When Does Stochastic Gradient Algorithm Work Well?

LAM M. NGUYEN

Department of Industrial and Systems Engineering, Lehigh University, USA

NAM H. NGUYEN

IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

DZUNG T. PHAN

IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

JAYANT R. KALAGNANAM

IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

KATYA SCHEINBERG

Department of Industrial and Systems Engineering, Lehigh University, USA

LEHIGH
U N I V E R S I T Y.

When Does Stochastic Gradient Algorithm Work Well?

Lam M. Nguyen   Nam H. Nguyen   Dzung T. Phan   Jayant R. Kalagnanam

Katya Scheinberg

January 20, 2018

Abstract

In this paper, we consider a general stochastic optimization problem which is often at the core of supervised learning, such as deep learning and linear classification. We consider a standard stochastic gradient descent (SGD) method with a fixed, large step size and propose a novel assumption on the objective function, under which this method has the improved convergence rates (to a neighborhood of the optimal solutions). We then empirically demonstrate that these assumptions hold for logistic regression and standard deep neural networks on classical data sets. Thus our analysis helps to explain when efficient behavior can be expected from the SGD method in training classification models and deep neural networks.

1 Introduction and Motivation

In this paper we are interested in analyzing behavior of the stochastic gradient algorithm when solving empirical and expected risk minimization problems. For the sake of generality we consider the following stochastic optimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ F(\mathbf{w}) = \mathbb{E}[f(\mathbf{w}; \xi)] \right\}, \tag{1}$$

where ξ is a random variable obeying some distribution.

In the case of empirical risk minimization with a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, ξ_i is a random variable that is defined by a single random sample $(\mathbf{x}, \mathbf{y})$ drawn uniformly from the training set. Then, by defining $f_i(\mathbf{w}) := f(\mathbf{w}; \xi_i)$ we write the empirical risk minimization as follows:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{w}) \right\}. \tag{2}$$

Lam M. Nguyen, Industrial and Systems Engineering, Lehigh University, PA, USA. Email: LamNguyen.MLTD@gmail.com

Nam H. Nguyen, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Email: nnguyen@us.ibm.com

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

Jayant R. Kalagnanam, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Email: jayant@us.ibm.com

Katya Scheinberg, Industrial and Systems Engineering, Lehigh University, PA, USA. Email: katyas@lehigh.edu. The work of this author is partially supported by NSF Grants CCF 16-18717 and CCF 17-40796

More generally ξ can be a random variable defined by a random subset of samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in I}$ drawn from the training set, in which case formulation (1) still applies to the empirical risk minimization. On the other hand, if ξ represents a sample or a set of samples drawn from the data distribution, then (1) represents the expected risk minimization.

Stochastic gradient descent (SGD), originally introduced in [10], has become the method of choice for solving not only (1) but also (2) when n is large. Theoretical justification for using SGD for machine learning problems is given, for example, in [1], where it is shown that, at least for convex problem, SGD is an optimal method for minimizing expected risk, which is the ultimate goal of learning. From the practical perspective SGD is often preferred to the standard gradient descent (GD) method simply because GD requires computation of a full gradient on each iteration, which, for example, in the case of deep neural networks (DNN), requires applying backpropagation for all n samples, which can be prohibitive.

Consequently, due to its simplicity in implementation and efficiency in dealing with large scale datasets, SGD has become by far the most common method for training deep neural networks and other large scale ML models. However, it is well known that SGD can be slow and unreliable in some practical applications as its behavior is strongly dependent on the chosen stepsize and on the variance of the stochastic gradients. While the method may provide fast initial improvement, it may slow down drastically after a few epochs and can even fail to move close enough to a solution for a fixed learning rate. To overcome this oscillatory behavior, several variants of SGD have been recently proposed. For example, methods such as AdaGrad [4], RMSProp [12], and Adam [6] adaptively select the stepsize for each component of $\mathbf{w}$. Other techniques include diminishing stepsize scheme [2] and variance reduction methods [11, 3, 5, 9]. These latter methods reduce the variance of the stochastic gradient estimates, by either computing a full gradient after a certain number of iterations or by storing the past gradients, both of which can be expensive. Moreover, these methods only apply to the finite sum problem (2) but not the general problem (1). On the other hand these methods enjoy faster convergence rates than that of SGD. For example, when $F(\mathbf{w})$ is strongly convex, convergence rates of the variance reduction methods (as well as that of GD itself) are linear, while for SGD it is only sublinear. While GD has to compute the entire gradient on *every* iteration, which makes it more expensive than the variance reduction methods, its convergence analysis allows for a much larger fixed stepsizes than those allowed in the variance reduction methods. In this paper we are particularly interested in addressing an observation: a simple SGD with a fixed, reasonably large, step size can have a fast convergence rate to some neighborhood of the optimal solutions, without resorting to additional procedures for variance reduction.

Let us consider an example of recovering a signal $\hat{\mathbf{w}} \in \mathbb{R}^2$ from n noisy observations $y_i = y_i^{\text{clean}} + e_i$ where $y_i^{\text{clean}} = (\mathbf{a}_i^\mathsf{T} \hat{\mathbf{w}})^2$. Here, $\mathbf{a}_i$'s are random vectors and e_i's are noise components. To recover $\hat{\mathbf{w}}$ from the observation vector $\mathbf{y}$, we solve a non-convex fourth-order polynomial minimization problem

$$\min_{\mathbf{w}} \left\{ F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (y_i - (\mathbf{a}_i^\mathsf{T} \mathbf{w})^2)^2 \right\}.$$

Note that there are at least two global solutions to this problem, which we denote w_* and $-w_*$. We consider two possible scenarios:

(i) *All* of the component functions $f_i(\mathbf{w}) = (y_i - (\mathbf{a}_i^\mathsf{T} \mathbf{w})^2)^2$ have *relatively small* gradients at both of the optimal solutions $\mathbf{w}_*$ and $-\mathbf{w}_*$ of the aggregate $F(\mathbf{w})$. In this case this means that w_* recovers a good fit for the observations y.

(ii) There are *many* indices i such that at the optimal solutions of $F(\mathbf{w})$, the associated gradients ∇f_i

are *large*. This happens when w_* does not provide a good fit, which can happen when the noise e_i is large.
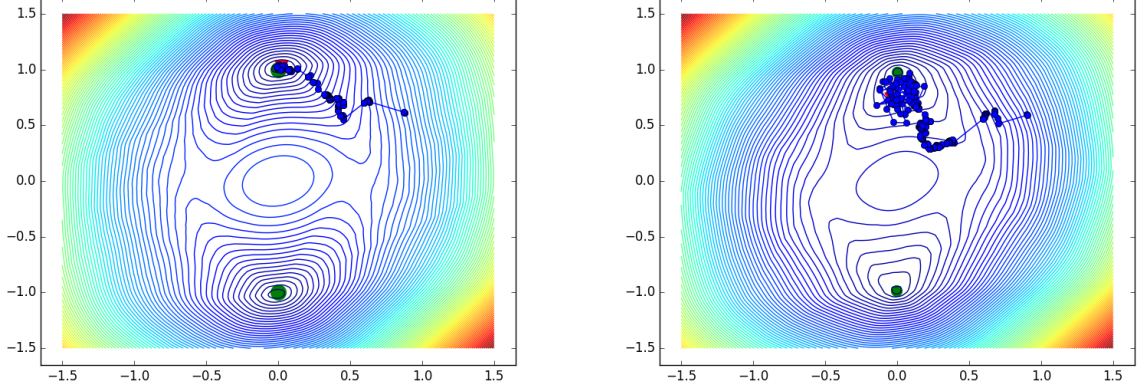

Figure 1: Stochastic Gradient Descent

We set $n = 100$ and generate these two scenarios by setting all the noise components e_i to be small (1% of the energy of $\mathbf{y}^{\text{clean}}$) for case (i) or setting only first 40 noise components to be large (25% of the energy of $\mathbf{y}^{\text{clean}}$) for case (ii). We can observe from Figure 1 that SGD algorithm converges to the optimal solution of $F(\mathbf{w})$ in case (i) depicted in the left figure; but fails to converge to the solution of F in case (ii) as shown in the right figure. The intuition behind this behavior is as follows. At every step of SGD, the iterate moves towards to the optimal solutions of the individual component function that has been randomly chosen on this iteration. If a majority of component functions f_i have their optimal solutions close to the optimum of the entire problem F, then SGD effectively acts as GD. On the other hand, if the optimal solutions of a lot of $f_i's$ are far from each other and from the overall optimum, then iterates of SGD wander randomly in the region around these individual optima, as shown on the right of Figure 1. Hence, SGD cannot work effectively in case (ii), unless we either reduce the learning rate or reduce the variance of the steps thus attaining more accurate gradient information.

In this paper we generalize this result for stochastic problem (1) under much weaker assumptions. In particular, we do not assume that the gradients *vanish* at the solution, but that they are bounded by some small constant. Moreover, we do not impose this property on *all* stochastic gradients, but assume that it holds with suitably large probability. We then show that SGD has fast convergence rates in the strongly convex, convex and nonconvex cases, until some accuracy is reached, where this accuracy is dictated by the behavior of the stochastic gradients at the optimal solution.

