A Unified Convergence Analysis for Shuffling-Type Gradient Methods

Lam M. Nguyen Quoc Tran-Dinh Dzung T. Phan Phuong Ha Nguyen
Marten van Dijk

February 18, 2020

Abstract

In this paper, we provide a unified convergence analysis for a class of shuffling-type gradient methods for solving a well-known finite-sum minimization problem commonly used in machine learning. This algorithm covers various variants such as randomized reshuffling, single shuffling, and cyclic/incremental gradient schemes. We consider two different settings: strongly convex and non-convex problems. Our main contribution consists of new non-asymptotic and asymptotic convergence rates for a general class of shuffling-type gradient methods to solve both non-convex and strongly convex problems. While our rate in the non-convex problem is new (i.e., not known yet under standard assumptions), the rate on the strongly convex case matches (up to a constant) the best-known results. However, unlike existing works in this direction, we only use standard assumptions such as smoothness and strong convexity. Finally, we empirically illustrate the effect of learning rates via a non-convex logistic regression and neural network training examples.

1 Introduction

The goal of this paper is to provide a unified analysis for a class of shuffling-type gradient methods to solve the following well-known finite sum minimization problem:

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

where $f(\cdot, i) : \mathbb{R}^d \to \mathbb{R}$ is a given smooth and possibly non-convex function for $i \in [n] := \{1, \cdots, n\}$. This problem covers a wide range of convex and non-convex models in machine learning and statistical

Lam M. Nguyen, IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Email: LamNguyen.MLTD@ibm.com

Quoc Tran-Dinh, Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Email: quoctd@email.unc.edu

Dzung T. Phan, IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Email: phandu@us.ibm.com

Phuong Ha Nguyen, Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. Email: phuongha.ntu@gmail.com

Marten van Dijk, Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. Email: marten.van_dijk@uconn.edu

learning, including, but not limited to, logistic regression, multi-kernel learning, conditional random fields, and neural networks. Especially, it covers *empirical risk minimization* as a special case. Very often, (1) lives in a high dimensional space, and/or it has a large number of components n. Hence, deterministic optimization methods relying on full gradients are usually inefficient to solve (1), see, e.g., [4, 40].

The stochastic gradient descent (SGD) method, originally introduced in [37], has been widely used to solve (1) due to its efficiency in dealing with large-scale problems in big data regimes. In the last fifteen years, there has been a tremendous progress of research in SGD, where numerous stochastic and randomized-based algorithms have been proposed, making it the most active research area in optimization as well as in machine learning. In addition, due to the revolution of deep learning, SGD for non-convex optimization in deep learning also becomes an extremely active research topic nowadays.

SGD is also a method of choice to solve the following common expectation minimization problem (1):

$$\min_{w\in\mathbb{R}^d}\Big\{F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(w;x,y)]\Big\}, \tag{2}$$

where $\mathcal{D}$ is some distribution. Note that (1) is completely deterministic, while (2) is a stochastic optimization formulation. The question is *whether we could take the advantage of the finite-sum structure of problem* (1) *to have a better convergence rate than that of* (2).

To solve (1), at each step, SGD chooses an index $i \in [n]$ uniformly at random, and updates the iterate as $w_{t+1} = w_t - \eta_t\nabla f(w_t; i_t)$ for $t = 0, 1, \cdots, K$, which is up-to n times "component gradient" cheaper than an iteration of a full gradient method with the updates $w_{t+1} = w_t - \eta_t\frac{1}{n}\sum_{i=1}^n f(w_t; i)$, where $\eta_t > 0$ is some learning rate at the t-th iteration. Although SGD was introduced in 1951, its convergence rate was investigated much later [35]. The convergence rate achieved by SGD for solving (2) in the strongly convex case is $\mathcal{O}(\frac{1}{K})$ [29, 35, 32, 34] and for finding a stationary point of (2) in the non-convex case is $\mathcal{O}(\frac{1}{\sqrt{K}})$ [13], where K is the total number of iterations. Since (1) can be viewed as a special case of (2), these rates also apply to (1).

Classical SGD for solving (1) relies on an i.i.d. sampling scheme to select component $f(\cdot, i)$ for updating the iterates w_t. We refer to this method as *standard SGD*. In practice, however, other mechanisms for selecting component $f(\cdot, i)$ such as *randomized shuffling technique* are more desirable to use for implementing stochastic gradient algorithms due to their implementation convenience and efficiency [5, 3, 19]. Unfortunately, convergence analysis for shuffling schemes and cyclic strategy is much more challenging than that of the standard SGD or its variants due to the lack of independence. Hitherto, there has been only a very limited number of theoretical works that show the convergence results of shuffling techniques, and mainly for the strongly convex case [15, 16, 38, 26].

In this paper, we conduct a study on convergence aspects of shuffling-type gradient methods for the general (non-convex) objective function as well as the strongly convex one. We provide a unified convergence analysis framework and apply it to different variants of shuffling schemes in both non-convex and strongly convex settings.

Contributions: Our main contribution consists of:

- We prove $\mathcal{O}(1/T^{2/3})$-convergence rate in epoch for constant step-sizes and $\tilde{\mathcal{O}}(1/T^{2/3})$-convergence rate for diminishing step-sizes for a general shuffling-type gradient method to solve the non-convex problem (1), where T is number of epochs. To the best of our knowledge, these are the first non-asymptotic rates for SGD with shuffling using both constant and diminishing learning rates under standard assumptions.

Table 1: Comparison of results in the strongly convex case: (1) Without Bounded Gradient, (2) Without Bounded Hessian, (3) Non-convex f_i, (4) Diminishing learning rate by epoch, (5) Convergence with probability one

Reference	Complexity	(1)	(2)	(3)	(4)	(5)
[15]	$\mathcal{O}(1/T^2)$	✗	✗	✗	✓	✗
[16]	$\mathcal{O}(1/T^2)$	✗	✗	✗	✗	✗
[26]	$\mathcal{O}(1/T^2)$	✗	✓	✗	✗	✗
This paper	$\mathcal{O}(1/T^2)$	✓	✓	✓	✓	✓

- We establish asymptotic convergence to a stationary point under diminishing learning rate scheme. We achieve the best performance among different variants with the learning rate $\eta_t = \mathcal{O}(\frac{1}{t^{1/3}})$ in both theory and practice. In fact, our learning rate is close to "scheduled" learning rate, i.e., it is a constant during each epoch update and decreases along the epochs.
- We prove $\mathcal{O}(1/T^2)$-convergence rate in epoch of our general shuffling-type gradient scheme for the strongly convex case without any "boundedness" assumptions. Different from existing works, our analysis does not require convexity of each component function.

For more details of comparison between our work and recent state-of-the-arts, see the **Comparison** paragraph.

Related work: Let us briefly review the most related works to our methods in this paper. The random shuffling method has been empirically studied in early works such as [5] and also discussed in [3]. Its cyclic variant, known as an incremental gradient method was proposed in [28], where the convergence analysis was given in [27] for a subgradient variant, and in [14] for gradient variants. These results are only for convex problems. Other incremental gradient variants can be found, e.g. in [9, 10] known as SAGA-based methods.

In [15], the authors showed that if T is large, the randomized shuffling gradient method asymptotically converges as $\mathcal{O}(1/T^2)$-rate under a proper stepsize. However, this rate was only shown for strongly convex problems with bounded gradient/sequence, smoothness, and Lipschitz Hessian. These assumptions all together are very unlikely to hold in practice. Under the same conditions, [16] improved the convergence rate to $\mathcal{O}(1/(nT)^2 + 1/T^3)$ non-asymptotically, but in the regime of $T/\log(T) \geq \mathcal{O}(n)$. Another related work is [26], which achieves $\tilde{\mathcal{O}}(1/(nT^2))$ convergence rates without Lipschitz Hessian when T is above the order of the condition number. Recently, an $\mathcal{O}(1/(nT)^2 + 1/(nT^3))$ lower bound was proved in [38] under the same assumptions as [16].

In [42], the authors replaced i.i.d. sampling scheme by a randomized shuffling strategy and established that variance reduced methods such as SAGA and SVRG still have linear converge for strongly convex problems but using a unusual energy function. Unfortunately, it is unclear how to transform such a criterion to standard convergence criterions such as loss residuals or solution distances. It has also been observed that Gradient Descent and variance reduction methods (e.g., SAG [22], SAGA [11], SVRG [17], and SARAH [33]) for solving (1) under strong strong convexity have linear convergence rates, but they are not efficient in practice due to full gradient evaluations.

In [39], a convergence rate to a neighborhood of the optimal value of an SGD variant without replacement sampling strategy were studied for general convex. Clearly, this type of convergence is different from ours, and requires n to be large to get a suitable bound. If problem (1) is generalized linear and strongly convex, then a faster non-asymptotic rate of $\mathcal{O}(\log(K)/K)$ was achieved. Another recent work is [25] which considers different distributed SGD variants with shuffling for strongly convex, general convex, and

non-convex problems. The authors could only show convergence to a neighborhood of an optimal solution or a stationary point as in [39]. In addition, the convergence rates are much slower than existing results for the strongly convex case, and also slower than ours, while requiring stronger assumptions.

Comparision: To the best of our knowledge, only [25, 24] studied convergence rates of Algorithm 1 for solving non-convex instances of (1). Whereas [25] only proves $\mathcal{O}(1/\sqrt{nT} + \log(n)/n)$-convergence rate to a neighborhood of a stationary point of a randomized reshuffling variant, [24] showed $\mathcal{O}(1/K^{1/2})$ convergence rate under bounded subgradients, weak convexity, and quadratic growth conditions for an incremental subgradient variant. Our convergence rate is $\mathcal{O}(1/T^{2/3})$ in epoch, which corresponds to $\mathcal{O}(n^{2/3}/K^{2/3})$ in the total of iterations, and hence is better than [24] and using different assumptions.

For the strongly convex case, Table 1 shows a comparison of results on the convergence rate for Algorithm 1 to solve (1) under the smooth and strongly convex setting. Here, we compare these methods in terms of required assumptions on F, learning rate, and convergence type. Perhaps, the best convergence rate was proved in [16] but under stronger assumptions than ours, which nearly matches the lower bound proved in [38]. However, [16] only covers certain regimes, while our result is quite general. Further comparison between random shuffling methods and SGD, GD, and other deterministic shuffling schemes for strongly convex problems can be found, e.g. in [16, 38].

2 The Shuffling-Type Gradient Algorithm

Shuffling-type gradient methods are widely used in practice due to their efficiency compared to standard SGD schemes [5]. These methods have been investigated in many recent papers, including [15, 16, 26].

In this paper, we analyze convergence rates for a class of shuffling-type gradient algorithms to solve (1) as described in Algorithm 1.

Algorithm 1 Shuffling-Type Gradient Scheme

Initialization: Choose an initial point $\tilde{w}_0 \in \mathbb{R}^d$;
for $t = 1, 2, \cdots, T$ **do**
  Set $w_0^{(t)} := \tilde{w}_{t-1}$;
  Generate any permutation $\sigma^{(t)}$ of $[n]$;
  for $i = 0, \cdots, n-1$ **do**
    Update $w_{i+1}^{(t)} := w_i^{(t)} - \eta_i^{(t)} \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1))$;
  end for
  Set $\tilde{w}_t := w_n^{(t)}$;
end for
Output: Choose $\widehat{w}_T$ uniformly randomly in $\{\tilde{w}_t\}_{t=1}^T$.

Note that $\sigma^{(t)}(j)$ is the j-th element of $\sigma^{(t)}$. Each outer iteration t of Algorithm 1 can be counted for one epoch. The inner loop updates the iterate sequence $\{w_i^{(t)}\}$ using only one component per iteration as in SGD by shuffling the objective components. Our analysis will be done epoch-wise. Here, the output $\widehat{w}_T$ can uniformly randomly be chosen from $\{\tilde{w}_t\}_{t=1}^T$, or can be chosen based on the best value of the loss F. As discussed in [13], the first option does not incur any additional cost by uniformly randomly generating an index $\widehat{T} \in \{1, \cdots, T\}$ a priori and running Algorithm 1 up to $\widehat{T}$ iterations instead of T. Depending on the choice of $\sigma^{(t)}$ we obtain different variants, especially the following methods:

- If $\sigma^{(t)} = \{1, 2, \cdots, n\}$ or some fixed permutation of $\{1, 2, \cdots, n\}$ for all epochs t, then Algorithm 1 is equivalent to a cyclic gradient method. This method can also be viewed as the incremental gradient scheme in [28].
- If $\sigma^{(t)}$ is randomly generated one time and repeatedly used at each iteration t, then Algorithm 1 becomes a single shuffling variant [38].
- If $\sigma^{(t)}$ is randomly generated at each epoch t, then Algorithm 1 reduces to a randomized reshuffling scheme, broadly used in practice [19].

These schemes have been studied, e.g. in [15, 16, 26], but their convergence analysis has mainly been investigated for the strongly convex case, and often under a strong set of assumptions.

Remark 1 (**Types of guarantee**). *Since we can choose permutations $\sigma^{(t)}$ either deterministically or randomly, our convergence and complexity bounds in the sequel will hold either deterministically or with probability* 1 *(w.p.1), respectively. Without loss of generality, we write these results in the context of* ***w.p.1****.*

3 Basic Assumptions and Mathematical Tools

Our analysis relies on the following basic assumptions. We first require a bounded below assumption on F.

Assumption 1. $F_* := \inf_{w \in \mathbb{R}^d} F(w) > -\infty$.

The L-smoothness assumption is fundamental in first-order methods, including SGD, and is expressed as follows:

Assumption 2 (L-smoothness). *$f(\cdot; i)$ is L-smooth for $\forall i \in [n]$, i.e., there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$, it holds that*

$$\|\nabla f(w; i) - \nabla f(w'; i)\| \leq L\|w - w'\|. \tag{3}$$

Assumption 2 implies that the objective function F is also L-smooth. Moreover, as shown in [30], for any $w, w' \in \mathbb{R}^d$, one has

$$F(w) \leq F(w') + \langle \nabla F(w'), w - w' \rangle + \frac{L}{2}\|w - w'\|^2. \tag{4}$$

We refer to Assumptions 1 and 2 as our basic assumptions which are required throughout the paper.

For the convenience of our analysis, we will consider the case where the learning rate within a single epoch is fixed. More specifically, at epoch t, let $\eta_t > 0$ be given, we consider the following form of learning rate in Algorithm 1:

$$\eta_i^{(t)} := \frac{\eta_t}{n}. \tag{5}$$

Then, we have the following update after each epoch:

$$w_n^{(t)} = w_0^{(t)} - \frac{\eta_t}{n} \sum_{i=0}^{n-1} \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1)). \tag{6}$$

The following lemmas provide key tools for our convergence analysis in the sequel (the proof is in Appendix).

