefore, with an appropriate choice of learning rate that guarantee $\left(1/\eta_t + C_1\eta_t^2\right) \leq 1/\eta_{t-1}$, we can unroll the recursion from Lemma 1. Thus we have our main result in the next Theorem.

Theorem 2. *Assume that Assumptions 1, 2, 3, and 5 hold. Let $\{\tilde{w}_t\}_{t=1}^T$ be the sequence generated by Algorithm 1 with the learning rate $\eta_i^{(t)} = \frac{\eta_t}{n}$ where $0 < \eta_t \leq \min\left\{\frac{n}{2M}, \frac{1}{2L}\right\}$. Let the number of iterations $T = \frac{\lambda}{\varepsilon^{3/2}}$ for some $\lambda > 0$ and $\varepsilon > 0$. Constants C_1, C_2, and C_3 are defined in (15) for any $\gamma > 0$. We further define $K = 1 + C_1 D^3 \varepsilon^{3/2}$ and specify the learning rate $\eta_t = K\eta_{t-1} = K^t\eta_0$ and $\eta_0 = \frac{D\sqrt{\varepsilon}}{K\exp(\lambda C_1 D^3)}$ such that $\frac{D\sqrt{\varepsilon}}{K} \leq \min\left\{\frac{n}{2M}, \frac{1}{2L}\right\}$ for some constant $D > 0$. Then we have*

$$\frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1}) - F_*] \leq \frac{K\exp(\lambda C_1 D^3)}{C_3 D\lambda}\|\tilde{w}_0 - w_*\|^2 \cdot \varepsilon + \frac{C_2}{C_3}\sigma_*^2, \tag{17}$$

where $F_* = \min_{w\in\mathbb{R}^d} F(w)$ and σ_*^2 is defined in (8).

Our analysis holds for arbitrarily constant values of the parameters γ, λ and D. In addition, we show our current analysis for every shuffling scheme. An interesting research question arises: whether the convergence results can be improved if one chooses to analyze a randomized shuffling scheme in this framework. However, we leave that question to future works.

Using Assumption 4, we can show the total complexity of Algorithm 1 for our setting.

Corollary 1. *Suppose that the conditions in Theorem 2 and Assumption 4 hold. Choose $C_1 D\lambda = 1$ and $\varepsilon = \hat{\varepsilon}/G$ such that $0 < \hat{\varepsilon} \leq G$ with the constants*

$$G = \frac{2C_1 D^2 e}{C_3}\|\tilde{w}_0 - w_*\|^2 + \frac{C_2 P}{C_3}, \text{ where}$$
$$\begin{cases} C_1 = \frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4L^2}{3M}, \\ C_2 = \frac{2}{M} + 1 + \frac{5}{6L^2} + \frac{8N}{3ML^2} + \frac{5}{12ML}, \\ C_3 = \frac{1}{L^2+1}\frac{\mu}{M}. \end{cases}$$

Then, the we need $T = \frac{\lambda G^{3/2}}{\hat{\varepsilon}^{3/2}}$ epochs to guarantee

$$\min_{1\leq t\leq T}[F(\tilde{w}_{t-1}) - F_*] \leq \frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1}) - F_*] \leq \hat{\varepsilon}.$$

Computational complexity. Our global convergence result in this Corollary holds for a fixed value of $\hat{\varepsilon}$ in Assumption 4. Thus, when $\varepsilon \to 0$, this assumption is equivalent to assuming $\sigma_*^2 = 0$. The total complexity of Corollary 1 is $\mathcal{O}\left(\frac{n}{\hat{\varepsilon}^{3/2}}\right)$. This rate matches the best known rate for unified sampling schemes for SGD in convex setting [Mishchenko et al., 2020, Nguyen et al., 2021]. However, our result holds for a broader class of functions that are possibly non-convex. Comparing to the non-convex setting, current literature [Mishchenko et al., 2020, Nguyen et al., 2021] also matches our rate to the order of $\hat{\varepsilon}$, however, we can only prove that SGD converges to a stationary point with a weaker criteria $\|\nabla F(w)\|^2 \leq \hat{\varepsilon}$ for general non-convex funtions. Table 1 shows these comparisons in various settings. Note that when using a randomized shuffling scheme, SGD often performs a better rate in terms of the data n in various settings with and without (strongly) convexity. For example, in strongly convex and/or PL setting, the convergence rate of RR is $\tilde{\mathcal{O}}(\sqrt{n}/\sqrt{\hat{\varepsilon}})$, which is better than unified schemes with $\tilde{\mathcal{O}}(n/\sqrt{\hat{\varepsilon}})$ [Ahn et al., 2020]. However, for a fair comparison, we do not report these results in Table 1 as our theoretical analysis is derived for unified shuffling scheme.

If we further assume that $L, M, N > 1$, the detailed complexity with respect to these constants is

$$\mathcal{O}\left(\frac{L^4(M+N)^{3/2}}{\mu^{3/2}} \cdot \frac{n}{\hat{\varepsilon}^{3/2}}\right).$$

We present all the detailed proofs in the Appendix. Our theoretical framework is new and adapted to the finite-sum minimization problem. Moreover, it utilizes the choice of shuffling sample schemes to show a better complexity in terms of $\hat{\varepsilon}$ than the complexity of vanilla i.i.d. sampling scheme.

Table 1: Comparisons of computational complexity (the number of individual gradient evaluations) needed by SGD algorithm to reach an $\hat{\varepsilon}$-accurate solution w that satisfies $F(w) - F(w_*) \leq \hat{\varepsilon}$ (or $\|\nabla F(w)\|^2 \leq \hat{\varepsilon}$ in the non-convex case).

