New Convergence Aspects of Stochastic Gradient Algorithms

Lam M. Nguyen* LamNguyen.MLTD@ibm.com
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA

Phuong Ha Nguyen* phuongha.ntu@gmail.com
Department of Electrical and Computer Engineering
University of Connecticut
Storrs, CT 06268, USA

Peter Richtárik Peter.Richtarik@ed.ac.uk
KAUST, KSA — Edinburgh, UK — MIPT, Russia

Katya Scheinberg katyas@lehigh.edu
Martin Takáč Takac.MT@gmail.com
Department of Industrial and Systems Engineering
Lehigh University
Bethlehem, PA 18015, USA

Marten van Dijk marten.van_dijk@uconn.edu
Department of Electrical and Computer Engineering
University of Connecticut
Storrs, CT 06268, USA

Editor: N/A

Abstract

The classical convergence analysis of SGD is carried out under the assumption that the norm of the stochastic gradient is uniformly bounded. While this might hold for some loss functions, it is violated for cases where the objective function is strongly convex. In Bottou et al. (2016), a new analysis of convergence of SGD is performed under the assumption that stochastic gradients are bounded with respect to the true gradient norm. We show that for stochastic problems arising in machine learning such bound always holds; and we also propose an alternative convergence analysis of SGD with diminishing learning rate regime, which results in more relaxed conditions than those in Bottou et al. (2016). We then move on the asynchronous parallel setting, and prove convergence of Hogwild! algorithm in the same regime in the case of diminished learning rate. It is well-known that SGD converges if a sequence of learning rates $\{\eta_t\}$ satisfies $\sum_{t=0}^{\infty} \eta_t \rightarrow \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$. We show the convergence of SGD for strongly convex objective function without using bounded gradient assumption when $\{\eta_t\}$ is a diminishing sequence and $\sum_{t=0}^{\infty} \eta_t \rightarrow \infty$. In other words, we extend the current state-of-the-art class of learning rates satisfying the convergence of SGD.

Keywords: Stochastic Gradient Algorithms, Asynchronous Stochastic Optimization, SGD, Hogwild!, bounded gradient

. * equally contributed. Corresponding author: Lam M. Nguyen

1. Introduction

We are interested in solving the following stochastic optimization problem

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

where ξ is a random variable obeying some distribution.

In the case of empirical risk minimization with a training set $\{(x_i, y_i)\}_{i=1}^n$, ξ_i is a random variable that is defined by a single random sample (x, y) pulled uniformly from the training set. Then, by defining $f_i(w) := f(w;\xi_i)$, empirical risk minimization reduces to

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

Problem (2) arises frequently in supervised learning applications Hastie et al. (2009). For a wide range of applications, such as linear regression and logistic regression, the objective function F is strongly convex and each f_i, $i \in [n]$, is convex and has Lipschitz continuous gradients (with Lipschitz constant L). Given a training set $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, the ℓ_2-regularized least squares regression model, for example, is written as (2) with $f_i(w) \overset{\text{def}}{=} (\langle x_i, w\rangle - y_i)^2 + \frac{\lambda}{2}\|w\|^2$. The ℓ_2-regularized logistic regression for binary classification is written with $f_i(w) \overset{\text{def}}{=} \log(1 + \exp(-y_i\langle x_i, w\rangle)) + \frac{\lambda}{2}\|w\|^2$, $y_i \in \{-1, 1\}$. It is well established by now that solving this type of problem by gradient descent (GD) Nesterov (2004); Nocedal and Wright (2006) may be prohibitively expensive and stochastic gradient descent (SGD) is thus preferable. Recently, a class of variance reduction methods Le Roux et al. (2012); Defazio et al. (2014); Johnson and Zhang (2013); Nguyen et al. (2017) has been proposed in order to reduce the computational cost. All these methods explicitly exploit the finite sum form of (2) and thus they have some disadvantages for very large scale machine learning problems and are not applicable to (1).

To apply SGD to the general form (1) one needs to assume existence of unbiased gradient estimators. This is usually defined as follows:

$$\mathbb{E}_\xi[\nabla f(w;\xi)] = \nabla F(w),$$

for any fixed w. Here we make an important observation: if we view (1) not as a general stochastic problem but as the expected risk minimization problem, where ξ corresponds to a random data sample pulled from a distribution, then (1) has an additional key property: for each realization of the random variable ξ, $f(w;\xi)$ is a convex function with Lipschitz continuous gradients. Notice that traditional analysis of SGD for general stochastic problem of the form (1) does not make any assumptions on individual function realizations. In this paper we derive convergence properties for SGD applied to (1) with these additional assumptions on $f(w;\xi)$ and also extend to the case when $f(w;\xi)$ are not necessarily convex.

Regardless of the properties of $f(w;\xi)$ we assume that F in (1) is strongly convex. We define the (unique) optimal solution of F as w_*.

Assumption 1 (μ-strongly convex) *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex, i.e., there exists a constant $\mu > 0$ such that $\forall w, w' \in \mathbb{R}^d$,*

$$F(w) - F(w') \geq \langle \nabla F(w'), (w - w')\rangle + \frac{\mu}{2}\|w - w'\|^2. \tag{3}$$

It is well-known in literature Nesterov (2004); Bottou et al. (2016) that Assumption 1 implies

$$2\mu[F(w) - F(w_*)] \leq \|\nabla F(w)\|^2 \ , \ \forall w \in \mathbb{R}^d. \tag{4}$$

The classical theoretical analysis of SGD assumes that the *stochastic gradients are uniformly bounded*, i.e. there exists a finite (fixed) constant $\sigma < \infty$, such that

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq \sigma^2 \ , \ \forall w \in \mathbb{R}^d \tag{5}$$

(see e.g. Shalev-Shwartz et al. (2007); Nemirovski et al. (2009); Recht et al. (2011); Hazan and Kale (2014); Rakhlin et al. (2012), etc.). However, this assumption is clearly false if F is strongly convex. Specifically, under this assumption together with strong convexity, $\forall w \in \mathbb{R}^d$, we have

$$\begin{aligned} 2\mu[F(w) - F(w_*)] &\overset{(4)}{\leq} \|\nabla F(w)\|^2 = \|\mathbb{E}[\nabla f(w;\xi)]\|^2 \\ &\leq \mathbb{E}[\|\nabla f(w;\xi)\|^2] \overset{(5)}{\leq} \sigma^2. \end{aligned}$$

Hence,

$$F(w) \leq \frac{\sigma^2}{2\mu} + F(w_*) \ , \ \forall w \in \mathbb{R}^d.$$

On the other hand strong convexity and $\nabla F(w_*) = 0$ imply

$$F(w) \geq \mu\|w - w_*\|^2 + F(w_*) \ , \ \forall w \in \mathbb{R}^d.$$

The last two inequalities are clearly in contradiction with each other for sufficiently large $\|w - w_*\|^2$.

Let us consider the following example: $f_1(w) = \frac{1}{2}w^2$ and $f_2(w) = w$ with $F(w) = \frac{1}{2}(f_1(w) + f_2(w))$. Note that F is strongly convex, while individual realizations are not necessarily so. Let $w_0 = 0$, for any number $t \geq 0$, with probability $\frac{1}{2^t}$ the steps of SGD algorithm for all $i < t$ are $w_{i+1} = w_i - \eta_i$. This implies that $w_t = -\sum_{i=1}^t \eta_i$ and since $\sum_{i=1}^\infty \eta_i = \infty$ then $|w_t|$ can be arbitrarily large for large enough t with probability $\frac{1}{2^t}$. Noting that for this example, $\mathbb{E}[\|\nabla f(w_t;\xi)\|^2] = \frac{1}{2}w_t^2 + \frac{1}{2}$, we see that $\mathbb{E}[\|\nabla f(w_t;\xi)\|^2]$ can also be arbitrarily large.

Recently, in the review paper Bottou et al. (2016), convergence of SGD for general stochastic optimization problem was analyzed under the following assumption: there exist constants M and N such that $\mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2] \leq M\|\nabla F(w_t)\|^2 + N$, where w_t, $t \geq 0$, are generated by the algorithm. This assumption does not contradict strong convexity, however, in general, constants M and N are unknown, while M is used to determine the learning rate η_t Bottou et al. (2016). In addition, the rate of convergence of the SGD algorithm depends on M and N. In this paper we show that under the smoothness assumption on individual realizations $f(w,\xi)$ it is possible to derive the bound $\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq M_0[F(w) - F(w_*)] + N$ with specific values of M_0, and N for $\forall w \in \mathbb{R}^d$, which in turn implies the bound $\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq M\|\nabla F(w)\|^2 + N$ with specific M, by strong convexity of F. We also note that, in Moulines and Bach (2011), the convergence of SGD without bounded

gradient assumption is studied. We then provide an alternative convergence analysis for SGD which shows convergence in expectation with a bound on learning rate which is larger than that in Bottou et al. (2016); Moulines and Bach (2011) by a factor of L/μ. We then use the new framework for the convergence analysis of SGD to analyze an asynchronous stochastic gradient method.

In Recht et al. (2011), an asynchronous stochastic optimization method called Hogwild! was proposed. Hogwild! algorithm is a parallel version of SGD, where each processor applies SGD steps independently of the other processors to the solution w which is shared by all processors. Thus, each processor computes a stochastic gradient and updates w without "locking" the memory containing w, meaning that multiple processors are able to update w at the same time. This approach leads to much better scaling of parallel SGD algorithm than a synchoronous version, but the analysis of this method is more complex. In Recht et al. (2011); Mania et al. (2015); De Sa et al. (2015) various variants of Hogwild! with a fixed step size are analyzed under the assumption that the gradients are bounded as in (5). In this paper, we extend our analysis of SGD to provide analysis of Hogwild! with diminishing step sizes and without the assumption on bounded gradients.

In a recent technical report Leblond et al. (2018) Hogwild! with fixed step size is analyzed without the bounded gradient assumption. We note that SGD with fixed step size only converges to a neighborhood of the optimal solution, while by analyzing the diminishing step size variant we are able to show convergence to the *optimal solution* with probability one. Both in Leblond et al. (2018) and in this paper, the version of Hogwild! with inconsistent reads and writes is considered.

It is well-known that SGD will converge if a sequence of learning rates $\{\eta_t\}$ satisfies the following conditions (1) $\sum_{t=0}^{\infty} \eta_t \to \infty$ and (2) $\sum_{t=0}^{\infty} \eta_t^2 < \infty$. As an important contribution of this paper, we show the convergence of SGD for strongly convex objective function without using bounded gradient assumption when $\{\eta_t\}$ is a diminishing sequence and $\sum_{t=0}^{\infty} \eta_t \to \infty$. In Moulines and Bach (2011), the authors also proved the convergence of SGD for $\{\eta_t = \mathcal{O}(1/t^q)\}, 0 < q \leq 1$, without using bounded gradient assumption and the second condition. Compared to Moulines and Bach (2011), we prove the convergence of SGD for $\{\eta_t = \mathcal{O}(1/t^q)\}$ which is $1/\mu$ times larger and our proposed class of learning rates satisfying the convergence of SGD is larger. Our proposed class of learning rates satisfying the convergence of SGD is larger than the current state-of-the art one.

We would like to highlight that this paper is originally from Nguyen et al. (2018) (Proceedings of the 35th International Conference on Machine Learning, 2018) but it presents a substantial extension by providing many new results for SGD and Hogwild!.

1.1 Contribution

We provide a new framework for the analysis of stochastic gradient algorithms in the strongly convex case under the condition of Lipschitz continuity of the individual function realizations, but **without requiring any bounds on the stochastic gradients**. Within this framework we have the following contributions:

- We prove the almost sure (w.p.1) convergence of SGD with diminishing step size. Our analysis provides a larger bound on the possible initial step size when compared to any previous analysis of convergence in expectation for SGD.

- We introduce a general recurrence for vector updates which has as its special cases (a) Hogwild! algorithm with diminishing step sizes, where each update involves all non-zero entries of the computed gradient, and (b) a position-based updating algorithm where each update corresponds to only one uniformly selected non-zero entry of the computed gradient.

- We analyze this general recurrence under inconsistent vector reads from and vector writes to shared memory (where individual vector entry reads and writes are atomic in that they cannot be interrupted by writes to the same entry) assuming that there exists a delay τ such that during the $(t+1)$-th iteration a gradient of a read vector w is computed which includes the aggregate of all the updates up to and including those made during the $(t-\tau)$-th iteration. In other words, τ controls to what extend past updates influence the shared memory.

  - Our upper bound for the expected convergence rate is sublinear, i.e., $O(1/t)$, and its precise expression allows comparison of algorithms (a) and (b) described above.
  - For SGD we can improve this upper bound by a factor 2 and also show that its initial step size can be larger.
  - We show that τ can be a function of t as large as $\approx \sqrt{t/\ln t}$ without affecting the asymptotic behavior of the upper bound; we also determine a constant T_0 with the property that, for $t \geq T_0$, higher order terms containing parameter τ are smaller than the leading $O(1/t)$ term. We give intuition explaining why the expected convergence rate is not more affected by τ. Our experiments confirm our analysis.
  - We determine a constant T_1 with the property that, for $t \geq T_1$, the higher order term containing parameter $\|w_0 - w_*\|^2$ is smaller than the leading $O(1/t)$ term.

- All the above contributions generalize to the non-convex setting where we do not need to assume that the component functions $f(w;\xi)$ are convex in w.

Compared to Nguyen et al. (2018), we have following new results:

- We prove the almost sure (w.p.1) convergence of Hogwild! with a diminishing sequence of learning rates $\{\eta_t\}$.

- We prove the convergence of SGD for diminishing sequences of learning rates $\{\eta_t\}$ with condition $\sum_{t=0}^{\infty} \eta_t \to \infty$. In other words, we extend the current state-of-the-art class of learning rates satisfying the convergence of SGD.

- We prove the convergence of SGD for our extended class of learning rates in batch model.

1.2 Organization

We analyse the convergence rate of SGD in Section 2 and introduce the general recursion and its analysis in Section 3. Section 4 studies the convergence of SGD for our extended class of learning rates. Experiments are reported in Section 5.

2. New Framework for Convergence Analysis of SGD

We introduce SGD algorithm in Algorithm 1.

Algorithm 1 Stochastic Gradient Descent (SGD) Method

Initialize: w_0
Iterate:
for $t = 0, 1, 2, \ldots$ **do**
  Choose a step size (i.e., learning rate) $\eta_t > 0$.
  Generate a random variable ξ_t.
  Compute a stochastic gradient $\nabla f(w_t; \xi_t)$.
  Update the new iterate $w_{t+1} = w_t - \eta_t \nabla f(w_t; \xi_t)$.
end for

The sequence of random variables $\{\xi_t\}_{t \geq 0}$ is assumed to be i.i.d.[1] Let us introduce our key assumption that each realization $\nabla f(w; \xi)$ is an L-smooth function.

Assumption 2 (L-smooth) *$f(w; \xi)$ is L-smooth for every realization of ξ, i.e., there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$,*

$$\|\nabla f(w; \xi) - \nabla f(w'; \xi)\| \leq L \|w - w'\|. \tag{6}$$

Assumption 2 implies that F is also L-smooth. Then, by the property of L-smooth function (in Nesterov (2004)), we have, $\forall w, w' \in \mathbb{R}^d$,

$$F(w) \leq F(w') + \langle \nabla F(w'), (w - w') \rangle + \frac{L}{2} \|w - w'\|^2. \tag{7}$$

The following additional convexity assumption can be made, as it holds for many problems arising in machine learning.

Assumption 3 *$f(w; \xi)$ is convex for every realization of ξ, i.e., $\forall w, w' \in \mathbb{R}^d$,*

$$f(w; \xi) - f(w'; \xi) \geq \langle \nabla f(w'; \xi), (w - w') \rangle.$$

We first derive our analysis under Assumptions 2, and 3 and then we derive weaker results under only Assumption 2.

2.1 Convergence With Probability One

As discussed in the introduction, under Assumptions 2 and 3 we can now derive a bound on $\mathbb{E}\|\nabla f(w; \xi)\|^2$.

Lemma 1 *Let Assumptions 2 and 3 hold. Then, for $\forall w \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla f(w; \xi)\|^2] \leq 4L[F(w) - F(w_*)] + N, \tag{8}$$

where $N = 2\mathbb{E}[\|\nabla f(w_*; \xi)\|^2]$; ξ is a random variable, and $w_* = \arg\min_w F(w)$.

1. Independent and identically distributed.

Using Lemma 1 and Super Martingale Convergence Theorem Bertsekas (2015) (Lemma 5 in the supplementary material), we can provide the sufficient condition for almost sure convergence of Algorithm 1 in the strongly convex case without assuming any bounded gradients.

Theorem 1 (Sufficient conditions for almost sure convergence) *Let Assumptions 1, 2 and 3 hold. Consider Algorithm 1 with a stepsize sequence such that*

$$0 < \eta_t \leq \frac{1}{2L} \ , \ \sum_{t=0}^{\infty} \eta_t = \infty \ \textit{and} \ \sum_{t=0}^{\infty} \eta_t^2 < \infty.$$

Then, the following holds w.p.1 (almost surely)

$$\|w_t - w_*\|^2 \to 0.$$

Note that the classical SGD proposed in Robbins and Monro (1951) has learning rate satisfying conditions

$$\sum_{t=0}^{\infty} \eta_t = \infty \text{ and } \sum_{t=0}^{\infty} \eta_t^2 < \infty$$

However, the original analysis is performed under the bounded gradient assumption, as in (5). In Theorem 1, on the other hand, we do not use this assumption, but instead assume Lipschitz smoothness and convexity of the function realizations, which does not contradict the strong convexity of $F(w)$.

The following result establishes a sublinear convergence rate of SGD.