We conjecture that success of SGD for training many machine learning models is the result of the associated optimization problems having this properties - most of the component gradients are suitably small at the solution. To verify this claim, we trained linear classifiers (via logistic regression) and standard neural networks on several well-known datasets and subsequently computed the fraction of individual gradients $\nabla f_i(\mathbf{w}_*)$ at the final solution $\mathbf{w}_*$ of F, that were small. The results show that more than 99% of component functions f_i have the vanishing gradient at $\mathbf{w}_*$. More numerical evidence is presented in the Section 3.

Hence we base our analysis on the following observation.

Main observation. *For many classification problems in supervised learning, majority of component func-*

3

tions f_i have small gradients at the optimal solution w_* (in the convex case) or at local minima of $F(\mathbf{w})$ (in the nonconvex case)

In this paper, based on this observation, we provide theoretical analysis of SGD under the assumption on the fraction of components with small gradient at the solution. Our analysis helps explain the good performance of SGD when applied to deep learning. We summarize the key contributions of the paper as follows.

- We conjecture that in many instances of empirical risk minimization and expected risk minimization SGD converges to a neighborhood of a stationary point of $F(\mathbf{w})$ such that the majority of component functions f_i have small gradients at that point. We verify numerically that this conjecture holds true for logistic regression and standard deep neural networks on a wide range of data sets.

- We formalize this conjecture as a condition under which we are able to establish improved convergence rates of SGD with fixed, large step size to a neighborhood of such stationary point when $F(\mathbf{w})$ is strongly convex, convex and nonconvex.

- Thus we establish that SGD converges fast to a neighborhood of the expected/empirical risk minimizer and that the size of the neighborhood is determined by some properties of the distribution of the stochastic gradient at the minimizer.

The remainder of the paper is organized as follows. The main convergence analysis for all three cases is carried out in Section 2. The computational evidence is presented in Section 3 and implications of our analysis and findings are summarized in Section 4. The proofs are presented in the Appendix.

2 Convergence Analyses of Stochastic Gradient Algorithms

In this section, we analyze the stochastic gradient descent algorithm (SGD) under a novel condition, based on the observations of the previous section, and derive improved convergence rates for the strongly convex, convex, and non-convex cases. We present each result in the form of a general theorem with the bound on a certain optimality measure (depending on the case), followed by the corollary where we demonstrate that improved convergence rate can be observed until this optimality measure becomes small. The rate and the threshold for optimality measure are dictated by the properties of the stochastic gradient at the solution.

First we introduce the basic definition of L-smoothness.

Definition 1. *A function ϕ is L-smooth if there exists a constant $L > 0$ such that*

$$\|\nabla\phi(\mathbf{w}) - \nabla\phi(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \ \forall \ \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{3}$$

For completeness, we state the SGD algorithm as Algorithm 1.

Let $\mathcal{F}_t = \sigma(\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t)$ be the σ-algebra generated by $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t$. We note that $\{\xi_{t,i}\}_{i=1}^{b}$ are independent of $\mathcal{F}_t$. Since $\{\xi_{t,i}\}_{i=1}^{b}$ are i.i.d.[1] with $\mathbb{E}[\nabla f(\mathbf{w}_t; \xi_{t,i})|\mathcal{F}_t] = \nabla F(\mathbf{w}_t)$, we have an unbiased estimate of gradient $\mathbb{E}[\mathbf{g}_t|\mathcal{F}_t] = \frac{1}{b}\sum_{i=1}^{b} \nabla F(\mathbf{w}_t) = \nabla F(\mathbf{w}_t)$.

---

[1]Independent and identically distributed random variables. We note from probability theory that if $X_1, \ldots, X_d$ are i.i.d. random variables then $g(X_1), \ldots, g(X_d)$ are also i.i.d. random variables if g is measurable function.

Algorithm 1 Stochastic Gradient Descent (SGD) Algorithm with fixed step size

---

Initialize $\mathbf{w}_0$, **choose** stepsize $\eta > 0$, and batch size b.
for $t = 0, 1, 2, \ldots$ **do**
    Generate random variables $\{\xi_{t,i}\}_{i=1}^b$ i.i.d. with $\mathbb{E}[\nabla f(\mathbf{w}_t; \xi_{t,i})|\mathcal{F}_t] = \nabla F(\mathbf{w}_t)$.
    Compute a stochastic gradient

$$\mathbf{g}_t = \frac{1}{b} \sum_{i=1}^b \nabla f(\mathbf{w}_t; \xi_{t,i}).$$

    Update the new iterate $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$.
end for

We now define the quantities that will be useful in our results.

Definition 2. *Let $\mathbf{w}_*$ be a stationary point of the objective function $F(\mathbf{w})$. For any given threshold $\epsilon > 0$, define*

$$p_\epsilon := \mathbb{P}\left\{ \|\mathbf{g}_*\|^2 \leq \epsilon \right\}, \tag{4}$$

where $\mathbf{g}_* = \frac{1}{b} \sum_{i=1}^b \nabla f(\mathbf{w}_*; \xi_i)$, as the probability that event $\|\mathbf{g}_*\|^2 \leq \epsilon$ happens for some i.i.d. random variables $\{\xi_i\}_{i=1}^b$. We also define

$$M_\epsilon := \mathbb{E}\left[\|\mathbf{g}_*\|^2 \mid \|\mathbf{g}_*\|^2 > \epsilon \right]. \tag{5}$$

The quantity p_ϵ measures the probability that event $\|\mathbf{g}_*\|^2 \leq \epsilon$ happens for some realizations of random variables ξ_i, $i = 1, \ldots, b$. Clearly, p_ϵ is bounded above by 1 and monotonically increasing with respect to ϵ. Quantity M_ϵ can be interpreted as the average bound of large components $\|\nabla f(\mathbf{w}_*; \xi)\|^2$. As we will see in our results below, quantities p_ϵ and M_ϵ appear in the convergence rate bound of the SGD algorithm. M_ϵ is also bounded above by $M_{max} = \max_\xi \|\nabla f(\mathbf{w}_*; \xi)\|^2$, which we assume is finite, hence in all our results we can replace M_ϵ by M_{max} if we want to eliminate its dependence on ϵ. On the other hand, the dependence of quantity p_ϵ on ϵ is key for our analysis. Based on the evidence shown in Section 3, we expect p_ϵ to be close to 1 for all but very small values of ϵ. We will derive our convergence rate bounds in terms of $\max\{\epsilon, 1 - p_\epsilon\}$. Clearly, as ϵ decreases, $1 - p_\epsilon$ increases and vice versa, but if there exists a small ϵ for which $1 - p_\epsilon \approx \epsilon$ then our results show convergence of SGD to an $\mathcal{O}(\epsilon)$ neighborhood of the solution, at an improved rate with respect to ϵ.

2.1 Convex objectives

In this section, we analyze the SGD method in the context of minimizing a convex objective function. We will bound the expected optimality gap at a given iterate in terms of the value of p_ϵ. First, we consider the case when F is strongly convex.

Definition 3. *A function ϕ is μ-strongly convex if there exists a constant $\mu > 0$ such that*

$$\phi(\mathbf{w}) - \phi(\mathbf{w}') \geq \nabla \phi(\mathbf{w}')^\top (\mathbf{w} - \mathbf{w}') + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|^2, \ \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{6}$$

Using this definition, we state the following result for the strongly convex case.

Theorem 1. *Suppose that $F(\mathbf{w})$ is μ-strongly convex and $f(\mathbf{w}; \xi)$ is L-smooth and convex for every realization of ξ. Consider Algorithm 1 with $\eta \leq \frac{1}{L}$. Then, for any $\epsilon > 0$*

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq (1 - \mu\eta(1 - \eta L))^t \|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu(1 - \eta L)} p_\epsilon \epsilon + \frac{2\eta}{\mu(1 - \eta L)}(1 - p_\epsilon) M_\epsilon, \quad (7)$$

where $\mathbf{w}_* = \arg\min_{\mathbf{w}} F(\mathbf{w})$, and p_ϵ and M_ϵ are defined in (4) and (5), respectively.

The main conclusion is stated in the following corollary.

Corollary 1. *For any ϵ such that $1 - p_\epsilon \leq \epsilon$, and for Algorithm 1 with $\eta \leq \frac{1}{2L}$, we have*

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq (1 - \mu\eta)^t \|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu}(1 + M_\epsilon)\epsilon.$$

Furthermore if $t \geq T$ for $T = \frac{1}{\mu\eta} \log\left(\frac{\mu\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 + M_\epsilon)\epsilon}\right)$, then

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq \frac{4\eta}{\mu}(1 + M_\epsilon)\epsilon. \quad (8)$$

Note that in Corollary 1 we assume that $\eta \leq \frac{1}{2L}$ instead of $\eta \leq \frac{1}{L}$ only to simplify the expressions. (The proof in detail is in the Appendix.) We conclude that under the assumption $1 - p_\epsilon \leq \epsilon$, Algorithm 1 has linear convergence rate in terms of any such ϵ.