Lemma 1. *Let $\{Y_t\}_{t\geq 1}$ be a nonnegative sequence in $\mathbb{R}$ and q be a positive number. For some $\alpha > 0, \beta \geq 0, \rho > 0$ and $D > 0$, let*

$$Y_{t+1} \leq (1 - \rho \cdot \eta_t)Y_t + D \cdot \eta_t^{q+1}, \tag{7}$$

with $\eta_t := \frac{\alpha}{(t+\beta)}$. Suppose that $\gamma > 0$ and $\lambda \geq q$ are given such that $Y_1 \leq \frac{\gamma}{(1+\beta)^q}$ and $\gamma(\rho\alpha - \lambda) \geq D\alpha^{q+1}$. Then

$$Y_t \leq \frac{\gamma}{(t+\beta)^q}. \tag{8}$$

Lemma 2. *Let $\{Y_t\}_{t\geq 1}$ and $\{Z_t\}_{t\geq 1}$ be two nonnegative sequences in $\mathbb{R}$ and m and q be two positive numbers such that $q > m$. For some $\rho > 0$ and $D > 0$, assume that*

$$Y_{t+1} \leq Y_t - \rho\eta_t^m \cdot Z_t + \eta_t^q \cdot D \tag{9}$$

where $\eta_t := \frac{\gamma}{(t+\beta)^\alpha}$ for some $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ such that $\alpha m \leq \frac{1}{2}$. Suppose that $Y_t \leq C + H\ln(t+\theta)$ for some $C > 0$, $H \geq 0$, $\theta > 0$, and $1 + \theta - \beta > (1 - \alpha m)e^{\frac{\alpha m}{1-\alpha m}}$ (where e is the natural number), for all $t \geq 1$. Then

$$\frac{1}{T}\sum_{t=1}^{T} Z_t \leq \frac{1}{T}\Big[\frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m} + \frac{C(T-1+\beta)^{\alpha m}}{2\rho\alpha m\gamma^m} + \frac{H(T-1+\beta)^{\alpha m}\ln(T+\theta)}{2\rho\alpha m\gamma^m}\Big] + \frac{D\gamma^{q-m}}{\rho}\cdot\frac{A(T)}{T}, \tag{10}$$

where

$$A(T) := \begin{cases} \ln(T+\beta) - \ln(\beta), & \text{if } \alpha(q-m) = 1 \\ \frac{(T+\beta)^{1-\alpha(q-m)}}{1-\alpha(q-m)}, & \text{otherwise.} \end{cases}$$

4 Convergence Analysis for Non-Convex Case

We provide convergence analysis for Algorithm 1 to solve non-convex smooth problem (1). The detailed proofs of our results are given in Appendix.

Assume that (1) satisfies the following assumption.

Assumption 3. *There exists a constant $G > 0$ such that $\forall w \in \mathbb{R}^d$, we have*

$$\|\nabla f(w; i)\|^2 \leq G^2, \quad \forall i \in [n]. \tag{11}$$

We emphasize that this assumption may not be appropriate for strongly convex problems, but it is often used in non-convex problems.

4.1 The General Case

Now, we state our first results on the non-convex case.

Theorem 1. *Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n} = \frac{\eta}{n}$, with $0 < \eta \leq \frac{1}{L}$. Then, under Assumptions 1, 2, and 3, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2}{T\eta}\big[F(\tilde{w}_0) - F_*\big] + \frac{L^2G^2}{3}\cdot\eta^2.$$

Assuming that L and G are known. Then, we can choose the following learning rate to get a concrete bound.

Corollary 1. *Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 and $\widehat{w}_T$ be its output. For a given tolerance $\epsilon > 0$, under the same conditions as in Theorem 1, if we choose the constant learning rate $\eta := \frac{\sqrt{\epsilon}}{LG}$, then to guarantee*

$$\mathbb{E}\left[\|\nabla F(\widehat{w}_T)\|^2\right] = \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \epsilon,$$

for (1)*, it requires $T := \left\lfloor 3LG[F(\tilde{w}_0) - F_*] \cdot \frac{1}{\epsilon^{3/2}} \right\rfloor$ outer iterations. As a result, the total number of gradient evaluations is at most $\mathcal{T}_w := \left\lfloor 3LG[F(\tilde{w}_0) - F_*] \cdot \frac{n}{\epsilon^{3/2}} \right\rfloor$.*

Remark 1. To have the same guarantee, the total complexity of a standard SGD is $\mathcal{O}(\frac{L_F\sigma^2}{\epsilon^2})$ under the condition

$$\mathbb{E}\left[\|\nabla f(w;\xi) - \nabla F(w)\|^2\right] \leq \sigma^2,$$

and L_F-smoothness when solving the following stochastic optimization problem

$$\min_{w\in\mathbb{R}^d}\left\{F(w) = \mathbb{E}_{\xi\sim D}[f(w;\xi)]\right\},$$

for a stochastic function $f : \mathbb{R}^d \times D \to \mathbb{R}$. Note that the standard SGD only requires F to be L_F-smooth while we impose the smoothness on individual realizations. Therefore, L_F and L may be different [13]. However, for a rough comparison, if $n < \mathcal{O}\left(\frac{L_F\sigma^2}{LG}\cdot\frac{1}{\epsilon^{1/2}}\right)$, then Algorithm 1 seems to have advantages over the standard SGD in the non-convex setting. From this point of view, it seems that Algorithm 1 is inefficient compared to SGD when n is large and the accuracy ϵ is low or moderate. However, we believe that Algorithm 1 allows more flexible strategy to choose $f(\cdot, i)$ rather than that i.i.d. sampling and our convergence analysis may be loose in that it does not take into account a tight dependence on n in our complexity bounds. We also note that our convergence guarantee is completely different from [25] as mentioned above. Nevertheless, Assumptions 2 and 3 hold for various applications in machine learning.

Corollary 2. *Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 and $\widehat{w}_T$ be its output. Under the same conditions as in Theorem 1, if we choose the constant learning rate $\eta = \frac{\gamma}{T^{1/3}}$ for some $\gamma > 0$, then*

$$\mathbb{E}\left[\|\nabla F(\widehat{w}_T)\|^2\right] = \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{R_0}{T^{2/3}},$$

where $R_0 := \frac{2[F(\tilde{w}_0)-F_*]}{\gamma} + \frac{\gamma^2L^2G^2}{3}$, and $\widehat{w}_T$ is the output.

Note that the total number of iterations is $K := nT$. Hence, if we express (2) in terms of K, then we have $\mathbb{E}\left[\|\nabla F(\widehat{w}_K)\|^2\right] \leq \frac{n^{2/3}R^0}{K^{2/3}}$. Clearly, this rate matches the recent results in [41, 8] up to a constant factor, but it is unclear to compare how the methods in those papers depend on n.

For general choices of diminishing learning rate, the following theorem characterizes an asymptotic convergence.

Theorem 2. *Suppose that Assumptions 1, 2, and 3 hold. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with diminishing learning rate $\eta_i^{(t)} = \frac{\eta_t}{n}$ such that*

$$\sum_{t=1}^{\infty} \eta_t = \infty \ \textit{and} \ \ \sum_{t=1}^{\infty} \eta_t^3 < \infty.$$

Then, w.p.1. (i.e. almost surely), we have

$$\liminf_{t\to\infty} \|\nabla F(\tilde{w}_{t-1})\|^2 = 0.$$

Now, let us vary the learning rate η_t to see how it affects the convergence rate bounds as stated in the following theorem.

Theorem 3. *Suppose that Assumptions 1, 2, and 3 hold. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n}$, where $\eta_t := \frac{\gamma}{(t+\beta)^\alpha} \leq \frac{1}{L}$, for some $\gamma > 0$, $\beta > 0$, and $\frac{1}{3} < \alpha < 1$. Let us define $C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3 L^2 G^2}{6(3\alpha-1)\beta^{3\alpha-1}} > 0$ be a given constant. Then, the following statements hold:*

- *If $\alpha = \frac{1}{2}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2(1+\beta)^{1/2}[F(\tilde{w}_0) - F_*]}{\gamma} \cdot \frac{1}{T} + \frac{2C}{\gamma}\left(\frac{(T-1+\beta)^{1/2}}{T}\right) + \frac{L^2G^2\gamma^2}{3}\left(\frac{\ln(T+\beta) - \ln(\beta)}{T}\right).$$

- *If $\alpha \neq \frac{1}{2}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2(1+\beta)^{\alpha}[F(\tilde{w}_0) - F_*]}{\gamma} \cdot \frac{1}{T} + \frac{C}{\alpha\gamma}\left(\frac{(T-1+\beta)^{\alpha}}{T}\right) + \frac{L^2G^2\gamma^2}{3(1-2\alpha)}\left(\frac{(T+\beta)^{1-2\alpha}}{T}\right).$$

Remark 2. In Theorem 3, if we choose $\alpha = \frac{1}{3} + \delta$ for some $0 < \delta < \frac{1}{6}$, then we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\tilde{w}_{t-1})\|^2 &\leq \frac{2(1+\beta)^{\frac{1}{3}+\delta}[F(\tilde{w}_0) - F_*]}{\gamma} \cdot \frac{1}{T} \\ &\quad + \frac{C}{\gamma(\frac{1}{3}+\delta)}\left(\frac{(T-1+\beta)^{\frac{1}{3}+\delta}}{T}\right) + \frac{L^2G^2\gamma^2}{1-6\delta}\left(\frac{(T+\beta)^{\frac{1}{3}-2\delta}}{T}\right) \\ &= \mathcal{O}\left(\frac{1}{T^{\frac{2}{3}-\delta}}\right),\end{aligned}$$

where $C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3 L^2 G^2}{18\delta\beta^{3\delta}} > 0$. Notice that the convergence rate for regular SGD is $\mathcal{O}\left(\frac{1}{T^{1/2}}\right)$.

For the special case $\alpha = \frac{1}{3}$, we have the following result.

Theorem 4. *Suppose that Assumptions 1, 2, and 3 hold. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n}$, where $\eta_t := \frac{\gamma}{(t+\beta)^{1/3}} \leq \frac{1}{L}$, for some $\gamma > 0$, and $\beta > 0$. Let us define*

$C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3 L^2 G^2}{6(1+\beta)} > 0$ *be a given constant. Then,*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 &\leq \frac{2(1+\beta)^{1/3}[F(\tilde{w}_0) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{3C}{\gamma}\left(\frac{(T-1+\beta)^{1/3}}{T}\right) \\ &\quad + \frac{\gamma^2 L^2 G^2}{2}\left(\frac{(T-1+\beta)^{1/3}\ln(T+1+\beta)}{T}\right) + L^2G^2\gamma^2\left(\frac{(T+\beta)^{1/3}}{T}\right) \\ &= \mathcal{O}\left(\frac{\ln(T)}{T^{2/3}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{T^{2/3}}\right).\end{aligned}$$

Remark 3. The choice of the learning rates in Theorems 3 and 4 is not necessarily dependent on the smoothness constant L. Since η_t is diminishing and L is finite, after some certain epoch, it is able to satisfy $\eta_t = \frac{\gamma}{(t+\beta)^\alpha} \leq \frac{1}{L}$. Theoretical results still hold by shifting the iteration indices. Therefore, the choices of γ, β, and α are quite flexible. However, if we choose these properly from the beginning, then this will give an advantage with respect to convergence in practice.

4.2 Convergence Under Gradient Dominance

The convergence rate of Algorithm 1 can be improved if the following gradient dominance condition holds.

Assumption 4. *F is said to be τ-gradient dominant if there exists a constant $\tau > 0$ such that $\forall w \in \mathbb{R}^d$, it holds that*

$$F(w) - F_* \leq \tau\|\nabla F(w)\|^2, \tag{12}$$

where F_* is the global minimum value of F on $\mathbb{R}^d$.

This assumption is well-known in literature (see e.g. [36, 31, 18]) and is weaker than strong convexity assumption. We can observe that every stationary point of the τ-gradient dominant function F is a global minimizer. However, such a function F does not necessarily need to be convex.

Theorem 5. *Suppose that Assumptions 1, 2, 3, and 4 hold for* (1)*. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} := \frac{\eta_t}{n}$ for solving* (1)*. Let η_t be updated as $\eta_t := \frac{\alpha}{t+\beta}$ for some $\alpha > 0$ and $\beta \geq 0$. Assume further that $\gamma > 0$ and $\lambda \geq 2$ are two constants such that $F(\tilde{w}_0) - F_* \leq \frac{\gamma}{(1+\beta)^2}$ and $\gamma(\alpha - 2\tau\lambda) \geq \frac{L^2G^2\tau}{3}\alpha^3$. Then, we have*

$$F(\tilde{w}_t) - F_* \leq \frac{\gamma}{(t+1+\beta)^2}, \quad \forall t \geq 0.$$

The following gives a concrete choice of parameters.

Corollary 3. *Suppose that conditions of Theorem 5 hold. Then, for any $\beta \geq 0$, if we choose γ as*

$$\gamma := \max\left\{\frac{125}{3}L^2G^2\tau^3, [F(\tilde{w}_0) - F_*](1+\beta)^2\right\}, \tag{13}$$

then, using the learning rate $\eta_t := \frac{5\tau}{t+\beta}$ in Algorithm 1, we have

$$F(\tilde{w}_t) - F_* \leq \frac{\gamma}{(t+1+\beta)^2},$$

where F_* is the global optimal value of (1)*.*

Note that [16] also provided $\mathcal{O}(1/T^2)$ convergence rate but under stronger assumptions, i.e., Lipschitz Hessian and $T/\log(T) > \mathcal{O}(n)$.

5 Convergence Analysis for Strong Convexity

We first analyze convergence under strong convexity assumption, and then move to the general convex case.

Let us recall the following assumptions imposed on (1).

Assumption 5 (μ-strong convexity)**.** *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, i.e., there exists a constant $\mu > 0$ such that $\forall w, w' \in \mathbb{R}^d$, it holds that*

$$F(w) \geq F(w') + \langle \nabla F(w'), w - w' \rangle + \frac{\mu}{2}\|w - w'\|^2. \tag{14}$$

It is well-known from the literature [30, 6] that Assumption 5 implies the existence and uniqueness of the optimal solution w_* of (1), and

$$F(w) - F(w_*) \leq \frac{1}{2\mu}\|\nabla F(w)\|^2 \ , \ \forall w \in \mathbb{R}^d. \tag{15}$$

It is important to note that Assumption 5 only requires F to be strongly convex, but some components $f(\cdot, i)$ can be non-convex.

Definition 1. *Define the following quantities*

$$N_i := \|\nabla f(w_*; i)\|^2 \ \ \forall i \in [n], \ \ \textit{and} \ \ N := \max_{i \in [n]} N_i. \tag{16}$$

Clearly, since n is finite, both N_i and N are finite for $i \in [n]$.

We prove the following result for the strongly convex case.

Theorem 6. *Assume that Assumptions 2 and 5 hold. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} := \frac{\eta_t}{n}$ to solve* (1)*. Let $\alpha > 0$ and $\beta > 0$ be chosen such that $\alpha = \frac{\mu}{2L^2}\beta$, and $\eta_t := \frac{\alpha}{t+\beta}$. Suppose that $\gamma > 0$ and $\lambda \geq 2$ are two constants such that $F(\tilde{w}_0) - F(w_*) \leq \frac{\gamma}{(1+\beta)^2}$ and $\gamma(\mu\alpha - 3\lambda) \geq 3(\mu^2 + L^2)N\alpha^3$ with N given in* (16)*. Then, we have*

$$F(\tilde{w}_t) - F(w_*) \leq \frac{\gamma}{(t + 1 + \beta)^2}, \quad \forall t \geq 0.$$

The following gives a specific choice of parameters.

