ISE

Industrial and Systems Engineering

SGD and Hogwild!
Convergence Without the Bounded Gradients Assumption

LAM M. NGUYEN
Department of Industrial and Systems Engineering, Lehigh University, USA

PHUONG HA NGUYEN
Department of Electrical and Computer Engineering, University of Connecticut, USA

MARTEN VAN DIJK
Department of Electrical and Computer Engineering, University of Connecticut, USA

PETER RICHTÁRIK
KAUST, KSA — Edinburgh, UK — MIPT, Russia

KATYA SCHEINBERG
Department of Industrial and Systems Engineering, Lehigh University, USA

MARTIN TAKÁČ
Department of Industrial and Systems Engineering, Lehigh University, USA

ISE Technical Report 18T-004

LEHIGH UNIVERSITY.

SGD and Hogwild! Convergence Without the Bounded Gradients Assumption

Lam M. Nguyen [1] **Phuong Ha Nguyen** [2] **Marten van Dijk** [2] **Peter Richtárik** [3] **Katya Scheinberg** [1]
Martin Takáč [1]

Abstract

Stochastic gradient descent (SGD) is the optimization algorithm of choice in many machine learning applications such as regularized empirical risk minimization and training deep neural networks. The classical analysis of convergence 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 always 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. Here we show that for stochastic problems arising in machine learning such bound always holds. Moreover, we propose an alternative convergence analysis of SGD with diminishing learning rate regime, which is results in more relaxed conditions that those in (Bottou et al., 2016). We then move on the asynchronous parallel setting, and prove convergence of the Hogwild! algorithm in the same regime, obtaining the first convergence results for this method in the case of diminished learning rate.

[1]Department of Industrial and Systems Engineering, Lehigh University, USA. [2]Department of Electrical and Computer Engineering, University of Connecticut, USA. [3]KAUST, KSA — Edinburgh, UK — MIPT, Russia. Correspondence to: Lam M. Nguyen <LamNguyen.MLTD@gmail.com>, Phuong Ha Nguyen <phuongha.ntu@gmail.com>, Marten van Dijk <marten.van_dijk@uconn.edu>, Peter Richtárik <Peter.Richtarik@ed.ac.uk>, Katya Scheinberg <katyas@lehigh.edu>, Martin Takáč <Takac.MT@gmail.com>.

Lam M. Nguyen was partially supported by NSF Grants CCF 16-18717. 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 supported by U.S. National Science Foundation, under award number NSF:CCF:1618717, NSF:CMMI:1663256 and NSF:CCF:1740796.

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 & 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 & 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) do not make any assumptions on individual function realizations. In this paper we will derive convergence properties for SGD applied to (1) with the additional assumptions on $f(w;\xi)$ and also show how they hold when functions $f(w;\xi)$ are not assumed convex.

Regardless of the properties of $f(w;\xi)$ in this paper we focus on the case of (1) when F 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 & 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 $\|w - w_*\|^2$.

Instead of assuming that (5) is bounded for all w, it is enough to assume $\mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2] \leq \sigma^2$, where w_t, $t \geq 0$, are the iterates generated by the algorithm. From our analysis above, it clearly implies an implicit assumption that the iterates of the algorithm remain in a bounded region around w_*. This condition is impossible to guarantee with probability one for the classical stochastic gradient method.

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 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) 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.

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 are the first to 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.

1.2. Organization

We analyse the convergence rate of SGD in Section 2 and introduce the general recursion and its analysis in Section 3. Experiments are reported in Section 4.

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

Note that $\mathcal{F}_t = \sigma(w_0, \xi_0, \ldots, \xi_{t-1})$ is the σ-algebra generated by $w_0, \xi_0, \ldots, \xi_{t-1}$, i.e., $\mathcal{F}_t$ contains all the information of $w_0, \ldots, w_t$. 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., $\forall w, w' \in \mathbb{R}^d$,*

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

[1] Independent and identically distributed.