The following theorem establishes convergence rate bound for Algorithm 1 when the strong convexity assumption on $F(\mathbf{w})$ is relaxed.

Theorem 2. *Suppose that $f(\mathbf{w}; \xi)$ is L-smooth and convex for every realization of ξ. Consider Algorithm 1 with $\eta < \frac{1}{L}$. Then for any $\epsilon > 0$, we have*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 - \eta L)t} + \frac{\eta}{(1 - \eta L)} p_\epsilon \epsilon + \frac{\eta M_\epsilon}{(1 - \eta L)}(1 - p_\epsilon), \quad (9)$$

where $\mathbf{w}_*$ is any optimal solution of $F(\mathbf{w})$, and p_ϵ and M_ϵ are defined in (4) and (5), respectively.

Again, the convergence rate of SGD is governed by the initial solution and quantities p_ϵ and M_ϵ. Hence we have the following corollary.

Corollary 2. *If $f(\mathbf{w}; \xi)$ is L-smooth and convex for every realization of ξ, then for any ϵ such that $1 - p_\epsilon \leq \epsilon$, and $\eta \leq \frac{1}{2L}$, it holds that*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\eta t} + 2\eta(1 + M_\epsilon)\epsilon.$$

Hence, if $t \geq T$ for $T = \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{(2\eta^2)(1 + M_\epsilon)\epsilon}$, we have

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq 4\eta(1 + M_\epsilon)\epsilon. \quad (10)$$

Similarly to the strongly convex case, under the key assumption that $1 - p_\epsilon \leq \epsilon$, we show that Algorithm 1 achieves $\mathcal{O}(\epsilon)$ optimality gap, in expectation, in $\mathcal{O}(1/\epsilon)$ iterations. In Corollary 2 we again assume that $\eta \leq \frac{1}{2L}$ instead of $\eta < \frac{1}{L}$ only to simplify the expressions and to replace $\frac{1}{1 - \eta L}$ term with 2 in the complexity bound.

2.2 Nonconvex objectives

In this section, we establish expected complexity bound for Algorithm 1 when applied to nonconvex objective functions. This setting includes deep neural networks in which the cost function is a sum of nonconvex function components. Despite the nonconvexity of the objective, it has been observed that deep neural networks can be trained fairly quickly by SGD algorithms. It has also been observed that after reaching certain accuracy, the SGD algorithm may slow down dramatically.

For the analysis of the nonconvex case, we need to make an assumption on the rate of change in the gradients near all local solutions, or at least those to which iterates $\mathbf{w}_t$ generated by the algorithm may converge.

Assumption 1. *We assume that there exists a constant $N > 0$, such that for any sequence of iterates $\mathbf{w}_0$, $\mathbf{w}_1$, ..., $\mathbf{w}_t$ of any realization of Algorithm 1, there exists a stationary point $\mathbf{w}_*$ of $F(\mathbf{w})$ (possibly dependent on that sequence) such that*

$$\frac{1}{t+1} \sum_{k=0}^{t} \left(\mathbb{E} \left[\left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_k; \xi_{k,i}) - \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_*; \xi_{k,i}) \right\|^2 \Big| \mathcal{F}_k \right] \right) \leq N \frac{1}{t+1} \sum_{k=0}^{t} \|\nabla F(\mathbf{w}_k)\|^2, \quad (11)$$

where the expectation is taken over random variables $\xi_{k,i}$ conditioned on $\mathcal{F}_t = \sigma(\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t)$, which is the σ-algebra generated by $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t$. Let $\mathcal{W}_*$ denote the set of all such stationary points $\mathbf{w}_*$, determined by the constant N and by realizations $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t$.

This assumption is made for any realization $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t$ and states that the average squared norm of the difference between the stochastic gradient directions computed by Algorithm 1 at $\mathbf{w}_t$ and the same stochastic gradient computed at $\mathbf{w}_*$, over any t iterations, is proportional to the average true squared gradient norm. If $\mathbf{w}_*$ is a stationary point for all $f(\mathbf{w}; \xi_{k,i})$, in other words, $\nabla f(\mathbf{w}_*; \xi_{k,i}) = 0$ for all realizations of $\xi_{k,i}$, then Assumption 1 simply states that all stochastic gradients have the same average expected rate of growth as the true gradient, as the iterates get further away from $\mathbf{w}_*$. Notice that $\mathbf{w}_*$ may not be a stationary point for all $f(\mathbf{w}; \xi_{k,i})$, hence Assumption 1 bounds the average expected rate of *change* of the stochastic gradients in terms of the rate of change of the true gradient. In the next section we will demonstrate numerically that Assumption 1 holds for problems of training deep neural networks.

We also need to slightly modify Definition 2.

Definition 4. *Let $\mathbf{g}_* = \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_*; \xi_i)$ for some i.i.d. random variables $\{\xi_i\}_{i=1}^{b}$. For any given threshold $\epsilon > 0$, define*

$$p_\epsilon := \inf_{\mathbf{w}_* \in \mathcal{W}_*} \mathbb{P} \left\{ \|\mathbf{g}_*\|^2 \leq \epsilon \right\}, \quad (12)$$

where the infimum is taken over the set $\mathcal{W}_*$ defined in Assumption 1. Similarly, we also define

$$M_\epsilon := \sup_{\mathbf{w}_* \in \mathcal{W}_*} \mathbb{E} \left[\|\mathbf{g}_*\|^2 \mid \|\mathbf{g}_*\|^2 > \epsilon \right]. \quad (13)$$

We know that p_ϵ and M_ϵ defined as above exist since $p_\epsilon \geq 0$ and $M_\epsilon \leq M_{max}$. This time, if we assume that $1 - p_\epsilon \leq \epsilon$ for some reasonably small ϵ, this implies that for *all* stationary points of $F(\mathbf{w})$ that appear in Assumption 1 a large fraction of stochastic gradients have small norm at those points. Essentially, $\mathcal{W}_*$ consists of stationary points to which different realization of SGD iterates converge.

Theorem 3. *Let Assumption 1 hold for some $N > 0$. Suppose that F is L-smooth. Consider Algorithm 1 with $\eta < \frac{1}{LN}$. Then, for any $\epsilon > 0$, we have*

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq \frac{[F(\mathbf{w}_0) - F^*]}{\eta(1 - L\eta N)(t+1)} + \frac{L\eta}{(1 - L\eta N)}\epsilon + \frac{L\eta M_\epsilon}{(1 - L\eta N)}(1 - p_\epsilon),$$

where F^* is any lower bound of F; and p_ϵ and M_ϵ are defined in (12) and (13) respectively.

Corollary 3. *Let Assumption 1 hold and p_ϵ and M_ϵ be defined as in (12) and (13). For any ϵ such that $1 - p_\epsilon \leq \epsilon$, and for $\eta \leq \frac{1}{2LN}$, we have*

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq \frac{2[F(\mathbf{w}_0) - F^*]}{\eta(t+1)} + 2L\eta(1 + M_\epsilon)\epsilon.$$

Hence, if $t \geq T$ for $T = \frac{[F(\mathbf{w}_0) - F^*]}{(L\eta^2)(1 + M_\epsilon)\epsilon}$, we have

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq 4L\eta(1 + M_\epsilon)\epsilon.$$

3 Numerical Experiments

The purpose of this section is to numerically validate our assumptions on p_ϵ as defined in Definition 2. We wish to show that there exists a small ϵ satisfying

$$1 - p_\epsilon \approx \epsilon. \tag{14}$$

For our numerical experiments, we consider the finite sum minimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^d}\left\{F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}f_i(\mathbf{w})\right\}. \tag{15}$$

Definition 5. *Let $\mathbf{w}_*$ be a stationary point of the objective function $F(\mathbf{w})$. For any given threshold $\epsilon > 0$, define the set $\mathcal{S}_\epsilon$ and its complement $\mathcal{B}_\epsilon$*

$$\mathcal{S}_\epsilon := \left\{i : \|\nabla f_i(\mathbf{w}_*)\|^2 \leq \epsilon\right\} \qquad \text{and} \qquad \mathcal{B}_\epsilon := [n]\backslash\mathcal{S}_\epsilon.$$

We also define the quantity $p_\epsilon := \frac{|\mathcal{S}_\epsilon|}{n}$ that measures the size of the set $\mathcal{S}_\epsilon$ and the upper bound M_ϵ

$$\frac{1}{|\mathcal{B}_\epsilon|}\sum_{i \in \mathcal{B}_\epsilon}\|\nabla f_i(\mathbf{w}_*)\|^2 \leq M_\epsilon.$$

3.1 Logistic Regression for Convex Case

We consider ℓ_2-regularized logistic regression problems with

$$f_i(\mathbf{w}) = \log(1 + \exp(-y_i\mathbf{x}_i^\top\mathbf{w})) + \frac{\lambda}{2}\|\mathbf{w}\|^2,$$

where the penalty parameter λ is set to $1/n$, a widely-used value in the literature [9]. We conducted experiments on popular datasets `covtype, ijcnn1, w8a, a9a, mushrooms, phishing, skin_nonskin` from the LIBSVM website [2] and `ijcnn2` [3]. The optimal solution $\mathbf{w}_*$ of the convex problem (15) is found by using the full-batch `L-BFGS` method [7] with the stopping criterion $\|\nabla F(\mathbf{w}_*)\|_2 \leq 10^{-12}$. We then ran Algorithm 1 using the learning rate $\eta = 10^{-1}$ and the batch-size $b = 1$ and 100 epochs. The final solution given by the SGD algorithm is denoted by $\mathbf{w}_{SGD}$. We report the value of p_ϵ defined in Definition 5 expressed in percentage form for different values of ϵ.