Corollary 4. *Suppose that the conditions in Theorem 6 hold. Let α, β, and γ be chosen as*

$$\begin{cases} \alpha & := \frac{12L^2+\mu^2}{2L^2\mu}, \\ \beta & := \frac{12L^2+\mu^2}{\mu^2}, \\ C & := \frac{3(\mu^2+L^2)(12L^2+\mu^2)^3 N}{4L^4\mu^5} \\ \gamma & := \max\left\{C, (1+\beta)^2\left[F(\tilde{w}_0) - F(w_*)\right]\right\}. \end{cases} \tag{17}$$

Then, we have

$$F(\tilde{w}_t) - F(w_*) \leq \frac{\gamma}{(t+1+\beta)^2}.$$

Given a tolerance $\epsilon > 0$, to guarantee $F(\tilde{w}_t) - F(w_*) \leq \epsilon$, the total number of gradient evaluations is at most $\mathcal{O}(\frac{n}{\sqrt{\epsilon}})$.

Since the total number of iterations is $K := nT$, if we write the convergence rates in terms of K, then we have $F(\tilde{w}_T) - F_* \leq \mathcal{O}\left(\frac{n^2}{K^2}\right)$, which is worse than the one in [16]. However, as mentioned, our assumptions are much weaker than those in [16]. In addition, Algorithm 1 covers much broader class of algorithms compared to [16].

6 Numerical Experiments

We provide two representative numerical experiments to show the benefit of the learning rate $\eta_t = \frac{\gamma}{t^{1/3}}$ for non-convex problems. This choice corresponds to $\alpha = \frac{1}{3}$ for the best convergence performance as given in Theorem 4. We experimented with 10 runs and reported the average results

6.1 Non-Convex Logistic Regression Example

We consider the following binary classification problem with non-convex loss widely used in the literature:

$$\min_{w\in\mathbb{R}^d} \Big\{ F(w) = \frac{1}{n}\sum_{i=1}^{n} \Big[\log(1+\exp(-y_i x_i^T w)) + \frac{\lambda}{2}\sum_{j=1}^{d} \frac{w_j^2}{1+w_j^2} \Big] \Big\}, \tag{18}$$

where $\{(x_i, y_i)\}_{i=1}^n$ is a set of training examples, and $\lambda > 0$ is a given regularization parameter.

We conducted experiments to demonstrate the advantage in performance of Algorithm 1 on the classification data set *w8a* ($n = 49,749$ training data) from LIBSVM [7]. Since we only care about the non-convexity of each f_i instead of statistical properties, we simply choose $\lambda = 0.01$, but other values also work.

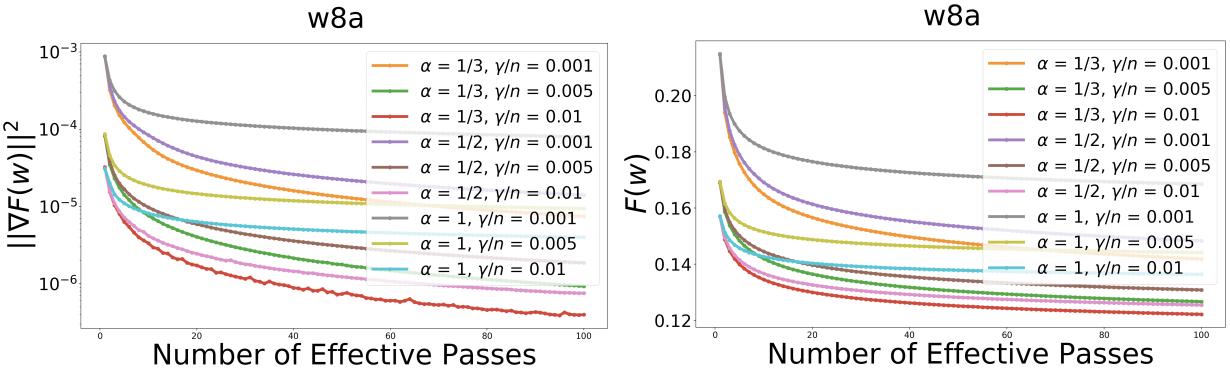

Figure 1: A comparison of $F(w)$ and $\|\nabla F(w)\|^2$ (starting from the 2nd epoch) for the non-convex logistic regression problem (18) on different values of α and γ/n using the *w8a* dataset.

We apply Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n}$ to solve (18), where $\eta_t = \frac{\gamma}{(t+\beta)^\alpha}$ and $\sigma^{(t)}$ is generated randomly to obtain an SGD variant with randomized reshuffling strategy. Figure 1 shows the comparisons of the

squared norm $\|\nabla F(w)\|^2$ of gradient and the value $F(w)$ of the objective function on different values of $\alpha = \{1/3, 1/2, 1\}$ and $\gamma/n = \{0.001, 0.005, 0.01\}$, respectively, on the data set *w8a*. As predicted by our theory, the choice of $\alpha = 1/3$ generally performs better than others. Combining with $\gamma/n = 0.01$, this variant perform best on such a given dataset. Note that since we only plot w.r.t. epochs $t \geq 1$, the initial values in these plots are different.

If we use another dataset, *ijcnn1* ($n = 91,701$) from LIBSVM, then under the same setting of Algorithm 1, we obtain the result as in Figures 2.

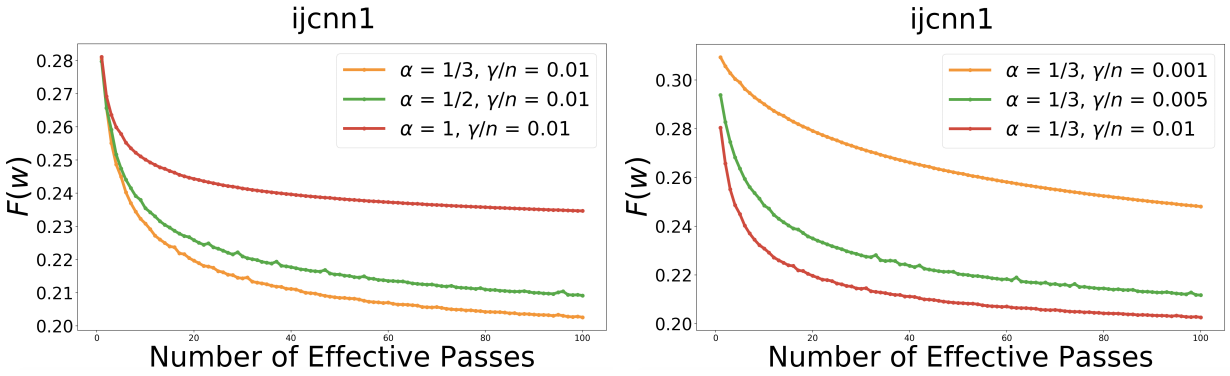

Figure 2: A comparison on $F(w)$ (starting from the 2nd epoch) of Algorithm 1 for solving (18) using different values of α and γ/n and the *ijcnn1* data set.

The first plot of this figure again shows that $\alpha = 1/3$ gives the best performance. Once, we fix $\alpha = 1/3$ and using different ratios γ/n, then as showed in the second plot, $\gamma/n = 0.01$ seems to work best.

6.2 Fully Connected Neural Network Example

Our second example is to test Algorithm 1 on a neural network. We perform this test on a neural network with two fully connected hidden layers of 300 and 100 nodes, followed by a fully connected output layer which feeds into the soft-max cross entropy loss. We use Tensorflow [1] to train this model on the well-known MNIST data set with $n = 60,000$ [23]. This data set has 10 classes corresponding to 10 soft-max output nodes in the network, and are normalized to interval $[0, 1]$ as a simple data pre-processing.

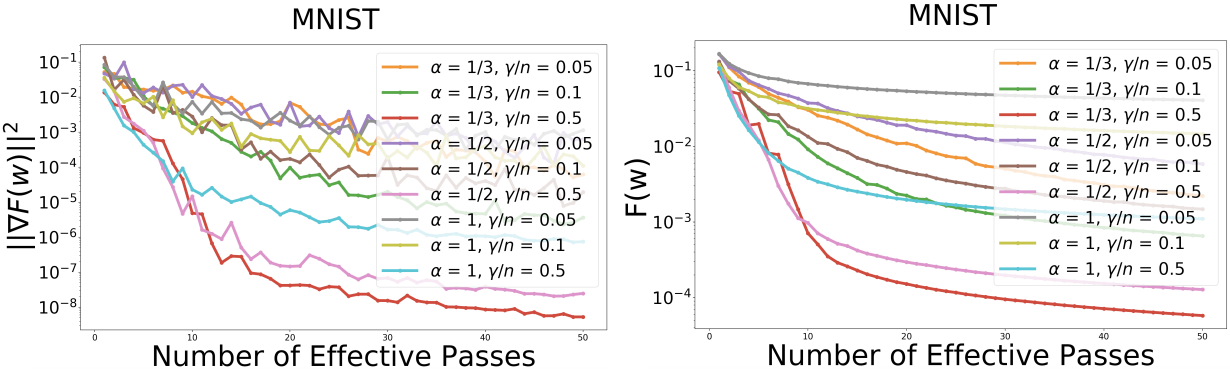

Figure 3: A comparison of $F(w)$ and $\|\nabla F(w)\|^2$ (from 2nd epoch) for the neural network training problem produced by Algorithm 1 on different values of α and γ/n using the *MNIST* dataset.

We apply Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n}$, where $\eta_t = \frac{\gamma}{(t+\beta)^\alpha}$ to solve this training problem. We repeatedly run the algorithm 10 times and report the average results in Figure 3. These plots compare the squared norm $\|\nabla F(w)\|^2$ of gradient and the value $(F(w))$ of the objective function on different values of $\alpha = \{1/3, 1/2, 1\}$ and $\gamma/n = \{0.05, 0.1, 0.5\}$, respectively, on the data set *MNIST*.

Now, we conduct another test on the CIFAR-10 dataset ($n = 50,000$ samples and 10 classes) [21]. We run Algorithm 1 with the same setting as in the previous test, then the results are plotted in Figure 4.

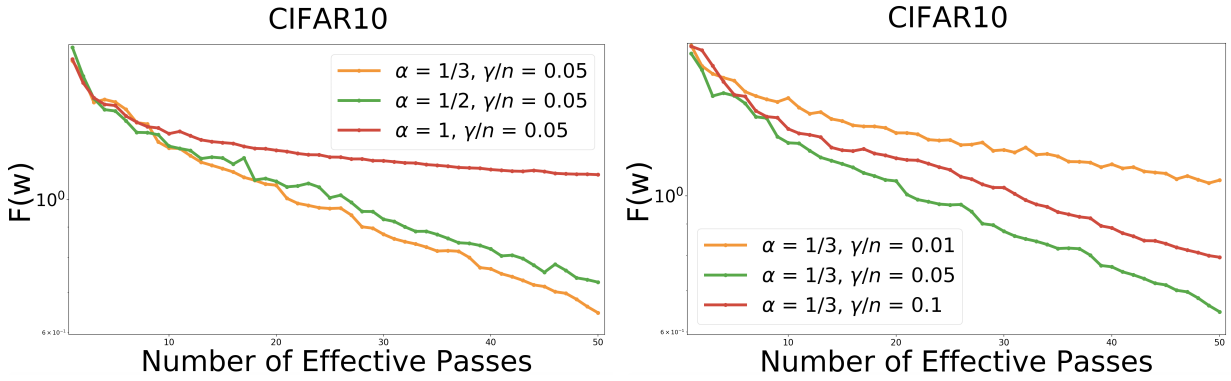

Figure 4: A comparisons on $F(w)$ of Algorithm 1 for the neural network training problem using different values of α and γ/n on the *CIFAR10* dataset.

We observe again from Figure 4 that $\alpha = 1/3$ works best when fixing $\gamma/n = 0.05$. Once we fix $\alpha = 1/3$ and test on γ/n, the ratio $\gamma/n = 0.1$ gives the best performance.

As empirically observed in [5] that randomized shuffling gradient methods often perform better than SGD. This behavior has been also observed in deep learning and other machine learning training tasks. In this paper, we only provide some evidence on the choice of learning rate guided by our theoretical results using only a randomized reshuffling strategy. We omit an intensive comparison between our methods and SGD as well as other shuffling strategies due to space limit.

7 Conclusion

We have conducted an intensive convergence analysis for a class of shuffling-type gradient methods for solving a finite-sum minimization problem. We have proved a non-asymptotic $\mathcal{O}(1/T^{2/3})$ convergence rate for our algorithm for solving non-convex problems under standard assumptions. We have also considered this rate in both constant and diminishing learning rates, and investigated an asymptotic convergence. To the best of our knowledge, this is the first work showing non-asymptotic $\mathcal{O}(1/T^{2/3})$ convergence rate for a wide class of shuffling-type gradient methods in non-convex settings. In the strongly convex setting, we have achieved the same order $\mathcal{O}(1/T^2)$ convergence rate under just strong convexity and smoothness, which is weaker than known result up to a constant factor of n, but our result uses much weaker assumptions than state-of-the-arts. We believe that our results would provide a unified analysis for shuffling-type algorithms using both randomized and deterministic sampling strategies, where it covers the well-known incremental gradient scheme as a special case. Our numerical examples on two non-convex problems have greatly verified our theoretical results. We believe that our analysis framework can be extended to study non-asymptotic convergence rates of SGD and its variants, including adaptive SGD ones such as Adam [20] and AdaGrad [12] under shuffling strategies.

Acknowledgements: The authors would like to thank Trang H. Tran for her valuable comments on some technical proofs. The work of Q. Tran-Dinh has partly been supported by the National Science Foundation (NSF), grant no. DMS-1619884, the Office of Naval Research (ONR), grant no. N00014-20-1-2088 (2020-2023), and The Statistical and Applied Mathematical Sciences Institute (SAMSI).

References

[1] 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. Software available from tensorflow.org.

[2] Dimitri P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey, 2015.

[3] L. Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

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

[5] Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. 2009.

[6] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[7] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[8] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. *arxiv:1905.10018*, 2019.

[9] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1646–1654, 2014.

[10] A. Defazio, T. Caetano, and J. Domke. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133, 2014.

[11] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[12] 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.

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

[14] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Convergence rate of incremental gradient and Newton methods. *arXiv preprint arXiv:1510.08562*, 2015.

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

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

[17] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[18] 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.

[19] Hiroyuki Kasai. SGDLibrary: A MATLAB library for stochastic optimization algorithms. *Journal of Machine Learning Research*, 18(215):1–5, 2018.

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

[21] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[22] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012.

[23] 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.

[24] X. Li, Z. Zhu, A. So, and J. D. Lee. Incremental methods for weakly convex optimization. *arXiv preprint arXiv:1907.11687*, 2019.