Settings	**References**	**Complexity**	**Shuffling Schemes**	**Global Solution**
Convex	Nemirovski et al. [2009], Shamir and Zhang [2013] (1)	$\mathcal{O}\left(\frac{\Delta_0^2+G^2}{\hat{\varepsilon}^2}\right)$	✗	✓
	Mishchenko et al. [2020], Nguyen et al. [2021] (2)	$\mathcal{O}\left(\frac{n}{\hat{\varepsilon}^{3/2}}\right)$	✓	✓
PL condition	Nguyen et al. [2021]	$\tilde{\mathcal{O}}\left(\frac{n\sigma^2}{\hat{\varepsilon}^{1/2}}\right)$	✓	✓
Star-convex related	Gower et al. [2021] (3)	$\mathcal{O}\left(\frac{1}{\hat{\varepsilon}^2}\right)$	✗	✓
Non-convex	Ghadimi and Lan [2013] (5)	$\mathcal{O}\left(\frac{\sigma^2}{\hat{\varepsilon}^2}\right)$	✗	✗
	Nguyen et al. [2021], Mishchenko et al. [2020] (5)	$\mathcal{O}\left(\frac{n\sigma}{\hat{\varepsilon}^{3/2}}\right)$	✓	✗
Our setting (non-convex)	**This paper, Corollary 1** (4)	$\mathcal{O}\left(\frac{n(N\vee 1)^{3/2}}{\hat{\varepsilon}^{3/2}}\right)$	✓	✓

(0) We note that the assumptions in this table are not comparable and we only show the roughly complexity in terms of $\hat{\varepsilon}$. In addition, to make fair comparisons, we only report the complexity of unified shuffling schemes.
(1) Standard results for SGD in convex literature often use a different set of assumptions from the one in this paper (e.g. bounded domain that $\|w - w_*\|^2 \leq \Delta_0$ for each iterate w and/or bounded gradient that $\mathbb{E}[\|\nabla f(w;i)\|] \leq G^2$). We report the corresponding complexity for a rough comparison.
(2) [Mishchenko et al., 2020] shows a bound for Incremental Gradient while [Nguyen et al., 2021] has a unified setting. We translate these results for unified shuffling schemes from these references to the convex setting.
(3) Since we cannot find a reference containing the convergence rate for vanilla SGD and star-convex functions, we adapt the reference Gower et al. [2021] here. This paper shows a result for L-smooth and quasar convex function with an additional Expected Residual (ER) assumption, which is weaker than assuming smoothness for $f(\cdot;i)$ and interpolation property. The star-convex results hold when the quasar-convex parameter is 1.
(4) Since we use a different set of assumptions than the other references, we only report the rough comparison in n, N and $\hat{\varepsilon}$, where N is the constant from Assumption 5 and $N \vee 1 = \max(N, 1)$. Note that $N = 0$ in the framework of star-smooth-convex function. In addition, we need $\sigma_*^2 = 0$ so that the complexity holds with arbitrary $\hat{\varepsilon}$. We explain the detailed complexity below and in the Appendix.
(5) Standard literature for SGD in non-convex setting assumes a bounded variance that $\mathbb{E}_i\left[\|f(w;i) - \nabla F(w)\|^2\right] \leq \sigma^2$, we report the rough comparison.

4 Numerical Experiments

In this section, we show some experiments for shuffling-type SGD algorithms to demonstrate our theoretical results of convergence to a global solution. Following the setting of Theorem 1, we consider the non-convex regression problem with squared loss function. We choose fully connected neural networks in our implementation. We experiment with different regression datasets: the Diabetes dataset from sklearn library [Efron et al., 2004, Pedregosa et al., 2011] with 442 samples in dimension 10; the Life expectancy dataset from WHO [Repository, 2016] with 1649 trainable data points and 19 features. In addition, we test with the California Housing data from StatLib repository [Repository, 1997, Pedregosa et al., 2011] with a training set of 16514 samples and 8 features.

For the small Diabetes dataset, we use the classic LeNet-300-100 model [LeCun et al., 1998]. For other larger datasets, we use similar fully connected neural networks with an additional starting layer of 900 neurons. We apply the randomized reshuffling scheme using PyTorch framework [Paszke et al., 2019]. This shuffling scheme is the common heuristic in training neural networks and is implemented in many deep learning platforms (e.g. TensorFlow, PyTorch, and Keras [Abadi et al., 2015, Paszke et al., 2019, Chollet et al., 2015]).

For each dataset $\{x_i, y_i\}$, we preprocess and modify the initial data lightly to guarantee the over-parameterized setting in our experiment. We do this by using a pre-training stage: firstly we use GD/SGD algorithm to find a weight w that yields a sufficiently small value for the loss function (for Diabetes dataset we train to 10^{-8} and for other datasets we train to 10^{-2}). Letting the input data x_i be fixed, we change the label data to $\hat{y}_i$ such that the weight w yields a small loss function $\mathcal{O}(\epsilon)$ for the optimization associated with data $\{x_i, \hat{y}_i\}$, and the distance between $\hat{y}_i$ and y_i is small. Then the modified data is ready for the next stage. We summarize the data (after modification) in our experiments in Table 2.

Table 2: Datasets used in our experiments

Data name	# Samples	# Features	Networks layers	Sources
Diabetes	442	10	300-100	Efron et al. [2004]
Life Expectancy	1649	19	900-300-100	Repository [2016]
California Housing	16514	8	900-300-100	Repository [1997]

For each dataset, we first tune the step size using a coarse grid search $[0.0001, 0.001, 0.01, 0.1, 1]$ for 100 epochs. Then, for example, if 0.01 performs the best, we test the second grid search $[0.002, 0.005, 0.01, 0.02, 0.05]$ for 5000 epochs. Finally, we progress to the training stage with 10^6 epochs and repeat that experiment for 5 random seeds. We report the average results with confidence intervals in Figure 1.

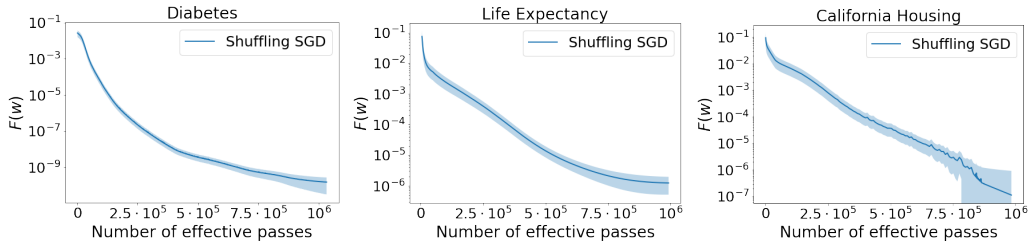

Figure 1: The train loss produced by Shuffling SGD algorithm for three datasets: Diabetes, Life Expectancy and California Housing.