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

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

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

Using Lemma 1 and Super Martingale Convergence Theorem (Bertsekas, 2015) (Lemma 4 in Appendix), 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*

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

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

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

Note that the classical SGD proposed in (Robbins & 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* $\forall w \in \mathbb{R}^d$,

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

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 (so, 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 (28) 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.

Appendix D proves the following theorem where

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

Theorem 5. *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 $4 \le \alpha_t \le \alpha$ 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 5 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 $\alpha \ge 4$.

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 Appendix D 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 Appendix D shows that we may allow τ to be a monotonic increasing function of t with

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

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 5.*

Experiments in Section 4 confirm our analysis.

Remark 4. *The higher order terms in the proof in Appendix D show that, as in Theorem 2, the expected convergence rate in Theorem 5 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 5 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*.*

3.3. Convergence Analysis without Convexity

In Appendix D we also show that the proof of Theorem 5 can easily be modified such that Theorem 5 with $E \geq \frac{4L\kappa\alpha D}{\mu}$ also holds in the non-convex case, i.e., we do not need Assumption 3. Note that this case is not analyzed in (Leblond et al., 2018).

Theorem 6. *Let Assumptions 1 and 2 hold. Then, we can conclude the statement of Theorem 5 with* $E \geq \frac{4L\kappa\alpha D}{\mu}$ *for* $\kappa = \frac{L}{\mu}$*.*

4. Numerical Experiments

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

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

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 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).

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

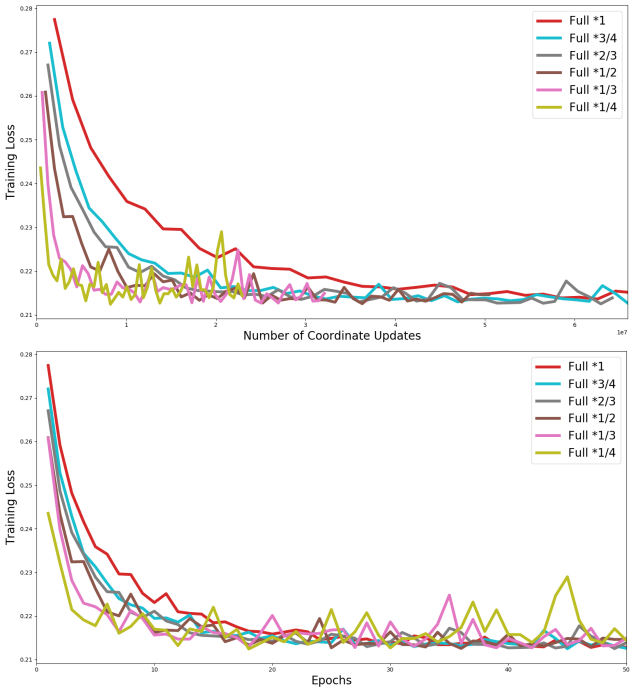

Figure 1: *ijcnn1* for different fraction of non-zero set

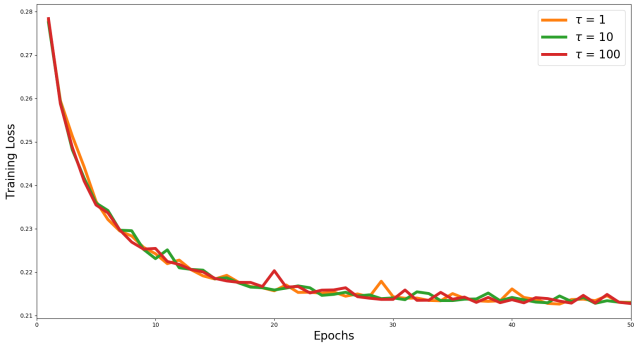

Figure 2: *ijcnn1* for different values of τ with the whole non-zero set

We experimented with 10 runs and reported the average results. We choose the step size based on Theorem 5, 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 $S^{\xi_t}_{u_t}$ 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.