Theorem 2 *Let Assumptions 1, 2 and 3 hold. Let $E = \frac{2\alpha L}{\mu}$ with $\alpha = 2$. Consider Algorithm 1 with a stepsize sequence such that $\eta_t = \frac{\alpha}{\mu(t+E)} \leq \eta_0 = \frac{1}{2L}$. The expectation $\mathbb{E}[\|w_t - w_*\|^2]$ is at most*

$$\frac{4\alpha^2 N}{\mu^2} \frac{1}{(t - T + E)}$$

for

$$t \geq T = \frac{4L}{\mu} \max\{\frac{L\mu}{N}\|w_0 - w_*\|^2, 1\} - \frac{4L}{\mu}.$$

2.2 Convergence Analysis without Convexity

In this section, we provide the analysis of Algorithm 1 without using Assumption 3, that is, $f(w;\xi)$ is not necessarily convex. We still do not need to impose the bounded stochastic gradient assumption, since we can derive an analogue of Lemma 1, albeit with worse constant in the bound.

Lemma 2 *Let Assumptions 1 and 2 hold. Then, for $\forall w \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq 4L\kappa[F(w) - F(w_*)] + N, \tag{9}$$

where $\kappa = \frac{L}{\mu}$ and $N = 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]$; ξ is a random variable, and $w_* = \arg\min_w F(w)$.

Based on the proofs of Theorems 1 and 2, we can easily have the following two results (Theorems 3 and 4).

Theorem 3 (Sufficient conditions for almost sure convergence) *Let Assumptions 1 and 2 hold. Then, we can conclude the statement of Theorem 1 with the definition of the step size replaced by* $0 < \eta_t \leq \frac{1}{2L\kappa}$ *with* $\kappa = \frac{L}{\mu}$.

Theorem 4 *Let Assumptions 1 and 2 hold. Then, we can conclude the statement of Theorem 2 with the definition of the step size replaced by* $\eta_t = \frac{\alpha}{\mu(t+E)} \leq \eta_0 = \frac{1}{2L\kappa}$ *with* $\kappa = \frac{L}{\mu}$ *and* $\alpha = 2$*, and all other occurrences of* L *in* E *and* T *replaced by* $L\kappa$.

We compare our result in Theorem 4 with that in Bottou et al. (2016) in the following remark.

Remark 1 *By strong convexity of* F*, Lemma 2 implies* $\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq 2\kappa^2\|\nabla F(w)\|^2 + N$*, for* $\forall w \in \mathbb{R}^d$*, where* $\kappa = \frac{L}{\mu}$ *and* $N = 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]$*. We can now substitute the value* $M = 2\kappa^2$ *into Theorem 4.7 in Bottou et al. (2016). We observe that the resulting initial learning rate in Bottou et al. (2016) has to satisfy* $\eta_0 \leq \frac{1}{2L\kappa^2}$ *while our results allows* $\eta_0 = \frac{1}{2L\kappa}$*. We are able to achieve this improvement by introducing Assumption 2, which holds for many ML problems.*

Recall that under Assumption 3, our initial learning rate is $\eta_0 = \frac{1}{2L}$ *(in Theorem 2). Thus Assumption 3 provides further improvement of the conditions on the learning rate.*

3. Asynchronous Stochastic Optimization aka Hogwild!

Hogwild! Recht et al. (2011) is an asynchronous stochastic optimization method where writes to and reads from vector positions in shared memory can be inconsistent (this corresponds to (13) as we shall see). However, as mentioned in Mania et al. (2015), for the purpose of analysis the method in Recht et al. (2011) performs single vector entry updates that are randomly selected from the non-zero entries of the computed gradient as in (12) (explained later) and requires the assumption of consistent vector reads together with the bounded gradient assumption to prove convergence. Both Mania et al. (2015) and De Sa et al. (2015) prove the same result for fixed step size based on the assumption of bounded stochastic gradients in the strongly convex case but now without assuming consistent vector reads and writes. In these works the fixed step size η must depend on σ from the bounded gradient assumption, however, one does not usually know σ and thus, we cannot compute a suitable η a-priori.

As claimed by the authors in Mania et al. (2015), they can eliminate the bounded gradient assumption in their analysis of Hogwild!, which however was only mentioned as a remark without proof. On the other hand, the authors of recent unpublished work Leblond et al. (2018) formulate and prove, without the bounded gradient assumption, a precise theorem about the convergence rate of Hogwild! of the form

$$\mathbb{E}[\|w_t - w_*\|^2] \leq (1-\rho)^t(2\|w_0 - w_*\|^2) + b,$$

where ρ is a function of several parameters but independent of the fixed chosen step size η and where b is a function of several parameters and has a linear dependency with respect to the fixed step size, i.e., $b = O(\eta)$.

In this section, we discuss the convergence of Hogwild! with **diminishing** stepsize where writes to and reads from vector positions in shared memory can be **inconsistent**. This is a slight modification of the original Hogwild! where the stepsize is fixed. In our analysis we also **do not use the bounded gradient assumption** as in Leblond et al. (2018). Moreover, (a) we focus on solving the **more general problem** in (1), while Leblond et al. (2018) considers the specific case of the "finite-sum" problem in (2), and (b) we show that our analysis generalizes to the **non-convex case**, i.e., we do not need to assume functions $f(w;\xi)$ are convex (we only require $F(w) = \mathbb{E}[f(w;\xi)]$ to be strongly convex) as opposed to the assumption in Leblond et al. (2018).

3.1 Recursion

We first formulate a general recursion for w_t to which our analysis applies, next we will explain how the different variables in the recursion interact and describe two special cases, and finally we present pseudo code of the algorithm using the recursion.

The recursion explains which positions in w_t should be updated in order to compute w_{t+1}. Since w_t is stored in shared memory and is being updated in a possibly non-consistent way by multiple cores who each perform recursions, the shared memory will contain a vector w whose entries represent a mix of updates. That is, before performing the computation of a recursion, a core will first read w from shared memory, however, while reading w from shared memory, the entries in w are being updated out of order. The final vector $\hat{w}_t$ read by the core represents an aggregate of a mix of updates in previous iterations.

The general recursion is defined as follows: For $t \geq 0$,

$$w_{t+1} = w_t - \eta_t d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t; \xi_t), \tag{10}$$

where

- $\hat{w}_t$ represents the vector used in computing the gradient $\nabla f(\hat{w}_t;\xi_t)$ and whose entries have been read (one by one) from an aggregate of a mix of previous updates that led to w_j, $j \leq t$, and
- the $S_{u_t}^{\xi_t}$ are diagonal 0/1-matrices with the property that there exist real numbers d_ξ satisfying
$$d_\xi \mathbb{E}[S_u^\xi | \xi] = D_\xi, \tag{11}$$
where the expectation is taken over u and D_ξ is the diagonal $0/1$ matrix whose 1-entries correspond to the non-zero positions in $\nabla f(w;\xi)$, i.e., the i-th entry of D_ξ's diagonal is equal to 1 if and only if there exists a w such that the i-th position of $\nabla f(w;\xi)$ is non-zero.

The role of matrix $S_{u_t}^{\xi_t}$ is that it filters which positions of gradient $\nabla f(\hat{w}_t;\xi_t)$ play a role in (10) and need to be computed. Notice that D_ξ represents the support of $\nabla f(w;\xi)$; by $|D_\xi|$ we denote the number of 1s in D_ξ, i.e., $|D_\xi|$ equals the size of the support of $\nabla f(w;\xi)$.

We will restrict ourselves to choosing (i.e., fixing a-priori) *non-empty* matrices S_u^ξ that "partition" D_ξ in D approximately "equally sized" S_u^ξ:

$$\sum_u S_u^\xi = D_\xi,$$

where each matrix S_u^ξ has either $\lfloor |D_\xi|/D \rfloor$ or $\lceil |D_\xi|/D \rceil$ ones on its diagonal. We uniformly choose one of the matrices $S_{u_t}^{\xi_t}$ in (10), hence, d_ξ equals the number of matrices S_u^ξ, see (11).

In other to explain recursion (10) we first consider two special cases. For $D = \bar{\Delta}$, where

$$\bar{\Delta} = \max_\xi \{|D_\xi|\}$$

represents the maximum number of non-zero positions in any gradient computation $f(w;\xi)$, we have that for all ξ, there are exactly $|D_\xi|$ diagonal matrices S_u^ξ with a single 1 representing each of the elements in D_ξ. Since $p_\xi(u) = 1/|D_\xi|$ is the uniform distribution, we have $\mathbb{E}[S_u^\xi|\xi] = D_\xi/|D_\xi|$, hence, $d_\xi = |D_\xi|$. This gives the recursion

$$w_{t+1} = w_t - \eta_t |D_\xi| [\nabla f(\hat{w}_t;\xi_t)]_{u_t}, \tag{12}$$

where $[\nabla f(\hat{w}_t;\xi_t)]_{u_t}$ denotes the u_t-th position of $\nabla f(\hat{w}_t;\xi_t)$ and where u_t is a uniformly selected position that corresponds to a non-zero entry in $\nabla f(\hat{w}_t;\xi_t)$.

At the other extreme, for $D = 1$, we have exactly one matrix $S_1^\xi = D_\xi$ for each ξ, and we have $d_\xi = 1$. This gives the recursion

$$w_{t+1} = w_t - \eta_t \nabla f(\hat{w}_t;\xi_t). \tag{13}$$

Recursion (13) represents Hogwild!. In a single-core setting where updates are done in a consistent way and $\hat{w}_t = w_t$ yields SGD.

Algorithm 2 gives the pseudo code corresponding to recursion (10) with our choice of sets S_u^ξ (for parameter D).

Algorithm 2 Hogwild! general recursion

Input: $w_0 \in \mathbb{R}^d$
for $t = 0, 1, 2, \ldots$ **in parallel do**
  read each position of shared memory w denoted by $\hat{w}_t$ **(each position read is atomic)**
  draw a random sample ξ_t and a random "filter" $S_{u_t}^{\xi_t}$
  for positions h where $S_{u_t}^{\xi_t}$ has a 1 on its diagonal **do**
    compute g_h as the gradient $\nabla f(\hat{w}_t;\xi_t)$ at position h
    add $\eta_t d_{\xi_t} g_h$ to the entry at position h of w in shared memory **(each position update is atomic)**
  end for
end for

3.2 Analysis

Besides Assumptions 1, 2, and for now 3, we assume the following assumption regarding a parameter τ, called the delay, which indicates which updates in previous iterations have certainly made their way into shared memory w.

Assumption 4 (Consistent with delay τ) *We say that shared memory is consistent with delay τ with respect to recursion (10) if, for all t, vector $\hat{w}_t$ includes the aggregate of the*

updates up to and including those made during the $(t-\tau)$-th iteration (where (10) defines the $(t+1)$-st iteration). Each position read from shared memory is atomic and each position update to shared memory is atomic (in that these cannot be interrupted by another update to the same position).

In other words in the $(t+1)$-th iteration, $\hat{w}_t$ equals $w_{t-\tau}$ plus some subset of position updates made during iterations $t-\tau, t-\tau+1, \ldots, t-1$. We assume that there exists a constant delay τ satisfying Assumption 4.

3.3 Convergence With Probability One

Appendix D proves the following theorem

Theorem 5 (Sufficient conditions for almost sure convergence for Hogwild!) *Let Assumptions 1, 2, 3 and 4 hold. Consider Hogwild! method described in Algorithm* (2) *with a stepsize sequence such that*

$$0 < \eta_t = \frac{1}{LD(2+\beta)(k+t)} < \frac{1}{4LD}, \beta > 0, k \geq 3\tau.$$

Then, the following holds w.p.1 (almost surely)

$$\|w_t - w_*\| \to 0.$$

3.4 Convergence in Expectation

The supplementary material proves the following theorem where

$$\bar{\Delta}_D \stackrel{\text{def}}{=} D \cdot \mathbb{E}[\lceil |D_\xi|/D \rceil].$$

Theorem 6 *Suppose Assumptions 1, 2, 3 and 4 and consider Algorithm 2 for sets S_u^ξ with parameter D. Let $\eta_t = \frac{\alpha_t}{\mu(t+E)}$ with $4 \leq \alpha_t \leq \alpha$ and $E = \max\{2\tau, \frac{4L\alpha D}{\mu}\}$. Then, the expected number of single vector entry updates after t iterations is equal to*

$$t' = t\bar{\Delta}_D/D$$

and expectations $\mathbb{E}[\|\hat{w}_t - w_*\|^2]$ and $\mathbb{E}[\|w_t - w_*\|^2]$ are at most

$$\frac{4\alpha^2 DN}{\mu^2}\frac{t}{(t+E-1)^2} + O\left(\frac{\ln t}{(t+E-1)^2}\right).$$

In terms of t', the expected number single vector entry updates after t iterations, $\mathbb{E}[\|\hat{w}_t - w_*\|^2]$ and $\mathbb{E}[\|w_t - w_*\|^2]$ are at most

$$\frac{4\alpha^2 \bar{\Delta}_D N}{\mu^2}\frac{1}{t'} + O\left(\frac{\ln t'}{t'^2}\right).$$

Remark 2 *In (12) $D = \bar{\Delta}$, hence, $\lceil |D_\xi|/D \rceil = 1$ and $\bar{\Delta}_D = \bar{\Delta} = \max_\xi\{|D_\xi|\}$. In (13) $D = 1$, hence, $\bar{\Delta}_D = \mathbb{E}[|D_\xi|]$. This shows that the upper bound in Theorem 6 is better for (13) with $D = 1$. If we assume no delay, i.e. $\tau = 0$, in addition to $D = 1$, then we obtain SGD. Theorem 2 shows that, measured in t', we obtain the upper bound*

$$\frac{4\alpha_{SGD}^2 \bar{\Delta}_D N}{\mu^2} \frac{1}{t'}$$

with $\alpha_{SGD} = 2$ as opposed to $\alpha \geq 4$.

With respect to parallelism, SGD assumes a single core, while (13) and (12) allow multiple cores. Notice that recursion (12) allows us to partition the position of the shared memory among the different processor cores in such a way that each partition can only be updated by its assigned core and where partitions can be read by all cores. This allows optimal resource sharing and could make up for the difference between $\bar{\Delta}_D$ for (12) and (13). We hypothesize that, for a parallel implementation, D equal to a fraction of $\bar{\Delta}$ will lead to best performance.

Remark 3 *Surprisingly, the leading term of the upper bound on the convergence rate is independent of delay τ. On one hand, one would expect that a more recent read which contains more of the updates done during the last τ iterations will lead to better convergence. When inspecting the second order term in the proof in the supplementary material, we do see that a smaller τ (and/or smaller sparsity) makes the convergence rate smaller. That is, asymptotically t should be large enough as a function of τ (and other parameters) in order for the leading term to dominate.*

Nevertheless, in asymptotic terms (for larger t) the dependence on τ is not noticeable. In fact, the supplementary material shows that we may allow τ to be a monotonic increasing function of t with

$$\frac{2L\alpha D}{\mu} \leq \tau(t) \leq \sqrt{t \cdot L(t)},$$

where $L(t) = \frac{1}{\ln t} - \frac{1}{(\ln t)^2}$ (this will make $E = \max\{2\tau(t), \frac{4L\alpha D}{\mu}\}$ also a function of t). The leading term of the convergence rate does not change while the second order terms increase to $O(\frac{1}{t \ln t})$. We show that, for

$$t \geq T_0 = \exp[2\sqrt{\Delta}(1 + \frac{(L+\mu)\alpha}{\mu})],$$

where $\Delta = \max_i \mathbb{P}(i \in D_\xi)$ measures sparsity, the higher order terms that contain $\tau(t)$ (as defined above) are at most the leading term.

Our intuition behind this phenomenon is that for large τ, all the last τ iterations before the t-th iteration use vectors $\hat{w}_j$ with entries that are dominated by the aggregate of updates that happened till iteration $t - \tau$. Since the average sum of the updates during the last τ iterations is equal to

$$-\frac{1}{\tau} \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_t) \tag{14}$$

and all $\hat{w}_j$ look alike in that they mainly represent learned information before the $(t-\tau)$-th iteration, (14) becomes an estimate of the expectation of (14), i.e.,

$$\sum_{j=t-\tau}^{t-1} \frac{-\eta_j}{\tau} \mathbb{E}[d_{\xi_j} S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_t)] = \sum_{j=t-\tau}^{t-1} \frac{-\eta_j}{\tau} \nabla F(\hat{w}_j). \tag{15}$$

This looks like GD which in the strong convex case has convergence rate $\leq c^{-t}$ for some constant $c > 1$. This already shows that larger τ could help convergence as well. However, estimate (14) has estimation noise with respect to (15) which explains why in this thought experiment we cannot attain c^{-t} but can only reach a much smaller convergence rate of e.g. $O(1/t)$ as in Theorem 6.

Experiments in Section 5 confirm our analysis.

Remark 4 *The higher order terms in the proof in the supplementary material show that, as in Theorem 2, the expected convergence rate in Theorem 6 depends on $\|w_0 - w_*\|^2$. The proof shows that, for*

$$t \geq T_1 = \frac{\mu^2}{\alpha^2 N D} \|w_0 - w_*\|^2,$$

the higher order term that contains $\|w_0 - w_*\|^2$ is at most the leading term. This is comparable to T in Theorem 2 for SGD.

Remark 5 *Step size $\eta_t = \frac{\alpha_t}{\mu(t+E)}$ with $4 \leq \alpha_t \leq \alpha$ can be chosen to be fixed during periods whose ranges exponentially increase. For $t + E \in [2^h, 2^{h+1})$ we define $\alpha_t = \frac{4(t+E)}{2^h}$. Notice that $4 \leq \alpha_t < 8$ which satisfies the conditions of Theorem 6 for $\alpha = 8$. This means that we can choose*

$$\eta_t = \frac{\alpha_t}{\mu(t+E)} = \frac{4}{\mu 2^h}$$

as step size for $t + E \in [2^h, 2^{h+1})$. This choice for η_t allows changes in η_t to be easily synchronized between cores since these changes only happen when $t + E = 2^h$ for some integer h. That is, if each core is processing iterations at the same speed, then each core on its own may reliably assume that after having processed $(2^h - E)/P$ iterations the aggregate of all P cores has approximately processed $2^h - E$ iterations. So, after $(2^h - E)/P$ iterations a core will increment its version of h to $h + 1$. This will introduce some noise as the different cores will not increment their h versions at exactly the same time, but this only happens during a small interval around every $t + E = 2^h$. This will occur rarely for larger h.