For California Housing data, Shuffling SGD fluctuates toward the end of the training process. Nevertheless, for all three datasets it converges steadily to a small value of loss function. In summary, this experiment confirms our theoretical guarantee that demonstrates a convergence to global solution for shuffling-type SGD algorithm in neural network settings.

5 Conclusion

In this paper, we study the global convergence property for shuffling-type SGD methods. We consider a relaxed set of assumptions in the framework of star-smooth-convex functions and show the total complexity of $\mathcal{O}(\frac{n}{\hat{\epsilon}^{3/2}})$ to reach an $\hat{\epsilon}$-accurate global solution. This result matches the previous computational complexity of unified shuffling methods in convex settings. Our theoretical framework utilizes the choice of shuffling sample schemes for finite-sum minimization problems in machine learning. We provide discussions on the relations of our framework and the well-known over-parameterized settings, as well as current literature on the star-convexity class of functions. In addition, we show the connections to neural network architectures and discuss how these learning models fit into our optimization frameworks. Potential research questions arising from our paper include practical network designs and relaxed theoretical settings that support the global convergence of Shuffling SGD methods. Moreover, the global convergence framework for other stochastic gradient methods [Duchi et al., 2011, Kingma and Ba, 2014] and variance reduction methods [Nguyen et al., 2017, Beznosikov and Takáč, 2021] with shuffling sampling schemes is also an interesting direction.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17526–17535. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/cb8acb1dc9821bf74e6ca9068032d623-Paper.pdf.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/allen-zhu19a.html.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/arora19a.html.

Aleksandr Beznosikov and Martin Takáč. Random-reshuffled sarah does not need a full gradient computations. *arXiv preprint arXiv:2111.13322*, 2021.

Johan Bjorck, Anmol Kabra, Kilian Q. Weinberger, and Carla Gomes. Characterizing the loss landscape in non-negative matrix factorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6768–6776, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16836.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.*, 60(2):223–311, 2018.

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms, 2009.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJ33wwxRb.

Francois Chollet et al. Keras. *GitHub*, 2015. URL https://github.com/fchollet/keras.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM, 2008. doi: 10.1145/1390156.1390177. URL https://doi.org/10.1145/1390156.1390177.

Soham De, Abhay Yadav, David Jacobs, and Tom Goldstein. Automated Inference with Adaptive Batches. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1504–1513. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/de17a.html.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019a. URL http://proceedings.mlr.press/v97/du19c.html.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=S1eK3i09YQ.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Diabetes dataset, 2004. URL https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html.

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

Yoav Goldberg, Graeme Hirst, Yang Liu, and Meng Zhang. Neural network methods for natural language processing. *Comput. Linguistics*, 44(1), 2018. doi: 10.1162/COLI_r_00312. URL https://doi.org/10.1162/COLI_r_00312.

Ian J. Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay D. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6082.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, Oct 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01440-w. URL http://dx.doi.org/10.1007/s10107-019-01440-w.

Jeff Haochen and Suvrit Sra. Random shuffling beats sgd after finite epochs. In *International Conference on Machine Learning*, pages 2624–2633. PMLR, 2019.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018. URL http://jmlr.org/papers/v19/16-465.html.

Kaiming He and Jian Sun. Convolutional neural networks at constrained time cost. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5353–5360. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299173. URL https://doi.org/10.1109/CVPR.2015.7299173.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on Learning Theory*, pages 1894–1938. PMLR, 2020.

Jikai Jin. On the convergence of first order methods for quasar-convex optimization. *arXiv preprint arXiv:2010.04937*, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing.

Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jasper C.H. Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614, 2016. doi: 10.1109/FOCS.2016.71.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1306–1314. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/loizou21a.html.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3325–3334. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/ma18a.html.

Si Yi Meng, Sharan Vaswani, Issam Hadj Laradji, Mark Schmidt, and Simon Lacoste-Julien. Fast and furious convergence: Stochastic second order methods under interpolation. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1375–1386. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/meng20a.html.

Konstantin Mishchenko, Ahmed Khaled Ragab Bayoumi, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004. ISBN 1-4020-7553-7.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org, 2017.

Lam M. Nguyen, Quoc Tran-Dinh, Dzung T. Phan, Phuong Ha Nguyen, and Marten van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021. URL http://jmlr.org/papers/v22/20-1238.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Global Health Observatory Data Repository. Life expectancy and healthy life expectancy, 2016. URL https://apps.who.int/gho/data/view.main.SDG2016LEXREGv?lang=en.

StatLib Repository. California housing, 1997. URL https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html.

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

Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

Karthik Abinav Sankararaman, Soham De, Zheng Xu, W Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *International conference on machine learning*, pages 8469–8479. PMLR, 2020.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition, 2013.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/shamir13.html.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19(1):2822–2878, January 2018. ISSN 1532-4435.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/vaswani19a.html.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2053–2062, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks, 2018.

On the Convergence to a Global Solution of Shuffling-Type Gradient Algorithms Supplementary Material, NeurIPS 2023

A Theoretical settings: Proof of Theorem 1 and Lemma 2

A.1 Proof of Theorem 1

Proof. Let us use the notation $f(w;i) = \phi_i(h(w;i)) = \frac{1}{2}\|h(w;i) - y^{(i)}\|^2$. We consider an architecture $h(w;i)$ with w be the vectorized weight and h consists of a final bias layer b:

$$h(w;i) = W^T z(\theta;i) + b \in \mathbb{R}^c,$$

where $w = \textbf{vec}(\{\theta, W, b\})$ and $z(\theta;i)$ are some inner architecture, which can be chosen arbitrarily.

Firstly, we compute the gradient of $f(\cdot;i)$ with respect to $b \in \mathbb{R}^c$. For $j = 1, \ldots, c$, we have

$$\frac{\partial f(w;i)}{\partial b_j} = \frac{\partial \phi_i(h(w;i))}{\partial b_j} = \sum_{k=1}^{c} \frac{\partial h(w;i)_k}{\partial b_j} \cdot \frac{\partial \phi_i(x)}{\partial x_k}\Big|_{x=h(w;i)} = \frac{\partial \phi_i(x)}{\partial x_j}\Big|_{x=h(w;i)}, \ i = 1, \ldots, n. \tag{18}$$

The last equality follows since $\frac{\partial h(w;i)_k}{\partial b_j} = 0$ for every $k \neq j$ and $\frac{\partial h(w;i)_k}{\partial b_j} = 1$ for $k = j$. In other words, it is the identity matrix.