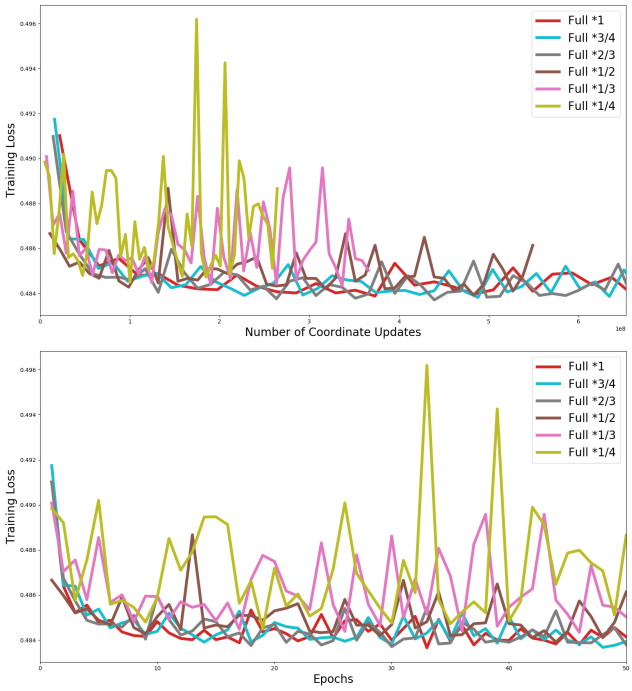

Figure 3: *covtype* for different fraction of non-zero set

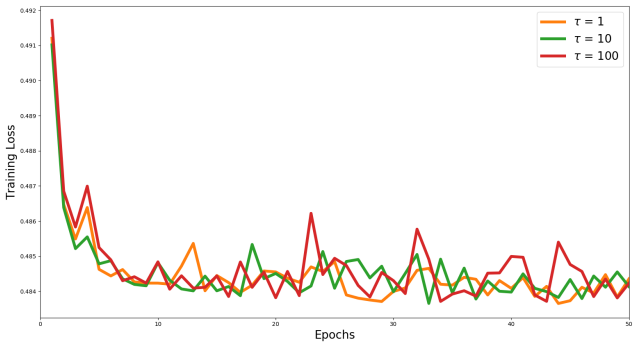

Figure 4: *covtype* for different values of τ with the whole non-zero set

5. Conclusion

We have provided a new framework for 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 realizations, but without requiring any bounds on the stochastic gradients. Our

framework shows almost sure convergence of SGD and provides sublinear upper bounds for the expected convergence rate of a general recursion which includes Hogwild! for inconsistent reads and writes as a special case. Our framework provides new intuition which will help understanding convergence as observed in practice.

References

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

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

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

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

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

Hazan, Elad and Kale, Satyen. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512, 2014. URL http://jmlr.org/papers/v15/hazan14a.html.

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

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

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

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

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

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

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

Nguyen, Lam, Nguyen, Nam, Phan, Dzung, Kalagnanam, Jayant, and Scheinberg, Katya. When does stochastic gradient algorithm work well? *arXiv:1801.06159*, 2018.

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

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

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

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

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

Yu, H. Some proof details for asynchronous stochastic approximation algorithms, 2011.

Appendix

A. Review of Useful Theorems

Lemma 3 (Generalization of the result in (Johnson & Zhang, 2013))**.** *Let Assumptions 1 and 3 hold. Then, $\forall w \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \leq 2L[F(w) - F(w_*)], \tag{16}$$