3.5 Convergence Analysis without Convexity

In the supplementary material, we also show that the proof of Theorem 6 can easily be modified such that Theorem 6 with $E \geq \frac{4L\kappa\alpha D}{\mu}$ also holds in the non-convex case of the component functions, i.e., we do not need Assumption 3. Note that this case is not analyzed in Leblond et al. (2018).

Theorem 7 *Let Assumptions 1 and 2 hold. Then, we can conclude the statement of Theorem 6 with $E \geq \frac{4L\kappa\alpha D}{\mu}$ for $\kappa = \frac{L}{\mu}$.*

Theorem 8 (Sufficient conditions for almost sure convergence) *Let Assumptions 1 and 2 hold. Then, we can conclude the statement of Theorem 5 with the definition of the step size replaced by* $0 < \eta_t = \frac{1}{LD\kappa(2+\beta)(k+t)}$ *with* $\kappa = \frac{L}{\mu}$.

4. Convergence of Large Stepsizes

In Robbins and Monro (1951), the authors proved the convergence of SGD for step size sequences $\{\eta_t\}$ satisfying conditions

$$\sum_{t=0}^{\infty} \eta_t = \infty \text{ and } \sum_{t=0}^{\infty} \eta_t^2 < \infty.$$

In Moulines and Bach (2011), the authors studied the expected convergence rates for another class of step sizes of $\mathcal{O}(1/t^p)$ where $0 < p \leq 1$. This class has many large step sizes in comparison with Robbins and Monro (1951), for example $\eta_t = 1/t^p$ where $0 < p < 1/2$. In this section, we prove that SGD will converge without using bounded gradient assumption if $\{\eta_t\}$ is a diminishing sequence and $\sum_{t=0}^{\infty} \eta_t \to \infty$. Compared to Moulines and Bach (2011), we prove the convergence of SGD for step sizes $\eta_t = \mathcal{O}(1/t^q)$ which is $1/\mu$ times larger. Our proposed class is much larger than the classes in Robbins and Monro (1951) and Moulines and Bach (2011). The proofs of all theorems and lemmas in this section are provided in Appendix D.6.

4.1 Convergence of Large Stepsizes

Theorem 9 *Let Assumptions 1, 3 and 2 hold. Consider Algorithm 1 with a step size sequence such that:* $\eta_t \leq \frac{1}{2L}$, $\eta_t \to 0$, $\frac{d}{dt}\eta_t \leq 0$ *and* $\sum_{t=0}^{\infty} \eta_t \to \infty$. *Then,*

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \to 0.$$

Theorem 9 only discusses about the convergence of SGD for the given step size sequence $\{\eta_t\}$ above. The expected convergence rate of SGD with the setup in Theorem 9 is analysed in Theorem 10.

Theorem 10 *Let Assumptions 1, 3 and 2 hold. Consider Algorithm 1 with a step size sequence such that* $\eta_t \leq \frac{1}{2L}$, $\eta_t \to 0$, $\frac{d}{dt}\eta_t \leq 0$, *and* $\sum_{t=0}^{\infty} \eta_t \to \infty$. *Then* $\mathbb{E}[\|w_{t+1} - w_*\|^2]$ *is at most*

$$N \exp(n(0)) 2n(M^{-1}(\ln[\frac{n(t+1)}{n(0)}] + M(t+1))) + \exp(-M(t+1))[\exp(M(1))n^2(0)N + \mathbb{E}[\|w_0 - w_*\|^2]],$$

where $n(t) = \mu\eta_t$ *and* $M(t) = \int_{x=0}^{t} n(x)dx$.

As shown later (i.e., see (53)), we have

$$\mathbb{E}[\|w_t - w_*\|^2] \leq AC(t) + B\exp(-M(t)),$$

where A and B are constants, $C(t)$ is defined in (16). We show that an alternative proof for the convergence of SGD with the setup above based on the study of $C(t)$ can be developed.

Lemma 3 *Let*

$$C(t) = \exp(-M(t)) \int_{x=0}^{t} \exp(M(x)) n^2(x) dx, \tag{16}$$

where $\frac{d}{dx} M(x) = n(x)$ *and the function* $n(x)$ *satisfies the following conditions:*

1. $\frac{d}{dx} n(x) < 0$,
2. $\frac{d}{dx} n(x)$ *is continuous.*

Then, there is a moment T *such as for all* $t > T$, $C(t) > n(t)$.

Proof

We take the derivative of $C(t)$, i.e.,

$$\begin{aligned} \frac{d}{dt} C(t) &= -\exp(-M(t)) n(t) \int_{x=0}^{t} \exp(M(x)) n^2(x) dx + \exp(-M(t)) \exp(M(t)) n^2(t) \\ &= n(t)[n(t) - C(t)] \end{aligned}$$

This shows that

$$C(t) \text{ is decreasing if and only if } C(t) > n(t).$$

Initially $C(0) = 0$ and $n(0) > 0$, hence, $C(t)$ starts increasing from $t \geq 0$. Since $n(t)$ decreases for all $t \geq 0$, we know that there must exist a first cross over point x:

- There exists a value x such that $C(t)$ increases for $0 \leq t < x$, and
- $C(x) = n(x)$ with derivative $dC(t)/dt|_{t=x} = 0$.

Since $n(x)$ has a derivative < 0, we know that $C(t) > n(t)$ immediately after x. Suppose that $C(y) = n(y)$ for some $y > x$ with $C(t) > n(t)$ for $x < t < y$. This implies that $dC(t)/dt|_{t=y} = 0$ and since $dC(t)/dt$ is continuous

$$C(y - \epsilon) = C(y) + O(\epsilon^2).$$

Also,

$$n(y - \epsilon) = n(y) - \epsilon dn(t)/dt|_{t=y} + O(\epsilon^2).$$

Since $dn(t)/dt|_{t=y} < 0$, we know that there exists an ϵ small enough (close to 0) such that

$$C(y - \epsilon) < n(y - \epsilon).$$

This contradicts $C(t) > n(t)$ for $x < t < y$. We conclude that there does not exist a $y > x$ such that $C(y) = n(y)$:

- For $t > x$, $C(t) > n(t)$ and $C(t)$ is strictly decreasing.

Conclusion. For any given $n(t)$, there exists a time T such that $C(t) < n(t)$ for all $t \in [0, T)$, $C(T) = n(T)$ and then $C(t) > n(t)$ when $t \in (T, \infty]$. Note that $C(t)$ is always bigger then zero. As proved above, $C(t)$ always decreases after T and thus, $C(t)$ converges to zero when t goes to infinity. ■

As proved in Lemma 3, $C(t) \to 0$ when $t \to \infty$. Moreover $\exp(-M(t)) \to 0$ when $t \to \infty$ because $\sum_{t=0}^{\infty} n(t) \to \infty$. Based on two these results, we conclude that $\mathbb{E}[\|w_t - w_*\|^2] \to 0$ when $t \to \infty$. This is alternative proof for the convergence of SGD as shown in Theorem 9.

Theorem 11 *Among all stepsizes $\eta_{q,t} = 1/(K+t)^q$ where $q > 0$, K is a constant such that $\eta_{q,t} \leq \frac{1}{2L}$, with the stepsize $\eta_{1,t} = 1/(2L+t)$, SGD algorithm enjoys the fastest convergence.*

4.2 Convergence of Large Stepsizes in Batch Mode

We define

$$\mathcal{F}_t = \sigma(w_0, \xi_0', u_0, \ldots, \xi_{t-1}', u_{t-1}),$$

where

$$\xi_i' = (\xi_{i1}, \ldots, \xi_{ik_i}).$$

We consider the following general algorithm with the following gradient updating rule:

$$w_{t+1} = w_t - \eta_t d_{\xi_t'} S_{u_t}^{\xi_t'} \nabla f(w_t; \xi_t'), \tag{17}$$

where $f(w_t; \xi_t') = \frac{1}{k_t} \sum_{i=1}^{k_t} f(w_t; \xi_{ti})$.

Theorem 12 *Let Assumptions 1, 2 and 3 hold, $\{\eta_t\}$ is a diminishing sequence with conditions $\sum_{t=0}^{\infty} \eta_t \to \infty$ and $0 < \eta_t \leq \frac{1}{2LD}$ for all $t \geq 0$. Then, the sequence $\{w_t\}$ converges to w_* where*

$$w_{t+1} = w_t - \eta_t d_{\xi_t'} S_{u_t}^{\xi_t'} \nabla f(w_t; \xi_t').$$

The proof of Theorem 12 is provided in Appendix D.7.

5. Numerical Experiments

For our numerical experiments, we consider the finite sum minimization problem in (2). We consider ℓ_2-regularized logistic regression problems with

$$f_i(w) = \log(1 + \exp(-y_i \langle x_i, w \rangle)) + \frac{\lambda}{2} \|w\|^2,$$

where the penalty parameter λ is set to $1/n$, a widely-used value in literature Le Roux et al. (2012).

We conducted experiments on a single core for Algorithm 2 on two popular datasets `ijcnn1` ($n = 91,701$ training data) and `covtype` ($n = 406,709$ training data) from the LIBSVM[2] website. Since we are interested in the expected convergence rate with respect to the number of iterations, respectively number of single position vector updates, we do not need

2. http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/

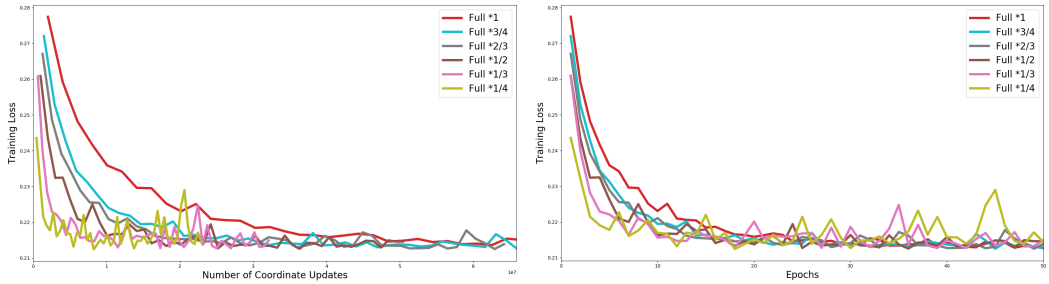

Figure 1: *ijcnn1* for different fraction of non-zero set

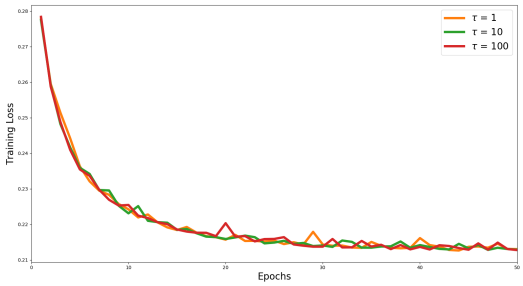

Figure 2: *ijcnn1* for different τ with the whole non-zero set

a parallelized multi-core simulation to confirm our analysis. The impact of efficient resource scheduling over multiple cores leads to a performance improvement complementary to our analysis of (10) (which, as discussed, lends itself for an efficient parallelized implementation). We experimented with 10 runs and reported the average results. We choose the step size based on Theorem 6, i.e, $\eta_t = \frac{4}{\mu(t+E)}$ and $E = \max\{2\tau, \frac{16LD}{\mu}\}$. For each fraction $v \in \{1, 3/4, 2/3, 1/2, 1/3, 1/4\}$ we performed the following experiment: In Algorithm 2 we choose each "filter" matrix $S_{u_t}^{\xi_t}$ to correspond with a random subset of size $v|D_{\xi_t}|$ of the non-zero positions of D_{ξ_t} (i.e., the support of the gradient corresponding to ξ_t). In addition we use $\tau = 10$. For the two datasets, Figures 1 and 3 plot the training loss for each fraction with $\tau = 10$. The top plots have t', the number of coordinate updates, for the horizontal axis. The bottom plots have the number of epochs, each epoch counting n iterations, for the horizontal axis. The results show that each fraction shows a sublinear expected convergence rate of $O(1/t')$; the smaller fractions exhibit larger deviations but do seem to converge faster to the minimum solution.

In Figures 2 and 4, we show experiments with different values of $\tau \in \{1, 10, 100\}$ where we use the whole non-zero set of gradient positions (i.e., $v = 1$) for the update. Our analysis states that, for $t = 50$ epochs times n iterations per epoch, τ can be as large as $\sqrt{t \cdot L(t)} = 524$ for `ijcnn1` and 1058 for `covtype`. The experiments indeed show that $\tau \leq 100$ has little effect on the expected convergence rate.

6. Conclusion

We have provided the analysis of stochastic gradient algorithms with diminishing step size in the strongly convex case under the condition of Lipschitz continuity of the individual function

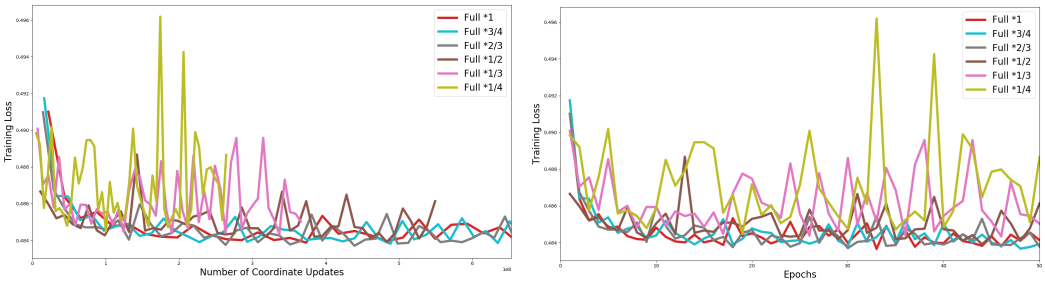

Figure 3: *covtype* for different fraction of non-zero set

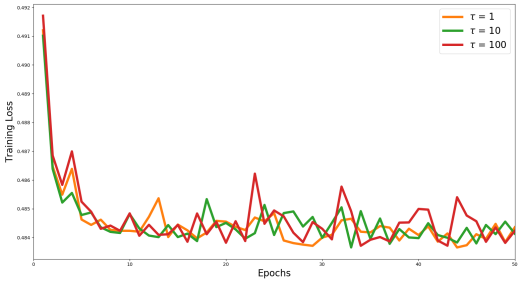

Figure 4: *covtype* for different τ with the whole non-zero set

realizations, but without requiring any bounds on the stochastic gradients. We showed almost sure convergence of SGD and provided sublinear upper bounds for the expected convergence rate of a general recursion which includes Hogwild! for inconsistent reads and writes as a special case. We also provided new intuition which will help understanding convergence as observed in practice.

Acknowledgement

Phuong Ha Nguyen and Marten van Dijk were supported in part by AFOSR MURI under award number FA9550-14-1-0351. Katya Scheinberg was partially supported by NSF Grants CCF 16-18717 and CCF 17-40796. Martin Takáč was partially supported by the U.S. National Science Foundation, under award number NSF:CCF:1618717, NSF:CMMI:1663256 and NSF:CCF:1740796

References

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

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.

Christopher M De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild-style algorithms. In *Advances in neural information processing systems*, pages 2674–2682, 2015.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edition, 2009.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15: 2489–2512, 2014.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

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.

Remi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *arXiv:1801.03749*, 2018.

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. *arXiv preprint arXiv:1507.06970*, 2015.

Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 451–459. Curran Associates, Inc., 2011.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, January 2009. ISSN 1052-6234. doi: 10.1137/070704277.

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

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. *ICML*, 2017.

Lam M. Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. SGD and Hogwild! convergence without the bounded gradients assumption. *ICML*, 2018.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*. icml.cc / Omnipress, 2012.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 693–701. Curran Associates, Inc., 2011.

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

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 807–814, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273598.

Appendix A. Review of Useful Theorems

Lemma 4 (Generalization of the result in Johnson and Zhang (2013)) *Let Assumptions 2 and 3 hold. Then, $\forall w \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \leq 2L[F(w) - F(w_*)], \tag{18}$$

where ξ is a random variable, and $w_* = \arg\min_w F(w)$.

Lemma 5 (Bertsekas (2015)) *Let Y_k, Z_k, and W_k, $k = 0, 1, \ldots$, be three sequences of random variables and let $\{\mathcal{F}_k\}_{k\geq 0}$ be a filtration, that is, σ-algebras such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Suppose that:*

- *The random variables Y_k, Z_k, and W_k are nonnegative, and $\mathcal{F}_k$-measurable.*
- *For each k, we have $\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq Y_k - Z_k + W_k$.*
- *There holds, w.p.1,*

$$\sum_{k=0}^{\infty} W_k < \infty.$$

Then, we have, w.p.1,

$$\sum_{k=0}^{\infty} Z_k < \infty \text{ and } Y_k \to Y \geq 0.$$

Appendix B. Proofs of Lemmas 1 and 2

B.1 Proof of Lemma 1

Lemma 1. *Let Assumptions 2 and 3 hold. Then, for $\forall w \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq 4L[F(w) - F(w_*)] + N,$$

where $N = 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]$; ξ is a random variable, and $w_* = \arg\min_w F(w)$.
Proof Note that

$$\|a\|^2 = \|a - b + b\|^2 \leq 2\|a - b\|^2 + 2\|b\|^2, \tag{19}$$

$$\Rightarrow \frac{1}{2}\|a\|^2 - \|b\|^2 \leq \|a - b\|^2. \tag{20}$$

Hence,

$$\begin{aligned}\frac{1}{2}\mathbb{E}[\|\nabla f(w;\xi)\|^2] - \mathbb{E}[\|\nabla f(w_*;\xi)\|^2] &= \mathbb{E}\left[\frac{1}{2}\|\nabla f(w;\xi)\|^2 - \|\nabla f(w_*;\xi)\|^2\right] \\ &\overset{(20)}{\leq} \mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \\ &\overset{(18)}{\leq} 2L[F(w) - F(w_*)] \end{aligned} \tag{21}$$

Therefore,

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \overset{(19)(21)}{\leq} 4L[F(w)-F(w_*)] + 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2].$$

■

B.2 Proof of Lemma 2

Lemma 2. *Let Assumptions 1 and 2 hold. Then, for* $\forall w \in \mathbb{R}^d$,

$$\mathbb{E}\|\nabla f(w;\xi)\|^2 \leq 4L\kappa[F(w)-F(w_*)] + N,$$

where $\kappa = \frac{L}{\mu}$ *and* $N = 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]$*;* ξ *is a random variable, and* $w_* = \arg\min_w F(w)$.