Let us denote that $f_i^* = \min_w f(w;i)$ and $\phi_i^* = \min_x \phi_i(x)$. We prove the following statement for $\mu = 1$:

$$\|\nabla_w f(w;i)\|^2 \geq \|\nabla_x \phi_i(x)|_{x=h(w;i)}\|^2 \geq 2\mu[\phi_i(h(w;i)) - \phi_i^*] \geq 2\mu[f(w;i) - f_i^*],$$

for every $w \in \mathbb{R}^d$, and $i = 1, \ldots, n$.

We begin with the first inequality:

$$\begin{aligned}\|\nabla_w f(w;i)\|^2 = \sum_{j=1}^{d} \Big(\frac{\partial f(w;i)}{\partial w_j}\Big)^2 &\geq \sum_{j=d-c+1}^{d} \Big(\frac{\partial f(w;i)}{\partial w_j}\Big)^2 = \sum_{j=1}^{c} \Big(\frac{\partial f(w;i)}{\partial b_j}\Big)^2 \\ &\overset{(18)}{=} \sum_{j=1}^{c} \Big(\frac{\partial \phi_i(x)}{\partial x_j}\Big|_{x=h(w;i)}\Big)^2 = \|\nabla_x \phi_i(x)|_{x=h(w;i)}\|^2.\end{aligned}$$

Now let us prove the PL condition for each function $\phi_i(x)$, i.e., there exists a constant $\mu > 0$ such that:

$$\|\nabla_x \phi_i(x)\|^2 \geq 2\mu[\phi_i(x) - \phi_i^*] \ \forall x \in \mathbb{R}^c, \ i = 1, \ldots, n.$$

Recall the squared loss that $\phi_i(x) = \frac{1}{2}\|x - y^{(i)}\|^2$ and $\nabla_x \phi_i(x) = x - y^{(i)}$. We can see that the constant $\mu = 1$ satisfies the following inequality for every $x \in \mathbb{R}^c, \ i = 1, \ldots, n$:

$$\|\nabla_x \phi_i(x)\|^2 = \|x - y^{(i)}\|^2 = 2\frac{1}{2}\|x - y^{(i)}\|^2 = 2\mu\phi_i(x) \geq 2\mu[\phi_i(x) - \phi_i^*],$$

where the last inequality follows since $\phi_i^* \geq 0$.

The PL condition for ϕ_i directly implies the second inequality. The last inequality follows from the facts that $f(w;i) = \phi_i(h(w;i))$ and $f_i^* = \min_w f_i \geq \min_x \phi_i(x) = \phi_i^*$. Hence, Theorem 1 is proved. □

A.2 Statement and Proof of Lemma 2

Lemma 2. *The function g is star-M-smooth-convex with respect to $\hat{w}$ for some constant $M > 0$ if one of the two following conditions holds:*

1. *g is L-smooth and convex.*
2. *g is L-smooth and g is star-convex with respect to $\hat{w}$.*

Proof. The first statement of Lemma 2 follows directly from equation (11). We have that g is star-M-smooth-convex with respect to any reference point and $M = L$.

Now we proceed to the second statement. From the star-convex property of g with respect to $\hat{w}$, we have

$$g(w) - g(\hat{w}) \leq \langle \nabla g(w), w - \hat{w} \rangle, \ \forall w \in \mathbb{R}^d,$$

and $\nabla g(\hat{w}) = 0$ since $\hat{w}$ is the global minimizer of g. On the other hand, g is L-smooth and we have

$$g(\hat{w}) \leq g\left(w - \frac{1}{L}\nabla g(w)\right) \leq g(w) - \frac{1}{2L}\|\nabla g(w)\|^2,$$

which is equivalent to $\|\nabla g(w)\|^2 \leq 2L[g(w) - g(\hat{w})]$, $i \in [n]$. Since $\nabla g(\hat{w}) = 0$, $i \in [n]$, we have for $\forall w \in \mathbb{R}^d$

$$\|\nabla g(w) - \nabla g(\hat{w})\|^2 \leq 2L[g(w) - g(\hat{w})] \overset{(12)}{\leq} 2L\langle \nabla g(w) - \nabla g(\hat{w}), w - w_* \rangle.$$

This is a star-M-smooth-convex function as in Definition 2 with $M = 2L$. □

B Preliminary results for SGD Shuffling Algorithm

In this section, we present the preliminary results for Algorithm 1. Firstly, from the choice of learning rate $\eta_i^{(t)} := \frac{\eta_t}{n}$ and the update $w_{i+1}^{(t)} := w_i^{(t)} - \eta_i^{(t)} \nabla f(w_i^{(t)}; \pi^{(t)}(i+1))$ in Algorithm 1, for $i \in [n]$, we have

$$w_i^{(t)} = w_{i-1}^{(t)} - \frac{\eta_t}{n}\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i)) = w_0^{(t)} - \frac{\eta_t}{n}\sum_{j=0}^{i-1} \nabla f(w_j^{(t)}; \pi^{(t)}(j+1)). \tag{19}$$

Hence,

$$w_0^{(t+1)} = w_n^{(t)} = w_0^{(t)} - \frac{\eta_t}{n}\sum_{j=0}^{n-1} \nabla f(w_j^{(t)}; \pi^{(t)}(j+1)). \tag{20}$$

Next, we refer to a Lemma in [Nguyen et al., 2021] to bound the updates of shuffling SGD algorithms.