As we can see from Table 1 that $\epsilon = 10^{-3}$ satisfies (14) for all cases. For datasets `covtype, ijcnn1, ijcnn2, phishing` and `skin_nonskin`, ϵ can take a smaller value 10^{-4}. The small value for ϵ indicates that SGD with a fixed step size can converge to a small neighborhood of the optimal solution of F. The success of using SGD is illustrated, optimality gaps $F(\mathbf{w}_{SGD}) - F(\mathbf{w}_*)$ are small in our experiments.

Table 1: Percentage of f_i with small gradient value for different threshold ϵ (Logistic Regression)

Datasets	$F(\mathbf{w}_{SGD}) - F(\mathbf{w}_*)$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-4}$	$\epsilon = 10^{-5}$	$\epsilon = 10^{-6}$	Train accuracy
covtype	$5 \cdot 10^{-4}$	100%	100%	100%	99.9995%	54.9340%	0.7562
ijcnn1	$1 \cdot 10^{-4}$	100%	100%	100%	96.8201%	89.0197%	0.9219
ijcnn2	$2 \cdot 10^{-4}$	100%	100%	100%	99.2874%	90.4565%	0.9228
w8a	$8 \cdot 10^{-5}$	100%	99.9899%	99.4231%	98.3557%	92.7818%	0.9839
a9a	$4 \cdot 10^{-3}$	100%	100%	84.0945%	58.5824%	40.0909%	0.8491
mushrooms	$3 \cdot 10^{-5}$	100%	100%	99.9261%	98.7568%	94.4239%	1.0000
phishing	$2 \cdot 10^{-4}$	100%	100%	100%	89.9231%	73.8128%	0.9389
skin_nonskin	$4 \cdot 10^{-5}$	100%	100%	100%	99.6331%	91.3730%	0.9076

We compare convergence rates of SGD (learning rate $\eta = 0.1 < \frac{1}{2L}$) with `SVRG` [5] and `L-BFGS` [7] as shown in Figure 2. We can observe that SGD has better performance than SVRG and L-BFGS in the beginning until it achieves $\mathcal{O}(\epsilon)$ accuracy, for the value of ϵ consistent to what is indicated in Table 1. We note that the values of M_ϵ for all datasets should not exceed 10^{-2} according to Table 1.

3.2 Neural Networks for Nonconvex Case

For experiments with nonconvex problems we train DNNs using two standard network architectures: feed forward network (FFN) and convolutional neural network (CNN). Configuration of FNN includes 2 dense layers each containing 256 neurons followed by a ReLU activation. The output layer consists of c neurons with the softmax activation where c is the number of classes. For CNN, we configure the network to have 2 convolutional layers followed by 2 dense layers. Convolutional layers contain a convolutional operator followed by a ReLU activation and then a max pooling. The number of filters of both the convolutional operators are set to 64 and the associated filter sizes are 5×5. Number of neurons in dense layers are 384 and 192, respectively, and the activation used in these layers is again ReLU. Throughout the simulations, we use popular datasets which include `MNIST` [4] (60000 training data images of size 28×28 contained in 10 classes), `SVHN` [5] (73257 training images of size 32×32 contained in 10 classes), `CIFAR10` (50000 training color images of size 32×32 contained in 10 classes), and `CIFAR100` [6] (50000 training color images of size 32×32 contained in 100 classes).

---

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[3] http://mlbench.org/repository/data/viewslug/ijcnn1/

[4] http://yann.lecun.com/exdb/mnist/

[5] http://ufldl.stanford.edu/housenumbers/

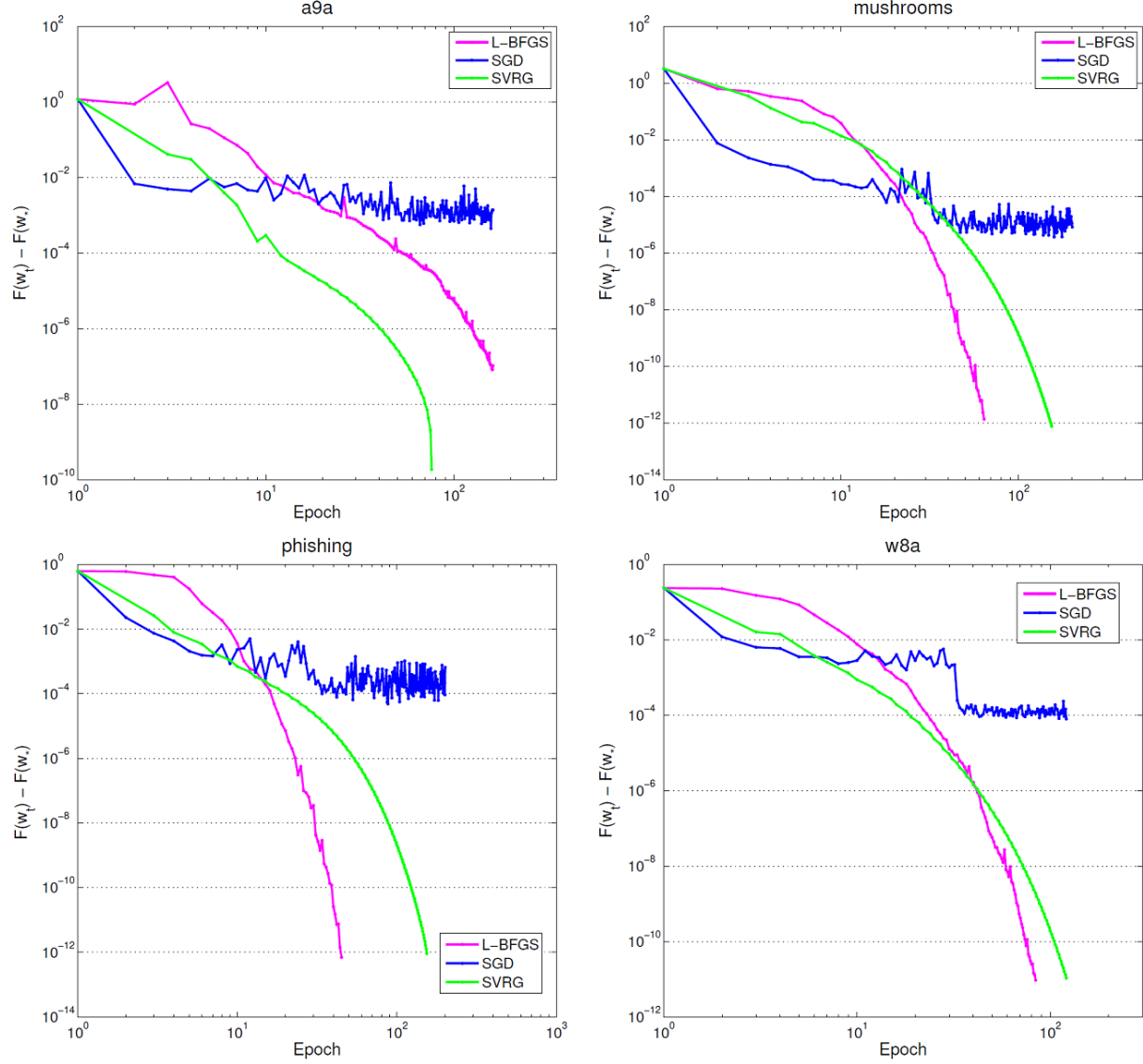[6] https://www.cs.toronto.edu/ kriz/cifar.html

Figure 2: The convergence comparisons of SGD, SVRG, and L-BFGS

We trained the networks by the popular Adam algorithm with a minibatch of size 32 and reported the values of p_ϵ at the last iteration $\mathbf{w}_{Adam}$. In all our experiments, we did not apply batch normalization and dropout techniques during the training. Since the problem of interest is nonconvex, multiple local minima could exist. We experimented with 10 seeds and reported the minimum result (minimum of the percentage of component functions with small gradient value). Table 2 shows the values of p_ϵ in terms of percentage for different thresholds ϵ. As is clear from the table, p_ϵ is close to 1 for a sufficiently small ϵ. It confirms that the majority of component functions f_i has negligible gradients at the final solution of F.