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

[26] 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.

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

[28] A. Nedic and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM J. on Optim.*, 12(1):109–138, 2001.

[29] 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.

[30] Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004.

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

[32] Lam Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. SGD and Hogwild! convergence without the bounded gradients assumption. In *Proceedings of the 35th International Conference on Machine Learning-Volume 80*, pages 3747–3755, 2018.

[33] 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.

[34] Lam M. Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takáč, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019.

[35] B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

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

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

[38] Itay Safran and Ohad Shamir. How good is sgd with random shuffling? *arXiv preprint arXiv:1908.00045*, 2018.

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

[40] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. MIT Press, 2012.

[41] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for stochastic composite nonconvex optimization. *Preprint: UNC-STOR 07.10.2019*, 2019.

[42] B. Ying, K. Yuan, and A. H. Sayed. Convergence of variance-reduced stochastic learning under random reshuffling. *arXiv preprint arXiv:1708.01383*, 2(3):6, 2017.

Appendix

A Some Key Lemmas for Convergence Analysis

Appendix proves some key lemmas that will be used for our convergence analysis of the entire paper.

A.1 General Frameworks

Let us first prove two general elementary lemmas used for our convergence analysis in the sequel.

Lemma 1. *Let $\{Y_t\}_{t\geq 1}$ be a nonnegative sequence in $\mathbb{R}$ and q be a positive number. For some $\alpha > 0, \beta \geq 0, \rho > 0$ and $D > 0$, let*

$$Y_{t+1} \leq (1 - \rho \cdot \eta_t)Y_t + D \cdot \eta_t^{q+1}, \tag{19}$$

with $\eta_t := \frac{\alpha}{(t+\beta)}$. Suppose that $\gamma > 0$ and $\lambda \geq q$ are given such that $Y_1 \leq \frac{\gamma}{(1+\beta)^q}$ and $\gamma(\rho\alpha - \lambda) \geq D\alpha^{q+1}$. Then

$$Y_t \leq \frac{\gamma}{(t+\beta)^q}. \tag{20}$$

Proof. First, let us consider $\psi(\tau) := 1 + \lambda\tau(1+\tau)^q - (1+\tau)^q$ with $\tau \geq 0$. Clearly $\psi(0) = 0$, and $\psi'(\tau) = (1+\tau)^{q-1}\left[\lambda - q + \lambda\tau(1+q)\right] \geq 0$ for all $\tau \geq 0$ provided that $\lambda \geq q$. As a consequence, $\psi(\tau) \geq \psi(0) = 0$ for all $\tau \geq 0$. This fact leads to $(1+\tau)^q \leq \frac{1}{1-\tau\lambda}$ for $0 \leq \tau < \frac{1}{\lambda}$. Therefore, by using $\tau := \frac{1}{k+\beta} < \frac{1}{\lambda}$ into the last inequality, we obtain $(1 + \frac{1}{k+\beta})^q \leq \frac{k+\beta}{k+\beta-\lambda}$, which is equivalent to $(k+\beta)^{q+1} \geq (k+\beta+1)^q(k+\beta-\lambda)$ for $\lambda \geq q$ and $\lambda < k+\beta$. For $\lambda \geq k+\beta$, it is trivial that $(k+\beta)^{q+1} \geq (k+\beta+1)^q(k+\beta-\lambda)$ for any $\beta \geq 0$. Therefore, we have $(k+\beta)^{q+1} \geq (k+\beta+1)^q(k+\beta-\lambda)$ for $\lambda \geq q$.

Now, we prove (20) by induction. For $t = 1$, (20) becomes $Y_1 \leq \frac{\gamma}{(1+\beta)^q}$ which is exactly our initial condition. Suppose that (20) holds for all $t \leq k$, that is $Y_t \leq \frac{\gamma}{(t+\beta)^q}$. Now, we show that it holds for $t := k+1$. Indeed, from (19) and $\eta_k := \frac{\alpha}{(k+\beta)}$, we have

$$\begin{aligned}
Y_{k+1} &\overset{(19)}{\leq} (1 - \rho \cdot \eta_k)Y_k + \eta_k^{q+1} \cdot D \\
&\overset{(20)}{\leq} \left(1 - \frac{\rho\alpha}{(k+\beta)}\right)\frac{\gamma}{(k+\beta)^q} + \frac{D\alpha^{q+1}}{(k+\beta)^{q+1}} \\
&= \left(\frac{k+\beta-\rho\alpha}{(k+\beta)^{q+1}}\right)\gamma + \frac{D\alpha^{q+1}}{(k+\beta)^{q+1}} \\
&= \left(\frac{k+\beta-\lambda}{(k+\beta)^{q+1}}\right)\gamma - \left(\frac{\rho\alpha-\lambda}{(k+\beta)^{q+1}}\right)\gamma + \frac{D\alpha^{q+1}}{(k+\beta)^{q+1}} \\
&\leq \frac{\gamma}{(k+1+\beta)^q},
\end{aligned}$$

where the last inequality follows from the initial condition $\gamma(\rho\alpha - \lambda) \geq D\alpha^{q+1}$, which is equivalent to $-\gamma(\rho\alpha - \lambda) + D\alpha^{q+1} \leq 0$, and the fact that $(k+\beta)^{q+1} \geq (k+\beta+1)^q(k+\beta-\lambda)$ for $\lambda \geq q$ as proved above. Finally, by the induction argument, we conclude that (20) holds for all $t \geq 1$. □

We first prove the following elementary results in Lemma 3 which will be used in the proof of Lemma 2.

Lemma 3. *The following statements hold:*

(a) *For any $0 \leq \nu \leq \frac{1}{2}$ and $s > 0$, we have*

$$(s+1)^\nu - s^\nu \leq \frac{1}{2s^{1-\nu}}. \tag{21}$$

(b) *For any $c > 0$, $\theta > 0$, $\beta > 0$, and $1 + \theta - \beta > ce^{\frac{1-c}{c}}$, the function $f(t) := \frac{\ln(t+1+\theta)}{(t+\beta)^c}$ is monotonically decreasing on $[0, +\infty)$.*

(c) *Suppose that f is a real-valued and monotonically decreasing function on $[a, +\infty)$ such that $f(x) \geq 0$ for all $x \in [a, +\infty)$. Then, for any choice of $N \geq a$ and $t \geq N$, we have*

$$\sum_{i=N+1}^{t} f(i) \leq \int_N^t f(x)dx. \tag{22}$$

Proof. (a) If $2\nu \leq 1$, then $\left(\frac{s+1}{s}\right)^{1-2\nu} \geq 1$, which is equivalent to $\frac{s+1}{s} \geq \left(\frac{s+1}{s}\right)^{2\nu}$. This leads to $(s+1)^\nu s^{1-\nu} - s^\nu (s+1)^{1-\nu} \leq 0$. Hence, we have

$$(s+1)^\nu - s^\nu = \frac{1 + (s+1)^\nu s^{1-\nu} - s^\nu (s+1)^{1-\nu}}{(s+1)^{1-\nu} + s^{1-\nu}} \leq \frac{1}{(s+1)^{1-\nu} + s^{1-\nu}} \leq \frac{1}{2s^{1-\nu}},$$

which proves (21).

(b) Our goal is to show that $f'(t) < 0$ for all $t \geq 0$. We can directly compute $f'(t)$ as

$$f'(t) = (t+\beta)^{-c-1}\left[1 - \frac{1+\theta-\beta}{t+1+\theta} - c \cdot \ln(t+1+\theta)\right] = (t+\beta)^{-c-1} g(t+1+\theta),$$

where $g(\tau) := 1 - \frac{1+\theta-\beta}{\tau} - c\ln(\tau)$. We consider $g(\tau)$ for $\tau > 0$. It is obvious to show that $g'(\tau) = \frac{1+\theta-\beta}{\tau^2} - \frac{c}{\tau} = \frac{(1+\theta-\beta)-c\tau}{\tau^2}$ and $g''(\tau) = \frac{c\tau - 2(1+\theta-\beta)}{\tau^3}$. Hence, $g'(\tau) = 0$ at $\tau^* := \frac{1+\theta-\beta}{c}$ and $g''(\tau^*) = -\frac{c^3}{(1+\theta-\beta)^2} < 0$. Consequently, g attains its maximum at τ^*, and by the condition that $1 + \theta - \beta > ce^{\frac{1-c}{c}}$, we have

$$g(\tau) \leq g(\tau^*) = 1 - c - c\ln\left(\frac{1+\theta-\beta}{c}\right) < 0.$$

Since $f'(t) = (t+\beta)^{-c-1} g(t+1+\theta)$, where $(t+\beta)^{-c-1} > 0$ for any $t \geq 0$ and c, we have $f'(t) < 0$ for all $t \geq 0$. Hence, f is monotonically decreasing on $[0, +\infty)$.

(c) If f is monotonically decreasing and nonnegative, then $f(i) \leq \int_{i-1}^{i} f(x)dx$ for any i. Hence, by summing this inequality from $i = N+1$ to t, we have

$$\sum_{i=N+1}^{t} f(i) \leq \sum_{i=N+1}^{t} \int_{i-1}^{i} f(x)dx = \int_N^t f(x)dx,$$

which proves (22). □

Lemma 2. *Let $\{Y_t\}_{t\geq 1}$ and $\{Z_t\}_{t\geq 1}$ be two nonnegative sequences in $\mathbb{R}$ and m and q be two positive numbers such that $q > m$. For some $\rho > 0$ and $D > 0$, assume that*

$$Y_{t+1} \leq Y_t - \rho\eta_t^m \cdot Z_t + \eta_t^q \cdot D \tag{23}$$

where $\eta_t := \frac{\gamma}{(t+\beta)^\alpha}$ for some $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ such that $\alpha m \leq \frac{1}{2}$. Suppose that $Y_t \leq C + H\ln(t+\theta)$ for some $C > 0$, $H \geq 0$, $\theta > 0$, and $1 + \theta - \beta > (1-\alpha m)e^{\frac{\alpha m}{1-\alpha m}}$ (where e is the natural number), for all $t \geq 1$. Then, we have

$$\frac{1}{T}\sum_{t=1}^{T} Z_t \leq \frac{1}{T}\left[\frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m} + \frac{C(T-1+\beta)^{\alpha m}}{2\rho\alpha m\gamma^m} + \frac{H(T-1+\beta)^{\alpha m}\ln(T+\theta)}{2\rho\alpha m\gamma^m}\right] + \frac{D\gamma^{q-m}}{\rho}\cdot\frac{A(T)}{T}, \tag{24}$$

where

$$A(T) := \begin{cases} \ln(T+\beta) - \ln(\beta), & \text{if } \alpha(q-m) = 1 \\ \dfrac{(T+\beta)^{1-\alpha(q-m)}}{1-\alpha(q-m)}, & \text{otherwise.} \end{cases}$$

Proof. From the recursive inequality (23) and $\eta_t := \frac{\gamma}{(t+\beta)^\alpha}$, we have

$$Z_t \leq \frac{1}{\rho\eta_t^m}(Y_t - Y_{t+1}) + \frac{D\eta_t^{q-m}}{\rho} = \frac{(t+\beta)^{\alpha m}}{\rho\gamma^m}(Y_t - Y_{t+1}) + \frac{D\gamma^{q-m}}{\rho}\cdot\frac{1}{(t+\beta)^{\alpha(q-m)}}.$$

Next, using (21) from Lemma 3(a) with $s := t+\beta$ and $\nu := m\alpha$ we have

$$(t+\beta+1)^{\alpha m} - (t+\beta)^{\alpha m} \leq \frac{1}{2(t+\beta)^{1-\alpha m}}, \tag{25}$$

because we assume $m\alpha \leq \frac{1}{2}$. Summing the first inequality from $t = 1, \cdots, T$ and taking average, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} Z_t &\leq \frac{1}{\rho\gamma^m}\cdot\frac{1}{T}\sum_{t=1}^{T}(t+\beta)^{\alpha m}(Y_t - Y_{t+1}) + \frac{D\gamma^{q-m}}{\rho}\cdot\frac{1}{T}\sum_{t=1}^{T}\frac{1}{(t+\beta)^{\alpha(q-m)}} \\
&= \frac{1}{\rho\gamma^m}\cdot\frac{1}{T}\left[(1+\beta)^{\alpha m}Y_1 - (T+\beta)^{\alpha m}Y_{T+1}\right] + \frac{1}{\rho\gamma^m}\cdot\frac{1}{T}\sum_{t=1}^{T-1}\left((t+1+\beta)^{\alpha m} - (t+\beta)^{\alpha m}\right)Y_{t+1} \\
&\quad + \frac{D\gamma^{q-m}}{\rho}\cdot\frac{1}{T}\sum_{t=1}^{T}\frac{1}{(t+\beta)^{\alpha(q-m)}} \\
&\overset{(25)}{\leq} \frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m}\cdot\frac{1}{T} + \frac{1}{2\rho\gamma^m}\cdot\frac{1}{T}\sum_{t=1}^{T-1}\frac{C + H\ln(t+1+\theta)}{(t+\beta)^{1-\alpha m}} + \frac{D\gamma^{q-m}}{\rho}\cdot\frac{1}{T}\sum_{t=1}^{T}\frac{1}{(t+\beta)^{\alpha(q-m)}} \\
&\overset{(22)}{\leq} \frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m}\cdot\frac{1}{T} + \frac{C}{2\rho\gamma^m}\cdot\frac{1}{T}\int_{t=0}^{T-1}\frac{dt}{(t+\beta)^{1-\alpha m}} + \frac{H}{2\rho\gamma^m}\cdot\frac{1}{T}\int_{t=0}^{T-1}\frac{\ln(t+1+\theta)}{(t+\beta)^{1-\alpha m}}dt \\
&\quad + \frac{D\gamma^{q-m}}{\rho}\cdot\frac{1}{T}\int_{t=0}^{T}\frac{dt}{(t+\beta)^{\alpha(q-m)}},
\end{aligned}$$

where the second inequality follows since $0 \leq Y_t \leq C + H\ln(t+\theta)$ for some $C > 0$, $H \geq 0$, and $\theta > 0$, for all $t \geq 1$, and $\alpha m \leq \frac{1}{2}$. The last inequality follows since $\frac{\ln(t+1+\theta)}{(t+\beta)^{1-\alpha m}}$ is nonnegative and monotonically

decreasing on $[0,\infty)$ according to Lemma 3(b) with $1-\alpha m \geq \frac{1}{2} > 0$ and $1+\theta-\beta > (1-\alpha m)e^{\frac{\alpha m}{1-\alpha m}}$, and both $\frac{1}{(t+\beta)^{1-\alpha m}}$ and $\frac{1}{(t+\beta)^{\alpha(q-m)}}$ are also nonnegative and monotonically decreasing on $[0,\infty)$. Note that

$$\begin{aligned}\int_{t=0}^{T-1} \frac{\ln(t+1+\theta)}{(t+\beta)^{1-\alpha m}} dt &= \frac{1}{\alpha m}(t+\beta)^{\alpha m}\ln(t+1+\theta)\Big|_{t=0}^{T-1} - \frac{1}{\alpha m}\int_{t=0}^{T-1} \frac{(t+\beta)^{\alpha m}}{(t+1+\theta)} dt \\ &\leq \frac{1}{\alpha m}(T-1+\beta)^{\alpha m}\ln(T+\theta).\end{aligned}$$