Lemma 3 (Lemma 5 in Nguyen et al. [2021])**.** *Suppose that Assumption 2 holds for (1). Let $\{w_i^{(t)}\}$ be generated by Algorithm 1 with the learning rate $\eta_i^{(t)} := \frac{\eta_t}{n} > 0$ for a given positive sequence $\{\eta_t\}$. If $0 < \eta_t \leq \frac{1}{2L}$ for all $t \geq 1$, we have*

$$\frac{1}{n}\sum_{j=0}^{n-1} \|w_j^{(t)} - w_*\|^2 \leq 4\|w_0^{(t)} - w_*\|^2 + 8\sigma_*^2 \cdot \eta_t^2, \tag{21}$$

$$\frac{1}{n}\sum_{j=0}^{n-1} \|w_j^{(t)} - w_0^{(t)}\|^2 \leq \eta_t^2 \cdot \frac{8L^2}{3}\|w_0^{(t)} - w_*\|^2 \ + \ \frac{16L^2\sigma_*^2}{3} \cdot \eta_t^4 \ + \ 2\sigma_*^2 \cdot \eta_t^2. \tag{22}$$

Now considering the term $\|w_n^{(t)} - w_0^{(t)}\|^2$, we get that

$$\|w_n^{(t)} - w_0^{(t)}\|^2 \overset{(20)}{\leq} \frac{\eta_t^2}{n}\left\|\frac{1}{n}\sum_{j=0}^{n-1} \nabla f(w_j^{(t)}; \pi^{(t)}(j+1))\right\|^2$$

$$
\begin{aligned}
&= \frac{\eta_t^2}{n}\left\|\frac{1}{n}\sum_{j=0}^{n-1}(\nabla f(w_j^{(t)};\pi^{(t)}(j+1)) - \nabla f(w_*;\pi^{(t)}(j+1)))\right\|^2 \\
&\le \frac{\eta_t^2}{n}\frac{1}{n}\sum_{j=0}^{n-1}\left\|\nabla f(w_j^{(t)};\pi^{(t)}(j+1)) - \nabla f(w_*;\pi^{(t)}(j+1))\right\|^2 \\
&\overset{(2)}{\le} \frac{L^2\eta_t^2}{n}\frac{1}{n}\sum_{j=0}^{n-1}\|w_j^{(t)} - w_*\|^2 \\
&\overset{(21)}{\le} \frac{4L^2\eta_t^2}{n}\|w_0^{(t)} - w_*\|^2 + \frac{8L^2\eta_t^4}{n}\sigma_*^2.
\end{aligned}
$$

We further have

$$
\begin{aligned}
\frac{1}{n}\sum_{j=0}^{n}\|w_j^{(t)} - w_0^{(t)}\|^2 &= \frac{1}{n}\sum_{j=0}^{n-1}\|w_j^{(t)} - w_0^{(t)}\|^2 + \frac{1}{n}\|w_n^{(t)} - w_0^{(t)}\|^2 \\
&\le \eta_t^2\cdot\frac{8L^2}{3}\|w_0^{(t)} - w_*\|^2 \;+\; \frac{16L^2\sigma_*^2}{3}\cdot\eta_t^4 \;+\; 2\sigma_*^2\cdot\eta_t^2 \\
&\quad + \frac{4L^2\eta_t^2}{n}\|w_0^{(t)} - w_*\|^2 + \frac{8L^2\eta_t^4}{n}\sigma_*^2. \qquad (23)
\end{aligned}
$$

C Main results: Proofs of Lemma 4, Lemma 1, Theorem 2, and Corollary 1

C.1 Proof of Lemma 4

Lemma 4. *Let $\{w_i^{(t)}\}_{t=1}^T$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n}$, with $0 < \eta_t \le \frac{n}{2M}$ for $\eta_t \le \frac{1}{2L}$. Then, under Assumptions 1, 2, and 5, we have*

$$
\|w_0^{(t+1)} - w_*\|^2 \le \left(1 + B_1\eta_t^3\right)\|w_0^{(t)} - w_*\|^2 - \frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))\|^2 + B_2\eta_t\sigma_*^2, \qquad (24)
$$

where

$$
\begin{cases} B_1 = \frac{8L^2}{3} + \frac{14NL^2}{M}, \\ B_2 = \frac{2}{M} + 1 + \frac{5}{6L^2} + \frac{8N}{3ML^2}. \end{cases} \qquad (25)
$$

Proof. We start with Assumption 5. Using the inequality $\frac{1}{2}\|a\|^2 - \|b\|^2 \le \|a-b\|^2$, we have for $t = 1,\dots,T$ and $i = 1,\dots,n$:

$$
\begin{aligned}
&\frac{1}{2}\|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))\|^2 - \|\nabla f(w_*;\pi^{(t)}(i))\|^2 \\
&\quad\le \|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i)) - \nabla f(w_*;\pi^{(t)}(i))\|^2 \\
&\quad\overset{(13)}{\le} M\langle\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i)) - \nabla f(w_*;\pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle + N\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\quad= M\langle\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle - M\langle\nabla f(w_*;\pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle \\
&\qquad + N\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2,
\end{aligned}
$$

This statement is equivalent to

$$
\begin{aligned}
-\langle\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle \le &-\frac{1}{2M}\|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))\|^2 + \frac{1}{M}\|\nabla f(w_*;\pi^{(t)}(i))\|^2 \\
&- \langle\nabla f(w_*;\pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle
\end{aligned}
$$

$$+ \frac{N}{M}\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2, \tag{26}$$

For any $w_* \in W^*$, from the update (19) we have,

$$\begin{aligned}
\|w_i^{(t)} - w_*\|^2 &\overset{(19)}{=} \|w_{i-1}^{(t)} - w_*\|^2 - \frac{2\eta_t}{n}\langle \nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle + \frac{\eta_t^2}{n^2}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 \\
&\overset{(26)}{\leq} \|w_{i-1}^{(t)} - w_*\|^2 - \frac{2\eta_t}{2Mn}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 + \frac{2\eta_t}{Mn}\|\nabla f(w_*; \pi^{(t)}(i))\|^2 \\
&\quad - \frac{2\eta_t}{n}\langle \nabla f(w_*; \pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle + \frac{2\eta_t N}{Mn}\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\quad + \frac{\eta_t^2}{n^2}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 \\
&\overset{(a)}{\leq} \|w_{i-1}^{(t)} - w_*\|^2 - \frac{\eta_t}{2Mn}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 + \frac{2\eta_t}{Mn}\|\nabla f(w_*; \pi^{(t)}(i))\|^2 \\
&\quad - \frac{2\eta_t}{n}\langle \nabla f(w_*; \pi^{(t)}(i)), w_{i-1}^{(t)} - w_*\rangle + \frac{2\eta_t N}{Mn}\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&= \|w_{i-1}^{(t)} - w_*\|^2 - \frac{\eta_t}{2Mn}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 + \frac{2\eta_t}{Mn}\|\nabla f(w_*; \pi^{(t)}(i))\|^2 \\
&\quad - \frac{2\eta_t}{n}\langle \nabla f(w_*; \pi^{(t)}(i)), w_{i-1}^{(t)} - w_0^{(t)}\rangle - \frac{2\eta_t}{n}\langle \nabla f(w_*; \pi^{(t)}(i)), w_0^{(t)} - w_*\rangle \\
&\quad + \frac{2\eta_t N}{Mn}\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\overset{(b)}{\leq} \|w_{i-1}^{(t)} - w_*\|^2 - \frac{\eta_t}{2Mn}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 + \frac{2\eta_t}{Mn}\|\nabla f(w_*; \pi^{(t)}(i))\|^2 \\
&\quad + \frac{\eta_t}{n}\|\nabla f(w_*; \pi^{(t)}(i))\|^2 + \frac{\eta_t}{n}\|w_{i-1}^{(t)} - w_0^{(t)}\|^2 \\
&\quad - \frac{2\eta_t}{n}\langle \nabla f(w_*; \pi^{(t)}(i)), w_0^{(t)} - w_*\rangle + \frac{2\eta_t N}{Mn}\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2,
\end{aligned}$$