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

Lemma 4 ((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,

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

B. Proofs of Lemmas

B.1. Proof of Lemma 1

Lemma 1. *Let Assumptions 1, 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

$$\|a\|^2 = \|a - b + b\|^2 \leq 2\|a - b\|^2 + 2\|b\|^2, \tag{17}$$

$$\Rightarrow \frac{1}{2}\|a\|^2 - \|b\|^2 \leq \|a - b\|^2. \tag{18}$$

Hence,

$$\begin{aligned}\frac{1}{2}\mathbb{E}[\|\nabla f(w;\xi)\|^2] - \mathbb{E}[\|\nabla f(w_*;\xi)\|^2] &= \mathbb{E}\left[\frac{1}{2}\|\nabla f(w;\xi)\|^2 - \|\nabla f(w_*;\xi)\|^2\right] \\ &\overset{(18)}{\leq} \mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \\ &\overset{(21)}{\leq} 2L[F(w) - F(w_*)] \end{aligned} \tag{19}$$

Therefore,

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \overset{(17)(19)}{\leq} 4L[F(w) - F(w_*)] + 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]. \qquad \square$$

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,

$$\begin{aligned}\frac{1}{2}\mathbb{E}[\|\nabla f(w;\xi)\|^2] - \mathbb{E}[\|\nabla f(w_*;\xi)\|^2] &= \mathbb{E}\left[\frac{1}{2}\|\nabla f(w;\xi)\|^2 - \|\nabla f(w_*;\xi)\|^2\right] \\ &\overset{(18)}{\leq} \mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \\ &\overset{(6)}{\leq} L^2\|w - w_*\|^2 \\ &\overset{(3)}{\leq} \frac{2L^2}{\mu}[F(w) - F(w_*)] = 2L\kappa[F(w) - F(w_*)]. \end{aligned} \tag{20}$$

Therefore,

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \overset{(17)(20)}{\leq} 4L\kappa[F(w) - F(w_*)] + 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2].$$

□

B.3. Proof of Lemma 3

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

$$\mathbb{E}[\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2] \leq 2L[F(w) - F(w_*)], \tag{21}$$

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

Proof. The proof for the finite-sum problem is originally from (Johnson & Zhang, 2013) while this proof for general problem can be found in (Nguyen et al., 2018). Given any ξ, for all $w \in \mathbb{R}^d$, consider

$$h(w;\xi) := f(w;\xi) - f(w_*;\xi) - \nabla f(w_*;\xi)^T(w - w_*).$$

Since $h(w;\xi)$ is convex by w and $\nabla h(w_*;\xi) = 0$, we have $h(w_*;\xi) = \min_w h(w;\xi)$. Hence,

$$\begin{aligned} 0 = h(w_*;\xi) &\leq \min_\eta [h(w - \eta\nabla h(w;\xi);\xi)] \\ &\overset{(7)}{\leq} \min_\eta \left[h(w;\xi) - \eta\|\nabla h(w;\xi)\|^2 + \frac{L\eta^2}{2}\|\nabla h(w;\xi)\|^2\right] \\ &= h(w;\xi) - \frac{1}{2L}\|\nabla h(w;\xi)\|^2. \end{aligned}$$

Hence,

$$\|\nabla f(w;\xi) - \nabla f(w_*;\xi)\|^2 \leq 2L[f(w;\xi) - f(w_*;\xi) - \nabla f(w_*;\xi)^T(w - w_*)].$$

Taking the expectation with respect to ξ, we have

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

□

B.4. Proof of Lemma 4

Before proving Lemma 4, we first introduce the following lemma in (Yu, 2011).

Lemma 5 ((Yu, 2011)). *Let $\{X_k\}$, $\{\beta_k\}$, and $\{Y_k\}$, $k = 0, 1, \ldots$, be three sequences of nonnegative random variables adapted to $\{\mathcal{F}_k\}$ and satisfying*

$$\mathbb{E}[X_{k+1}|\mathcal{F}_k] \leq (1 + \beta_k)X_k + Y_k \ , \ k \geq 0.$$

Then $\{X_k\}$ converges w.p.1 (a.s.) to a finite limit on the event $\{\sum_{k=0}^{\infty} \beta_k < \infty, \sum_{k=0}^{\infty} Y_k < \infty\}$.