Therefore, we consider two cases:

- If $\alpha(q-m)=1$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} Z_t &\leq \frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m}\cdot\frac{1}{T} + \frac{C}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}-\beta^{\alpha m}}{T}\right) \\ &\quad + \frac{H}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}\ln(T+\theta)}{T}\right) + \frac{D\gamma^{q-m}}{\rho}\left(\frac{\ln(T+\beta)-\ln(\beta)}{T}\right) \\ &\leq \frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m}\cdot\frac{1}{T} + \frac{C}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}}{T}\right) \\ &\quad + \frac{H}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}\ln(T+\theta)}{T}\right) + \frac{D\gamma^{q-m}}{\rho}\left(\frac{\ln(T+\beta)-\ln(\beta)}{T}\right).\end{aligned}$$

- If $\alpha(q-m)\neq 1$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} Z_t &\leq \frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m}\cdot\frac{1}{T} + \frac{C}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}-\beta^{\alpha m}}{T}\right) \\ &\quad + \frac{H}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}\ln(T+\theta)}{T}\right) \\ &\quad + \frac{D\gamma^{q-m}}{\rho(1-\alpha(q-m))}\left(\frac{(T+\beta)^{1-\alpha(q-m)}-\beta^{1-\alpha(q-m)}}{T}\right) \\ &\leq \frac{(1+\beta)^{\alpha m}Y_1}{\rho\gamma^m}\cdot\frac{1}{T} + \frac{C}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}}{T}\right) \\ &\quad + \frac{H}{2\rho\alpha m\gamma^m}\left(\frac{(T-1+\beta)^{\alpha m}\ln(T+\theta)}{T}\right) \\ &\quad + \frac{D\gamma^{q-m}}{\rho(1-\alpha(q-m))}\left(\frac{(T+\beta)^{1-\alpha(q-m)}}{T}\right).\end{aligned}$$

Here, the result is obtained by directly computing the integrals. Hence, (24) is proved. □

A.2 Key Lemmas for Convex Problems

This subsection provides three key lemmas for Algorithm 1 for convex problems. Note that $w_* = \arg\min_{w\in\mathbb{R}^d} F(w)$.

Lemma 4. *Suppose that Assumption 2 holds. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with the*

learning rate $\eta_i^{(t)} := \frac{\eta_t}{n} > 0$ *for a given positive sequence* $\{\eta_t\}$*. Then, we have*

$$\begin{aligned} \|w_i^{(t)} - w_0^{(t)}\|^2 &\leq \eta_t^2 \cdot \frac{2iL^2}{n^2} \sum_{j=0}^{i-1} \|w_j^{(t)} - w_*\|^2 + \eta_t^2 \cdot 2N. \\ \|w_i^{(t)} - w_*\|^2 &\leq 2\|w_i^{(0)} - w_*\|^2 + \frac{4L^2 i\eta_t^2}{n^2} \cdot \sum_{j=0}^{i-1} \|w_j^{(t)} - w_*\|^2 + 4N\eta_t^2. \end{aligned} \tag{26}$$

Proof. Since $\eta_i^{(t)} = \frac{\eta_t}{n}$, for $i \in [n]$, and for any $t \geq 1$, by the gradient update step, we have

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

Using the last expression, the optimality condition $\nabla F(w_*) = 0$ in (*a*), $(u+v)^2 \leq 2u^2 + 2v^2$ in (*b*), and the Cauchy-Schwarz inequality in (*c*), for $i \in [n]$, we can derive

$$\begin{aligned} & \|w_i^{(t)} - w_0^{(t)}\|^2 = \frac{\eta_t^2}{n^2} \Big\| \sum_{j=0}^{i-1} \nabla f(w_j^{(t)}; \sigma^{(t)}(j+1)) \Big\|^2 \\ = \quad & \frac{i^2 \cdot \eta_t^2}{n^2} \Big\| \frac{1}{i} \sum_{j=0}^{i-1} \nabla f(w_j^{(t)}; \sigma^{(t)}(j+1)) \Big\|^2 \\ \overset{(a)}{=} \quad & \frac{i^2 \cdot \eta_t^2}{n^2} \Big\| \frac{1}{i} \sum_{j=0}^{i-1} \nabla f(w_j^{(t)}; \sigma^{(t)}(j+1)) - \frac{1}{i} \sum_{j=0}^{i-1} \nabla f(w_*; \sigma^{(t)}(j+1)) - \frac{1}{i} \sum_{j=i}^{n-1} \nabla f(w_*; \sigma^{(t)}(j+1)) \Big\|^2 \\ \overset{(b)}{\leq} \quad & \frac{2i^2 \cdot \eta_t^2}{n^2} \Big\| \frac{1}{i} \sum_{j=0}^{i-1} \nabla f(w_j^{(t)}; \sigma^{(t)}(j+1)) - \frac{1}{i} \sum_{j=0}^{i-1} \nabla f(w_*; \sigma^{(t)}(j+1)) \Big\|^2 \\ & + \frac{2i^2 \cdot \eta_t^2}{n^2} \Big\| \frac{1}{i} \sum_{j=i}^{n-1} \nabla f(w_*; \sigma^{(t)}(j+1)) \Big\|^2 \\ \overset{(c)}{\leq} \quad & \frac{2i^2 \cdot \eta_t^2}{n^2} \cdot \frac{1}{i} \cdot \sum_{j=0}^{i-1} \|\nabla f(w_j^{(t)}; \sigma^{(t)}(j+1)) - \nabla f(w_*; \sigma^{(t)}(j+1))\|^2 \\ & + \frac{2i^2 \cdot \eta_t^2}{n^2} \cdot \frac{(n-i)}{i^2} \cdot \sum_{j=i}^{n-1} \|\nabla f(w_*; \sigma^{(t)}(j+1))\|^2 \\ \overset{(3),(16)}{\leq} \quad & \frac{2L^2 i \cdot \eta_t^2}{n^2} \sum_{j=0}^{i-1} \|w_j^{(t)} - w_*\|^2 + \frac{2(n-i) \cdot \eta_t^2}{n^2} \sum_{j=i}^{n-1} N_{\sigma^{(t)}(j+1)} \\ \overset{(16)}{\leq} \quad & \frac{2L^2 i \cdot \eta_t^2}{n^2} \sum_{j=0}^{i-1} \|w_j^{(t)} - w_*\|^2 + \frac{2(n-i) \cdot \eta_t^2}{n^2} (n-i)N, \qquad (\text{using } N_{\sigma^{(t)}(j+1)} \leq N) \\ \leq \quad & \eta_t^2 \cdot \frac{2iL^2}{n^2} \sum_{j=0}^{i-1} \|w_j^{(t)} - w_*\|^2 + \eta_t^2 \cdot 2N. \end{aligned}$$

This is exactly the first inequality of (26). By induction, for $i \in [n]$ the last inequality leads to

$$\|w_i^{(t)} - w_*\|^2 \leq 2\|w_0^{(t)} - w_*\|^2 + 2\|w_i^{(t)} - w_0^{(t)}\|^2$$

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

which proves the second estimate of (26). □

Lemma 5. *Under the same conditions as in Lemma 4, for any $t \geq 1$, if $0 < \eta_t \leq \frac{1}{2L}$, then we have*

$$\frac{1}{n} \sum_{i=0}^{n-1} i \sum_{j=0}^{i-1} \|w_j^{(t)} - w_*\|^2 \leq \frac{4}{3} n^2 \|w_0^{(t)} - w_*\|^2 + \frac{8N}{3} n^2 \eta_t^2. \tag{27}$$

Proof. For notational simplicity, let us denote $W_i^{(t)} := \|w_i^{(t)} - w_*\|^2$, $A := \frac{2L^2}{n^2}$, and $B := 2N$. By Lemma 4, for $j \in [n]$, and for any $t \geq 1$ we have

$$W_j^{(t)} \leq 2W_0^{(t)} + 2A\eta_t^2 \cdot j \sum_{k=0}^{j-1} W_k^{(t)} + \eta_t^2 \cdot 2B.$$

Summing up this inequality from $j = 0$ to $j = i - 1$ with $i \geq 1$, we have

$$\sum_{j=0}^{i-1} W_j^{(t)} = W_0^{(t)} + \sum_{j=1}^{i-1} W_j^{(t)} \leq W_0^{(t)} + 2(i-1)W_0^{(t)} + 2B(i-1)\eta_t^2 + 2A\eta_t^2 \left(\sum_{j=1}^{i-1} j \sum_{k=0}^{j-1} W_k^{(t)} \right). \tag{28}$$

By convention, we have $\sum_{j=h}^{k} g_j = 0$ for all $h > k$. Moreover, since $j \leq i - 1 < i$ and $W_k^{(t)} \geq 0$, we have

$$\sum_{j=1}^{i-1} j \sum_{k=0}^{j-1} W_k^{(t)} \leq \sum_{j=1}^{i-1} j \sum_{k=0}^{i-1} W_k^{(t)} \leq \frac{i^2}{2} \sum_{k=0}^{i-1} W_k^{(t)}.$$

Using this inequality into (28), we can further derive

$$\sum_{k=0}^{i-1} W_k^{(t)} \leq W_0^{(t)} + 2(i-1)W_0^{(t)} + 2B(i-1)\eta_t^2 + 2A\eta_t^2 \cdot \frac{i^2}{2} \sum_{k=0}^{i-1} W_k^{(t)}.$$

Rearranging this inequality by moving the last term from the RHS to the LHS and then dividing both sides by $(1 - \eta_t^2 A i^2) > 0$ we arrive at

$$\sum_{k=0}^{i-1} W_k^{(t)} \leq \left(\frac{2i-1}{1 - Ai^2\eta_t^2} \right) W_0^{(t)} + \left(\frac{2B(i-1)}{1 - Ai^2\eta_t^2} \right) \cdot \eta_t^2.$$

Since $\eta_t^2 \leq \frac{1}{4L^2} \leq \frac{n^2}{4L^2 i^2}$ for $i \in [n]$, we have $Ai^2\eta_t^2 \leq \frac{n^2}{4L^2 i^2} \cdot i^2 \cdot \frac{2L^2}{n^2} = \frac{1}{2}$, we can upper bound the last inequality as

$$\sum_{k=0}^{i-1} W_k^{(t)} \leq 2(2i-1)W_0^{(t)} + 4B(i-1)\eta_t^2 < 4iW_0^{(t)} + 4Bi \cdot \eta_t^2. \tag{29}$$

Note that $\sum_{i=0}^{n-1} i^2 = \frac{2n^3 - 3n^2 + n}{6} \leq \frac{n^3}{3}$, using this inequality and (29), we can show that

$$\frac{1}{n} \sum_{i=0}^{n-1} i \sum_{k=0}^{i-1} W_k^{(t)} < \frac{4W_0^{(t)}}{n} \sum_{i=0}^{n-1} i^2 + \frac{4B\eta_t^2}{n} \sum_{i=0}^{n-1} i^2 \leq \frac{4n^2}{3} W_0^{(t)} + \frac{4Bn^2}{3} \cdot \eta_t^2,$$

which proves (27) by substituting back $W_i^{(t)} := \|w_i^{(t)} - w_*\|^2$ and $B := 2N$. □

Lemma 6. *Under the same conditions as in Lemma 4, we have*

$$\frac{1}{n}\sum_{i=0}^{n-1}\|w_i^{(t)}-w_0^{(t)}\|^2 \le \eta_t^2\cdot\frac{8L^2}{3}\|w_0^{(t)}-w_*\|^2+\frac{16L^2N}{3}\cdot\eta_t^4+2N\cdot\eta_t^2. \tag{30}$$

Proof. From the first inequality of (26) and (27), we can derive

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1}\|w_i^{(t)}-w_0^{(t)}\|^2 &\overset{(26)}{\le} \eta_t^2\cdot\frac{2L^2}{n^2}\left(\frac{1}{n}\sum_{i=0}^{n-1}i\cdot\sum_{j=0}^{i-1}\|w_j^{(t)}-w_*\|^2\right)+2N\cdot\eta_t^2\\ &\overset{(27)}{\le} \eta_t^2\cdot\frac{2L^2}{n^2}\left(\frac{4}{3}n^2\|w_0^{(t)}-w_*\|^2+\frac{2}{3}n^2\cdot 4N\eta_t^2\right)+2N\cdot\eta_t^2\\ &= \frac{8L^2}{3}\cdot\eta_t^2\cdot\|w_0^{(t)}-w_*\|^2+\frac{16L^2N}{3}\cdot\eta_t^4+2N\cdot\eta_t^2,\end{aligned}$$

which is exactly (30). □

B Convergence Analysis for Non-Convex Case

In this section, we provide the full proofs of the results in Section 4 of the main text.