where (a) follows since $\eta_t \leq \frac{n}{2M}$ and (b) follows by the inequality $2\langle a, b\rangle \leq \|a\|^2\|b\|^2$.

Note that $\frac{1}{n}\sum_{i=1}^{n}\langle \nabla f(w_*; \pi^{(t)}(i)), w_0^{(t)} - w_*\rangle = \langle \nabla F(w_*), w_0^{(t)} - w_*\rangle = 0$ since w_* is a global solution of F. Now we sum the derived statement for $i = 1, \ldots, n$ and get

$$\begin{aligned}
\|w_n^{(t)} - w_*\|^2 &\leq \|w_0^{(t)} - w_*\|^2 - \frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 \\
&\quad + \eta_t\left(\frac{2}{M} + 1\right)\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_*; \pi^{(t)}(i))\|^2 + \frac{\eta_t}{n}\sum_{i=1}^{n}\|w_{i-1}^{(t)} - w_0^{(t)}\|^2 \\
&\quad + \frac{2N\eta_t}{M}\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\overset{(8),(22)}{\leq} \|w_0^{(t)} - w_*\|^2 - \frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2 \\
&\quad + \left(\frac{2}{M} + 1\right)\eta_t\sigma_*^2 + \frac{8L^2\eta_t^3}{3}\|w_0^{(t)} - w_*\|^2 + \frac{16L^2\eta_t^5}{3}\sigma_*^2 + 2\eta_t^3\sigma_*^2 \\
&\quad + \frac{2N\eta_t}{M}\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\overset{(23)}{\leq} \|w_0^{(t)} - w_*\|^2 - \frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))\|^2
\end{aligned}$$

$$+\left(\frac{2}{M}+1\right)\eta_t\sigma_*^2+\frac{8L^2\eta_t^3}{3}\|w_0^{(t)}-w_*\|^2+\frac{16L^2\eta_t^5}{3}\sigma_*^2+2\eta_t^3\sigma_*^2$$
$$+\frac{16NL^2\eta_t^3}{3M}\|w_0^{(t)}-w_*\|^2+\frac{32NL^2\eta_t^5}{3M}\sigma_*^2+\frac{4N\eta_t^3}{M}\sigma_*^2$$
$$+\frac{8NL^2\eta_t^3}{Mn}\|w_0^{(t)}-w_*\|^2+\frac{16NL^2\eta_t^5}{Mn}\sigma_*^2,$$

where we apply the derivations from Lemma 3. Now noting that $\eta_t \le \frac{1}{2L}$, $n \le 1$ and rearranging the terms we get:

$$\|w_n^{(t)}-w_*\|^2 \le \|w_0^{(t)}-w_*\|^2-\frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))\|^2$$
$$+\left(\frac{8L^2}{3}+\frac{16NL^2}{3M}+\frac{8NL^2}{M}\right)\eta_t^3\|w_0^{(t)}-w_*\|^2$$
$$+\left(\frac{2}{M}+1+\frac{1}{3L^2}+\frac{1}{2L^2}+\frac{2N}{3ML^2}+\frac{N}{ML^2}+\frac{N}{ML^2}\right)\eta_t\sigma_*^2$$

Since $w_n^{(t)}=w_0^{(t+1)}=\tilde{w}_t$, we have the desired result in (24). □

C.2 Proof of Lemma 1

Proof. From (24) where B_1 and B_2 are defined in (25), we have

$$\|w_0^{(t+1)}-w_*\|^2$$
$$\le \left(1+B_1\eta_t^3\right)\|w_0^{(t)}-w_*\|^2-\frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))\|^2+B_2\eta_t\sigma_*^2$$
$$\overset{(a)}{\le}\left(1+B_1\eta_t^3\right)\|w_0^{(t)}-w_*\|^2-\frac{\gamma}{\gamma+1}\frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_0^{(t)};\pi^{(t)}(i))\|^2$$
$$+\frac{\eta_t\gamma}{2M}\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))-\nabla f(w_0^{(t)};\pi^{(t)}(i))\|^2+B_2\eta_t\sigma_*^2$$
$$\overset{(2)}{\le}\left(1+B_1\eta_t^3\right)\|w_0^{(t)}-w_*\|^2-\frac{\gamma}{\gamma+1}\frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_0^{(t)};\pi^{(t)}(i))\|^2$$
$$+\frac{\eta_t\gamma L^2}{2M}\frac{1}{n}\sum_{i=1}^n\|w_{i-1}^{(t)}-w_0^{(t)}\|^2+B_2\sigma_*^2$$
$$\overset{(22)}{\le}\left(1+B_1\eta_t^3\right)\|w_0^{(t)}-w_*\|^2-\frac{\gamma}{\gamma+1}\frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_0^{(t)};\pi^{(t)}(i))\|^2$$
$$+\frac{\eta_t^3\gamma L^2}{2M}\left(\frac{8L^2}{3}\|w_0^{(t)}-w_*\|^2+\frac{16L^2\sigma_*^2}{3}\cdot\eta_t^2+2\sigma_*^2\right)+B_2\eta_t\sigma_*^2$$
$$=\left(1+B_1\eta_t^3+\frac{4\eta_t^3\gamma L^4}{3M}\right)\|w_0^{(t)}-w_*\|^2+\left(B_2+\frac{\eta_t^2\gamma L^2}{M}+\frac{8\eta_t^4\gamma L^4}{3M}\right)\eta_t\sigma_*^2$$
$$-\frac{\gamma}{\gamma+1}\frac{\eta_t}{2M}\frac{1}{n}\sum_{i=1}^n\|\nabla f(w_0^{(t)};\pi^{(t)}(i))\|^2$$
$$\overset{(5)}{\le}\left(1+\eta_t^3\left(B_1+\frac{4\gamma L^4}{3M}\right)\right)\|w_0^{(t)}-w_*\|^2+\left(B_2+\frac{\eta_t^2\gamma L^2}{M}+\frac{8\eta_t^4\gamma L^4}{3M}\right)\eta_t\sigma_*^2$$
$$-\frac{\gamma}{\gamma+1}\frac{2\mu\eta_t}{2M}\frac{1}{n}\sum_{i=1}^n[f(w_0^{(t)};\pi^{(t)}(i))-f_i^*]$$
$$\overset{(3)}{\le}\left(1+\eta_t^3\left(B_1+\frac{4\gamma L^4}{3M}\right)\right)\|w_0^{(t)}-w_*\|^2+\left(B_2+\frac{\eta_t^2\gamma L^2}{M}+\frac{8\eta_t^4\gamma L^4}{3M}\right)\eta_t\sigma_*^2$$