Proof. We define

$$X'_k = \frac{X_k}{\prod_{i=0}^{k-1}(1+\beta_i)} \ , \ Y'_k = \frac{Y_k}{\prod_{i=0}^{k}(1+\beta_i)} \ , \ k \geq 1,$$
$$\text{with } X'_0 = X_0 \text{ and } Y'_0 = Y_0.$$

Then,

$$\mathbb{E}[X'_{k+1}|\mathcal{F}_k] = \frac{1}{\prod_{i=0}^{k}(1+\beta_i)}\mathbb{E}[X_{k+1}|\mathcal{F}_k] \leq X'_k + Y'_k \ , \ k \geq 0,$$

where the first equality is from the fact that $\mathbb{E}[XY|\mathcal{F}] = X\mathbb{E}[Y|\mathcal{F}]$, if $X \in \mathcal{F}$. We then define

$$U_k = X'_k - \sum_{i=0}^{k-1} Y'_i \text{ and } U_0 = X_0.$$

Then,

$$\mathbb{E}[U_{k+1}|\mathcal{F}_k] = \mathbb{E}[X'_{k+1}|\mathcal{F}_k] - \sum_{i=0}^{k} Y'_i \leq U_k \ , \ k \geq 0.$$

Hence, $\{U_k\}$ is supermartingale with respect to $\{\mathcal{F}_k\}$. Let $a > 0$ and

$$\nu_a = \begin{cases} \min\left\{k \geq 0 : \sum_{i=0}^{k} Y'_i > a\right\} & \text{if } \nu_a \notin \emptyset, \\ \infty & \text{otherwise} \end{cases}$$

Define $L_k = (a + U_k)I_{\{\nu_a > k\}}$. (Note that $I_{\{\nu_a > k\}} = 1$ if $\nu_a > k$ and $I_{\{\nu_a > k\}} = 0$ otherwise.) We can see that

$$L_k = (a + U_k)I_{\{\nu_a > k\}} = \left(X'_k + \left(a - \sum_{i=0}^{k-1} Y'_i\right)\right) I_{\{\nu_a > k\}} \geq 0 \ , \ k \geq 0.$$

Since

$$(a + U_{k+1})I_{\{\nu_a = k+1\}} = \left(X'_{k+1} + \left(a - \sum_{i=0}^{k} Y'_i\right)\right) I_{\{\nu_a = k+1\}} \geq 0,$$

so we have

$$\begin{aligned} \mathbb{E}[(a + U_{k+1})I_{\{\nu_a > k+1\}}|\mathcal{F}_k] &\leq \mathbb{E}[(a + U_{k+1})I_{\{\nu_a > k+1\}} + (a + U_{k+1})I_{\{\nu_a = k+1\}}|\mathcal{F}_k] \\ &= \mathbb{E}[(a + U_{k+1})I_{\{\nu_a > k\}}|\mathcal{F}_k] \\ &\leq (a + U_k)I_{\{\nu_a > k\}}. \end{aligned}$$

Hence, $\mathbb{E}[L_{k+1}|\mathcal{F}_k] \leq L_k$, which implies that $\{L_k\} = \{(a + U_k)I_{\{\nu_a > k\}}\}$ is a non-negative supermartingale with respect to $\{\mathcal{F}_k\}$ and so converges w.p.1 to a finite limit.

This means that $\{U_k\}$ converges w.p.1 on the event $\{\nu_a = \infty\}$, that is, also converges w.p.1 on the event $\{\sum_{k=0}^{\infty} Y'_k \leq a\}$. Since a is arbitrary, it follows that $\{U_k\}$ converges w.p.1 on the event $\{\sum_{k=0}^{\infty} Y'_k < \infty\}$.

Thus, $\{X'_k\}$ converges w.p.1 on the event $\{\sum_{k=0}^{\infty} Y'_k < \infty\}$, and so $\{X'_k\}$ converges w.p.1 on the event $\{\sum_{k=0}^{\infty} Y_k < \infty\}$ since $0 \leq Y'_k \leq Y_k$, $k \geq 0$.