Table 2: Percentage of f_i with small gradient value for different threshold ϵ (Neural Networks)

Datasets	Architecture	$\|\nabla F(\mathbf{w}_{Adam})\|^2$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-5}$	$\epsilon = 10^{-7}$	Train accuracy	N	M
MNIST	**FFN**	$1.3 \cdot 10^{-15}$	100%	100%	99.99%	1.0000	6500	$2.1 \cdot 10^{-8}$
SVHN	**FFN**	$3.5 \cdot 10^{-3}$	99.94%	99.92%	99.91%	0.9997	12000	500
MNIST	**CNN**	$1.6 \cdot 10^{-17}$	100%	100%	100%	1.0000	6083	$6.4 \cdot 10^{-8}$
SVHN	**CNN**	$8.1 \cdot 10^{-7}$	99.99%	99.98%	99.96%	0.9999	8068	0.18
CIFAR10	**CNN**	$5.1 \cdot 10^{-20}$	100%	100%	100%	1.0000	1205	$8.7 \cdot 10^{-14}$
CIFAR100	**CNN**	$5.5 \cdot 10^{-2}$	99.50%	99.45%	99.42%	0.9988	984	3000

The value of N is the estimation of N in (11), which is shown in Section 3.3. We note that for some datasets and network structures, Adam did not converge to a real local solution (SVHN-FFN and CIFAR100-CNN) and Table 2 shows only an approximation of the behavior at the local solution.

3.3 Nonconvex assumption verification

This section shows how to estimate N. We are proving some numerical experiments to verify Assumption 1. Let us define

$$r_t = \frac{\frac{1}{t+1} \sum_{k=0}^{t} \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_*)\|^2 \right)}{\frac{1}{t+1} \sum_{k=0}^{t} \|F(\mathbf{w}_k)\|^2}$$

We show two plots to see behaviors of r_t for MNIST (FFN) and CIFAR10 (CNN) (others are reported in Table 2. We can observe from Figure 3 that r_t is bounded above by a constant. (Note that $r_t \leq N$.)
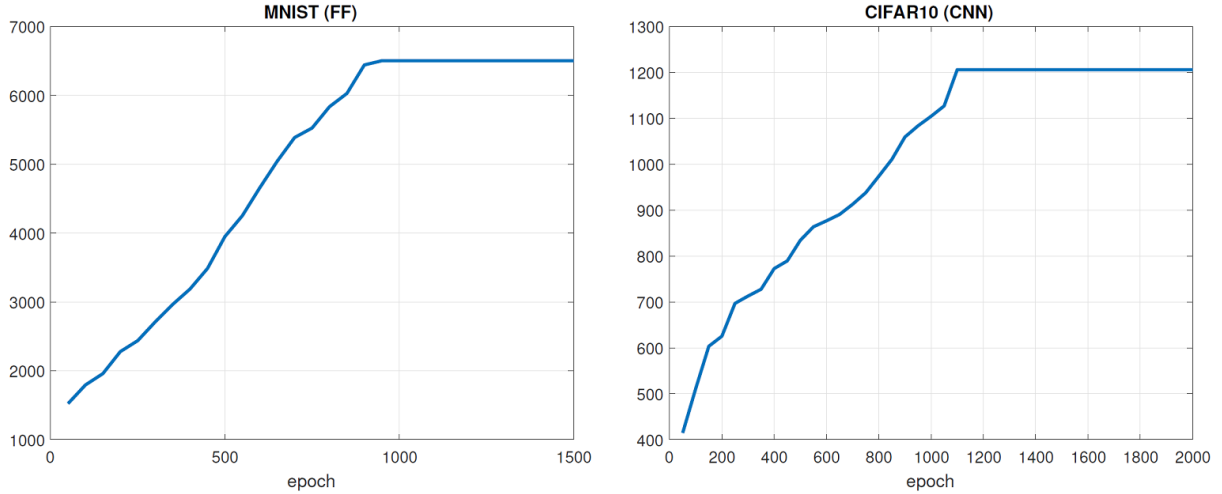

Figure 3: The behaviors of r_t

4 Conclusions

We have demonstrated that based on the behavior of the stochastic gradient estimates at or near the stationary points, SGD with fixed step size converges with the same rate as full gradient descent of the variance reduction methods, until it reaches the accuracy where the variance in the stochastic gradient estimates starts to dominate and prevents further convergence. In particular out assumption is that $1 - \epsilon$ fraction of the stochastic gradient estimates have squared norm below ϵ at the solution. Note ϵ can be made arbitrarily small by increasing the minibatch size b. Indeed we have the following lemma

Lemma 1. *Let $\xi_1, \ldots, \xi_b$ be i.i.d. with $\mathbb{E}[\nabla f(\mathbf{w}; \xi_i)] = \nabla F(\mathbf{w})$, $i = 1, \ldots, b$, for all $\mathbf{w} \in \mathbb{R}^d$. Then,*

$$\mathbb{E}\left[\left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}; \xi_i) - \nabla F(\mathbf{w}) \right\|^2 \right] = \frac{\mathbb{E}[\|\nabla f(\mathbf{w}; \xi_1)\|^2] - \|\nabla F(\mathbf{w})\|^2}{b}. \tag{16}$$

11

It is easy to see that by choosing large b the relation $1 - p_\epsilon \leq \epsilon$ can be achieved for smaller values of ϵ. In the limit for arbitrarily small ϵ we recover full gradient method and its convergence behavior.

References

[1] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 161–168, USA, 2007. Curran Associates Inc.

[2] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.

[3] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

[4] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[5] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[7] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

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

[9] L. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. *ICML*, 2017.

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

[11] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, pages 1–30, 2016.

[12] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. Technical report, 2012.

Appendix

Useful Lemmas

Let $\{\xi_i\}_{i=1}^b$ be i.i.d. random variables with $\mathbb{E}[f(\mathbf{w}; \xi_i)] = F(\mathbf{w})$. From Definition 2, we have

$$\mathbb{E}\left[\|\mathbf{g}_*\|^2\right] = \mathbb{E}\left[\|\mathbf{g}_*\|^2 \mid \|\mathbf{g}_*\|^2 \leq \epsilon\right] \cdot \mathbb{P}\left\{\|\mathbf{g}_*\|^2 \leq \epsilon\right\} + \mathbb{E}\left[\|\mathbf{g}_*\|^2 \mid \|\mathbf{g}_*\|^2 > \epsilon\right] \cdot \mathbb{P}\left\{\mathbf{g}_*\|^2 > \epsilon\right\}$$
$$\leq p_\epsilon \epsilon + (1 - p_\epsilon) M_\epsilon, \tag{17}$$

where $\mathbf{g}_* = \frac{1}{b} \sum_{i=1}^b \nabla f(\mathbf{w}_*; \xi_i)$.

Lemma 2 ([8]). *Suppose that ϕ is L-smooth. Then,*

$$\phi(\mathbf{w}) \leq \phi(\mathbf{w}') + \nabla \phi(\mathbf{w}')^\top (\mathbf{w} - \mathbf{w}') + \frac{L}{2} \|\mathbf{w} - \mathbf{w}'\|^2, \ \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{18}$$

Lemma 3 ([8]). *Suppose that ϕ is L-smooth and convex. Then,*

$$(\nabla \phi(\mathbf{w}) - \nabla \phi(\mathbf{w}'))^\top (\mathbf{w} - \mathbf{w}') \geq \frac{1}{L} \|\nabla \phi(\mathbf{w}) - \nabla \phi(\mathbf{w}')\|^2, \ \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{19}$$

Lemma 4 ([8]). *Suppose that ϕ is L-smooth and convex. Then,*

$$\|\nabla \phi(\mathbf{w})\|^2 \leq 2L(\phi(\mathbf{w}) - \phi(\mathbf{w}_*)), \forall \mathbf{w} \in \mathbb{R}^d, \tag{20}$$

where $\mathbf{w}_* = \arg\min_{\mathbf{w}} \phi(\mathbf{w})$.

Lemma 5 ([8]). *Suppose that ϕ is μ-strongly convex. Then,*

$$2\mu[\phi(\mathbf{w}) - \phi(\mathbf{w}_*)] \leq \|\nabla \phi(\mathbf{w})\|^2, \ \forall \mathbf{w} \in \mathbb{R}^d, \tag{21}$$

where $\mathbf{w}_* = \arg\min_{\mathbf{w}} \phi(\mathbf{w})$.

Lemma 6 ([5]). *Suppose that $f(\mathbf{w}; \xi)$ is L-smooth and convex for every realization of ξ. Then,*

$$\mathbb{E}[\|\nabla f(\mathbf{w}; \xi) - \nabla f(\mathbf{w}_*; \xi)\|^2] \leq 2L[F(\mathbf{w}) - F(\mathbf{w}_*)], \ \forall \mathbf{w} \in \mathbb{R}^d, \tag{22}$$

where ξ is a random variable, and $\mathbf{w}_* = \arg\min_{\mathbf{w}} F(\mathbf{w})$.