$$
\begin{aligned}
&- \frac{\gamma}{\gamma+1}\frac{\mu\eta_t}{M}[F(w_0^{(t)}) - F_*] \\
&\overset{(b)}{\leq} \left(1 + \eta_t^3\left(B_1 + \frac{4\gamma L^4}{3M}\right)\right)\|w_0^{(t)} - w_*\|^2 + \left(B_2 + \frac{\gamma}{4M} + \frac{\gamma}{6M}\right)\eta_t\sigma_*^2 \\
&\quad - \frac{\gamma}{\gamma+1}\frac{\mu\eta_t}{M}[F(w_0^{(t)}) - F_*],
\end{aligned}
$$

where (a) follows since $-\|b\|^2 \leq \gamma\|a-b\|^2 - \frac{\gamma}{\gamma+1}\|a\|^2$ for any $\gamma > 0$ and (b) follows since $\eta_t \leq \frac{1}{2L}$. Since $w_0^{(t+1)} = \tilde{w}_t$, we obtain the desired result in (14). □

C.3 Proof of Theorem 2

Proof. For $t = 1, \ldots, T = \frac{\lambda}{\varepsilon^{3/2}}$ for some $\lambda > 0$

$$
\begin{aligned}
\eta_t &= (1 + C_1 D^3 \varepsilon^{3/2})\eta_{t-1} = (1 + C_1 D^3 \varepsilon^{3/2})^t \eta_0 \leq (1 + C_1 D^3 \varepsilon^{3/2})^T \eta_0 \\
&= (1 + C_1 D^3 \varepsilon^{3/2})^{\lambda/\varepsilon^{3/2}} \eta_0 = (1 + C_1 D^3 \varepsilon^{3/2})^{\lambda/\varepsilon^{3/2}} \frac{D\sqrt{\varepsilon}}{(1 + C_1 D^3 \varepsilon^{3/2})\exp(\lambda C_1 D^3)} \\
&\leq \frac{D\sqrt{\varepsilon}}{(1 + C_1 D^3 \varepsilon^{3/2})} \leq \min\left\{\frac{n}{2M}, \frac{1}{2L}\right\},
\end{aligned} \tag{27}
$$

since $(1+x)^{1/x} \leq e$, $x > 0$. From (14), we have

$$
[F(\tilde{w}_{t-1}) - F_*] \leq \frac{1}{C_3}\left(\frac{1}{\eta_t} + C_1\eta_t^2\right)\|\tilde{w}_{t-1} - w_*\|^2 - \frac{1}{C_3\eta_t}\|\tilde{w}_t - w_*\|^2 + \frac{C_2}{C_3}\sigma_*^2. \tag{28}
$$

We proceed to prove the following inequality for $t = 1, \ldots, T$,

$$
\frac{1}{\eta_t} + C_1\eta_t^2 \leq \frac{1}{\eta_{t-1}}. \tag{29}
$$

From (27), and $\eta_t = K\eta_{t-1}$ where $K = (1 + C_1 D^3 \varepsilon^{3/2})$, we have

$$
\begin{aligned}
C_1\eta_t^2 &= C_1 K^2 \eta_{t-1}^2 = C_1 K^2 \frac{\eta_{t-1}^3}{\eta_{t-1}} \\
&\leq C_1 K^2 \frac{D^3\varepsilon^{3/2}}{K^3\eta_{t-1}} = \frac{C_1 D^3 \varepsilon^{3/2}}{K\eta_{t-1}} && \text{since } \eta_{t-1} \leq \frac{D\sqrt{\varepsilon}}{K} \\
&= \frac{K-1}{K}\frac{1}{\eta_{t-1}} = \frac{1}{\eta_{t-1}} - \frac{1}{K\eta_{t-1}} && \text{since } K = (1 + C_1 D^3 \varepsilon^{3/2}) \\
&= \frac{1}{\eta_{t-1}} - \frac{1}{K\eta_{t-1}} = \frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}, && \text{since } \eta_t = K\eta_{t-1}.
\end{aligned}
$$

for $t = 1, \ldots, T$. Hence, from (28), we have

$$
\begin{aligned}
[F(\tilde{w}_{t-1}) - F_*] &\leq \frac{1}{C_3}\left(\frac{1}{\eta_t} + C_1\eta_t^2\right)\|\tilde{w}_{t-1} - w_*\|^2 - \frac{1}{C_3\eta_t}\|\tilde{w}_t - w_*\|^2 + \frac{C_2}{C_3}\sigma_*^2 \\
&\leq \frac{1}{C_3\eta_{t-1}}\|\tilde{w}_{t-1} - w_*\|^2 - \frac{1}{C_3\eta_t}\|\tilde{w}_t - w_*\|^2 + \frac{C_2}{C_3}\sigma_*^2.
\end{aligned}
$$