B.1 Proofs of Theorem 1, Corollary 1, and Corollary 2: Convergence Analysis with Constant Stepsize

Theorem 1. *Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)}=\frac{\eta_t}{n}=\frac{\eta}{n}$, with $\eta_t\le\frac{1}{L}$. Then, under Assumptions 1, 2, and 3, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \le \frac{2}{T\eta}\big[F(\tilde{w}_0)-F_*\big]+\frac{L^2G^2}{3}\cdot\eta^2. \tag{31}$$

Proof. First, from the update $w_{i+1}^{(t)}:=w_i^{(t)}-\eta_i^{(t)}\nabla f(w_i^{(t)};\sigma^{(t)}(i+1))$ in Algorithm 1 with $\eta_i^{(t)}:=\frac{\eta_t}{n}$, for $i\in[n]$, we have

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

Hence, for $i\in[n]$, using this expression and (11) in Assumption 3, we can bound

$$\|w_i^{(t)}-w_0^{(t)}\|^2=\frac{\eta_t^2}{n^2}\Big\|\sum_{j=0}^{i-1}\nabla f(w_j^{(t)};\sigma^{(t)}(j+1))\Big\|^2=\frac{i^2\cdot\eta_t^2}{n^2}\Big\|\frac{1}{i}\sum_{j=0}^{i-1}\nabla f(w_j^{(t)};\sigma^{(t)}(j+1))\Big\|^2\overset{(11)}{\le}\frac{i^2\cdot\eta_t^2}{n^2}G^2. \tag{32}$$

Since F is L-smooth, we can derive

$$
\begin{aligned}
F(w_0^{(t+1)}) &\overset{(4)}{\leq} F(w_0^{(t)}) + \nabla F(w_0^{(t)})^T(w_0^{(t+1)} - w_0^{(t)}) + \frac{L}{2}\|w_0^{(t+1)} - w_0^{(t)}\|^2 \\
&\overset{(6)}{=} F(w_0^{(t)}) - \eta_t \nabla F(w_0^{(t)})^T \left(\frac{1}{n}\sum_{i=0}^{n-1} \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1))\right) + \frac{L\eta_t^2}{2}\left\|\frac{1}{n}\sum_{i=0}^{n-1} \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1))\right\|^2 \\
&\overset{(a)}{=} F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{\eta_t}{2}\left\|\nabla F(w_0^{(t)}) - \frac{1}{n}\sum_{i=0}^{n-1} \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1))\right\|^2 \\
&\quad - \frac{\eta_t}{2}(1 - L\eta_t)\left\|\frac{1}{n}\sum_{i=0}^{n-1} \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1))\right\|^2 \\
&\overset{(b)}{\leq} F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{\eta_t}{2}\left\|\frac{1}{n}\sum_{i=0}^{n-1} \nabla f(w_0^{(t)}; \sigma^{(t)}(i+1)) - \frac{1}{n}\sum_{i=0}^{n-1} \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1))\right\|^2 \\
&\overset{(c)}{\leq} F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{\eta_t}{2n}\sum_{i=0}^{n-1}\left\|\nabla f(w_0^{(t)}; \sigma^{(t)}(i+1)) - \nabla f(w_i^{(t)}; \sigma^{(t)}(i+1))\right\|^2 \\
&\overset{(3)}{\leq} F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{L^2\eta_t}{2n}\sum_{i=0}^{n-1}\|w_i^{(t)} - w_0^{(t)}\|^2, \qquad (33)
\end{aligned}
$$

where (a) follows from $u^T v = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u-v\|^2)$, (b) follows from the fact that $\eta_t \leq \frac{1}{L}$, and (c) is from the Cauchy-Schwarz inequality. Hence, using (32) and following the same argument as (33), we can derive

$$
\begin{aligned}
F(w_0^{(t+1)}) &\overset{(32)}{\leq} F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{L^2\eta_t}{2}\frac{1}{n}\sum_{i=0}^{n-1}\frac{i^2\cdot\eta_t^2}{n^2}G^2 \\
&\leq F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{L^2G^2}{6}\cdot\eta_t^3,
\end{aligned}
$$

where we use $\sum_{i=0}^{n-1} i^2 \leq \frac{n^3}{3}$ in the last inequality. Note that $\tilde{w}_t = w_0^{(t+1)}$ and $\tilde{w}_{t-1} = w_0^{(t)}$ in Algorithm 1, the last estimate becomes

$$
F(\tilde{w}_t) \leq F(\tilde{w}_{t-1}) - \frac{\eta_t}{2}\|\nabla F(\tilde{w}_{t-1})\|^2 + \frac{L^2G^2}{6}\cdot\eta_t^3. \qquad (34)
$$

Using $\eta_t := \eta$ into (34) and rearranging its result, we end up with

$$
\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2}{\eta}\big[F(\tilde{w}_{t-1}) - F(\tilde{w}_t)\big] + \frac{L^2G^2}{3}\cdot\eta^2.
$$

Summing the last inequality from $t = 1, \cdots, T$ and taking average, we finally obtain

$$
\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2}{T\eta}\big[F(\tilde{w}_0) - F_*\big] + \frac{L^2G^2}{3}\cdot\eta^2,
$$

which is exactly (31). □

Corollary 1. *Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 and $\widehat{w}_T$ be its output. For given tolerance $\epsilon > 0$, under the same conditions as in Theorem 1, if we choose the constant learning rate $\eta := \frac{\sqrt{\epsilon}}{LG}$, then to guarantee*

$$\mathbb{E}\left[\|\nabla F(\widehat{w}_T)\|^2\right] = \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \epsilon,$$

for (1)*, it requires $T := \left\lfloor 3LG[F(\tilde{w}_0) - F_*] \cdot \frac{1}{\epsilon^{3/2}} \right\rfloor$ outer iterations. As a result, the total number of gradient evaluations is at most $\mathcal{T}_w := \left\lfloor 3LG[F(\tilde{w}_0) - F_*] \cdot \frac{n}{\epsilon^{3/2}} \right\rfloor$.*

Proof. Given $\epsilon > 0$, to guarantee $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \epsilon$, by using (31) in Theorem 1, we impose

$$\frac{2}{T\eta}\big[F(\tilde{w}_0) - F_*\big] + \frac{L^2G^2}{3}\cdot\eta^2 \leq \epsilon.$$

Using $\eta = \frac{\sqrt{\epsilon}}{LG}$ into this equation, we can easily get

$$\frac{2LG}{T\sqrt{\epsilon}}\big[F(\tilde{w}_0) - F_*\big] \leq \frac{2\epsilon}{3} \quad \Rightarrow \quad T \geq 3LG\big[F(\tilde{w}_0) - F_*\big]\cdot\frac{1}{\epsilon^{3/2}}$$

Rounding this expression we get $\mathcal{T} := \left\lfloor 3LG[F(\tilde{w}_0) - F_*] \cdot \frac{1}{\epsilon^{3/2}} \right\rfloor$. As a result, the total number of gradient evaluations is $\mathcal{T}_w := n\mathcal{T} = \left\lfloor 3LG[F(\tilde{w}_0) - F_*] \cdot \frac{n}{\epsilon^{3/2}} \right\rfloor$. □

Corollary 2. *Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 and $\widehat{w}_T$ be its output. Under the same conditions as in Theorem 1, if we choose the constant learning rate $\eta := \frac{\gamma}{T^{1/3}}$ for some $\gamma > 0$, then*

$$\mathbb{E}\left[\|\nabla F(\widehat{w}_T)\|^2\right] = \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{1}{T^{2/3}}\left[\frac{2\big[F(\tilde{w}_0) - F_*\big]}{\gamma} + \frac{\gamma^2L^2G^2}{3}\right].$$

Proof. Substituting $\eta = \frac{\gamma}{T^{1/3}}$ into (31) of Theorem 1, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2}{T\eta}\big[F(\tilde{w}_0) - F_*\big] + \eta^2\frac{L^2G^2}{3} = \frac{1}{T^{2/3}}\left[\frac{2[F(\tilde{w}_0) - F_*]}{\gamma} + \frac{\gamma^2L^2G^2}{3}\right],$$

which is exactly our desired estimate. □

B.2 Proof of Theorem 2: Asymptotic Convergence with Diminishing Step-Size

To establish the results in this section, we will use the following lemma from [2].

Lemma 7 ([2])**.** *Let $\{Y_t\}_{t\geq 0}$, $\{Z_t\}_{t\geq 0}$, and $\{W_t\}_{t\geq 0}$ be three sequences of random variables. Let $\{\mathcal{F}_t\}_{t\geq 0}$ be a filtration, that is, σ-algebras such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all $t \geq 0$. Suppose that the following conditions hold:*

(i) *The random variables Y_t, Z_t, and W_t are nonnegative, and $\mathcal{F}_t$-measurable;*
(ii) *For each $t \geq 0$, we have $\mathbb{E}\left[Y_{t+1} \mid \mathcal{F}_t\right] \leq Y_t - Z_t + W_t$;*
(iii) *With probability 1, it holds that $\sum_{t=0}^{\infty} W_t < \infty$.*

Then, w.p.1, we have

$$\sum_{t=0}^{\infty} Z_t < \infty \quad \textit{and} \quad Y_t \to Y \geq 0.$$

Using Lemma 7 we prove Theorem 2 in the main text as follows.

Theorem 2. *Suppose Assumptions 1, 2, and 3 hold. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with diminishing learning rate $\eta_i^{(t)} := \frac{\eta_t}{n}$ such that*

$$\sum_{t=1}^{\infty} \eta_t = \infty \textit{ and } \sum_{t=1}^{\infty} \eta_t^3 < \infty.$$

Then, w.p.1. (i.e. almost surely), the following limit holds:

$$\liminf_{t \to \infty} \|\nabla F(\tilde{w}_{t-1})\| = 0.$$

Proof. First, following the same argument as in the proof of (34) of Theorem 1, we have

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \frac{\eta_{t+1}}{2}\|\nabla F(\tilde{w}_t)\|^2 + \frac{L^2 G^2}{6} \cdot \eta_{t+1}^3.$$

Let us define $\mathcal{F}_t = \sigma(\tilde{w}_0, \cdots, \tilde{w}_t)$ be the σ-algebra generated by $\tilde{w}_0, \cdots, \tilde{w}_t$. Then, for $t \geq 0$, the last inequality implies

$$\mathbb{E}\left[[F(\tilde{w}_{t+1}) - F_*] \mid \mathcal{F}_t\right] \leq [F(\tilde{w}_t) - F_*] - \frac{\eta_{t+1}}{2}\|\nabla F(\tilde{w}_t)\|^2 + \frac{L^2 G^2}{6} \cdot \eta_{t+1}^3.$$

Let us define $Y_t := [F(\tilde{w}_t) - F_*] \geq 0$, $Z_t := \frac{\eta_{t+1}}{2}\|\nabla F(\tilde{w}_t)\|^2 \geq 0$ and $W_t := \frac{L^2 G^2}{6} \cdot \eta_{t+1}^3$. Then, the first condition (i) of Lemma 7 holds. Moreover, the last inequality shows that $\mathbb{E}\left[Y_{t+1} \mid \mathcal{F}_t\right] \leq Y_t - Z_t + W_t$, which means that the condition (ii) of Lemma 7 holds. Since $\sum_{t=1}^{\infty} \eta_t^3 < \infty$, we have $\sum_{t=0}^{\infty} W_t < \infty$, which fulls fill the condition (iii) of Lemma 7. Then, by applying Lemma 7, we obtain w.p.1 that

$$F(\tilde{w}_t) - F_* \to Y \geq 0 \quad \text{and} \quad \sum_{t=0}^{\infty} \frac{\eta_{t+1}}{2}\|\nabla F(\tilde{w}_t)\|^2 < \infty.$$

We prove $\liminf_{t \to \infty} \|\nabla F(\tilde{w}_{t-1})\| = 0$ w.p.1. by contradiction. Indeed, we assume that there exist $\epsilon > 0$ and $t_0 \geq 0$ such that $\|\nabla F(\tilde{w}_t)\|^2 \geq \epsilon$ for $\forall t \geq t_0$. In this case, since $\sum_{t=0}^{\infty} \eta_t = \infty$, we have

$$\infty > \sum_{t=t_0}^{\infty} \frac{\eta_{t+1}}{2}\|\nabla F(\tilde{w}_t)\|^2 \geq \frac{\epsilon}{2} \sum_{t=t_0}^{\infty} \eta_{t+1} = \infty.$$

This is a contradiction. As a result, w.p.1., we have $\liminf_{k \to \infty} \|\nabla F(\tilde{w}_k)\|^2 = 0$, or equivalently, it holds that $\liminf_{k \to \infty} \|\nabla F(\tilde{w}_k)\| = 0$. □

B.3 Convergence Analysis with Difference Learning Rates

This subsection provides convergence analysis for general choice of learning rate.

Theorem 3. *Suppose that Assumptions 1, 2, and 3 hold. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n}$, where $\eta_t := \frac{\gamma}{(t+\beta)^\alpha} \leq \frac{1}{L}$, for some $\gamma > 0$, $\beta > 0$, and $\frac{1}{3} < \alpha < 1$. Let us define $C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3 L^2 G^2}{6(3\alpha-1)\beta^{3\alpha-1}} > 0$. Then, the following statements hold:*

- *If $\alpha = \frac{1}{2}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2(1+\beta)^{1/2}[F(\tilde{w}_0) - F_*]}{\gamma} \cdot \frac{1}{T} + \frac{2C}{\gamma}\left(\frac{(T-1+\beta)^{1/2}}{T}\right) + \frac{L^2G^2\gamma^2}{3}\left(\frac{\ln(T+\beta) - \ln(\beta)}{T}\right).$$

- *If $\alpha \neq \frac{1}{2}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2(1+\beta)^{\alpha}[F(\tilde{w}_0) - F_*]}{\gamma} \cdot \frac{1}{T} + \frac{C}{\alpha\gamma}\left(\frac{(T-1+\beta)^{\alpha}}{T}\right) + \frac{L^2G^2\gamma^2}{3(1-2\alpha)}\left(\frac{(T+\beta)^{1-2\alpha}}{T}\right).$$

Proof. By (34), we have

$$F(\tilde{w}_t) \leq F(\tilde{w}_{t-1}) - \frac{\eta_t}{2}\|\nabla F(\tilde{w}_{t-1})\|^2 + \eta_t^3 \frac{L^2G^2}{6} \leq F(\tilde{w}_{t-1}) + \eta_t^3 \frac{L^2G^2}{6}.$$

Notice that since $\eta_t = \frac{\gamma}{(t+\beta)^\alpha}$, summing up this inequality from $t = 1$ to $t = k \geq 1$, we have

$$\begin{aligned} F(\tilde{w}_k) &\leq F(\tilde{w}_0) + \frac{L^2G^2}{6}\sum_{t=1}^{k}\eta_t^3 = F(\tilde{w}_0) + \frac{L^2G^2}{6}\sum_{t=1}^{k}\frac{\gamma^3}{(t+\beta)^{3\alpha}} \\ &\overset{(22)}{\leq} F(\tilde{w}_0) + \frac{\gamma^3L^2G^2}{6}\int_{t=0}^{k}\frac{dt}{(t+\beta)^{3\alpha}} = F(\tilde{w}_0) + \frac{\gamma^3L^2G^2}{6}\left[-\frac{(t+\beta)^{-(3\alpha-1)}}{3\alpha-1}\Big|_{t=0}^{k}\right] \\ &\leq F(\tilde{w}_0) + \frac{\gamma^3L^2G^2}{6(3\alpha-1)\beta^{3\alpha-1}}. \end{aligned}$$

Here, we use the fact that $\frac{1}{(t+\beta)^{3\alpha}}$ is nonnegative and monotonically decreasing on $[0, +\infty)$ and $\frac{1}{3} < \alpha < 1$. Subtracting F_* to both sides, for $t \geq 1$, we have

$$F(\tilde{w}_t) - F_* \leq [F(\tilde{w}_0) - F_*] + \frac{\gamma^3L^2G^2}{6(3\alpha-1)\beta^{3\alpha-1}}. \tag{35}$$

On the other hand, subtracting F_* to both sides of (34), we have

$$F(\tilde{w}_t) - F_* \leq [F(\tilde{w}_{t-1}) - F_*] - \frac{\eta_t}{2}\|\nabla F(\tilde{w}_{t-1})\|^2 + \frac{L^2G^2}{6}\cdot\eta_t^3. \tag{36}$$

Now, let us define $Y_t := F(\tilde{w}_{t-1}) - F_* \geq 0$, $Z_t := \|\nabla F(\tilde{w}_{t-1})\|^2 \geq 0$, for $t \geq 1$, $\rho := \frac{1}{2}$, and $D := \frac{L^2G^2}{6}$. The estimate (36) becomes

$$Y_{t+1} \leq Y_t - \rho\eta_t Z_t + D\eta_t^3.$$

Let us define $C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3 L^2 G^2}{6(3\alpha-1)\beta^{3\alpha-1}} > 0$. By (35), we have $Y_t \leq C$ (note that $H = 0$), $t \geq 1$. Applying Lemma 2 with $q = 3$ and $m = 1$. we conclude that

- If $\alpha = \frac{1}{2}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2(1+\beta)^{1/2}[F(\tilde{w}_0) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{2C}{\gamma}\left(\frac{(T-1+\beta)^{1/2}}{T}\right) + \frac{L^2G^2\gamma^2}{3}\left(\frac{\ln(T+\beta) - \ln(\beta)}{T}\right).$$

- If $\alpha \neq \frac{1}{2}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 \leq \frac{2(1+\beta)^{\alpha}[F(\tilde{w}_0) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{C}{\alpha\gamma}\left(\frac{(T-1+\beta)^{\alpha}}{T}\right) + \frac{L^2G^2\gamma^2}{3(1-2\alpha)}\left(\frac{(T+\beta)^{1-2\alpha}}{T}\right).$$

This completes the proof. □

Remark 4. *In Theorem 3, if we choose $\alpha = \frac{1}{3} + \delta$ for some $0 < \delta < \frac{1}{6}$, then we have*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 &\leq \frac{2(1+\beta)^{\frac{1}{3}+\delta}[F(\tilde{w}_0) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{C}{\gamma(\frac{1}{3}+\delta)}\left(\frac{(T-1+\beta)^{\frac{1}{3}+\delta}}{T}\right) \\ &\quad + \frac{L^2G^2\gamma^2}{1-6\delta}\left(\frac{(T+\beta)^{\frac{1}{3}-2\delta}}{T}\right) \\ &= \mathcal{O}\left(\frac{1}{T^{\frac{2}{3}-\delta}}\right),\end{aligned}$$

where $C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3 L^2 G^2}{18\delta\beta^{3\delta}} > 0$. Note that the convergence rate for regular SGD is $\mathcal{O}\left(\frac{1}{T^{1/2}}\right)$.