Proof. Given any ξ, for all $\mathbf{w} \in \mathbb{R}^d$, consider

$$h(\mathbf{w}; \xi) := f(\mathbf{w}; \xi) - f(\mathbf{w}_*; \xi) - \nabla f(\mathbf{w}_*; \xi)^\top (\mathbf{w} - \mathbf{w}_*).$$

Since $h(\mathbf{w}; \xi)$ is convex by $\mathbf{w}$ and $\nabla h(\mathbf{w}_*; \xi) = 0$, we have $h(\mathbf{w}_*; \xi) = \min_{\mathbf{w}} h(\mathbf{w}; \xi)$. Hence,

$$0 = h(\mathbf{w}_*; \xi) \leq \min_{\eta} [h(\mathbf{w} - \eta \nabla h(\mathbf{w}; \xi); \xi)]$$
$$\overset{(18)}{\leq} \min_{\eta} \left[h(\mathbf{w}; \xi) - \eta \|\nabla h(\mathbf{w}; \xi)\|^2 + \frac{L\eta^2}{2} \|\nabla h(\mathbf{w}; \xi)\|^2 \right]$$
$$= h(\mathbf{w}; \xi) - \frac{1}{2L} \|\nabla h(\mathbf{w}; \xi)\|^2.$$

Hence,

$$\|\nabla f(\mathbf{w}; \xi) - \nabla f(\mathbf{w}_*; \xi)\|^2 \le 2L[f(\mathbf{w}; \xi) - f(\mathbf{w}_*; \xi) - \nabla f(\mathbf{w}_*; \xi)^\top (\mathbf{w} - \mathbf{w}_*)].$$

Taking the expectation with respect to ξ, we have

$$\mathbb{E}[\|\nabla f(\mathbf{w}; \xi) - \nabla f(\mathbf{w}_*; \xi)\|^2] \le 2L[F(\mathbf{w}) - F(\mathbf{w}_*)]. \qquad \square$$

Proof of Lemma 1

Lemma 1. *Let* $\xi_1, \ldots, \xi_b$ *be i.i.d. with* $\mathbb{E}[\nabla f(\mathbf{w}; \xi_i)] = \nabla F(\mathbf{w})$, $i = 1, \ldots, b$, *for all* $\mathbf{w} \in \mathbb{R}^d$. *Then,*

$$\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}; \xi_i) - \nabla F(\mathbf{w})\right\|^2\right] = \frac{\mathbb{E}[\|\nabla f(\mathbf{w}; \xi_1)\|^2] - \|\nabla F(\mathbf{w})\|^2}{b}. \tag{23}$$

Proof. We are going to use mathematical induction to prove the result. With $b = 1$, it is easy to see

$$\mathbb{E}\left[\|\nabla f(\mathbf{w}; \xi_1) - \nabla F(\mathbf{w})\|^2\right] = \mathbb{E}[\|\nabla f(\mathbf{w}; \xi_1)\|^2] - 2\|\nabla F(\mathbf{w})\|^2 + \|\nabla F(\mathbf{w})\|^2$$
$$= \mathbb{E}[\|\nabla f(\mathbf{w}; \xi_1)\|^2] - \|\nabla F(\mathbf{w})\|^2.$$

Let assume that it is true with $b = m - 1$, we are going to show it is also true with $b = m$. We have

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f(\mathbf{w}; \xi_i) - \nabla F(\mathbf{w})\right\|^2\right]$$

$$= \mathbb{E}\left[\left\|\frac{\sum_{i=1}^{m-1}\nabla f(\mathbf{w}; \xi_i) - (m-1)\nabla F(\mathbf{w}) + (\nabla f(\mathbf{w}; \xi_m) - \nabla F(\mathbf{w}))}{m}\right\|^2\right]$$

$$= \frac{1}{m^2}\left(\mathbb{E}\left[\left\|\sum_{i=1}^{m-1}\nabla f(\mathbf{w}; \xi_i) - (m-1)\nabla F(\mathbf{w})\right\|^2\right] + \mathbb{E}\left[\|\nabla f(\mathbf{w}; \xi_m) - \nabla F(\mathbf{w})\|^2\right]\right)$$

$$+ \frac{1}{m}\mathbb{E}\left[2\left(\sum_{i=1}^{m-1}\nabla f(\mathbf{w}; \xi_i) - (m-1)\nabla F(\mathbf{w})\right)^\top (\nabla f(\mathbf{w}; \xi_m) - \nabla F(\mathbf{w}))\right]$$

$$= \frac{1}{m^2}\left(\mathbb{E}\left[\left\|\sum_{i=1}^{m-1}\nabla f(\mathbf{w}; \xi_i) - (m-1)\nabla F(\mathbf{w})\right\|^2\right] + \mathbb{E}\left[\|\nabla f(\mathbf{w}; \xi_m) - \nabla F(\mathbf{w})\|^2\right]\right)$$

$$= \frac{1}{m^2}\left((m-1)\mathbb{E}[\|\nabla f(\mathbf{w}; \xi_1)\|^2] - (m-1)\|\nabla F(\mathbf{w})\|^2 + \mathbb{E}[\|\nabla f(\mathbf{w}; \xi_m)\|^2] - \|\nabla F(\mathbf{w})\|^2\right)$$

$$= \frac{1}{m}\left(\mathbb{E}[\|\nabla f(\mathbf{w}; \xi_1)\|^2] - \|\nabla F(\mathbf{w})\|^2\right).$$

The third and the last equalities follow since $\xi_1, \ldots, \xi_b$ be i.i.d. with $\mathbb{E}[\nabla f(\mathbf{w}; \xi_i)] = \nabla F(\mathbf{w})$. Therefore, the desired result is achieved. $\qquad \square$

Proof of Theorem 1

Theorem 1. *Suppose that $F(\mathbf{w})$ is μ-strongly convex and $f(\mathbf{w};\xi)$ is L-smooth and convex for every realization of ξ. Consider Algorithm 1 with $\eta \leq \frac{1}{L}$. Then, for any $\epsilon > 0$*

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq (1 - \mu\eta(1 - \eta L))^t\|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu(1-\eta L)}p_\epsilon\epsilon + \frac{2\eta}{\mu(1-\eta L)}(1 - p_\epsilon)M_\epsilon,$$

where $\mathbf{w}_* = \arg\min_{\mathbf{w}} F(\mathbf{w})$, and p_ϵ and M_ϵ are defined in (4) and (5), respectively.

Proof. We have

$$\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 = \|\mathbf{w}_t - \eta\mathbf{g}_t - \mathbf{w}_*\|^2$$
$$= \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta\mathbf{g}_t^\top(\mathbf{w}_t - \mathbf{w}_*) + \eta^2\|\mathbf{g}_t\|^2$$
$$= \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_t;\xi_{t,i})^\top(\mathbf{w}_t - \mathbf{w}_*) + \eta^2\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_t;\xi_{t,i})\right\|^2$$
$$\leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_t;\xi_{t,i})^\top(\mathbf{w}_t - \mathbf{w}_*) + 2\eta^2\left\|\frac{1}{b}\sum_{i=1}^{b}(\nabla f(\mathbf{w}_t;\xi_{t,i}) - \nabla f(\mathbf{w}_*;\xi_{t,i}))\right\|^2$$
$$+ 2\eta^2\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2$$
$$\leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_t;\xi_{t,i})^\top(\mathbf{w}_t - \mathbf{w}_*) + 2\eta^2\frac{1}{b}\sum_{i=1}^{b}\|\nabla f(\mathbf{w}_t;\xi_{t,i}) - \nabla f(\mathbf{w}_*;\xi_{t,i})\|^2$$
$$+ 2\eta^2\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2 \tag{24}$$
$$\overset{(19)}{\leq} \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_t;\xi_{t,i})^\top(\mathbf{w}_t - \mathbf{w}_*) + 2\eta^2 L\frac{1}{b}\sum_{i=1}^{b}(\nabla f(\mathbf{w}_t;\xi_{t,i}) - \nabla f(\mathbf{w}_*;\xi_{t,i}))^\top(\mathbf{w}_t - \mathbf{w}_*)$$
$$+ 2\eta^2\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2.$$

Hence, by taking the expectation, conditioned on $\mathcal{F}_t = \sigma(\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t)$ (which is the σ-algebra generated by $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t$), we have

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2|\mathcal{F}_t] \leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta(1 - \eta L)\nabla F(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_*) + 2\eta^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2\bigg|\mathcal{F}_t\right]$$

$$\overset{(6)}{\leq} (1 - \mu\eta(1 - \eta L))\|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta(1 - \eta L)[F(\mathbf{w}_t) - F(\mathbf{w}_*)]$$

$$+ 2\eta^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2\right]$$

$$\overset{\eta \leq 1/L,(17)}{\leq} (1 - \mu\eta(1 - \eta L))\|\mathbf{w}_t - \mathbf{w}_*\|^2 + 2\eta^2 p_\epsilon\epsilon + 2\eta^2(1 - p_\epsilon)M_\epsilon.$$

The first inequality follows since

$$\mathbb{E}\left[\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_t;\xi_{t,i})\Big|\mathcal{F}_t\right] = \mathbb{E}\left[\frac{1}{b}\sum_{i=1}^{b}\left(\nabla f(\mathbf{w}_t;\xi_{t,i}) - \nabla f(\mathbf{w}_*;\xi_{t,i})\right)\Big|\mathcal{F}_t\right] = \nabla F(\mathbf{w}_t).$$

We note in the second equality that $\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2\Big|\mathcal{F}_t\right] = \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2\right]$ since $\xi_{t,i}$ is independent of $\mathcal{F}_t$. By taking the expectation for both sides of the above equation, we obtain

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2] \le (1 - \mu\eta(1 - \eta L))\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] + 2\eta^2 p_\epsilon\epsilon + 2\eta^2(1 - p_\epsilon)M_\epsilon.$$

Hence, we conclude

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2] \le (1 - \mu\eta(1 - \eta L))^{t+1}\|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu(1 - \eta L)}p_\epsilon\epsilon + \frac{2\eta}{\mu(1 - \eta L)}(1 - p_\epsilon)M_\epsilon.$$

$\square$

Proof of Theorem 2

Theorem 2. *Suppose that $f(\mathbf{w};\xi)$ is L-smooth and convex for every realization of ξ. Consider Algorithm 1 with $\eta < \frac{1}{L}$. Then for any $\epsilon > 0$, we have*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \le \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 - \eta L)t} + \frac{\eta}{(1 - \eta L)}p_\epsilon\epsilon + \frac{\eta M_\epsilon}{(1 - \eta L)}(1 - p_\epsilon),$$

where $\mathbf{w}_*$ is any optimal solution of $F(\mathbf{w})$, and p_ϵ and M_ϵ are defined in (4) and (5), respectively.