Hence, $\{X_k\}$ converges w.p.1 on the event $\{\sum_{k=0}^{\infty} Y_k < \infty, \prod_{k=0}^{\infty}(1+\beta_k) < \infty\}$. We know that $1 + x \leq e^x$, then

$$\prod_{k=0}^{\infty}(1+\beta_k) \leq \prod_{k=0}^{\infty} e^{\beta_k} = \exp\left(\sum_{k=0}^{\infty} \beta_k\right).$$

Therefore, $\{X_k\}$ converges w.p.1 to a finite limit on the event $\{\sum_{k=0}^{\infty} Y_k < \infty, \sum_{k=0}^{\infty} \beta_k < \infty\}$. □

We then use the above result of Lemma 5 to prove Lemma 4. The original proof can be found in (Yu, 2011).

Lemma 4. *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,

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

Proof. Since for each k, $Z_k \geq 0$ and $\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq Y_k - Z_k + W_k$, we have

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq Y_k + W_k.$$

Applying Lemma 5 with $\beta_k = 0$ for all k, we have that $\{Y_k\}$ converges w.p.1 to a finite limit on the event $\{\sum_{k=0}^{\infty} W_k < \infty\}$. Under the assumption, we know that $\{\sum_{k=0}^{\infty} W_k < \infty\}$ holds w.p.1. Therefore, $Y_k \to Y \geq 0$, w.p.1.

To show $\sum_{k=0}^{\infty} Z_k < \infty$ w.p.1, consider

$$M_k = Y_k + \sum_{i=0}^{k-1} Z_i \text{ and } M_0 = Y_0.$$

We have

$$\mathbb{E}[M_{k+1}|\mathcal{F}_k] = \mathbb{E}[Y_{k+1}|\mathcal{F}_k] + \sum_{i=0}^{k} Z_i \leq Y_k - Z_k + W_k + \sum_{i=0}^{k} Z_i = M_k + W_k,$$

where the first equality is from the fact that $\mathbb{E}[X|\mathcal{F}] = X$, if $X \in \mathcal{F}$. Hence, applying Lemma 5 with $\beta_k = 0$ for all k, we have that $M_k \to M \geq 0$, w.p.1. Therefore, $\sum_{k=0}^{\infty} Z_k = M - Y < \infty$, w.p.1. This completes the proof. □

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

$$\mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2|\mathcal{F}_t] \leq 4L[F(w_t) - F(w_*)] + N, \tag{22}$$

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,

$$\begin{aligned}
\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t] &= \mathbb{E}[\|w_t - \eta_t\nabla f(w_t;\xi_t) - w_*\|^2|\mathcal{F}_t] \\
&= \|w_t - w_*\|^2 - 2\eta_t\langle \nabla F(w_t), (w_t - w_*)\rangle + \eta_t^2\mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2|\mathcal{F}_t] \\
&\overset{(3)(22)}{\leq} \|w_t - w_*\|^2 - \mu\eta_t\|w_t - w_*\|^2 - 2\eta_t[F(w_t) - F(w_*)] + 4L\eta_t^2[F(w_t) - F(w_*)] + \eta_t^2 N \\
&= \|w_t - w_*\|^2 - \mu\eta_t\|w_t - w_*\|^2 - 2\eta_t(1 - 2L\eta_t)[F(w_t) - F(w_*)] + \eta_t^2 N \\
&\leq \|w_t - w_*\|^2 - \mu\eta_t\|w_t - w_*\|^2 + \eta_t^2 N.
\end{aligned}$$

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

$$\mathbb{E}[\|w_{t+1} - w_*\|^2|\mathcal{F}_t] \leq \|w_t - w_*\|^2 - \mu\eta_t\|w_t - w_*\|^2 + \eta_t^2 N. \tag{23}$$

Since $\sum_{t=0}^{\infty}\eta_t^2 N < \infty$, we could apply Lemma 4. 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 (23), 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

$$\mathbb{E}[\|w_t - w_*\|^2] \leq \frac{N}{\mu^2} G \frac{1}{(t+E)}, \tag{24}$$

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 (24) (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 (24) by induction in t.