For the extreme case where $\alpha = \frac{1}{3}$, we have the following result.

Theorem 4. *Suppose that Assumptions 1, 2, and 3 hold. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} = \frac{\eta_t}{n}$, where $\eta_t := \frac{\gamma}{(t+\beta)^{1/3}} \leq \frac{1}{L}$, for some $\gamma > 0$, and $\beta > 0$, and $C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3 L^2 G^2}{6(1+\beta)}$. Then, the following bound holds:*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 &\leq \frac{2(1+\beta)^{1/3}[F(\tilde{w}_0) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{3C}{\gamma}\left(\frac{(T-1+\beta)^{1/3}}{T}\right) \\ &\quad + \frac{\gamma^2L^2G^2}{2}\left(\frac{(T-1+\beta)^{1/3}\ln(T+1+\beta)}{T}\right) + L^2G^2\gamma^2\left(\frac{(T+\beta)^{1/3}}{T}\right) \\ &= \mathcal{O}\left(\frac{\ln(T)}{T^{2/3}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{T^{2/3}}\right).\end{aligned}$$

Proof. By (34), we have

$$F(\tilde{w}_t) \leq F(\tilde{w}_{t-1}) - \frac{\eta_t}{2}\|\nabla F(\tilde{w}_{t-1})\|^2 + \eta_t^3 \frac{L^2G^2}{6} \leq F(\tilde{w}_{t-1}) + \eta_t^3 \frac{L^2G^2}{6}.$$

Since $\eta_t = \frac{\gamma}{(t+\beta)^{1/3}}$, summing up this inequality from $t = 1$ to $t = k \geq 1$, we obtain

$$\begin{aligned} F(\tilde{w}_k) &\leq F(\tilde{w}_0) + \frac{L^2G^2}{6}\sum_{t=1}^{k}\eta_t^3 = F(\tilde{w}_0) + \frac{L^2G^2}{6}\sum_{t=1}^{k}\frac{\gamma^3}{(t+\beta)} \\ &\overset{(22)}{\leq} F(\tilde{w}_0) + \frac{\gamma^3L^2G^2}{6(1+\beta)} + \frac{\gamma^3L^2G^2}{6}\int_{t=1}^{k}\frac{dt}{(t+\beta)} \leq F(\tilde{w}_0) + \frac{\gamma^3L^2G^2}{6(1+\beta)} + \frac{\gamma^3L^2G^2}{6}\ln(k+2+\beta). \end{aligned}$$

Here, we use the fact that $\frac{1}{t+\beta}$ is nonnegative and monotonically decreasing on $[0, +\infty)$. Subtracting F_* to both sides, for $t \geq 1$, we have

$$F(\tilde{w}_t) - F_* \leq [F(\tilde{w}_0) - F_*] + \frac{\gamma^3L^2G^2}{6(1+\beta)} + \frac{\gamma^3L^2G^2}{6}\ln(t+2+\beta). \tag{37}$$

Define $Y_t := F(\tilde{w}_{t-1}) - F_* \geq 0$, $Z_t := \|\nabla F(\tilde{w}_{t-1})\|^2 \geq 0$, for $t \geq 1$, $\rho := \frac{1}{2}$, and $D := \frac{L^2G^2}{6}$. The estimate (36) becomes

$$Y_{t+1} \leq Y_t - \rho\eta_t Z_t + D\eta_t^3.$$

Let us define $C := [F(\tilde{w}_0) - F_*] + \frac{\gamma^3L^2G^2}{6(1+\beta)} > 0$, $H := \frac{\gamma^3L^2G^2}{6} > 0$, and $\theta := 1 + \beta > 0$. Clearly, we have $1 + \theta - \beta = 2 > \frac{2}{3}e^{1/2}$. By (37), we have $Y_t \leq C + H\ln(t+\theta)$ for $t \geq 1$. Applying Lemma 2 with $q = 3$, $m = 1$, and $\alpha = \frac{1}{3}$. we conclude that

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\tilde{w}_{t-1})\|^2 &\leq \frac{2(1+\beta)^{1/3}[F(\tilde{w}_0) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{3C}{\gamma}\left(\frac{(T-1+\beta)^{1/3}}{T}\right) \\ &\quad + \frac{\gamma^2L^2G^2}{2}\left(\frac{(T-1+\beta)^{1/3}\ln(T+1+\beta)}{T}\right) + L^2G^2\gamma^2\left(\frac{(T+\beta)^{1/3}}{T}\right) \\ &= \mathcal{O}\left(\frac{\ln(T)}{T^{2/3}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{T^{2/3}}\right), \end{aligned}$$

which proves our main bound. □

B.4 Proofs of Theorem 5 and Corollary 3: Convergence under Gradient Dominance

Theorem 5. *Suppose that Assumptions 1, 2, 3, and 4 hold for* (1). *Let* $\{w_i^{(t)}\}$ *be the sequence generated by Algorithm 1 with* $\eta_i^{(t)} := \frac{\eta_t}{n}$ *for solving* (1). *Let* η_t *be updated as* $\eta_t := \frac{\alpha}{t+\beta}$ *for some* $\alpha > 0$ *and* $\beta \geq 0$. *Assume further that* $\gamma > 0$ *and* $\lambda \geq 2$ *are two constants such that* $F(\tilde{w}_0) - F_* \leq \frac{\gamma}{(1+\beta)^2}$ *and* $\gamma(\alpha - 2\tau\lambda) \geq \frac{L^2G^2\tau}{3}\alpha^3$. *Then, we have*

$$F(\tilde{w}_t) - F_* \leq \frac{\gamma}{(t+1+\beta)^2}, \quad \forall t \geq 0.$$

Proof. Using (34) of Theorem 1 and Assumption 4, we can derive

$$F(\tilde{w}_t) - F_* \leq F(\tilde{w}_{t-1}) - F_* - \frac{\eta_t}{2}\|\nabla F(\tilde{w}_{t-1})\|^2 + \frac{L^2G^2}{6}\cdot\eta_t^3$$
$$\overset{(12)}{\leq} \left(1 - \frac{\eta_t}{2\tau}\right)\left[F(\tilde{w}_{t-1}) - F_*\right] + \frac{L^2G^2}{6}\cdot\eta_t^3.$$

Let $Y_t := F(\tilde{w}_{t-1}) - F_* \geq 0$, $\rho := \frac{1}{2\tau}$, and $D := \frac{L^2G^2}{6} > 0$. We now verify that these quantities satisfy the conditions of Lemma 1 with $q = 2$. Indeed, the condition $\gamma(\alpha - 2\tau\lambda) \geq \frac{L^2G^2\tau}{3}\alpha^3$ is equivalent $\gamma(\rho\alpha - \lambda) \geq D\alpha^3$. We also have $Y_1 = F(\tilde{w}_0) - F_* \leq \frac{\gamma}{(1+\beta)^2}$. Hence, applying Lemma 1 with $q = 2$, w.p.1., we end up with $F(\tilde{w}_t) - F_* \leq \frac{\gamma}{(t+1+\beta)^2}$, which proves our theorem. □

Corollary 3. *Suppose that conditions of Theorem 5 hold. Then, for any $\beta \geq 0$, if we choose α and γ such that*

$$\alpha := 5\tau \quad \textit{and} \quad \gamma := \max\left\{\frac{125}{3}L^2G^2\tau^3, \left[F(\tilde{w}_0) - F_*\right](1+\beta)^2\right\}, \tag{38}$$

then, using the learning rate $\eta_t := \frac{\alpha}{t+\beta} = \frac{5\tau}{t+\beta}$ in Algorithm 1, we have

$$F(\tilde{w}_t) - F_* \leq \frac{\gamma}{(t+1+\beta)^2},$$

where F_* is the global optimal value of (1)*.*

Proof. For any $\beta \geq 0$, if we choose $\alpha = 5\tau$ and $\lambda = 2$, then the condition $\gamma(\alpha - 2\tau\lambda) \geq \frac{L^2G^2\tau}{3}\alpha^3$ in Theorem 5 becomes

$$\gamma \geq \frac{L^2G^2\tau}{3(\alpha - 2\tau\lambda)}\alpha^3 = \frac{L^2G^2\tau}{3(5\tau - 4\tau)}(5\tau)^3 = \frac{125}{3}L^2G^2\tau^3.$$

At the same time, we requires $Y_1 := [F(\tilde{w}_0) - F_*] \leq \frac{\gamma}{(1+\beta)^2}$ in Lemma 1, which is equivalent to $\gamma \geq [F(\tilde{w}_0) - F_*](1+\beta)^2$. Combining both conditions, we can choose γ as in (38). Consequently, w.p.1. we have $F(\tilde{w}_t) - F_* \leq \frac{\gamma}{(t+1+\beta)^2}$. □

C Convergence Analysis for Convex Case

In this section, we prove the full proof of the results in the main text of Section 5.

C.1 Proofs of Theorem 6 and Corollary 4: The Strongly Convex Case

We first prove the main result of Section 5.

Theorem 6. *Assume that Assumptions 2 and 5 hold for* (1)*. Let $\{w_i^{(t)}\}$ be the sequence generated by Algorithm 1 with $\eta_i^{(t)} := \frac{\eta_t}{n}$ for solving* (1)*. Let α and β be two positive constants such that $\alpha = \frac{\mu}{2L^2}\beta$, and the learning rate η_t be updated as $\eta_t := \frac{\alpha}{t+\beta}$. Assume further that $\gamma > 0$ and $\lambda \geq 2$ are two constants such*

that $F(\tilde{w}_0) - F(w_*) \leq \frac{\gamma}{(1+\beta)^2}$ *and* $\gamma(\mu\alpha - 3\lambda) \geq 3(\mu^2 + L^2)N\alpha^3$ *with* $N := \max_{1\leq i\leq n} \|\nabla f(w_*; i)\|^2$. *Then, we have*

$$F(\tilde{w}_t) - F(w_*) \leq \frac{\gamma}{(t+1+\beta)^2}, \quad \forall t \geq 0.$$

Proof. Using (32) and following the same argument as (33) in the proof of Theorem 1, we can derive

$$\begin{aligned}
F(w_0^{(t+1)}) &\leq F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{L^2\eta_t}{2}\frac{1}{n}\sum_{i=0}^{n-1}\|w_i^{(t)} - w_0^{(t)}\|^2 \\
&\overset{(30)}{\leq} F(w_0^{(t)}) - \frac{\eta_t}{2}\|\nabla F(w_0^{(t)})\|^2 + \frac{L^2\eta_t}{2}\left(\eta_t^2 \cdot \frac{8L^2}{3}\|w_0^{(t)} - w_*\|^2 + \eta_t^4 \cdot \frac{16L^2N}{3} + \eta_t^2 \cdot 2N\right) \\
&\overset{(15),(14)}{\leq} F(w_0^{(t)}) - \mu\eta_t\left[F(w_0^{(t)}) - F(w_*)\right] + \eta_t^3\frac{8L^4}{3\mu}\left[F(w_0^{(t)}) - F(w_*)\right] \\
&\quad + \frac{8L^4N}{3} \cdot \eta_t^5 + L^2N \cdot \eta_t^3, \qquad (39)
\end{aligned}$$

Subtracting $F(w_*)$ from both sides of (39), we can further derive

$$F(w_0^{(t+1)}) - F(w_*) \leq \left[1 - \eta_t\left(\mu - \frac{8L^4}{3\mu}\eta_t^2\right)\right]\left[F(w_0^{(t)}) - F(w_*)\right] + \eta_t^3\left(\eta_t^2\frac{8L^4N}{3} + L^2N\right). \qquad (40)$$

Now, assume that $0 < \eta_t \leq \sqrt{\frac{3}{8}}\frac{\mu}{L^2}$. Then, one can show that

$$\mu - \frac{8L^4}{3\mu}\eta_t^2 \geq \mu - \frac{2}{3}\mu = \frac{\mu}{3} > 0 \quad \text{and} \quad \frac{8L^4N}{3}\eta_t^2 \leq \frac{8L^4N}{3} \cdot \frac{3}{8}\frac{\mu^2}{L^4} = \mu^2N.$$

Using these bounds into (40), we can further upper bound it as

$$F(w_0^{(t+1)}) - F(w_*) \leq \left(1 - \frac{\mu}{3}\eta_t\right)\left[F(w_0^{(t)}) - F(w_*)\right] + \eta_t^3\left(\mu^2N + L^2N\right). \qquad (41)$$

Next, we note that we have imposed $\eta_t \leq \min\left\{\frac{1}{2L}, \sqrt{\frac{3}{8}}\frac{\mu}{L^2}\right\}$ due to the Lemma 5. To simplify the choice of η_t, we can impose a stricter condition $\eta_t \leq \frac{\mu}{2L^2}$. The condition $\alpha = \frac{\mu\beta}{2L^2}$ and the update rule $\eta_t = \frac{\alpha}{t+\beta}$ guarantee this condition.