Proof. If ϕ is convex, then

$$\phi(\mathbf{w}) - \phi(\mathbf{w}') \ge \nabla\phi(\mathbf{w}')^\top(\mathbf{w} - \mathbf{w}'), \quad \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \tag{25}$$

We have

$$\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 = \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_t; \xi_{t,i})^\top (\mathbf{w}_t - \mathbf{w}_*) + \eta^2 \left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_t; \xi_{t,i}) \right\|^2$$

$$\overset{(25)}{\leq} \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta \frac{1}{b} \sum_{i=1}^{b} \left(f(\mathbf{w}_t; \xi_{t,i}) - f(\mathbf{w}_*; \xi_{t,i}) \right) + \eta^2 \left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_t; \xi_{t,i}) \right\|^2$$

$$\leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta \frac{1}{b} \sum_{i=1}^{b} \left(f(\mathbf{w}_t; \xi_{t,i}) - f(\mathbf{w}_*; \xi_{t,i}) \right) + 2\eta^2 \left\| \frac{1}{b} \sum_{i=1}^{b} (\nabla f(\mathbf{w}_t; \xi_{t,i}) - \nabla f(\mathbf{w}_*; \xi_{t,i})) \right\|^2$$

$$+ 2\eta^2 \left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_*; \xi_{t,i}) \right\|^2$$

$$\leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta \frac{1}{b} \sum_{i=1}^{b} \left(f(\mathbf{w}_t; \xi_{t,i}) - f(\mathbf{w}_*; \xi_{t,i}) \right) + 2\eta^2 \frac{1}{b} \sum_{i=1}^{b} \|\nabla f(\mathbf{w}_t; \xi_{t,i}) - \nabla f(\mathbf{w}_*; \xi_{t,i})\|^2$$

$$+ 2\eta^2 \left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_*; \xi_{t,i}) \right\|^2$$

$$\overset{(19)}{\leq} \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta \frac{1}{b} \sum_{i=1}^{b} \left(f(\mathbf{w}_t; \xi_{t,i}) - f(\mathbf{w}_*; \xi_{t,i}) \right) + 2\eta^2 L \frac{1}{b} \sum_{i=1}^{b} (\nabla f(\mathbf{w}_t; \xi_{t,i}) - \nabla f(\mathbf{w}_*; \xi_{t,i}))^\top (\mathbf{w}_t - \mathbf{w}_*)$$

$$+ 2\eta^2 \left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_*; \xi_{t,i}) \right\|^2 .$$

Hence, taking the expectation with condition on $\mathcal{F}_t = \sigma(\mathbf{w}_0, \mathbf{w}_1, \dots, \mathbf{w}_t)$ yields

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 | \mathcal{F}_t] \leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta [F(\mathbf{w}_t) - F(\mathbf{w}_*)] + 2\eta^2 L \nabla F(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}_*)$$

$$+ 2\eta^2 \mathbb{E} \left[\left\| \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_*; \xi_{t,i}) \right\|^2 \Big| \mathcal{F}_t \right]$$

$$\overset{(25),(23),(17)}{\leq} \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta(1 - \eta L)[F(\mathbf{w}_t) - F(\mathbf{w}_*)] + 2\eta^2 p_\epsilon \epsilon + 2\eta^2 (1 - p_\epsilon) M_\epsilon.$$

Taking the expectation for both sides of the above equation yields

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2] \leq \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] - 2\eta(1 - \eta L)\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] + 2\eta^2 p_\epsilon \epsilon + 2\eta^2 (1 - p_\epsilon) M_\epsilon.$$

With $\eta < \frac{1}{L}$, one obtains

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq \frac{1}{2\eta(1 - \eta L)} \left(\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] - \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2] \right)$$

$$+ \frac{\eta}{(1 - \eta L)} p_\epsilon \epsilon + \frac{\eta M_\epsilon}{(1 - \eta L)} (1 - p_\epsilon).$$

By summing from $k = 0, \ldots, t$ and averaging, we have

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[F(\mathbf{w}_k) - F(\mathbf{w}_*)] \leq \frac{1}{2\eta(1-\eta L)(t+1)}\|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{\eta}{(1-\eta L)}p_\epsilon\epsilon + \frac{\eta M_\epsilon}{(1-\eta L)}(1-p_\epsilon).$$

Since $\mathbb{E}[F(\mathbf{w}_k)]$ is a non-increasing function on k, the sum on the left hand side is larger than $t+1$ times its last element. Hence,

$$\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*)] \leq \frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[F(\mathbf{w}_k) - F(\mathbf{w}_*)]$$

$$\leq \frac{1}{2\eta(1-\eta L)(t+1)}\|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{\eta}{(1-\eta L)}p_\epsilon\epsilon + \frac{\eta M_\epsilon}{(1-\eta L)}(1-p_\epsilon).$$

$\square$

Proof of Theorem 3

Theorem 3. *Let Assumption 1 hold for some $N > 0$. Suppose that F is L-smooth. Consider Algorithm 1 with $\eta < \frac{1}{LN}$. Then, for any $\epsilon > 0$, we have*

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq \frac{[F(\mathbf{w}_0) - F^*]}{\eta\,(1 - L\eta N)\,(t+1)} + \frac{L\eta}{(1 - L\eta N)}\epsilon + \frac{L\eta M_\epsilon}{(1 - L\eta N)}(1-p_\epsilon),$$

where F^* is any lower bound of F; and p_ϵ and M_ϵ are defined in (12) *and* (13) *respectively.*

Proof. Let us assume that, there exists a local minima $\mathbf{w}_*$ of $F(\mathbf{w})$. We have

$$\mathbb{E}[F(\mathbf{w}_{t+1})|\mathcal{F}_t] = \mathbb{E}[F(\mathbf{w}_t - \eta\mathbf{g}_t)|\mathcal{F}_t] \overset{(18)}{\leq} F(\mathbf{w}_t) - \eta\|\nabla F(\mathbf{w}_t)\|^2 + \frac{L\eta^2}{2}\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_t;\xi_{t,i})\right\|^2 \Big| \mathcal{F}_t\right]$$

$$\leq F(\mathbf{w}_t) - \eta\|\nabla F(\mathbf{w}_t)\|^2 + L\eta^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}(\nabla f(\mathbf{w}_t;\xi_{t,i}) - \nabla f(\mathbf{w}_*;\xi_{t,i}))\right\|^2 \Big| \mathcal{F}_t\right]$$

$$+ L\eta^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{t,i})\right\|^2 \Big| \mathcal{F}_t\right]$$

$$\leq F(\mathbf{w}_t) - \eta\|\nabla F(\mathbf{w}_t)\|^2 + L\eta^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}(\nabla f(\mathbf{w}_t;\xi_{t,i}) - \nabla f(\mathbf{w}_*;\xi_{t,i}))\right\|^2 \Big| \mathcal{F}_t\right]$$

$$+ L\eta^2\epsilon + L\eta^2(1-p_\epsilon)M_\epsilon.$$

By summing from $k = 0, \ldots, t$ and averaging, we have

$$\frac{1}{t+1} \sum_{k=0}^{t} \mathbb{E}[F(\mathbf{w}_{k+1})|\mathcal{F}_k] \leq \frac{1}{t+1} \sum_{k=0}^{t} F(\mathbf{w}_k) - \eta \frac{1}{t+1} \sum_{k=0}^{t} \|\nabla F(\mathbf{w}_k)\|^2$$

$$+ L\eta^2 \frac{1}{t+1} \sum_{k=0}^{t} \left(\mathbb{E}\left[\left\| \frac{1}{b} \sum_{i=1}^{b} (\nabla f(\mathbf{w}_k; \xi_{k,i}) - \nabla f(\mathbf{w}_*; \xi_{k,i})) \right\|^2 \middle| \mathcal{F}_k \right] \right)$$

$$+ L\eta^2 \epsilon + L\eta^2 (1 - p_\epsilon) M_\epsilon$$

$$\overset{(11)}{\leq} \frac{1}{t+1} \sum_{k=0}^{t} F(\mathbf{w}_k) - \eta (1 - L\eta N) \frac{1}{t+1} \sum_{k=0}^{t} \|\nabla F(\mathbf{w}_k)\|^2$$