Notice that the induction proof of (24) 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 (24)

$$\begin{aligned}\frac{L\mu}{N}\mathbb{E}[\|w_T - w_*\|^2] &\leq \frac{L\mu}{N}\frac{N}{\mu^2}\frac{G}{(T+E)} \\ &= \frac{L}{\mu}\frac{4\max\{\frac{L\mu}{N}\mathbb{E}[\|w_0 - w_*\|^2], 1\}}{\frac{4L}{\mu}\max\{\frac{L\mu}{N}\mathbb{E}[\|w_0 - w_*\|^2], 1\}} = 1. \end{aligned} \tag{25}$$

Applying (24)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 (25). For any given w_0, we prove the theorem. □

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

$$d_\xi \mathbb{E}[S_u^\xi | \xi] = D_\xi, \tag{26}$$

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

$$\begin{aligned}&\mathbb{E}[|\langle \nabla f(w_1;\xi_1), \nabla f(w_2;\xi_2)\rangle|]\\ &\leq \frac{\sqrt{\Delta}}{2}\left(\mathbb{E}[\|\nabla f(w_1;\xi_1)\|^2] + \mathbb{E}[\|\nabla f(w_2;\xi_2)\|^2]\right)\end{aligned} \tag{27}$$

(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 (27) 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

$$w_{t+1} = w_t - \eta_t d_{\xi_t} S^{\xi_t}_{u_t} \nabla f(\hat{w}_t; \xi_t), \tag{28}$$

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 (28) 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 (28) 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,

$$\hat{w}_t = w_{t-\tau} - \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} \Sigma_{t,j} S^{\xi_j}_{u_j} \nabla f(\hat{w}_j; \xi_j). \tag{29}$$

Notice that the recursion (28) implies

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

By combining (30) and (29) we obtain

$$w_t - \hat{w}_t = - \sum_{j=t-\tau}^{t-1} \eta_j d_{\xi_j} (I - \Sigma_{t,j}) S^{\xi_j}_{u_j} \nabla f(\hat{w}_j; \xi_j), \tag{31}$$

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

$$\mathcal{F}_t = \sigma(w_0, \xi_1, u_1, \sigma_1, \ldots, \xi_{t-1}, u_{t-1}, \sigma_{t-1}),$$

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*

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

and

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

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 (26).

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:

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

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 (31),

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

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 (27) to the inner products implies

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

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 6: 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

$$\mathbb{E}[\|\nabla f(\hat{w}_j;\xi_j)\|^2]\leq 4L\mathbb{E}[F(\hat{w}_j)-F(w_*)]+N \tag{32}$$

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,*

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

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 6:

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

By (3) and (7), we have

$$-\langle\nabla F(\hat{w}_t),(\hat{w}_t-w_*)\rangle\leq-[F(\hat{w}_t)-F(w_*)]-\frac{\mu}{2}\|\hat{w}_t-w_*\|^2,\text{ and} \tag{33}$$

$$-\langle\nabla F(\hat{w}_t),(w_t-\hat{w}_t)\rangle\leq F(\hat{w}_t)-F(w_t)+\frac{L}{2}\|\hat{w}_t-w_t\|^2 \tag{34}$$

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

$$\begin{aligned}&\overset{(33)(34)}{\leq} \|w_t-w_*\|^2+2\eta_t[F(\hat{w}_t)-F(w_t)]+L\eta_t\|\hat{w}_t-w_t\|^2-2\eta_t[F(\hat{w}_t)-F(w_*)]-\mu\eta_t\|\hat{w}_t-w_*\|^2\\&\quad+\eta_t^2D\mathbb{E}[\|\nabla f(\hat{w}_t;\xi_t)\|^2|\mathcal{F}_t]\\&=\|w_t-w_*\|^2-2\eta_t[F(w_t)-F(w_*)]+L\eta_t\|\hat{w}_t-w_t\|^2-\mu\eta_t\|\hat{w}_t-w_*\|^2+\eta_t^2D\mathbb{E}[\|\nabla f(\hat{w}_t;\xi_t)\|^2|\mathcal{F}_t].\end{aligned}$$