Proof Analogous to the proof of Lemma 1, we have

Hence,

$$\begin{aligned}
\frac{1}{2}\mathbb{E}[\|\nabla f(w;\xi)\|^2] - \mathbb{E}[\|\nabla f(w_*;\xi)\|^2] &= \mathbb{E}\left[\frac{1}{2}\|\nabla f(w;\xi)\|^2 - \|\nabla f(w_*;\xi)\|^2\right] \\
&\overset{(20)}{\leq} \mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \\
&\overset{(6)}{\leq} L^2\|w-w_*\|^2 \\
&\overset{(3)}{\leq} \frac{2L^2}{\mu}[F(w)-F(w_*)] = 2L\kappa[F(w)-F(w_*)]. \quad (22)
\end{aligned}$$

Therefore,

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \overset{(19)(22)}{\leq} 4L\kappa[F(w)-F(w_*)] + 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2].$$

■

Appendix C. Analysis for Algorithm 1

In this Section, we provide the analysis of Algorithm 1 under Assumptions 1, 2, and 3.

We note that if $\{\xi_i\}_{i\geq 0}$ are i.i.d. random variables, then $\mathbb{E}[\|\nabla f(w_*;\xi_0)\|^2] = \cdots = \mathbb{E}[\|\nabla f(w_*;\xi_t)\|^2]$. We have the following results for Algorithm 1.

Theorem 1 (Sufficient condition for almost sure convergence). *Let Assumptions 1, 2 and 3 hold. Consider Algorithm 1 with a stepsize sequence such that*

$$0 < \eta_t \leq \frac{1}{2L}\ ,\ \sum_{t=0}^{\infty}\eta_t = \infty \text{ and } \sum_{t=0}^{\infty}\eta_t^2 < \infty.$$

Then, the following holds w.p.1 (almost surely)

$$\|w_t - w_*\|^2 \to 0.$$

Proof Let $\mathcal{F}_t = \sigma(w_0, \xi_0, \ldots, \xi_{t-1})$ be the σ-algebra generated by $w_0, \xi_0, \ldots, \xi_{t-1}$, i.e., $\mathcal{F}_t$ contains all the information of $w_0, \ldots, w_t$. Note that $\mathbb{E}[\nabla f(w_t; \xi_t)|\mathcal{F}_t] = \nabla F(w_t)$. By Lemma 1, we have

$$\mathbb{E}[\|\nabla f(w_t; \xi_t)\|^2|\mathcal{F}_t] \leq 4L[F(w_t) - F(w_*)] + N, \tag{23}$$

where $N = 2\mathbb{E}[\|\nabla f(w_*; \xi_0)\|^2] = \cdots = 2\mathbb{E}[\|\nabla f(w_*; \xi_t)\|^2]$ since $\{\xi_i\}_{i\geq 0}$ are i.i.d. random variables. Note that $w_{t+1} = w_t - \eta_t \nabla f(w_t; \xi_t)$. Hence,

$$\begin{aligned}
\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t] &= \mathbb{E}[\|w_t - \eta_t \nabla f(w_t; \xi_t) - w_*\|^2|\mathcal{F}_t] \\
&= \|w_t - w_*\|^2 - 2\eta_t \langle \nabla F(w_t), (w_t - w_*)\rangle + \eta_t^2 \mathbb{E}[\|\nabla f(w_t; \xi_t)\|^2|\mathcal{F}_t] \\
&\overset{(3)(23)}{\leq} \|w_t - w_*\|^2 - \mu\eta_t \|w_t - w_*\|^2 - 2\eta_t[F(w_t) - F(w_*)] \\
&+ 4L\eta_t^2[F(w_t) - F(w_*)] + \eta_t^2 N \\
&= \|w_t - w_*\|^2 - \mu\eta_t \|w_t - w_*\|^2 - 2\eta_t(1 - 2L\eta_t)[F(w_t) - F(w_*)] + \eta_t^2 N \\
&\leq \|w_t - w_*\|^2 - \mu\eta_t \|w_t - w_*\|^2 + \eta_t^2 N.
\end{aligned}$$

The last inequality follows since $0 < \eta_t \leq \frac{1}{2L}$. Therefore,

$$\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t] \leq \|w_t - w_*\|^2 - \mu\eta_t \|w_t - w_*\|^2 + \eta_t^2 N. \tag{24}$$

Since $\sum_{t=0}^{\infty} \eta_t^2 N < \infty$, we could apply Lemma 5. Then, we have w.p.1,

$$\|w_t - w_*\|^2 \to W \geq 0,$$
$$\text{and } \sum_{t=0}^{\infty} \mu\eta_t \|w_t - w_*\|^2 < \infty.$$

We want to show that $\|w_t - w_*\|^2 \to 0$, w.p.1. Proving by contradiction, we assume that there exist $\epsilon > 0$ and t_0, s.t. $\|w_t - w_*\|^2 \geq \epsilon$ for $\forall t \geq t_0$. Hence,

$$\sum_{t=0}^{\infty} \mu\eta_t \|w_t - w_*\|^2 \geq \mu\epsilon \sum_{t=0}^{\infty} \eta_t = \infty.$$

This is a contradiction. Therefore, $\|w_t - w_*\|^2 \to 0$ w.p.1. ■

Theorem 2. *Let Assumptions 1, 2 and 3 hold. Let* $E = \frac{2\alpha L}{\mu}$ *with* $\alpha = 2$. *Consider Algorithm 1 with a stepsize sequence such that* $\eta_t = \frac{\alpha}{\mu(t+E)} \leq \eta_0 = \frac{1}{2L}$. *The expectation* $\mathbb{E}[\|w_t - w_*\|^2]$ *is at most*

$$\frac{4\alpha^2 N}{\mu^2} \frac{1}{(t - T + E)}$$

for $t \geq T = \frac{4L}{\mu} \max\{\frac{L\mu}{N}\|w_0 - w_*\|^2, 1\} - \frac{4L}{\mu}$.

Proof Using the beginning of the proof of Theorem 1, taking the expectation to (24), with $0 < \eta_t \leq \frac{1}{2L}$, we have

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq (1 - \mu\eta_t)\mathbb{E}[\|w_t - w_*\|^2] + \eta_t^2 N.$$

We first show that

$$\mathbb{E}[\|w_t - w_*\|^2] \leq \frac{N}{\mu^2} G \frac{1}{(t+E)}, \tag{25}$$

where $G = \max\{I, J\}$, and

$$I = \frac{E\mu^2}{N} \mathbb{E}[\|w_0 - w_*\|^2] > 0,$$
$$J = \frac{\alpha^2}{\alpha - 1} > 0.$$

We use mathematical induction to prove (25) (this trick is based on the idea from Bottou et al. (2016)). Let $t = 0$, we have

$$\mathbb{E}[\|w_0 - w_*\|^2] \leq \frac{NG}{\mu^2 E},$$

which is obviously true since $G \geq \frac{E\mu^2}{N}\|w_0 - w_*\|^2$.

Suppose it is true for t, we need to show that it is also true for $t+1$. We have

$$\begin{aligned}
\mathbb{E}[\|w_{t+1} - w_*\|^2] &\leq \left(1 - \frac{\alpha}{t+E}\right) \frac{NG}{\mu^2(t+E)} + \frac{\alpha^2 N}{\mu^2(t+E)^2} \\
&= \left(\frac{t+E-\alpha}{\mu^2(t+E)^2}\right) NG + \frac{\alpha^2 N}{\mu^2(t+E)^2} \\
&= \left(\frac{t+E-1}{\mu^2(t+E)^2}\right) NG - \left(\frac{\alpha - 1}{\mu^2(t+E)^2}\right) NG + \frac{\alpha^2 N}{\mu^2(t+E)^2}.
\end{aligned}$$

Since $G \geq \frac{\alpha^2}{\alpha - 1}$,

$$-\left(\frac{\alpha - 1}{\mu^2(t+E)^2}\right) NG + \frac{\alpha^2 N}{\mu^2(t+E)^2} \leq 0.$$

This implies

$$\begin{aligned}
\mathbb{E}[\|w_{t+1} - w_*\|^2] &\leq \left(\frac{t+E-1}{\mu^2(t+E)^2}\right) NG \\
&= \left(\frac{(t+E)^2 - 1}{(t+E)^2}\right) \frac{NG}{\mu^2(t+E+1)} \\
&\leq \frac{NG}{\mu^2(t+E+1)}.
\end{aligned}$$

This proves (25) by induction in t.

Notice that the induction proof of (25) holds more generally for $E \geq \frac{2\alpha L}{\mu}$ with $\alpha > 1$ (this is sufficient for showing $\eta_t \leq \frac{1}{2L}$. In this more general interpretation we can see that the convergence rate is minimized for I minimal, i.e., $E = \frac{2\alpha L}{\mu}$ and for this reason we have fixed E as such in the theorem statement.

Notice that

$$G = \max\{I, J\} = \max\{\frac{2\alpha L \mu}{N} \mathbb{E}[\|w_0 - w_*\|^2], \frac{\alpha^2}{\alpha - 1}\}.$$

We choose $\alpha = 2$ such that η_t only depends on known parameters μ and L. For this α we obtain

$$G = 4 \max\{\frac{L\mu}{N}\mathbb{E}[\|w_0 - w_*\|^2], 1\}.$$

For $T = \frac{4L}{\mu}\max\{\frac{L\mu}{N}\mathbb{E}[\|w_0 - w_*\|^2], 1\} - \frac{4L}{\mu}$, we have that according to (25)

$$\begin{aligned}\frac{L\mu}{N}\mathbb{E}[\|w_T - w_*\|^2] &\leq \frac{L\mu}{N}\frac{N}{\mu^2}\frac{G}{(T+E)} \\ &= \frac{L}{\mu}\frac{4\max\{\frac{L\mu}{N}\mathbb{E}[\|w_0 - w_*\|^2], 1\}}{\frac{4L}{\mu}\max\{\frac{L\mu}{N}\mathbb{E}[\|w_0 - w_*\|^2], 1\}} = 1. \end{aligned} \tag{26}$$

Applying (25)with w_T as starting point rather than w_0 gives, for $t \geq \max\{T, 0\}$,

$$\mathbb{E}[\|w_t - w_*\|^2] \leq \frac{N}{\mu^2}G\frac{1}{(t - T + E)},$$

where G is now equal to

$$4 \max\{\frac{L\mu}{N}\mathbb{E}[\|w_T - w_*\|^2], 1\},$$

which equals 4, see (26). For any given w_0, we prove the theorem. ■

Appendix D. Analysis for Algorithm 2

D.1 Recurrence and Notation

We introduce the following notation: For each ξ, we define $D_\xi \subseteq \{1, \ldots, d\}$ as the set of possible non-zero positions in a vector of the form $\nabla f(w; \xi)$ for some w. We consider a fixed mapping from $u \in U$ to subsets $S_u^\xi \subseteq D_\xi$ for each possible ξ. In our notation we also let D_ξ represent the diagonal $d \times d$ matrix with ones exactly at the positions corresponding to D_ξ and with zeroes elsewhere. Similarly, S_u^ξ also denotes a diagonal matrix with ones at the positions corresponding to D_ξ.

We will use a probability distribution $p_\xi(u)$ to indicate how to randomly select a matrix S_u^ξ. We choose the matrices S_u^ξ and distribution $p_\xi(u)$ so that there exist d_ξ such that

$$d_\xi \mathbb{E}[S_u^\xi | \xi] = D_\xi, \tag{27}$$

where the expectation is over $p_\xi(u)$.

We will restrict ourselves to choosing *non-empty* sets S_u^ξ that partition D_ξ in D approximately equally sized sets together with uniform distributions $p_\xi(u)$ for some fixed D. So, if $D \leq |D_\xi|$, then sets have sizes $\lfloor |D_\xi|/D \rfloor$ and $\lceil |D_\xi|/D \rceil$. For the special case $D > |D_\xi|$ we have exactly $|D_\xi|$ singleton sets of size 1 (in our definition we only use non-empty sets).

For example, for $D = \bar{\Delta}$, where

$$\bar{\Delta} = \max_\xi\{|D_\xi|\}$$

represents the maximum number of non-zero positions in any gradient computation $f(w;\xi)$, we have that for all ξ, there are exactly $|D_\xi|$ singleton sets S_u^ξ representing each of the elements in D_ξ. Since $p_\xi(u) = 1/|D_\xi|$ is the uniform distribution, we have $\mathbb{E}[S_u^\xi|\xi] = D_\xi/|D_\xi|$, hence, $d_\xi = |D_\xi|$. As another example at the other extreme, for $D = 1$, we have exactly one set $S_1^\xi = D_\xi$ for each ξ. Now $p_\xi(1) = 1$ and we have $d_\xi = 1$.

We define the parameter

$$\bar{\Delta}_D \stackrel{\text{def}}{=} D \cdot \mathbb{E}[\lceil |D_\xi|/D \rceil],$$

where the expectation is over ξ. We use $\bar{\Delta}_D$ in the leading asymptotic term for the convergence rate in our main theorem. We observe that

$$\bar{\Delta}_D \leq \mathbb{E}[|D_\xi|] + D - 1$$

and $\bar{\Delta}_D \leq \bar{\Delta}$ with equality for $D = \bar{\Delta}$.

For completeness we define

$$\Delta \stackrel{\text{def}}{=} \max_i \mathbb{P}\left(i \in D_\xi\right).$$

Let us remark, that $\Delta \in (0, 1]$ measures the probability of collision. Small Δ means that there is a small chance that the support of two random realizations of $\nabla f(w;\xi)$ will have an intersection. On the other hand, $\Delta = 1$ means that almost surely, the support of two stochastic gradients will have non-empty intersection.

With this definition of Δ it is an easy exercise to show that for iid ξ_1 and ξ_2 in a finite-sum setting (i.e., ξ_i and ξ_2 can only take on a finite set of possible values) we have

$$\begin{aligned}&\mathbb{E}[|\langle \nabla f(w_1;\xi_1), \nabla f(w_2;\xi_2)\rangle|] \\ &\leq \frac{\sqrt{\Delta}}{2}\left(\mathbb{E}[\|\nabla f(w_1;\xi_1)\|^2] + \mathbb{E}[\|\nabla f(w_2;\xi_2)\|^2]\right)\end{aligned} \tag{28}$$

(see Proposition 10 in Leblond et al. (2018)). We notice that in the non-finite sum setting we can use the property that for any two vectors a and b, $\langle a, b\rangle \leq (\|a\|^2 + \|b\|^2)/2$ and this proves (28) with Δ set to $\Delta = 1$. In our asymptotic analysis of the convergence rate, we will show how Δ plays a role in non-leading terms – this, with respect to the leading term, it will not matter whether we use $\Delta = 1$ or Δ equal the probability of collision (in the finite sum case).

We have

$$w_{t+1} = w_t - \eta_t d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t;\xi_t), \tag{29}$$

where $\hat{w}_t$ represents the vector used in computing the gradient $\nabla f(\hat{w}_t;\xi_t)$ and whose entries have been read (one by one) from an aggregate of a mix of previous updates that led to w_j, $j \leq t$. Here, we assume that

- updating/writing to vector positions is atomic, reading vector positions is atomic, and
- there exists a "delay" τ such that, for all t, vector $\hat{w}_t$ includes all the updates up to and including those made during the $(t-\tau)$-th iteration (where (29) defines the $(t+1)$-st iteration).

Notice that we do **not assume consistent reads and writes of vector positions**. We only assume that up to a "delay" τ all writes/updates are included in the values of positions that are being read.

According to our definition of τ, in (29) vector $\hat{w}_t$ represents an inconsistent read with entries that contain all of the updates made during the 1st to $(t-\tau)$-th iteration. Furthermore each entry in $\hat{w}_t$ includes some of the updates made during the $(t-\tau+1)$-th iteration up to t-th iteration. Each entry includes its own subset of updates because writes are inconsistent. We model this by "masks" $\Sigma_{t,j}$ for $t-\tau \leq j \leq t-1$. A mask $\Sigma_{t,j}$ is a diagonal 0/1-matrix with the 1s expressing which of the entry updates made in the $(j+1)$-th iteration are included in $\hat{w}_t$. That is,

$$\hat{w}_t = w_{t-\tau} - \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} \Sigma_{t,j} S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j). \tag{30}$$

Notice that the recursion (29) implies

$$w_t = w_{t-\tau} - \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j). \tag{31}$$

By combining (31) and (30) we obtain

$$w_t - \hat{w}_t = - \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} (I - \Sigma_{t,j}) S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j), \tag{32}$$

where I represents the identity matrix.

D.2 Main Analysis

We first derive a couple lemmas which will help us deriving our main bounds. In what follows let Assumptions 1, 2, 3 and 4 hold for all lemmas. We define

$$\mathcal{F}_t = \sigma(w_0, \xi_0, u_0, \sigma_0, \ldots, \xi_{t-1}, u_{t-1}, \sigma_{t-1}),$$

where

$$\sigma_{t-1} = (\Sigma_{t,t-\tau}, \ldots, \Sigma_{t,t-1}).$$

When we subtract τ from, for example, t and write $t-\tau$, we will actually mean $\max\{t-\tau, 0\}$.

Lemma 6 *We have*

$$\mathbb{E}[\|d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t, \xi_t] \leq D \|\nabla f(\hat{w}_t; \xi_t)\|^2$$

and

$$\mathbb{E}[d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t; \xi_t) | \mathcal{F}_t] = \nabla F(\hat{w}_t).$$

Proof For the first bound, if we take the expectation of $\|d_{\xi_t} S^{\xi_t}_{u_t} \nabla f(\hat{w}_t; \xi_t)\|^2$ with respect to u_t, then we have (for vectors x we denote the value if its i-th position by $[x]_i$)

$$\mathbb{E}[\|d_{\xi_t} S^{\xi_t}_{u_t} \nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t, \xi_t] = d^2_{\xi_t} \sum_u p_{\xi_t}(u) \|S^{\xi_t}_u \nabla f(\hat{w}_t; \xi_t)\|^2 = d^2_{\xi_t} \sum_u p_{\xi_t}(u) \sum_{i \in S^{\xi_t}_u} [\nabla f(\hat{w}_t; \xi_t)]^2_i$$

$$= d_{\xi_t} \sum_{i \in D_{\xi_t}} [\nabla f(\hat{w}_t; \xi_t)]^2_i = d_{\xi_t} \|f(\hat{w}_t; \xi_t)\|^2 \leq D \|\nabla f(\hat{w}_t; \xi_t)\|^2,$$

where the transition to the second line follows from (27).