Now, let us define $Y_t := F(w_0^{(t)}) - F(w_*) = F(\tilde{w}_{t-1}) - F(w_*) \geq 0$, $\rho := \frac{\mu}{3}$, and $D := \mu^2N + L^2N$. The estimate (41) becomes

$$Y_{t+1} \leq (1 - \rho \cdot \eta_t)Y_t + D\eta_t^3.$$

Moreover, the condition $\gamma(\rho\alpha - \lambda) \geq D\alpha^3$ of Lemma 1 holds with $q = 2$ and $Y_1 = F(\tilde{w}_0) - F(w_*) \leq \frac{\gamma}{(1+\beta)^2}$. Applying this lemma we conclude that w.p.1, it holds that $Y_{t+1} = F(\tilde{w}_t) - F(w_*) \leq \frac{\gamma}{(t+1+\beta)^2}$. □

Corollary 4. *Suppose that the conditions in Theorem 6 hold. Let* α, β, *and* γ *be chosen as*

$$\begin{cases}
\alpha &:= \dfrac{12L^2 + \mu^2}{2L^2\mu}, \\
\beta &:= \dfrac{12L^2 + \mu^2}{\mu^2}, \\
\gamma &:= \max\left\{\dfrac{3(\mu^2 + L^2)(12L^2 + \mu^2)^3N}{4L^4\mu^5}, \dfrac{(12L^2 + 2\mu^2)^2}{\mu^4}\left[F(\tilde{w}_0) - F(w_*)\right]\right\}.
\end{cases} \qquad (42)$$

Then, we have

$$F(\tilde{w}_t) - F(w_*) \leq \frac{\gamma}{(t+1+\beta)^2}.$$

Proof. Let us choose $\lambda = 2$ for simplicity. Since $\alpha = \frac{\mu\beta}{2L^2}$, the condition $\gamma(\mu\alpha - 6) \geq 3(\mu^2 + L^2)N\alpha^3$ is equivalent to

$$\gamma \geq \frac{3(\mu^2+L^2)N\alpha^3}{\mu\alpha - 6} = \frac{3(\mu^2+L^2)N\mu^3\beta^3}{8L^6(\frac{\mu^2\beta}{2L^2} - 6)} = \frac{3(\mu^2+L^2)N\mu^3\beta^3}{4L^4(\mu^2\beta - 12L^2)}.$$

Let us choose $\beta := \frac{12L^2}{\mu^2} + 1 = \frac{12L^2+\mu^2}{\mu^2}$ as in the second line of (42). Then, the last condition becomes

$$\gamma \geq \frac{3(\mu^2+L^2)N\mu^3\beta^3}{4L^4(\mu^2\beta - 12L^2)} = \frac{3(\mu^2+L^2)(12L^2+\mu^2)^3N}{4L^4\mu^5}.$$

One the other hand, the condition $F(\tilde{w}_0) - F(w_*) \leq \frac{\gamma}{(1+\beta)^2}$ implies that

$$\gamma \geq \big[F(\tilde{w}_0) - F(w_*)\big](1+\beta)^2 = \frac{(12L^2+2\mu^2)^2}{\mu^4}\big[F(\tilde{w}_0) - F(w_*)\big].$$

Combining these two conditions, we can choose γ as in the last line of (42). Due to the choice of β, we have $\alpha = \frac{\mu\beta}{2L^2} = \frac{12L^2+\mu^2}{2L^2\mu}$ as in the first line of (42). Hence, all the conditions of Theorem 6 hold, leading to the last conclusion of this corollary. □

D (Regular) Stochastic Gradient Descent (SGD) Method

As a side result of our fundamental lemma, Lemma 2, we show in this section that we can apply the analysis framework of Lemma 2 to obtain convergence rate results for the regular SGD algorithm.

To keep it more general, we consider the stochastic optimization problem with respect to some distribution $\mathcal{D}$ as in (2):

$$\min_{w\in\mathbb{R}^d} \Big\{F(w) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(w;\xi)]\Big\}, \tag{43}$$

where ∇f is a unbiased gradient estimator of ∇F, i.e.,

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(w;\xi)] = \nabla F(w),\ \forall w \in \mathbb{R}^d.$$

The standard SGD method without mini-batch for solving (43) can be described as in Algorithm 2.

To analyze convergence rate of Algorithm 2, we assume that problem (43) satisfies Assumptions 1, 6, and 7.

Assumption 6 (L-weaker smooth)**.** *The objective function F of* (43) *satisfies, $\forall w, w' \in \mathbb{R}^d$,*

$$F(w) \leq F(w') + \langle \nabla F(w'), w - w'\rangle + \frac{L}{2}\|w - w'\|^2. \tag{44}$$

Algorithm 2 Stochastic Gradient Descent (SGD) Method (without mini-batch)

Initialize: Choose an initial point $w_1 \in \mathbb{R}^d$.
for $t = 1, 2, \cdots$ **do**
  Generate a realization of a random variable ξ_t and evaluate a stochastic gradient $\nabla f(w_t; \xi_t)$;
  Choose a step size (i.e., a learning rate) $\eta_t > 0$ (specified later);
  Update the new iterate $w_{t+1} := w_t - \eta_t \nabla f(w_t; \xi_t)$;
end for

Assumption 7. *For* (43)*, there exists a constant* $\sigma \in (0, +\infty)$ *such that* $\forall w \in \mathbb{R}^d$*, we have*

$$\mathbb{E}\left[\|\nabla f(w;\xi) - \nabla F(w)\|^2\right] \le \sigma^2. \tag{45}$$

We prove our first result for Algorithm 2 to solve (43) in the following theorem.

Theorem 7. *Assume that Assumptions 1, 6, and 7 hold for* (43)*. Let* $\{w_t\}$ *be the sequence generated by Algorithm 2 with* $0 < \eta_t := \frac{\gamma}{(t+\beta)^\alpha} \le \frac{1}{L}$ *for some* $\gamma > 0$*,* $\beta > 0$*, and* $\frac{1}{2} < \alpha < 1$*, and* $C := [F(w_1) - F_*] + \frac{L\sigma^2\gamma^2}{2(2\alpha-1)\beta^{2\alpha-1}} > 0$*. Then, the following bound holds:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(w_t)\|^2\right] \le \frac{2(1+\beta)^\alpha[F(w_1) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{C}{\alpha\gamma}\left(\frac{(T-1+\beta)^\alpha}{T}\right) + \frac{L\sigma^2\gamma}{(1-\alpha)}\left(\frac{(T+\beta)^{1-\alpha}}{T}\right). \tag{46}$$

Proof. Let $\mathcal{F}_t = \sigma(w_1, \cdots, w_t)$ be the σ-algebra generated by $w_1, \cdots, w_t$. From L-smooth property of F, we have

$$\begin{aligned}
\mathbb{E}[F(w_{t+1})|\mathcal{F}_t] &\le F(w_t) - \eta_t\|\nabla F(w_t)\|^2 + \frac{\eta_t^2 L}{2}\mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2|\mathcal{F}_t] \\
&= F(w_t) - \eta_t\left(1 - \frac{\eta_t L}{2}\right)\|\nabla F(w_t)\|^2 + \frac{\eta_t^2 L}{2}\mathbb{E}[\|\nabla f(w_t;\xi_t) - \nabla F(w_t)\|^2|\mathcal{F}_t] \\
&\overset{(45)}{\le} F(w_t) - \frac{\eta_t}{2}\|\nabla F(w_t)\|^2 + \frac{\eta_t^2 L}{2}\sigma^2,
\end{aligned}$$

where the first equality follows since $\mathbb{E}[\|\nabla f(w_t;\xi_t) - \nabla F(w_t)\|^2|\mathcal{F}_t] = \mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2|\mathcal{F}_t] - \|\nabla F(w_t)\|^2$; and the last inequality follows since F has bounded variance. Note that $\eta_t\left(1 - \frac{\eta_t L}{2}\right) \ge \frac{\eta_t}{2}$ since $0 < \eta_t \le \frac{1}{L}$. Subtracting F_* and taking the expectation to both sides, we have

$$\begin{aligned}
\mathbb{E}\left[F(w_{t+1}) - F_*\right] &\le \mathbb{E}\left[F(w_t) - F_*\right] - \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla F(w_t)\|^2\right] + \frac{\eta_t^2 L\sigma^2}{2}, \\
&\le \mathbb{E}\left[F(w_t) - F_*\right] + \frac{\eta_t^2 L\sigma^2}{2}.
\end{aligned} \tag{47}$$

Hence, taking the sum from $t = 1, \cdots, k$ to both sides, for $k \ge 1$, we have

$$\begin{aligned}
\mathbb{E}\left[F(w_{k+1}) - F_*\right] &\le \mathbb{E}\left[F(w_1) - F_*\right] + \frac{L\sigma^2}{2}\sum_{t=1}^{k}\eta_t^2 = [F(w_1) - F_*] + \frac{L\sigma^2}{2}\sum_{t=1}^{k}\frac{\gamma^2}{(t+\beta)^{2\alpha}} \\
&\overset{(22)}{\le} [F(w_1) - F_*] + \frac{L\sigma^2\gamma^2}{2}\int_{t=0}^{k}\frac{dt}{(t+\beta)^{2\alpha}}.
\end{aligned} \tag{48}$$

If $\frac{1}{2} < \alpha < 1$, then for $k \geq 1$

$$\mathbb{E}\left[F(w_{k+1}) - F_*\right] \leq \left[F(w_1) - F_*\right] + \frac{L\sigma^2\gamma^2}{2(2\alpha-1)\beta^{2\alpha-1}}, \tag{49}$$

Now, let us define $Y_t := \mathbb{E}\left[F(w_t) - F_*\right] \geq 0$, $Z_t := \mathbb{E}\left[\|\nabla F(w_t)\|^2\right] \geq 0$, for $t \geq 1$, $\rho := \frac{1}{2}$, and $D := \frac{L\sigma^2}{2}$. The estimate (47) becomes

$$Y_{t+1} \leq Y_t - \rho\eta_t Z_t + D\eta_t^2.$$

Let us define $C := \left[F(w_1) - F_*\right] + \frac{L\sigma^2\gamma^2}{2(2\alpha-1)\beta^{2\alpha-1}} > 0$. By (49), we have $Y_t \leq C$ (note that $H = 0$), $t \geq 1$. Applying Lemma 2 with $q = 2$, $m = 1$, and $\frac{1}{2} < \alpha < 1$. we conclude that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(w_t)\|^2\right] \leq \frac{2(1+\beta)^{\alpha}[F(w_1) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{C}{\alpha\gamma}\left(\frac{(T-1+\beta)^{\alpha}}{T}\right)\\ + \frac{L\sigma^2\gamma}{(1-\alpha)}\left(\frac{(T+\beta)^{1-\alpha}}{T}\right),\end{aligned}$$

which proves (46). □

Remark 5. *In Theorem 7, if we choose $\alpha = \frac{1}{2} + \delta$ for some $0 < \delta < \frac{1}{2}$, then we have*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(w_t)\|^2\right] &\leq \frac{2(1+\beta)^{\frac{1}{2}+\delta}[F(w_1) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{C}{\gamma(\frac{1}{2}+\delta)}\left(\frac{(T-1+\beta)^{\frac{1}{2}+\delta}}{T}\right)\\ &\quad + \frac{L\sigma^2\gamma}{(\frac{1}{2}-\delta)}\left(\frac{(T+\beta)^{\frac{1}{2}-\delta}}{T}\right).\\ &= \mathcal{O}\left(\frac{1}{T^{\frac{1}{2}-\delta}}\right),\end{aligned}$$

where $C := \left[F(w_1) - F_*\right] + \frac{L\sigma^2\gamma^2}{4\delta\beta^{4\delta}} > 0$.

Theorem 8. *Assume that Assumptions 1, 6, and 7 hold for* (43)*. Let $\{w_t\}$ be the sequence generated by Algorithm 2 with the step-size $0 < \eta_t := \frac{\gamma}{(t+\beta)^{1/2}} \leq \frac{1}{L}$ for some $\gamma > 0$ and $\beta > 0$, and $C := \left[F(w_1) - F_*\right] + \frac{L\sigma^2\gamma^2}{2(1+\beta)} > 0$. Then, the following bound holds:*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(w_t)\|^2\right] &\leq \frac{2(1+\beta)^{1/2}[F(w_1) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{2C}{\gamma}\left(\frac{(T-1+\beta)^{1/2}}{T}\right)\\ &\quad + L\gamma\sigma^2\left(\frac{(T-1+\beta)^{1/2}\ln(T+1+\beta)}{T}\right) + 2L\gamma\sigma^2\left(\frac{(T+\beta)^{1/2}}{T}\right)\\ &= \quad \mathcal{O}\left(\frac{\ln(T)}{T^{1/2}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{T^{1/2}}\right).\end{aligned} \tag{50}$$

Proof. If $\alpha = \frac{1}{2}$, by (48), we have

$$\mathbb{E}\left[F(w_{k+1}) - F_*\right] \leq \mathbb{E}\left[F(w_1) - F_*\right] + \frac{L\sigma^2}{2}\sum_{t=1}^{k}\eta_t^2 = \left[F(w_1) - F_*\right] + \frac{L\sigma^2}{2}\sum_{t=1}^{k}\frac{\gamma^2}{(t+\beta)}$$
$$\overset{(22)}{\leq} \left[F(w_1) - F_*\right] + \frac{L\sigma^2\gamma^2}{2(1+\beta)} + \frac{L\sigma^2\gamma^2}{2}\int_{t=1}^{k}\frac{dt}{(t+\beta)^{2\alpha}}.$$

Hence, for $k \geq 1$, we have

$$\mathbb{E}\left[F(w_{k+1}) - F_*\right] \leq \left[F(w_1) - F_*\right] + \frac{L\sigma^2\gamma^2}{2(1+\beta)} + \frac{L\sigma^2\gamma^2}{2}\ln(k+2+\beta), \tag{51}$$

Define $Y_t := \mathbb{E}\left[F(w_t) - F_*\right] \geq 0$, $Z_t := \mathbb{E}\left[\left\|\nabla F(w_t)\right\|^2\right] \geq 0$ for $t \geq 1$, $\rho := \frac{1}{2}$, and $D := \frac{L\sigma^2}{2}$. The estimate (47) becomes

$$Y_{t+1} \leq Y_t - \rho\eta_t Z_t + D\eta_t^2.$$

Let us define $C := \left[F(w_1) - F_*\right] + \frac{L\sigma^2\gamma^2}{2(1+\beta)} > 0$, $H := \frac{L\sigma^2\gamma^2}{2} > 0$, and $\theta := 1 + \beta > 0$. Clearly, $1 + \theta - \beta = 2 > \frac{1}{2}e$. By (37), we have $Y_t \leq C + H\ln(t+\theta)$ for $t \geq 1$. Applying Lemma 2 with $q = 2$, $m = 1$, and $\alpha = \frac{1}{2}$, we conclude that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(w_t)\right\|^2\right] \leq{}& \frac{2(1+\beta)^{1/2}[F(w_1) - F_*]}{\gamma}\cdot\frac{1}{T} + \frac{2C}{\gamma}\left(\frac{(T-1+\beta)^{1/2}}{T}\right)\\ &+ L\gamma\sigma^2\left(\frac{(T-1+\beta)^{1/2}\ln(T+1+\beta)}{T}\right) + 2L\gamma\sigma^2\left(\frac{(T+\beta)^{1/2}}{T}\right)\\ ={}& \mathcal{O}\left(\frac{\ln(T)}{T^{1/2}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{T^{1/2}}\right),\end{aligned}$$

which proves (50). □