Averaging the statement above for $t = 1, \ldots, T$, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1}) - F_*] &\leq \frac{1}{C_3\eta_0 T}\|\tilde{w}_0 - w_*\|^2 + \frac{C_2}{C_3}\sigma_*^2 \\
&\overset{(a)}{=} \frac{K\exp(\lambda C_1 D^3)}{C_3 D\sqrt{\varepsilon}}\frac{\varepsilon^{3/2}}{\lambda}\|\tilde{w}_0 - w_*\|^2 + \frac{C_2}{C_3}\sigma_*^2 \\
&= \frac{K\exp(\lambda C_1 D^3)}{C_3 D\lambda}\|\tilde{w}_0 - w_*\|^2 \cdot \varepsilon + \frac{C_2}{C_3}\sigma_*^2,
\end{aligned}
$$

where (a) follows since $\eta_0 = \frac{D\sqrt{\varepsilon}}{K\exp(\lambda C_1 D^3)}$ and $T = \frac{\lambda}{\varepsilon^{3/2}}$. □

C.4 Proof of Corollary 1

Proof. Choose $\gamma = \frac{1}{L^2}$, we have

$$\begin{cases} C_1 = \frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4L^2}{3M}, \\ C_2 = \frac{2}{M} + 1 + \frac{5}{6L^2} + \frac{8N}{3ML^2} + \frac{5}{12ML}, \\ C_3 = \frac{1}{L^2+1}\frac{\mu}{M}. \end{cases}$$

Note that $K = 1 + C_1 D^3 \varepsilon^{3/2}$ and $C_1 D^3 = 1/\lambda$, we get that $K = 1 + 1/T \leq 2$. We continue from the statement of Theorem 2 and the choice $C_1 D^3 \lambda = 1$:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1}) - F_*] &\leq \frac{K\exp(\lambda C_1 D^3)}{C_3 D\lambda}\|\tilde{w}_0 - w_*\|^2 \cdot \varepsilon + \frac{C_2}{C_3}\sigma_*^2 \\ &\leq \frac{2}{C_1 D^3 \lambda} \cdot \frac{C_1 D^2 \exp(\lambda C_1 D^3)}{C_3} \cdot \|\tilde{w}_0 - w_*\|^2 \cdot \varepsilon + \frac{C_2}{C_3}\sigma_*^2 && \text{since } K \leq 2 \\ &\leq \frac{2C_1 D^2 e}{C_3}\|\tilde{w}_0 - w_*\|^2 \cdot \varepsilon + \frac{C_2}{C_3}\sigma_*^2 && \text{since } C_1 D^3 \lambda = 1 \\ &\leq \frac{2C_1 D^2 e}{C_3}\|\tilde{w}_0 - w_*\|^2 \cdot \varepsilon + \frac{C_2}{C_3} P\varepsilon && \text{equation (9)} \\ &\leq \left(\frac{2C_1 D^2 e}{C_3}\|\tilde{w}_0 - w_*\|^2 + \frac{C_2 P}{C_3}\right)\varepsilon = G\varepsilon \end{aligned}$$

with

$$G = \frac{2C_1 D^2 e}{C_3}\|\tilde{w}_0 - w_*\|^2 + \frac{C_2 P}{C_3}.$$

Let $0 < \varepsilon \leq 1$ and choose $\hat{\varepsilon} = G\varepsilon$. Then the number of iterations T is

$$\begin{aligned} T &= \frac{\lambda}{\varepsilon^{3/2}} = \frac{\lambda G^{3/2}}{\hat{\varepsilon}^{3/2}} \\ &= \frac{1}{\hat{\varepsilon}^{3/2} C_1 D^3}\left(\frac{2C_1 D^2 e\|\tilde{w}_0 - w_*\|^2 + C_2 P}{C_3}\right)^{3/2} \\ &= \frac{1}{\hat{\varepsilon}^{3/2}} \cdot \frac{1}{C_1 D^3 C_3^{3/2}}\left(2C_1 D^2 e\|\tilde{w}_0 - w_*\|^2 + C_2 P\right)^{3/2} \\ &= \frac{1}{\hat{\varepsilon}^{3/2}} \cdot \frac{\left(2D^2 e\|\tilde{w}_0 - w_*\|^2\left(\frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4L^2}{3M}\right) + \left(\frac{2+M}{M} + \frac{5}{6L^2} + \frac{8N}{3ML^2} + \frac{5}{12ML}\right)P\right)^{3/2}}{\left(\frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4L^2}{3M}\right) D^3 \left(\frac{1}{L^2+1}\frac{\mu}{M}\right)^{3/2}} \end{aligned}$$

to guarantee

$$\min_{1\leq t\leq T}[F(\tilde{w}_{t-1}) - F_*] \leq \frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1}) - F_*] \leq \hat{\varepsilon}.$$

Hence, the total complexity (number of individual gradient computations needed to reach $\hat{\varepsilon}$ accuracy) is $\mathcal{O}\left(\frac{n}{\hat{\varepsilon}^{3/2}}\right)$.

If we further assume that $L, M, N > 1$:

$$\begin{aligned} T &= \frac{1}{\hat{\varepsilon}^{3/2}} \cdot \frac{\left(2D^2 e\|\tilde{w}_0 - w_*\|^2\left(\frac{8ML^2}{3} + 14NL^2 + \frac{4L^2}{3}\right) + \left(2 + M + \frac{5M}{6L^2} + \frac{8N}{3L^2} + \frac{5}{12L}\right)P\right)^{3/2}}{\left(\frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4L^2}{3M}\right) D^3 \left(\frac{\mu}{L^2+1}\right)^{3/2}} \\ &\leq \frac{1}{\hat{\varepsilon}^{3/2}} \cdot \left(\mathcal{O}((M+N)L^2)\right)^{3/2} \cdot \mathcal{O}\left(1/L^2\right) \cdot \left(\frac{L^2+1}{\mu}\right)^{3/2} \end{aligned}$$

$$= \mathcal{O}\left(\frac{L^4(M+N)^{3/2}}{\mu^{3/2}} \cdot \frac{1}{\hat{\varepsilon}^{3/2}}\right)$$

and the complexity is $\mathcal{O}\left(\frac{L^4(M+N)^{3/2}}{\mu^{3/2}} \cdot \frac{n}{\hat{\varepsilon}^{3/2}}\right)$. □