For the second bound, if we take the expectation of $d_{\xi_t} S^{\xi_t}_{u_t} \nabla f(\hat{w}_t; \xi_t)$ wrt u_t, then we have:

$$\mathbb{E}[d_{\xi_t} S^{\xi_t}_{u_t} \nabla f(\hat{w}_t; \xi_t) | \mathcal{F}_t, \xi_t] = d_{\xi_t} \sum_u p_{\xi_t}(u) S^{\xi_t}_u \nabla f(\hat{w}_t; \xi_t) = D_{\xi_t} \nabla f(\hat{w}_t; \xi_t) = \nabla f(\hat{w}_t; \xi_t),$$

and this can be used to derive

$$\mathbb{E}[d_{\xi_t} S^{\xi_t}_{u_t} f(\hat{w}_t; \xi_t) | \mathcal{F}_t] = \mathbb{E}[\mathbb{E}[d_{\xi_t} S^{\xi_t}_{u_t} f(\hat{w}_t; \xi_t) | \mathcal{F}_t, \xi_t] | \mathcal{F}_t] = \nabla F(\hat{w}_t).$$

■

As a consequence of this lemma we derive a bound on the expectation of $\|w_t - \hat{w}_t\|^2$.

Lemma 7 *The expectation of $\|w_t - \hat{w}_t\|^2$ is at most*

$$\mathbb{E}[\|w_t - \hat{w}_t\|^2] \leq (1 + \sqrt{\Delta}\tau) D \sum_{j=t-\tau}^{t-1} \eta_j^2 (2L^2 \mathbb{E}[\|\hat{w}_j - w_*\|^2] + N).$$

Proof As shown in (32),

$$w_t - \hat{w}_t = - \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} (I - \Sigma_{t,j}) S^{\xi_j}_{u_j} \nabla f(\hat{w}_j; \xi_j).$$

This can be used to derive an expression for the square of its norm:

$$\begin{aligned} \|w_t - \hat{w}_t\|^2 &= \| \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} (I - \Sigma_{t,j}) S^{\xi_j}_{u_j} \nabla f(\hat{w}_j; \xi_j) \|^2 \\ &= \sum_{j=t-\tau}^{t-1} \|\eta_j d_{\xi_j} (I - \Sigma_{t,j}) S^{\xi_j}_{u_j} \nabla f(\hat{w}_j; \xi_j)\|^2 \\ &\quad + \sum_{i \neq j \in \{t-\tau, \ldots, t-1\}} \langle \eta_j d_{\xi_j} (I - \Sigma_{t,j}) S^{\xi_j}_{u_j} \nabla f(\hat{w}_j; \xi_j), \eta_i d_{\xi_i} (I - \Sigma_{t,j}) S^{\xi_i}_{u_i} \nabla f(\hat{w}_i; \xi_i) \rangle. \end{aligned}$$

Applying (28) to the inner products implies

$$\begin{aligned}\|w_t - \hat{w}_t\|^2 &\leq \sum_{j=t-\tau}^{t-1} \|\eta_j d_{\xi_j}(I - \Sigma_{t,j}) S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j)\|^2 \\ &\quad + \sum_{i \neq j \in \{t-\tau,\ldots,t-1\}} [\|\eta_j d_{\xi_j}(I - \Sigma_{t,j}) S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j)\|^2 \\ &\quad + \|\eta_i d_{\xi_i}(I - \Sigma_{t,j}) S_{u_i}^{\xi_i} \nabla f(\hat{w}_i; \xi_i)\|^2] \sqrt{\Delta}/2 \\ &= (1 + \sqrt{\Delta}\tau) \sum_{j=t-\tau}^{t-1} \|\eta_j d_{\xi_j}(I - \Sigma_{t,j}) S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j)\|^2 \\ &\leq (1 + \sqrt{\Delta}\tau) \sum_{j=t-\tau}^{t-1} \eta_j^2 \|d_{\xi_j} S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j)\|^2.\end{aligned}$$

Taking expectations shows

$$\mathbb{E}[\|w_t - \hat{w}_t\|^2] \leq (1 + \sqrt{\Delta}\tau) \sum_{j=t-\tau}^{t-1} \eta_j^2 \mathbb{E}[\|d_{\xi_j} S_{u_j}^{\xi_j} \nabla f(\hat{w}_j; \xi_j)\|^2].$$

Now, we can apply Lemma 15: We first take the expectation over u_j and this shows

$$\mathbb{E}[\|w_t - \hat{w}_t\|^2] \leq (1 + \sqrt{\Delta}\tau) \sum_{j=t-\tau}^{t-1} \eta_j^2 D \mathbb{E}[\|\nabla f(\hat{w}_j; \xi_j)\|^2].$$

From Lemma 1 we infer

$$\mathbb{E}[\|\nabla f(\hat{w}_j; \xi_j)\|^2] \leq 4L\mathbb{E}[F(\hat{w}_j) - F(w_*)] + N \tag{33}$$

and by L-smoothness, see Equation 7 with $\nabla F(w_*) = 0$,

$$F(\hat{w}_j) - F(w_*) \leq \frac{L}{2} \|\hat{w}_j - w_*\|^2.$$

Combining the above inequalities proves the lemma. ■

Together with the next lemma we will be able to start deriving a recursive inequality from which we will be able to derive a bound on the convergence rate.

Lemma 8 *Let $0 < \eta_t \leq \frac{1}{4LD}$ for all $t \geq 0$. Then,*

$$\mathbb{E}[\|w_{t+1} - w_*\|^2 | \mathcal{F}_t] \leq \left(1 - \frac{\mu\eta_t}{2}\right) \|w_t - w_*\|^2 + [(L + \mu)\eta_t + 2L^2\eta_t^2 D] \|\hat{w}_t - w_t\|^2 + 2\eta_t^2 DN.$$

Proof Since $w_{t+1} = w_t - \eta_t d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t; \xi_t)$, we have

$$\|w_{t+1} - w_*\|^2 = \|w_t - w_*\|^2 - 2\eta_t \langle d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t; \xi_t), (w_t - w_*) \rangle + \eta_t^2 \|d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t; \xi_t)\|^2.$$

We now take expectations over u_t and ξ_t and use Lemma 15:

$$\begin{aligned}
& \mathbb{E}[\|w_{t+1} - w_*\|^2 | \mathcal{F}_t] \\
\leq\ & \|w_t - w_*\|^2 - 2\eta_t \langle \nabla F(\hat{w}_t), (w_t - w_*) \rangle + \eta_t^2 D \mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t] \\
=\ & \|w_t - w_*\|^2 - 2\eta_t \langle \nabla F(\hat{w}_t), (w_t - \hat{w}_t) \rangle - 2\eta_t \langle \nabla F(\hat{w}_t), (\hat{w}_t - w_*) \rangle + \eta_t^2 D \mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t].
\end{aligned}$$

By (3) and (7), we have

$$-\langle \nabla F(\hat{w}_t), (\hat{w}_t - w_*) \rangle \leq -[F(\hat{w}_t) - F(w_*)] - \frac{\mu}{2} \|\hat{w}_t - w_*\|^2, \text{ and} \tag{34}$$

$$-\langle \nabla F(\hat{w}_t), (w_t - \hat{w}_t) \rangle \leq F(\hat{w}_t) - F(w_t) + \frac{L}{2} \|\hat{w}_t - w_t\|^2 \tag{35}$$

Thus, $\mathbb{E}[\|w_{t+1} - w_*\|^2 | \mathcal{F}_t]$ is at most

$$\begin{aligned}
& \overset{(34)(35)}{\leq} \|w_t - w_*\|^2 + 2\eta_t [F(\hat{w}_t) - F(w_t)] + L\eta_t \|\hat{w}_t - w_t\|^2 - 2\eta_t [F(\hat{w}_t) - F(w_*)] - \mu\eta_t \|\hat{w}_t - w_*\|^2 \\
& \quad + \eta_t^2 D \mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t] \\
& = \|w_t - w_*\|^2 - 2\eta_t [F(w_t) - F(w_*)] + L\eta_t \|\hat{w}_t - w_t\|^2 - \mu\eta_t \|\hat{w}_t - w_*\|^2 + \eta_t^2 D \mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t].
\end{aligned}$$

Since

$$-\|\hat{w}_t - w_*\|^2 = -\|(w_t - w_*) - (w_t - \hat{w}_t)\|^2 \overset{(20)}{\leq} -\frac{1}{2} \|w_t - w_*\|^2 + \|w_t - \hat{w}_t\|^2,$$

$\mathbb{E}[\|w_{t+1} - w_*\|^2 | \mathcal{F}_t, \sigma_t]$ is at most

$$(1 - \frac{\mu\eta_t}{2}) \|w_t - w_*\|^2 - 2\eta_t [F(w_t) - F(w_*)] + (L + \mu)\eta_t \|\hat{w}_t - w_t\|^2 + \eta_t^2 D \mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t].$$

We now use $\|a\|^2 = \|a - b + b\|^2 \leq 2\|a - b\|^2 + 2\|b\|^2$ for $\mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t]$ to obtain

$$\mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t] \leq 2\mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t) - \nabla f(w_t; \xi_t)\|^2 | \mathcal{F}_t] + 2\mathbb{E}[\|\nabla f(w_t; \xi_t)\|^2 | \mathcal{F}_t]. \tag{36}$$

By Lemma 1, we have

$$\mathbb{E}[\|\nabla f(w_t; \xi_t)\|^2 | \mathcal{F}_t] \leq 4L[F(w_t) - F(w_*)] + N. \tag{37}$$

Applying (6) twice gives

$$\mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t) - \nabla f(w_t; \xi_t)\|^2 | \mathcal{F}_t, \sigma_t] \leq L^2 \|\hat{w}_t - w_t\|^2$$

and together with (36) and (37) we obtain

$$\mathbb{E}[\|\nabla f(\hat{w}_t; \xi_t)\|^2 | \mathcal{F}_t] \leq 2L^2 \|\hat{w}_t - w_t\|^2 + 4L[F(w_t) - F(w_*)] + N.$$

Plugging this into the previous derivation yields

$$\begin{aligned}
\mathbb{E}[\|w_{t+1} - w_*\|^2 | \mathcal{F}_t] &\leq (1 - \frac{\mu\eta_t}{2}) \|w_t - w_*\|^2 - 2\eta_t [F(w_t) - F(w_*)] + (L + \mu)\eta_t \|\hat{w}_t - w_t\|^2 \\
&\quad + 2L^2 \eta_t^2 D \|\hat{w}_t - w_t\|^2 + 8L\eta_t^2 D[F(w_t) - F(w_*)] + 2\eta_t^2 DN \\
&= (1 - \frac{\mu\eta_t}{2}) \|w_t - w_*\|^2 + [(L + \mu)\eta_t + 2L^2 \eta_t^2 D] \|\hat{w}_t - w_t\|^2 \\
&\quad - 2\eta_t (1 - 4L\eta_t D)[F(w_t) - F(w_*)] + 2\eta_t^2 DN.
\end{aligned}$$

Since $\eta_t \leq \frac{1}{4LD}$, $-2\eta_t(1-4L\eta_t D)[F(w_t)-F(w_*)] \leq 0$ (we can get a negative upper bound by applying strong convexity but this will not improve the asymptotic behavior of the convergence rate in our main result although it would improve the constant of the leading term making the final bound applied to SGD closer to the bound of Theorem 2 for SGD),

$$\mathbb{E}[\|w_{t+1}-w_*\|^2|\mathcal{F}_t] \leq \left(1-\frac{\mu\eta_t}{2}\right)\|w_t-w_*\|^2 + [(L+\mu)\eta_t + 2L^2\eta_t^2 D]\|\hat{w}_t - w_t\|^2 + 2\eta_t^2 DN$$

and this concludes the proof. ■

Assume $0 < \eta_t \leq \frac{1}{4LD}$ for all $t \geq 0$. Then, after taking the full expectation of the inequality in Lemma 8, we can plug Lemma 7 into it which yields the recurrence

$$\begin{aligned}\mathbb{E}[\|w_{t+1}-w_*\|^2] \;\leq\; & \left(1-\frac{\mu\eta_t}{2}\right)\mathbb{E}[\|w_t-w_*\|^2] + \\ & [(L+\mu)\eta_t + 2L^2\eta_t^2 D](1+\sqrt{\Delta}\tau)D\sum_{j=t-\tau}^{t-1}\eta_j^2(2L^2\mathbb{E}[\|\hat{w}_j-w_*\|^2]+N) \\ & +2\eta_t^2 DN. \end{aligned} \tag{38}$$

This can be solved by using the next lemma. For completeness, we follow the convention that an empty product is equal to 1 and an empty sum is equal to 0, i.e.,

$$\prod_{i=h}^{k} g_i = 1 \text{ and } \sum_{i=h}^{k} g_i = 0 \text{ if } k < h. \tag{39}$$

Lemma 9 *Let Y_t, β_t and γ_t be sequences such that $Y_{t+1} \leq \beta_t Y_t + \gamma_t$, for all $t \geq 0$. Then,*

$$Y_{t+1} \leq (\sum_{i=0}^{t}[\prod_{j=i+1}^{t}\beta_j]\gamma_i) + (\prod_{j=0}^{t}\beta_j)Y_0. \tag{40}$$

Proof We prove the lemma by using induction. It is obvious that (40) is true for $t=0$ because $Y_1 \leq \beta_1 Y_0 + \gamma_1$. Assume as induction hypothesis that (40) is true for $t-1$. Since $Y_{t+1} \leq \beta_t Y_t + \gamma_t$,

$$\begin{aligned} Y_{t+1} &\leq \beta_t Y_t + \gamma_t \\ &\leq \beta_t[(\sum_{i=0}^{t-1}[\prod_{j=i+1}^{t-1}\beta_j]\gamma_i) + (\prod_{j=0}^{t-1}\beta_j)Y_0] + \gamma_t \\ &\overset{(39)}{=} (\sum_{i=0}^{t-1}\beta_t[\prod_{j=i+1}^{t-1}\beta_j]\gamma_i) + \beta_t(\prod_{j=0}^{t-1}\beta_j)Y_0 + (\prod_{j=t+1}^{t}\beta_j)\gamma_t \\ &= [(\sum_{i=0}^{t-1}[\prod_{j=i+1}^{t}\beta_j]\gamma_i) + (\prod_{j=t+1}^{t}\beta_j)\gamma_t] + (\prod_{j=0}^{t}\beta_j)Y_0 \\ &= (\sum_{i=0}^{t}[\prod_{j=i+1}^{t}\beta_j]\gamma_i) + (\prod_{j=0}^{t}\beta_j)Y_0. \end{aligned}$$

■

Applying the above lemma to (38) will yield the following bound.

Lemma 10 *Let* $\eta_t = \frac{\alpha_t}{\mu(t+E)}$ *with* $4 \leq \alpha_t \leq \alpha$ *and* $E = \max\{2\tau, \frac{4L\alpha D}{\mu}\}$. *Then, expectation* $\mathbb{E}[\|w_{t+1} - w_*\|^2]$ *is at most*

$$\frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left(\sum_{i=1}^{t}\left[4a_i(1+\sqrt{\Delta}\tau)[N\tau + 2L^2\sum_{j=i-\tau}^{i-1}\mathbb{E}[\|\hat{w}_j - w_*\|^2] + 2N\right]\right)$$
$$+ \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0 - w_*\|^2],$$

where $a_i = (L+\mu)\eta_i + 2L^2\eta_i^2 D$.