$$+ L\eta^2 \epsilon + L\eta^2 (1 - p_\epsilon) M_\epsilon.$$

Taking the expectation for the above equation, we have

$$\frac{1}{t+1} \sum_{k=0}^{t} \mathbb{E}[F(\mathbf{w}_{k+1})] \leq \frac{1}{t+1} \sum_{k=0}^{t} \mathbb{E}[F(\mathbf{w}_k)] - \eta (1 - L\eta N) \frac{1}{t+1} \sum_{k=0}^{t} \mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2]$$

$$+ L\eta^2 \epsilon + L\eta^2 (1 - p_\epsilon) M_\epsilon.$$

Hence, with $\eta < \frac{1}{LN}$, we have

$$\frac{1}{t+1} \sum_{k=0}^{t} \mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq \frac{[\mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}_{t+1})]]}{\eta(1 - L\eta N)(t+1)} + \frac{L\eta}{(1 - L\eta N)} \epsilon + \frac{L\eta M_\epsilon}{(1 - L\eta N)} (1 - p_\epsilon)$$

$$\leq \frac{[F(\mathbf{w}_0) - F^*]}{\eta(1 - L\eta N)(t+1)} + \frac{L\eta}{(1 - L\eta N)} \epsilon + \frac{L\eta M_\epsilon}{(1 - L\eta N)} (1 - p_\epsilon),$$

where F^* is any lower bound of F. $\qquad \square$

Proof of Corollary 1

Corollary 1. *For any ϵ such that $1 - p_\epsilon \leq \epsilon$, and for Algorithm 1 with $\eta \leq \frac{1}{2L}$, we have*

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq (1 - \mu\eta)^t \|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu}(1 + M_\epsilon)\epsilon.$$

Furthermore if $t \geq T$ for $T = \frac{1}{\mu\eta} \log \left(\frac{\mu\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 + M_\epsilon)\epsilon} \right)$, then

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq \frac{4\eta}{\mu}(1 + M_\epsilon)\epsilon.$$

19

Proof. Taking the expectation, conditioning on $\mathcal{F}_t = \sigma(\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t)$ to (24), we have

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 | \mathcal{F}_t] \leq \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta \nabla F(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}_*)$$

$$+ 2\eta^2 \mathbb{E}[\|\nabla f(\mathbf{w}_t; \xi_{t,1}) - \nabla f(\mathbf{w}_*; \xi_{t,1})\|] + 2\eta^2 \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*; \xi_{t,i})\right\|^2 \middle| \mathcal{F}_t\right]$$

$$\stackrel{(6),(22)}{\leq} (1 - \mu\eta)\|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta(1 - 2\eta L)[F(\mathbf{w}_t) - F(\mathbf{w}_*)]$$

$$+ 2\eta^2 \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*; \xi_{t,i})\right\|^2\right]$$

$$\stackrel{\eta \leq \frac{1}{2L}, (17)}{\leq} (1 - \mu\eta)\|\mathbf{w}_t - \mathbf{w}_*\|^2 + 2\eta^2 p_\epsilon \epsilon + 2\eta^2 (1 - p_\epsilon) M_\epsilon.$$

The first inequality follows since $\{\xi_{i,i}\}_{i=1}^{b}$ are i.i.d. random variables. Hence, we have

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2] \leq (1 - \mu\eta)^{t+1}\|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu}p_\epsilon \epsilon + \frac{2\eta}{\mu}(1 - p_\epsilon)M_\epsilon.$$

Therefore,

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq (1 - \mu\eta)^t\|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu}p_\epsilon \epsilon + \frac{2\eta}{\mu}(1 - p_\epsilon)M_\epsilon$$

$$\leq (1 - \mu\eta)^t\|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu}(1 + M_\epsilon)\epsilon,$$

where the last inequality follows since $1 - p_\epsilon \leq \epsilon$.

First, we would like to find a T such that

$$(1 - \mu\eta)^T\|\mathbf{w}_0 - \mathbf{w}_*\|^2 = \frac{2\eta}{\mu}(1 + M_\epsilon)\epsilon.$$

Taking log for both sides, we have

$$T\log(1 - \mu\eta) + \log\left(\|\mathbf{w}_0 - \mathbf{w}_*\|^2\right) = \log\left(\frac{2\eta}{\mu}(1 + M_\epsilon)\epsilon\right).$$

Hence,

$$T = -\frac{1}{\log(1 - \mu\eta)}\log\left(\frac{\mu\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 + M_\epsilon)\epsilon}\right) \leq \frac{1}{\mu\eta}\log\left(\frac{\mu\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 + M_\epsilon)\epsilon}\right),$$

where the last inequality follows since $-1/\log(1 - x) \leq 1/x$ for $0 < x \leq 1$. Hence, if $t \geq T$ for $T = \frac{1}{\mu\eta}\log\left(\frac{\mu\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 + M_\epsilon)\epsilon}\right)$, then

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq \frac{2\eta}{\mu}(1 + M_\epsilon)\epsilon + \frac{2\eta}{\mu}(1 + M_\epsilon)\epsilon = \frac{4\eta}{\mu}(1 + M_\epsilon)\epsilon.$$

$\square$

Proof of Corollary 2

Corollary 2. *If $f(\mathbf{w}; \xi)$ is L-smooth and convex for every realization of ξ, then for any ϵ such that $1 - p_\epsilon \leq \epsilon$, and $\eta \leq \frac{1}{2L}$, it holds that*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\eta t} + 2\eta \left(1 + M_\epsilon\right) \epsilon.$$

Hence, if $t \geq T$ for $T = \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{(2\eta^2)(1 + M_\epsilon)\epsilon}$, we have

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq 4\eta \left(1 + M_\epsilon\right) \epsilon.$$

Proof. By Theorem 2, with $\eta \leq \frac{1}{2L}$, we have

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] &\leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1 - \eta L)t} + \frac{\eta}{(1 - \eta L)}p_\epsilon\epsilon + \frac{\eta M_\epsilon}{(1 - \eta L)}(1 - p_\epsilon) \\
&\leq \frac{2\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta t} + 2\eta p_\epsilon\epsilon + 2\eta M_\epsilon(1 - p_\epsilon) \\
&\leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\eta t} + 2\eta(1 + M_\epsilon)\epsilon.
\end{aligned}$$

Similar to the proof of Corollary 1, we want to find a T such that

$$\frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\eta T} = 2\eta(1 + M_\epsilon)\epsilon.$$

It is easy to see that if $t \geq T$ for $T = \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{(2\eta^2)(1 + M_\epsilon)\epsilon}$, then

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq 2\eta(1 + M_\epsilon)\epsilon + 2\eta(1 + M_\epsilon)\epsilon = 4\eta(1 + M_\epsilon)\epsilon.$$

$\square$

Proof of Corollary 3

Corollary 3. *Let Assumption 1 hold and p_ϵ and M_ϵ be defined as in (12) and (13). For any ϵ such that $1 - p_\epsilon \leq \epsilon$, and for $\eta \leq \frac{1}{2LN}$, we have*

$$\frac{1}{t + 1} \sum_{k=0}^{t} \mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq \frac{2[F(\mathbf{w}_0) - F^*]}{\eta(t + 1)} + 2L\eta(1 + M_\epsilon)\epsilon.$$

Hence, if $t \geq T$ for $T = \frac{[F(\mathbf{w}_0) - F^*]}{(L\eta^2)(1 + M_\epsilon)\epsilon}$, we have

$$\frac{1}{t + 1} \sum_{k=0}^{t} \mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq 4L\eta(1 + M_\epsilon)\epsilon.$$

Proof. By Theorem 3, with $\eta \leq \frac{1}{2LN}$, we have

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq \frac{[F(\mathbf{w}_0) - F^*]}{\eta\,(1 - L\eta N)\,(t+1)} + \frac{L\eta}{(1 - L\eta N)}\epsilon + \frac{L\eta M_\epsilon}{(1 - L\eta N)}(1 - p_\epsilon)$$

$$\leq \frac{2[F(\mathbf{w}_0) - F^*]}{\eta(t+1)} + 2L\eta\epsilon + 2L\eta M_\epsilon(1 - p_\epsilon)$$

$$\leq \frac{2[F(\mathbf{w}_0) - F^*]}{\eta(t+1)} + 2L\eta(1 + M_\epsilon)\epsilon.$$

Similar to the proof of Corollaries 1 and 2, we want to find a T such that

$$\frac{2[F(\mathbf{w}_0) - F^*]}{\eta T} = 2L\eta(1 + M_\epsilon)\epsilon.$$

It is easy to see that if $t \geq T$ for $T = \frac{[F(\mathbf{w}_0) - F^*]}{(L\eta^2)(1 + M_\epsilon)\epsilon}$, then

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq 2L\eta(1 + M_\epsilon)\epsilon + 2L\eta(1 + M_\epsilon)\epsilon = 4L\eta(1 + M_\epsilon)\epsilon.$$

□