Proof Notice that we may use (38) because $\eta_t \leq \frac{1}{4LD}$ follows from $\eta_t = \frac{\alpha_t}{\mu(t+E)} \leq \frac{\alpha}{\mu(t+E)}$ combined with $E \geq \frac{4L\alpha D}{\mu}$. From (38) with $a_t = (L+\mu)\eta_t + 2L^2\eta_t^2 D$ and η_t being decreasing in t we infer

$$\begin{aligned}
&\mathbb{E}[\|w_{t+1} - w_*\|^2] \\
\leq\ & \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}[\|w_t - w_*\|^2] + a_t(1+\sqrt{\Delta}\tau)D\eta_{t-\tau}^2\sum_{j=t-\tau}^{t-1}(2L^2\mathbb{E}[\|\hat{w}_j - w_*\|^2] + N) + 2\eta_t^2 DN \\
=\ & \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}[\|w_t - w_*\|^2] + a_t(1+\sqrt{\Delta}\tau)D\eta_{t-\tau}^2[N\tau + 2L^2\sum_{j=t-\tau}^{t-1}\mathbb{E}[\|\hat{w}_j - w_*\|^2] + 2\eta_t^2 DN.
\end{aligned}$$

Since $E \geq 2\tau$, $\frac{1}{t-\tau+E} \leq \frac{2}{t+E}$. Hence, together with $\eta_{t-\tau} = \frac{\alpha_{t-\tau}}{\mu(t-\tau+E)} \leq \frac{\alpha}{\mu(t-\tau+E)}$ we have

$$\eta_{t-\tau}^2 \leq \frac{4\alpha^2}{\mu^2}\frac{1}{(t+E)^2}. \tag{41}$$

This translates the above bound into

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq \beta_t\mathbb{E}[\|w_t - w_*\|^2] + \gamma_t,$$

for

$$\begin{aligned}
&\beta_t = 1 - \frac{\mu\eta_t}{2}, \\
&\gamma_t = 4a_t(1+\sqrt{\Delta}\tau)D\frac{\alpha^2}{\mu^2}\frac{1}{(t+E)^2}[N\tau + 2L^2\sum_{j=t-\tau}^{t-1}\mathbb{E}[\|\hat{w}_j - w_*\|^2] + 2\eta_t^2 DN, where \\
&a_t = (L+\mu)\eta_t + 2L^2\eta_t^2 D.
\end{aligned}$$

Application of Lemma 9 for $Y_{t+1} = \mathbb{E}[\|w_{t+1} - w_*\|^2]$ and $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$ gives

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq \left(\sum_{i=0}^{t}\left[\prod_{j=i+1}^{t}\left(1 - \frac{\mu\eta_j}{2}\right)\right]\gamma_i\right) + \left(\prod_{j=0}^{t}\left(1 - \frac{\mu\eta_j}{2}\right)\right)\mathbb{E}[\|w_0 - w_*\|^2].$$

In order to analyze this formula, since $\eta_j = \frac{\alpha_j}{\mu(j+E)}$ with $\alpha_j \geq 4$, we have

$$1 - \frac{\mu\eta_j}{2} = 1 - \frac{\alpha_j}{2(j+E)} \leq 1 - \frac{2}{j+E},$$

Hence (we can also use $1 - x \leq e^{-x}$ which leads to similar results and can be used to show that our choice for η_t leads to the tightest convergence rates in our framework),

$$\begin{aligned}\prod_{j=i}^{t}\left(1 - \frac{\mu\eta_j}{2}\right) &\leq \prod_{j=i}^{t}\left(1 - \frac{2}{j+E}\right) = \prod_{j=i}^{t}\frac{j+E-2}{j+E} \\ &= \frac{i+E-2}{i+E}\frac{i+E-1}{i+E+1}\frac{i+E}{i+E+2}\frac{i+E+1}{i+E+3}\cdots\frac{t+E-3}{t+E-1}\frac{t+E-2}{t+E} \\ &= \frac{(i+E-2)(i+E-1)}{(t+E-1)(t+E)} \leq \frac{(i+E-1)^2}{(t+E-1)(t+E)} \leq \frac{(i+E)^2}{(t+E-1)^2}.\end{aligned}$$

From this calculation we infer that

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq \left(\sum_{i=0}^{t}\left[\frac{(i+E)^2}{(t+E-1)^2}\right]\gamma_i\right) + \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0 - w_*\|^2]. \tag{42}$$

Now, we substitute $\eta_i \leq \frac{\alpha}{\mu(i+E)}$ in γ_i and compute

$$\begin{aligned}&\frac{(i+E)^2}{(t+E-1)^2}\gamma_i \\ &= \frac{(i+E)^2}{(t+E-1)^2}4a_i(1+\sqrt{\Delta}\tau)D\frac{\alpha^2}{\mu^2}\frac{1}{(i+E)^2}[N\tau + 2L^2\sum_{j=i-\tau}^{i-1}\mathbb{E}[\|\hat{w}_j - w_*\|^2] + \frac{(i+E)^2}{(t+E-1)^2}2ND\frac{\alpha^2}{\mu^2(i+E)^2} \\ &= \frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left[4a_i(1+\sqrt{\Delta}\tau)[N\tau + 2L^2\sum_{j=i-\tau}^{i-1}\mathbb{E}[\|\hat{w}_j - w_*\|^2] + 2N\right].\end{aligned}$$

Substituting this in (42) proves the lemma. ■

As an immediate corollary we can apply the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ to $\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2]$ to obtain

$$\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \leq 2\mathbb{E}[\|\hat{w}_{t+1} - w_{t+1}\|^2] + 2\mathbb{E}[\|w_{t+1} - w_*\|^2], \tag{43}$$

which in turn can be bounded by the previous lemma together with Lemma 7:

$$\begin{aligned}\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \quad &\leq \quad 2(1+\sqrt{\Delta}\tau)D\sum_{j=t+1-\tau}^{t}\eta_j^2(2L^2\mathbb{E}[\|\hat{w}_j - w_*\|^2] + N) + \\ &\quad 2\frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left(\sum_{i=1}^{t}\left[4a_i(1+\sqrt{\Delta}\tau)[N\tau + 2L^2\sum_{j=i-\tau}^{i-1}\mathbb{E}[\|\hat{w}_j - w_*\|^2] + 2N\right]\right) + \\ &\quad \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0 - w_*\|^2].\end{aligned}$$

Now assume a decreasing sequence Z_t for which we want to prove that $\mathbb{E}[\|\hat{w}_t - w_*\|^2] \leq Z_t$ by induction in t. Then, the above bound can be used together with the property that Z_t and η_t are decreasing in t to show

$$\sum_{j=t+1-\tau}^{t} \eta_j^2(2L^2\mathbb{E}[\|\hat{w}_j - w_*\|^2] + N) \leq \tau\eta_{t-\tau}^2(2L^2 Z_{t+1-\tau} + N) \leq 4\tau\frac{\alpha^2}{\mu^2}\frac{1}{(t+E-1)^2}(2L^2 Z_{t+1-\tau} + N),$$

where the last inequality follows from (41), and

$$\sum_{j=i-\tau}^{i-1} \mathbb{E}[\|\hat{w}_j - w_*\|^2] \leq \tau Z_{i-\tau}.$$

From these inequalities we infer

$$\begin{aligned}
\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \quad \leq \quad & 8(1+\sqrt{\Delta}\tau)\tau D\frac{\alpha^2}{\mu^2}\frac{1}{(t+E-1)^2}(2L^2 Z_{t+1-\tau} + N) + \\
& 2\frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left(\sum_{i=1}^{t}\left[4a_i(1+\sqrt{\Delta}\tau)[N\tau + 2L^2\tau Z_{i-\tau}] + 2N\right]\right) + \\
& \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0 - w_*\|^2]. \qquad (44)
\end{aligned}$$

Even if we assume a constant $Z \geq Z_0 \geq Z_1 \geq Z_2 \geq \ldots$, we can get a first bound on the convergence rate of vectors $\hat{w}^t$: Substituting Z gives

$$\begin{aligned}
\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \quad \leq \quad & 8(1+\sqrt{\Delta}\tau)\tau D\frac{\alpha^2}{\mu^2}\frac{1}{(t+E-1)^2}(2L^2 Z + N) + \\
& 2\frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left(\sum_{i=1}^{t}\left[4a_i(1+\sqrt{\Delta}\tau)[N\tau + 2L^2\tau Z] + 2N\right]\right) + \\
& \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0 - w_*\|^2]. \qquad (45)
\end{aligned}$$

Since $a_i = (L+\mu)\eta_i + 2L^2\eta_i^2 D$ and $\eta_i \leq \frac{\alpha}{\mu(i+E)}$, we have

$$\begin{aligned}
\sum_{i=1}^{t} a_i &= (L+\mu)\sum_{i=1}^{t}\eta_i + 2L^2 D\sum_{i=1}^{t}\eta_i^2 \\
&\leq (L+\mu)\sum_{i=1}^{t}\frac{\alpha}{\mu(i+E)} + 2L^2 D\sum_{i=1}^{t}\frac{\alpha^2}{\mu^2(i+E)^2} \\
&\leq \frac{(L+\mu)\alpha}{\mu}\sum_{i=1}^{t}\frac{1}{i} + \frac{2L^2\alpha^2 D}{\mu^2}\sum_{i=1}^{t}\frac{1}{i^2} \\
&\leq \frac{(L+\mu)\alpha}{\mu}(1+\ln t) + \frac{L^2\alpha^2 D\pi^2}{3\mu^2}, \qquad (46)
\end{aligned}$$

where the last inequality is a property of the harmonic sequence $\sum_{i=1}^{t}\frac{1}{i} \leq 1 + \ln t$ and $\sum_{i=1}^{t}\frac{1}{i^2} \leq \sum_{i=1}^{\infty}\frac{1}{i^2} = \frac{\pi^2}{6}$.

Substituting (46) in (45) and collecting terms yields

$$\begin{aligned}
&\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \\
\leq\ & 2\frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left(2Nt + 4(1+\sqrt{\Delta}\tau)\tau[N+2L^2Z]\left\{\frac{(L+\mu)\alpha}{\mu}(1+\ln t) + \frac{L^2\alpha^2 D\pi^2}{3\mu^2+1}\right\}\right) + \\
& \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0 - w_*\|^2]. \qquad (47)
\end{aligned}$$

Notice that the asymptotic behavior in t is dominated by the term

$$\frac{4\alpha^2 DN}{\mu^2}\frac{t}{(t+E-1)^2}.$$

If we define Z_{t+1} to be the right hand side of (47) and observe that this Z_{t+1} is decreasing and a constant Z exists (since the terms with Z decrease much faster in t compared to the dominating term), then this Z_{t+1} satisfies the derivations done above and a proof by induction can be completed.

Our derivations prove our main result: The expected convergence rate of read vectors is

$$\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \leq \frac{4\alpha^2 DN}{\mu^2}\frac{t}{(t+E-1)^2} + O\left(\frac{\ln t}{(t+E-1)^2}\right).$$

We can use this result in Lemma 10 in order to show that the expected convergence rate $\mathbb{E}[\|w_{t+1} - w_*\|^2]$ satisfies the same bound.

We remind the reader, that in the $(t+1)$-th iteration at most $\leq \lceil |D_{\xi_t}|/D \rceil$ vector positions are updated. Therefore the expected number of single vector entry updates is at most $\bar{\Delta}_D/D$.

Theorem 6. *Suppose Assumptions 1, 2, 3 and 4 and consider Algorithm 2. Let* $\eta_t = \frac{\alpha_t}{\mu(t+E)}$ *with* $4 \leq \alpha_t \leq \alpha$ *and* $E = \max\{2\tau, \frac{4L\alpha D}{\mu}\}$. *Then,* $t' = t\bar{\Delta}_D/D$ *is the expected number of single vector entry updates after* t *iterations and expectations* $\mathbb{E}[\|\hat{w}_t - w_*\|^2]$ *and* $\mathbb{E}[\|w_t - w_*\|^2]$ *are at most*

$$\frac{4\alpha^2 DN}{\mu^2}\frac{t}{(t+E-1)^2} + O\left(\frac{\ln t}{(t+E-1)^2}\right).$$

D.3 Convergence without Convexity of Component Functions

For the non-convex case, L in (33) must be replaced by $L\kappa$ and as a result L^2 in Lemma 7 must be replaced by $L^2\kappa$. Also L in (37) must be replaced by $L\kappa$. We now require that $\eta_t \leq \frac{1}{4L\kappa D}$ so that $-2\eta_t(1 - 4L\kappa\eta_t D)[F(w_t) - F(w_*)] \leq 0$. This leads to Lemma 8 where no changes are needed except requiring $\eta_t \leq \frac{1}{4L\kappa D}$. The changes in Lemmas 7 and 8 lead to a Lemma 10 where we require $E \geq \frac{4L\kappa\alpha D}{\mu}$ and where in the bound of the expectation L^2 must be replaced by $L^2\kappa$. This perculates through to inequality (47) with a similar change finally leading to Theorem 7, i.e., Theorem 6 where we only need to strengthen the condition on E to $E \geq \frac{4L\kappa\alpha D}{\mu}$ in order to remove Assumption 3.

D.4 Sensitivity to τ

What about the upper bound's sensitivity with respect to τ? Suppose τ is not a constant but an increasing function of t, which also makes E a function of t:

$$\frac{2L\alpha D}{\mu} \leq \tau(t) \leq t \text{ and } E(t) = 2\tau(t).$$

In order to obtain a similar theorem we increase the lower bound on α_t to

$$12 \leq \alpha_t \leq \alpha.$$

This allows us to modify the proof of Lemma 10 where we analyse the product

$$\prod_{j=i}^{t} \left(1 - \frac{\mu\eta_j}{2}\right).$$

Since $\alpha_j \geq 12$ and $E(j) = 2\tau(j) \leq 2j$,

$$1 - \frac{\mu\eta_j}{2} = 1 - \frac{\alpha_j}{2(j + E(j))} \leq 1 - \frac{12}{2(j+2j)} = 1 - \frac{2}{j} \leq 1 - \frac{2}{j+1}.$$

The remaining part of the proof of Lemma 10 continues as before where constant E in the proof is replaced by 1. This yields instead of (42)

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq \left(\sum_{i=1}^{t} \left[\frac{(i+1)^2}{t^2}\right] \gamma_i\right) + \frac{4}{t^2}\mathbb{E}[\|w_0 - w_*\|^2].$$

We again substitute $\eta_i \leq \frac{\alpha}{\mu(i+E(i))}$ in γ_i, realize that $\frac{(i+1)}{(i+E(i))} \leq 1$, and compute

$$\begin{aligned}
&\frac{(i+1)^2}{t^2}\gamma_i \\
&= \frac{(i+1)^2}{t^2} 4a_i(1 + \sqrt{\Delta}\tau(i))D\frac{\alpha^2}{\mu^2}\frac{1}{(i+E(i))^2}[N\tau(i) + 2L^2 \sum_{j=i-\tau(i)}^{i-1} \mathbb{E}[\|\hat{w}_j - w_*\|^2] \\
&+ \frac{(i+1)^2}{t^2} 2ND\frac{\alpha^2}{\mu^2(i+E(i))^2} \\
&\leq \frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left[4a_i(1 + \sqrt{\Delta}\tau(i))[N\tau(i) + 2L^2 \sum_{j=i-\tau(i)}^{i-1} \mathbb{E}[\|\hat{w}_j - w_*\|^2] + 2N\right].
\end{aligned}$$

This gives a new Lemma 10:

Lemma 11 *Assume $\frac{2L\alpha D}{\mu} \leq \tau(t) \leq t$ with $\tau(t)$ monotonic increasing. Let $\eta_t = \frac{\alpha_t}{\mu(t+E(t))}$ with $12 \leq \alpha_t \leq \alpha$ and $E(t) = 2\tau(t)$. Then, expectation $\mathbb{E}[\|w_{t+1} - w_*\|^2]$ is at most*

$$\frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left(\sum_{i=1}^{t}\left[4a_i(1 + \sqrt{\Delta}\tau(i))[N\tau(i) + 2L^2 \sum_{j=i-\tau(i)}^{i-1} \mathbb{E}[\|\hat{w}_j - w_*\|^2] + 2N\right]\right) + \frac{4}{t^2}\mathbb{E}[\|w_0 - w_*\|^2],$$

where $a_i = (L + \mu)\eta_i + 2L^2\eta_i^2 D$.

Now we can continue the same analysis that led to Theorem 6 and conclude that there exists a constant Z such that, see (45),

$$\begin{aligned}\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] &\leq 8(1+\sqrt{\Delta}\tau(t))\tau(t)D\frac{\alpha^2}{\mu^2}\frac{1}{t^2}(2L^2Z+N) + \\ &\quad 2\frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left(\sum_{i=1}^{t}\left[4a_i(1+\sqrt{\Delta}\tau(i))[N\tau(i)+2L^2\tau(i)Z]+2N\right]\right) + \\ &\quad \frac{4}{t^2}\mathbb{E}[\|w_0 - w_*\|^2]. \end{aligned} \tag{48}$$

Let us assume

$$\tau(t) \leq \sqrt{t \cdot L(t)}, \tag{49}$$

where

$$L(t) = \frac{1}{\ln t} - \frac{1}{(\ln t)^2}$$

which has the property that the derivative of $t/(\ln t)$ is equal to $L(t)$. Now we observe

$$\begin{aligned}\sum_{i=1}^{t} a_i\tau(i)^2 &= \sum_{i=1}^{t}[(L+\mu)\eta_i + 2L^2\eta_i^2 D]\tau(i)^2 \leq \sum_{i=1}^{t}[(L+\mu)\frac{\alpha}{\mu i} + 2L^2\frac{\alpha^2}{\mu^2 i^2}D]\cdot iL(i) \\ &= \frac{(L+\mu)\alpha}{\mu}\sum_{i=1}^{t}L(i) + O(\ln t) = \frac{(L+\mu)\alpha}{\mu}\frac{t}{\ln t} + O(\ln t)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{t} a_i\tau(i) &= \sum_{i=1}^{t}[(L+\mu)\eta_i + 2L^2\eta_i^2 D]\tau(i) \leq \sum_{i=1}^{t}[(L+\mu)\frac{\alpha}{\mu i} + 2L^2\frac{\alpha^2}{\mu^2 i^2}D]\cdot\sqrt{i}) \\ &= O(\sum_{i=1}^{t}\frac{1}{\sqrt{i}}) = O(\sqrt{t}).\end{aligned}$$

Substituting both inequalities in (48) gives

$$\begin{aligned}\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] &\leq 8(1+\sqrt{\Delta}\tau(t))\tau(t)D\frac{\alpha^2}{\mu^2}\frac{1}{t^2}(2L^2Z+N) + \\ &\quad 2\frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left(2Nt + 4\sqrt{\Delta}[\frac{(L+\mu)\alpha}{\mu}\frac{t}{\ln t} + O(\ln t)][N+2L^2Z] + O(\sqrt{t})\right) + \\ &\quad \frac{4}{t^2}\mathbb{E}[\|w_0 - w_*\|^2] \\ &\leq 2\frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left(2Nt + 4\sqrt{\Delta}[(1+\frac{(L+\mu)\alpha}{\mu})\frac{t}{\ln t} + O(\ln t)][N+2L^2Z] + O(\sqrt{t})\right) + \\ &\quad \frac{4}{t^2}\mathbb{E}[\|w_0 - w_*\|^2]\end{aligned} \tag{50}$$

Again we define Z_{t+1} as the right hand side of this inequality. Notice that $Z_t = O(1/t)$, since the above derivation proves

$$\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \leq \frac{4\alpha^2 DN}{\mu^2}\frac{1}{t} + O(\frac{1}{t\ln t}).$$

Summarizing we have the following main lemma:

Lemma 12 *Let Assumptions 1, 2, 3 and 4 hold and consider Algorithm 2. Assume* $\frac{2L\alpha D}{\mu} \leq \tau(t) \leq \sqrt{t \cdot L(t)}$ *with* $\tau(t)$ *monotonic increasing. Let* $\eta_t = \frac{\alpha_t}{\mu(t+2\tau(t))}$ *with* $12 \leq \alpha_t \leq \alpha$. *Then, the expected convergence rate of read vectors is*

$$\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \quad \leq \quad \frac{4\alpha^2 DN}{\mu^2}\frac{1}{t} + O(\frac{1}{t\ln t}),$$

where $L(t) = \frac{1}{\ln t} - \frac{1}{(\ln t)^2}$. *The expected convergence rate* $\mathbb{E}[\|w_{t+1} - w_*\|^2]$ *satisfies the same bound.*

Notice that we can plug $Z_t = O(1/t)$ back into an equivalent of (44) where we may bound $Z_{i-\tau(i)} = O(1/(i-\tau(i))$ which replaces Z in the second line of (45). On careful examination this leads to a new upper bound (50) where the $2L^2 Z$ terms gets absorped in a higher order term. This can be used to show that, for

$$t \geq T_0 = \exp[2\sqrt{\Delta}(1 + \frac{(L+\mu)\alpha}{\mu})],$$

the higher order terms that contain $\tau(t)$ (as defined above) are at most the leading term as given in Lemma 12.

Upper bound (50) also shows that, for

$$t \geq T_1 = \frac{\mu^2}{\alpha^2 ND}\|w_0 - w_*\|^2,$$

the higher order term that contains $\|w_0 - w_*\|^2$ is at most the leading term.

D.5 Convergence of Hogwild! with probability 1

Lemma 13 *Let us consider the sequence* $w_0, w_1, w_2, \ldots, w_t, \ldots, w_n$ *generated by* (29)*:*

$$w_{t+1} = w_t - \eta_t d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t; \xi_t),$$

and define

$$m_t = \max_{0 \leq i \leq n, 0 \leq t' \leq t} \|\nabla f(w_{t'}; \xi_i)\|.$$

Then,

$$m_t \leq m_0 \exp(LD \sum_{i=0}^{t-1} \eta_i), \tag{51}$$

where $d_{\xi_t} \leq D$ *for all* ξ_t.

Proof

From the L-smoothness assumption (i.e., $\|\nabla f(w;\xi) - \nabla f(w';\xi)\| \leq L\|w - w'\|,\ \forall w, w' \in \mathbb{R}^d$), for any given ξ_i:

$$\begin{aligned}\|\nabla f(w_{t+1};\xi_i) - \nabla f(w_t;\xi_i)\| &\leq L\|w_{t+1} - w_t\| = L\|\eta_t d_{\xi_t} S_{u_t}^{\xi_t} \nabla f(\hat{w}_t;\xi_t)\| \\ &\leq LD\eta_t \|\nabla f(\hat{w}_t;\xi_t)\|.\end{aligned}$$

Since

$$m_t = \max_{0\leq i\leq n, 0\leq t'\leq t} \|\nabla f(w_{t'};\xi_i)\|,$$

we have

$$\|\nabla f(w_{t+1};\xi_i) - \nabla f(w_t;\xi_i)\| \leq LD\eta_t m_t$$

for any $i \in [n]$ and t.
Using the triangular inequality, we obtain

$$\begin{aligned}\|\nabla f(w_{t+1};\xi_i)\| &\leq \|\nabla f(w_t;\xi_i)\| + \|\nabla f(w_{t+1};\xi_i) - \nabla f(w_t;\xi_i)\| \\ &\overset{\forall i, \|\nabla f(w_t;\xi_i)\|\leq m_t}{\leq} m_t + LD\eta_t \|\nabla f(\hat{w}_t;\xi_t)\| \\ &\overset{\forall i, \|\nabla f_i(\hat{w}_t;\xi_i)\|\leq m_t}{\leq} (1+LD\eta_t)m_t.\end{aligned}$$

Moreover, the result above implies $m_{t+1} \leq (1+LD\eta_t)m_t$ and unrolling m_t yields

$$m_{t+1} \leq m_0 \prod_{i=0}^{t}(1+LD\eta_i).$$

For all $x \geq 0$, it is always true that $1 + x \leq \exp(x)$. Hence, we have

$$m_{t+1} \leq m_0 \prod_{i=0}^{t}(1+LD\eta_i) \leq m_0 \prod_{i=0}^{t}\exp(LD\eta_i) \leq m_0 \exp(LD[\sum_{i=0}^{t}\eta_i]).$$

■

Theorem 5 (Sufficient conditions for almost sure convergence for Hogwild!) *Let Assumptions 1, 2, 3 and 4 hold. Consider Hogwild! method described in Algorithm* (2) *with a stepsize sequence such that*

$$0 < \eta_t = \frac{1}{LD(2+\beta)(k+t)} < \frac{1}{4LD}, \beta > 0, k \geq 3\tau.$$

Then, the following holds w.p.1 (almost surely)

$$\|w_t - w_*\| \to 0.$$

Proof

As shown in Lemma 8, for $0 < \eta_t \leq \frac{1}{4LD}$, we have

$$\begin{aligned}\mathbb{E}[\|w_{t+1} - w_*\|^2 | \mathcal{F}_t] &\leq \left(1 - \frac{\mu\eta_t}{2}\right)\|w_t - w_*\|^2 + [(L+\mu)\eta_t + 2L^2\eta_t^2 D]\|\hat{w}_t - w_t\|^2 + 2\eta_t^2 DN \\ &= \|w_t - w_*\|^2 - \frac{\mu\eta_t}{2}\|w_t - w_*\|^2 + [(L+\mu)\eta_t + 2L^2\eta_t^2 D]\|\hat{w}_t - w_t\|^2 + 2\eta_t^2 DN.\end{aligned}$$

If we can show that $\sum_{t=0}^{\infty}[(L+\mu)\eta_t + 2L^2\eta_t^2 D]\|\hat{w}_t - w_t\|^2$ is finite, then it is straight forward to apply the proof technique from Theorem 1 to show that $\|w_t - w_*\|^2 \to 0$ w.p.1. From the proof of Lemma 7, we know $\|w_t - \hat{w}_t\|^2$ is at most

$$(1+\sqrt{\Delta}\tau)\sum_{j=t-\tau}^{t-1}\eta_j^2\|d_{\xi_j}S_{u_j}^{\xi_j}\nabla f(\hat{w}_j;\xi_j)\|^2 \leq (1+\sqrt{\Delta}\tau)D^2\sum_{j=t-\tau}^{t-1}\eta_j^2\|\nabla f(\hat{w}_j;\xi_j)\|^2$$
$$\leq (1+\sqrt{\Delta}\tau)D^2\tau m_t^2\eta_{t-\tau}^2.$$

Since $\eta_{t-\tau} = (1-\frac{\tau}{k+t-\tau})\eta_t < \frac{1}{2}\eta_t$ when $k \geq 3\tau$ for all $t \geq 0$, it yields $\|w_t - \hat{w}_t\|^2 < (1+\sqrt{\Delta}\tau)D^2\tau\frac{1}{4}\eta_t^2 m_t^2$. Hence $\sum_{t=0}^{\infty}[(L+\mu)\eta_t + 2L^2\eta_t^2 D]\|\hat{w}_t - w_t\|^2$ is at most

$$[(L+\mu)\eta_0 + 2L^2\eta_0^2 D](1+\sqrt{\Delta}\tau)D^2\tau\sum_{t=0}^{\infty}\eta_t^2 m_t^2.$$

Combining $m_t \leq m_0\exp(LD\sum_{i=0}^{t}\eta_i)$ (see (51)) and $\eta_i = \frac{1}{LD(2+\beta)(k+i)}$ yields

$$m_t \leq m_0\exp(\frac{1}{2+\beta}\sum_{i=1}^{t}\frac{1}{i}) \leq m_0\exp(\frac{1}{2+\beta}(1+\ln t)) \leq m_0\exp(\frac{1}{2+\beta})t^{\frac{1}{2+\beta}}.$$

The second inequality is a property of harmonic number $H_t = \sum_{i=1}^{t}\frac{1}{i} \leq 1+\ln t$. Hence,

$$\eta_t m_t \leq \frac{1}{L(2+\beta)t}m_0\exp(\frac{1}{2+\beta})t^{\frac{1}{2+\beta}} \leq \frac{m_0\exp(\frac{1}{2+\beta})}{L(2+\beta)}t^{\frac{-(1+\beta)}{2+\beta}}.$$

Hence, we obtain

$$(\eta_t m_t)^2 \leq [\frac{m_0\exp(\frac{1}{2+\beta})}{L(2+\beta)}]^2 t^{\frac{-(2+2\beta)}{2+\beta}} \leq [\frac{m_0\exp(\frac{1}{2+\beta})}{L(2+\beta)}]^2 t^{-(1+\rho)},$$

where $\rho = \frac{\beta}{2+\beta}$.

Due to the property of over-harmonic series, $\sum_{t=1}^{\infty}\frac{1}{t^{1+\rho}}$ converges for any $\rho > 0$. In other words, $\sum_{t=0}^{\infty}(\eta_t m_t)^2$ is finite or $\sum_{t=0}^{\infty}[(L+\mu)\eta_t + 2L^2\eta_t^2 D]\|\hat{w}_t - w_t\|^2$ is finite.

■

D.6 Convergence of Large Stepsizes

Theorem 9 *Let Assumptions 1, 3 and 2 hold. Consider Algorithm 1 with a stepsize sequence such that* $\eta_t \leq \frac{1}{2L}$, $\eta_t \to 0$, $\frac{d}{dt}\eta_t \leq 0$ *and* $\sum_{t=0}^{\infty}\eta_t \to \infty$. *Then,*

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \to 0.$$

Proof

As shown in (24)

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq (1 - \mu\eta_t)\mathbb{E}[\|w_t - w_*\|^2] + \eta_t^2 N,$$

when $\eta_t \leq \frac{1}{2L}$.

Let $Y_{t+1} = \mathbb{E}[\|w_{t+1} - w_*\|^2]$, $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$, $\beta_t = 1 - \mu\eta_t$ and $\gamma_t = \eta_t^2 N$. As proved in Lemma 9, if $Y_{t+1} \leq \beta_t Y_t + \gamma_t$, then

$$\begin{aligned} Y_{t+1} &\leq \sum_{i=0}^{t} [\prod_{j=i+1}^{t} \beta_j]\gamma_i + (\prod_{i=0}^{t} \beta_i)Y_1 \\ &= \sum_{i=0}^{t} [\prod_{j=i+1}^{t} (1 - \mu\eta_j)]\gamma_i + [\prod_{i=0}^{t}(1 - \mu\eta_i)]\mathbb{E}[\|w_1 - w_*\|^2] \end{aligned}$$

Let us define

$$n(j) = \mu\eta_j. \tag{52}$$

Since $1 - x \leq \exp(-x)$ for all $x \geq 0$,

$$\prod_{j=i+1}^{t} (1 - \mu\eta_j) \leq \exp(-\sum_{j=i+1}^{t} (\mu\eta_j)) = \exp(-\sum_{j=i+1}^{t} n(j)).$$

Furthermore, since $n(j)$ is decreasing in j, we have

$$\sum_{j=i+1}^{t} n(j) \geq \int_{x=i+1}^{t+1} n(x)dx.$$

With two inequalities above, we have:

$$\begin{aligned} Y_{t+1} &\leq \sum_{i=0}^{t} \exp(-\sum_{j=i+1}^{t} n(j))n^2(i)N + \exp(-\sum_{j=0}^{t} n(j))Y_0 \\ &\leq \sum_{i=0}^{t} \exp(-\int_{x=i+1}^{t+1} n(x)dx)n^2(i)N + \exp(-\int_{x=0}^{t+1} n(x)dx)Y_0 \\ &= \sum_{i=0}^{t} \exp(-[M(t+1) - M(i+1)])n^2(i)N + \exp(-M(t+1))Y_0, \end{aligned}$$

where

$$M(y) = \int_{x=0}^{y} n(x)dx \text{ and } \frac{d}{dy}M(y) = n(y).$$

We focus on

$$F = \sum_{i=0}^{t} \exp(-[M(t+1) - M(i+1)])n^2(i).$$

We notice that

$$F = \exp(-M(t+1)) \sum_{i=0}^{t} \exp(M(i+1))n^2(i).$$

We know that $\exp(M(x+1))$ increases and $n^2(x)$ decreases, hence, in the most general case either their product first decreases and then starts to increase or their product keeps on increasing. We first discuss the decreasing and increasing case. Let $a(x) = \exp(M(x+1))n^2(x)$ denote this product and let integer $j \geq 0$ be such that $a(0) \geq a(1) \geq \ldots \geq a(j)$ and $a(j) \leq a(j+1) \leq a(j+2) \leq \ldots$ (notice that $j = 0$ expresses the situation where $a(i)$ only increases). Function $a(x)$ for $x \geq 0$ is minimized for some value h in $[j, j+1)$. For $1 \leq i \leq j$, $a(i) \leq \int_{x=i-1}^{i} a(x)dx$, and for $j+1 \leq i$, $a(i) \leq \int_{x=i}^{i+1} a(x)dx$. This yields the upper bound

$$\begin{aligned}\sum_{i=0}^{t} a(i) &= a(0) + \sum_{i=1}^{j} a(i) + \sum_{i=j+1}^{t} a(i) \\ &\leq a(0) + \int_{x=0}^{j} a(x)dx + \int_{x=j+1}^{t+1} a(x)dx, \\ &\leq a(0) + \int_{x=0}^{t+1} a(x)dx.\end{aligned}$$

The same upper bound holds for the other case as well, i.e., if $a(i)$ is only decreasing. We conclude

$$F \leq \exp(-M(t+1))[\exp(M(2))n^2(0) + \int_{x=0}^{j+1} \exp(M(x+1))n^2(x)dx].$$

Combined with

$$M(x+1) = \int_{y=0}^{x+1} n(y)dy \leq \int_{y=0}^{x} n(y)dy + n(x) = M(x) + n(x)$$

we obtain

$$\begin{aligned}F &\leq \exp(-M(t+1))[\exp(M(2))n^2(0) + \int_{x=0}^{t+1} \exp(M(x))n^2(x)\exp(n(x))dx] \\ &\leq \exp(-M(t+1))[\exp(M(2))n^2(0) + \exp(n(0))\int_{x=0}^{t+1} \exp(M(x))n^2(x)dx].\end{aligned}$$

This gives

$$\begin{aligned}Y_{t+1} &\leq \exp(-M(t+1))[\exp(M(1))n^2(0) + \exp(n(0))\int_{x=0}^{t+1} \exp(M(x))n^2(x)dx]N \\ &\quad + \exp(-M(t+1))Y_0 \\ &= N\exp(n(0))C(t+1) + \exp(-M(t+1))[\exp(M(1))n^2(0)N + Y_0], \qquad (53)\end{aligned}$$

where

$$C(t) = \exp(-M(t))\int_{x=0}^{t} \exp(M(x))n^2(x)dx.$$

For $y \le t$, we derive (notice that $n(x)$ is decreasing)

$$\begin{aligned}
C(t) &= \exp(-M(t)) \int_{x=0}^{t} \exp(M(x)) n^2(x) dx \\
&= \exp(-M(t)) \int_{x=0}^{y} \exp(M(x)) n^2(x) dx + \exp(-M(t)) \int_{x=y}^{t} \exp(M(x)) n^2(x) dx \\
&\le \exp(-M(t)) \int_{x=0}^{y} \exp(M(x)) n^2(x) dx + \exp(-M(t)) \int_{x=y}^{t} \exp(M(x)) n(x) n(y) dx \\
&= \exp(-M(t)) \int_{x=0}^{y} \exp(M(x)) n^2(x) dx + \exp(-M(t)) n(y) \int_{x=y}^{t} \exp(M(x)) n(x) dx \\
&= \exp(-M(t)) \int_{x=0}^{y} \exp(M(x)) n^2(x) dx + n(y)[1 - \exp(-M(t)) \exp(M(y))] \\
&\le \exp(-M(t)) \int_{x=0}^{y} \exp(M(x)) n^2(x) dx + n(y).
\end{aligned}$$

Let $\epsilon > 0$. Since $n(y) \to 0$ as $y \to \infty$, there exists a y such that $n(y) \le \epsilon/2$. Since $M(t) \to \infty$ as $t \to \infty$, $\exp(-M(t)) \to 0$ as $t \to \infty$. Hence, there exists a T such that for $t \ge T$, $\exp(-M(t)) \int_{x=0}^{y} \exp(M(x)) n^2(x) dx \le \epsilon/2$. This implies $C(t) \le \epsilon$ for $t \ge T$. This proves $C(t) \to 0$ as $t \to \infty$, and we conclude $Y_t \to 0$ as $t \to \infty$.

■

Theorem 10 *Let Assumptions 1, 3 and 2 hold. Consider Algorithm 1 with a stepsize sequence such that* $\eta_t \le \frac{1}{2L}$, $\eta_t \to 0$, $\frac{d}{dt}\eta_t \le 0$, *and* $\sum_{t=0}^{\infty} \eta_t \to \infty$. *Then,* $\mathbb{E}[\|w_{t+1} - w_*\|^2]$ *is at most*

$$N \exp(n(0)) 2n(M^{-1}(\ln[\frac{n(t+1)}{n(0)}] + M(t+1))) + \exp(-M(t+1))[\exp(M(1)) n^2(0) N + \mathbb{E}[\|w_0 - w_*\|^2]],$$

where $n(t) = \mu\eta_t$ *and* $M(t) = \int_{x=0}^{t} n(x) dx$.

Proof Now, we are compute the convergence rate of $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$ for a given $M(t)$. We have shown that $C(t) \le \exp(-M(t)) \int_{x=0}^{y} \exp(M(x)) n^2(x) dx + n(y)$. We are interested in the following problem: finding the largest $y \le t$ such as

$$\exp(-M(t))[\int_{x=0}^{y} \exp(M(x)) n^2(x) dx] \le n(y).$$

The solution is equal to

$$y = \sup\{y \le t : \exp(-M(t))[\int_{x=0}^{y} \exp(M(x)) n^2(x) dx] \le n(y)\}.$$

Since $M(x)$ always decreases,

$$
\begin{aligned}
y &\geq \sup\{y \leq t : \exp(-M(t))[\int_{x=0}^{y} \exp(M(x))n(x)n(0)dx] \leq n(y)\} \\
&= \sup\{y \leq t : \exp(-M(t))n(0)[\exp(M(y)) - \exp(M(0))] \leq n(y)\} \\
&= \sup\{y \leq t : \exp(M(y)) \leq \exp(M(0)) + \frac{n(y)}{n(0)} \exp(M(t))\} \\
&\geq \sup\{y \leq t : \exp(M(y)) \leq \exp(M(0)) + \frac{n(t)}{n(0)} \exp(M(t))\} \\
&\geq \sup\{y \leq t : \exp(M(y)) \leq \frac{n(t)}{n(0)} \exp(M(t))\} \\
&= \sup\{y \leq t : M(y) \leq \ln[\frac{n(t)}{n(0)}] + M(t)\} \\
&= M^{-1}(\ln[\frac{n(t)}{n(0)}] + M(t)),
\end{aligned}
$$

where $M^{-1}(t)$ exists for $t \in (0, n(0)]$ (since $M(y)$ strictly increases and maps into $(0, n(0)]$ for $y \geq 0$).

Therefore,

$$C(t) \leq 2n(M^{-1}(\ln[\frac{n(t)}{n(0)}] + M(t)))$$

and

$$Y_{t+1} \leq N \exp(n(0))2n(M^{-1}(\ln[\frac{n(t+1)}{n(0)}]+M(t+1)))+\exp(-M(t+1))[\exp(M(1))n^2(0)N+Y_0].$$

■

Theorem 11 *Among all stepsizes $\eta_{q,t} = 1/(K+t)^q$ where $q > 0$, K is a constant such that $\eta_{q,t} \leq \frac{1}{2L}$, with the stepsize $\eta_{1,t} = 1/(2L+t)$, SGD algorithm enjoys the fastest convergence.*

Proof Since in (53), we have

$$\mathbb{E}[\|w_t - w_*\|^2] \leq AC(t) + B \exp(-M(t)),$$

where $A = N \exp(n(0))$ and $B = \exp(M(1))n^2(0)N + \mathbb{E}[\|w_0 - w_*\|^2]$. Let us denote $C_q(t) = C(t), n_q(t) = \mu/(K+t)^q$ where $\eta_{q,t} = 1/(K+t)^q$. It is obvious that $n_q(t) > n_1(t)$ for all t and $q < 1$. It implies for any $n_q(t)$ with $q < 1$, we have

$$\exp(-M(t)) < \exp(-\int_{x=0}^{t} \mu 1/(K+x)dx) < \exp(-\int_{x=0}^{t} \mu 1/xdx) < 1/t.$$

Therefore, we always have $\exp(-M(t)) < 1/t = n_1(t) < n_q(t) < C_q(t)$. Now, we consider the following cases. We are finding $n(t)$ such as $n(t) = C(t)/2$. We rewrite it as

$$C(t) = \exp(-M(t)) \int_{x=0}^{t} \exp(M(x))n^2(x)dx = 2n(t)$$

Taking derivatives of both sides, we have:

$$-n^2 = n(n-2n) = n(n-C) = \frac{d}{dt}C = 2\frac{d}{dt}n$$

This is solved for $1/(at)$: $-1/(a^2t^2) = -2/(at^2)$ Hence, $a = 1/2$ and $n(t) = 2/t$. It means, $C_q(t) > C_1(t)$ and thus, the stepsize $\eta_{1,t} = 1/(K+t)$ enjoys the fastest convergence. ■

D.7 Convergence of Large Stepsizes in Batch Mode

We first derive a couple lemmas which will help us deriving our main bounds. In what follows let Assumptions 1, 2 and 3 hold for all lemmas.

Lemma 14 *Let us define* $f(w;(\xi_1,\ldots,\xi_k)) = \frac{1}{k}\sum_{i=1}^k f(w;\xi_i)$*, then we have the following properties:*

$$\mathbb{E}[f(w;(\xi_1,\ldots,\xi_k))] = F(w),$$

$$\mathbb{E}[\|\nabla f(w_*;(\xi_1,\ldots,\xi_k))\|^2] = \frac{\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]}{k},$$

and

$$\mathbb{E}[\|\nabla f(w;(\xi_1,\ldots,\xi_k))\|^2] \leq 4L[F(w)-F(w_*)] + \frac{N}{k}$$

Proof

The expectation of $f(w;(\xi_1,\ldots,\xi_k))$ is equal to

$$\mathbb{E}[f(w;(\xi_1,\ldots,\xi_k))] = \frac{1}{k}\sum_{i=1}^k \mathbb{E}[f(w;\xi_i)] = F(w). \tag{54}$$

Now we write $\mathbb{E}[\|\nabla f(w_*;(\xi_1,\ldots,\xi_k))\|^2]$ as $\mathbb{E}[\sum_{j=1}^d(\frac{1}{k}\sum_{i=1}^k[\nabla f(w_*;\xi_i)]_j)^2]$. It is equal to

$$\begin{aligned}&\mathbb{E}[\sum_{j=1}^d\{\frac{1}{k^2}\sum_{i=1}^k[\nabla f(w_*;\xi_i)]_j^2 + \frac{1}{k}\sum_{i_0\neq i_1}[\nabla f(w_*;\xi_{i_0})]_j[\nabla f(w_*;\xi_{i_1})]_j\}]\\ =&\mathbb{E}[\sum_{j=1}^d\frac{1}{k^2}\sum_{i=1}^k[\nabla f(w_*;\xi_i)]_j^2] + \mathbb{E}[\sum_{j=1}^d\frac{1}{k}\sum_{i_0\neq i_1}[\nabla f(w_*;\xi_{i_0})]_j[\nabla f(w_*;\xi_{i_1})]_j].\end{aligned}$$

The first term $\mathbb{E}[\sum_{j=1}^d\frac{1}{k^2}\sum_{i=1}^k[\nabla f(w_*;\xi_i)]_j^2]$ is equal to

$$\begin{aligned}&= \sum_{j=1}^d\frac{1}{k^2}\sum_{i=1}^k\mathbb{E}[[\nabla f(w_*;\xi_i)]_j^2] = \frac{1}{k^2}\sum_{i=1}^k\mathbb{E}[\sum_{j=1}^d[\nabla f(w_*;\xi_i)]_j^2] = \frac{1}{k^2}\sum_{i=1}^k\mathbb{E}[\|\nabla f(w_*;\xi_i)\|^2]\\ &= \frac{1}{k}E[\|\nabla f(w_*;\xi)\|^2].\end{aligned}$$

The second term $\mathbb{E}[\sum_{j=1}^{d}\frac{1}{k}\sum_{i_0\neq i_1}[\nabla f(w_*;\xi_{i_0})]_j[\nabla f(w_*;\xi_{i_1})]_j]$ is equal to

$$\sum_{j=1}^{d}\frac{1}{k}\sum_{i_0\neq i_1}\mathbb{E}[[\nabla f(w_*;\xi_{i_0})]_j]\cdot\mathbb{E}[[\nabla f(w_*;\xi_{i_1})]_j]=\sum_{j=1}^{d}\frac{1}{k}\sum_{i_0\neq i_1}[\mathbb{E}[\nabla f(w_*;\xi_{i_0})]]_j\cdot[\mathbb{E}[\nabla f(w_*;\xi_{i_1})]]_j.$$

Note that $\mathbb{E}[\nabla f(w_*;\xi_{i_0})]=\nabla\mathbb{E}[f(w_*;\xi_i)]=\nabla F(w_*)=0$. It means that the second term is equal to 0 and then we obtain

$$\mathbb{E}[\|\nabla f(w_*;(\xi_1,\ldots,\xi_k))\|^2]=\frac{\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]}{k}. \tag{55}$$

We have the following fact:

$$\begin{aligned}\|\nabla f(w;(\xi_1,\ldots,\xi_k))-\nabla f(w_*;(\xi_1,\ldots,\xi_k))\|^2&=\frac{1}{k^2}[\|\sum_{i=1}^{k}(\nabla f(w;\xi_i)-\nabla f(w_*;\xi_i))\|^2]\\&\leq\frac{1}{k}\sum_{i=1}^{k}\|\nabla f(w;\xi_i)-\nabla f(w_*;\xi_i)\|^2\\&=\|\nabla f(w;\xi_i)-\nabla f(w_*;\xi_i)\|^2.\end{aligned}$$

Since $\mathbb{E}[\|\nabla f(w;\xi)-\nabla f(w_*;\xi)\|^2]\leq 2L[F(w)-F(w_*)]$ (see (18)), we obtain

$$\mathbb{E}[\|\nabla f(w;(\xi_1,\ldots,\xi_k))-\nabla f(w_*;(\xi_1,\ldots,\xi_k))\|^2]\leq 2L[F(w)-F(w_*)].$$

With similar argument in Lemma 1, we can derive

$$\begin{aligned}\mathbb{E}[\|\nabla f(w;(\xi_1,\ldots,\xi_k))\|^2]&\leq 4L[F(w)-F(w_*)]+2\mathbb{E}[\|\nabla f(w_*;(\xi_1,\ldots,\xi_k))\|^2]\\&\overset{(55)}{\leq}4L[F(w)-F(w_*)]+2\frac{\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]}{k}\\&=4L[F(w)-F(w_*)]+\frac{N}{k},\end{aligned} \tag{56}$$

where $N=2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]$.

■

We define

$$\mathcal{F}_t=\sigma(w_0,\xi_0',u_0,\ldots,\xi_{t-1}',u_{t-1}),$$

where

$$\xi_i'=(\xi_{i1},\ldots,\xi_{ik_i}).$$

We consider the following general algorithm with the following gradient updating rule:

$$w_{t+1}=w_t-\eta_t d_{\xi_t'}S_{u_t}^{\xi_t'}\nabla f(w_t;\xi_t'), \tag{57}$$

where $f(w_t;\xi_t')=\frac{1}{k_t}\sum_{i=1}^{k_t}f(w_t;\xi_{ti})$.

Lemma 15 *We have*

$$\mathbb{E}[\|d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)\|^2 | \mathcal{F}_t, \xi'_t] \leq D \|\nabla f(w_t;\xi'_t)\|^2$$

and

$$\mathbb{E}[d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)|\mathcal{F}_t] = \nabla F(w_t).$$

Proof For the first bound, if we take the expectation of $\|d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)\|^2$ with respect to u_t, then we have (for vectors x we denote the value if its i-th position by $[x]_i$)

$$\mathbb{E}[\|d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)\|^2 | \mathcal{F}_t, \xi'_t] = d^2_{\xi'_t} \sum_u p_{\xi'_t}(u) \|S^{\xi'_t}_u \nabla f(w_t;\xi'_t)\|^2 = d^2_{\xi'_t} \sum_u p_{\xi'_t}(u) \sum_{i \in S^{\xi'_t}_u} [\nabla f(w_t;\xi'_t)]^2_i$$

$$= d_{\xi'_t} \sum_{i \in D_{\xi'_t}} [\nabla f(w_t;\xi'_t)]^2_i = d_{\xi'_t} \|f(w_t;\xi'_t)\|^2 \leq D \|\nabla f(w_t;\xi'_t)\|^2,$$

where the transition to the second line follows from (27).

For the second bound, if we take the expectation of $d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)$ wrt u_t, then we have:

$$\mathbb{E}[d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)|\mathcal{F}_t, \xi'_t] = d_{\xi'_t} \sum_u p_{\xi'_t}(u) S^{\xi'_t}_u \nabla f(w_t;\xi'_t) = \nabla f(w_t;\xi'_t),$$

and this can be used to derive

$$\mathbb{E}[d_{\xi'_t} S^{\xi'_t}_{u_t} f(w_t;\xi'_t)|\mathcal{F}_t] = \mathbb{E}[\mathbb{E}[d_{\xi'_t} S^{\xi'_t}_{u_t} f(w_t;\xi'_t)|\mathcal{F}_t, \xi'_t]|\mathcal{F}_t] = \mathbb{E}[\nabla f(w_t;\xi'_t)] = \nabla F(w_t).$$

The last equality comes from (54). ■

Lemma 16 *Let Assumptions 1, 2 and 3 hold,* $0 < \eta_t \leq \frac{1}{2LD}$ *for all* $t \geq 0$. *Then,*

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq (1 - \mu\eta_t)\|w_t - w_*\|^2 + \eta_t^2 \frac{ND}{k_t}.$$

Proof Since $w_{t+1} = w_t - \eta_t d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)$, we have

$$\|w_{t+1} - w_*\|^2 = \|w_t - w_*\|^2 - 2\eta_t \langle d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t), (w_t - w_*)\rangle + \eta_t^2 \|d_{\xi'_t} S^{\xi'_t}_{u_t} \nabla f(w_t;\xi'_t)\|^2.$$

We now take expectations over u_t and ξ_t and use Lemmas 15 and 14:

$$\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t] \tag{58}$$

$$\leq \|w_t - w_*\|^2 - 2\eta_t \langle \nabla F(w), w_t - w_*\rangle + \eta_t^2 D \mathbb{E}[\|\nabla f(w_t;\xi'_t)\|^2|\mathcal{F}_t] \tag{59}$$

By (1), we have

$$-\langle \nabla F(w), w_t - w_*\rangle \leq -[F(w) - F(w_*)] - \mu/2\|w_t - w_*\|^2 \tag{60}$$

Thus, $\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t]$ is at most

$$\overset{(60)}{\leq} \|w_t - w_*\|^2 - \mu\eta_t\|w_t - w_*\|^2 - 2\eta_t[F(w) - F(w_*)] + \eta_t^2 D\mathbb{E}[\|\nabla f(w_t;\xi_t')\|^2|\mathcal{F}_t].$$

Since $\mathbb{E}[\|\nabla f(w_t;\xi_t')\|^2|\mathcal{F}_t] \leq 4L[F(w) - F(w_*)] + \frac{N}{k_t}$ (see (56)), $\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t]$ is at most

$$\overset{(56)}{\leq} (1 - \mu\eta_t)\|w_t - w_*\|^2 - 2\eta_t(1 - 2\eta_t LD)[F(w) - F(w_*)] + \eta_t^2\frac{ND}{k_t}.$$

Using the condition $\eta_t \leq \frac{1}{2LD}$ yields the lemma.

■

As shown above,

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq (1 - \mu\eta_t)\mathbb{E}[\|w_t - w_*\|^2] + \frac{\eta_t^2 ND}{k_t},$$

when $\eta_t \leq \frac{1}{2LD}$.

Let $Y_{t+1} = \mathbb{E}[\|w_{t+1} - w_*\|^2]$, $Y_t = \mathbb{E}[\|w_t - w_*\|^2]$, $\beta_t = 1 - \mu\eta_t$ and $\gamma_t = \frac{\eta_t^2 ND}{k_t}$. As proved in Lemma 9, if $Y_{t+1} \leq \beta_t Y_t + \gamma_t$, then

$$\begin{aligned} Y_{t+1} &\leq \sum_{i=0}^{t}[\prod_{j=i+1}^{t}\beta_j]\gamma_i + (\prod_{i=0}^{t}\beta_i)Y_0 \\ &= \sum_{i=0}^{t}[\prod_{j=i+1}^{t}(1 - \mu\eta_j)]\gamma_i + [\prod_{i=0}^{t}(1 - \mu\eta_i)]\mathbb{E}[\|w_0 - w_*\|^2] \end{aligned}$$

Let us define $n(j) = \mu n_j$ and $M(y) = \int_{x=0}^{y} n(x)dx$ as in Section 4. We also obtain

$$C(t) = \exp(-M(t))\int_{x=0}^{t}\exp(M(x))\frac{n^2(x)}{k(x)}dx.$$

Theorem 12 *Let Assumptions 1, 2 and 3 hold, $\{\eta_t\}$ is a diminishing sequence with conditions $\sum_{t=0}^{\infty}\eta_t \to \infty$ and $0 < \eta_t \leq \frac{1}{2LD}$ for all $t \geq 0$. Then, the sequence $\{w_t\}$ converges to w_* where*

$$w_{t+1} = w_t - \eta_t d_{\xi_t'} S_{u_t}^{\xi_t'}\nabla f(w_t;\xi_t').$$

Proof To prove the convergence of w_t, we only need to prove the convergence of

$$C(t) = \exp(-M(t))\int_{x=0}^{t}\exp(M(x))\frac{n^2(x)}{k(x)}dx.$$

Let denote T the total number of gradient computations and define $K(t) = \int_{x=1}^{t} k(x)dx = T$. From, here we have $t = K^{-1}(T)$ and $\frac{dK(x)}{dx} = k(x)$. Now, we define $y = K(x)$ or

$x = K^{-1}(y)$ and $dy = k(x)dx$. We write

$$\begin{aligned} C(t) &= \exp(-M(t)) \int_{x=0}^{t} \exp(M(x)) \frac{n^2(x)}{k(x)} dx \\ &= \exp(-M(K^{-1}(T))) \int_{K(0)}^{K^{-1}(T)} \exp(M(K^{-1}(y))) \frac{n^2(x)}{k^2(x)} k(x) dx \\ &\leq \exp(-M(K^{-1}(T))) \int_{1}^{K^{-1}(T)} \exp(M(K^{-1}(y))) \frac{n^2(x)}{k^2(x)} k(x) dx. \end{aligned}$$

The last inequality is based on the fact that $K(0) \geq 1$.

Let us define $n'(x) = \frac{n(x)}{k(x)}$ and using the fact that $dy = k(x)dx$, we obtain

$$\begin{aligned} C(t) &= C(K^{-1}(T)) \\ &\leq \exp(-M(K^{-1}(T))) \int_{1}^{K^{-1}(T)} \exp(M(K^{-1}(y)))[n'(K^{-1}(y))]^2 dy. \end{aligned}$$

Since $K^{-1}(K(x)) = x$ or $\frac{dK^{-1}(K(x))}{dx} = 1/\frac{dK(x)}{dx} = 1/k(x)$, we have the $\frac{dM(K^{-1}(y))}{dy} = \frac{n(K^{-1}(y))}{k(K^{-1}(y))} = n'(K^{-1}(y))$. Hence, by denoting

$$\begin{aligned} C'(t) &= C(K^{-1}(t)), \\ n'(x) &= \frac{n(x)}{k(x)}, \\ M'(x) &= M(K^{-1}(x)), \end{aligned}$$

we can convert the general problem into the problem in Section D.6. This implies that the analysis of $C(t)$ in Section D.6 can directly apply to analyze $C(K^{-1}(T))$. Since we already proved the convergence of $C(t)$ in Section D.6, we obtain the theorem. ■