Since

$$-\|\hat{w}_t-w_*\|^2=-\|(w_t-w_*)-(w_t-\hat{w}_t)\|^2\overset{(18)}{\leq}-\frac{1}{2}\|w_t-w_*\|^2+\|w_t-\hat{w}_t\|^2,$$

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

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

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

$$\mathbb{E}[\|\nabla f(\hat{w}_t;\xi_t)\|^2|\mathcal{F}_t]\leq 2\mathbb{E}[\|\nabla f(\hat{w}_t;\xi_t)-\nabla f(w_t;\xi_t)\|^2|\mathcal{F}_t]+2\mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2|\mathcal{F}_t]. \tag{35}$$

By Lemma 1, we have

$$\mathbb{E}[\|\nabla f(w_t;\xi_t)\|^2|\mathcal{F}_t]\leq 4L[F(w_t)-F(w_*)]+N. \tag{36}$$

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 (35) and (36) 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

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

Since $\eta_t\leq\frac{1}{4LD}$, $-2\eta_t(1-4L\eta_tD)[F(w_t)-F(w_*)]\leq 0$ (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),

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

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

$$\begin{aligned}\mathbb{E}[\|w_{t+1}-w_*\|^2]\quad\leq\quad&\left(1-\frac{\mu\eta_t}{2}\right)\mathbb{E}[\|w_t-w_*\|^2]+\\&[(L+\mu)\eta_t+2L^2\eta_t^2D](1+\sqrt{\Delta}\tau)D\sum_{j=t-\tau}^{t-1}\eta_j^2(2L^2\mathbb{E}[\|\hat{w}_j-w_*\|^2]+N)+2\eta_t^2DN.\end{aligned} \tag{37}$$

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.,

$$\prod_{i=h}^{k}g_i=1\text{ and }\sum_{i=h}^{k}g_i=0\text{ if }k<h. \tag{38}$$

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,*

$$Y_{t+1} \leq (\sum_{i=0}^{t}[\prod_{j=i+1}^{t}\beta_j]\gamma_i) + (\prod_{j=0}^{t}\beta_j)Y_0. \tag{39}$$

Proof. We prove the lemma by using induction. It is obvious that (39) is true for $t = 0$ because $Y_1 \leq \beta_1 Y_0 + \gamma_1$. Assume as induction hypothesis that (39) is true for $t-1$. Since $Y_{t+1} \leq \beta_t Y_t + \gamma_t$,

$$\begin{aligned}
Y_{t+1} &\leq \beta_t Y_t + \gamma_t \\
&\leq \beta_t[(\sum_{i=0}^{t-1}[\prod_{j=i+1}^{t-1}\beta_j]\gamma_i) + (\prod_{j=0}^{t-1}\beta_j)Y_0] + \gamma_t \\
&\overset{(38)}{=} (\sum_{i=0}^{t-1}\beta_t[\prod_{j=i+1}^{t-1}\beta_j]\gamma_i) + \beta_t(\prod_{j=0}^{t-1}\beta_j)Y_0 + (\prod_{j=t+1}^{t}\beta_j)\gamma_t \\
&= [(\sum_{i=0}^{t-1}[\prod_{j=i+1}^{t}\beta_j]\gamma_i) + (\prod_{j=t+1}^{t}\beta_j)\gamma_t] + (\prod_{j=0}^{t}\beta_j)Y_0 \\
&= (\sum_{i=0}^{t}[\prod_{j=i+1}^{t}\beta_j]\gamma_i) + (\prod_{j=0}^{t}\beta_j)Y_0.
\end{aligned}$$

□

Applying the above lemma to (37) 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 (37) 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 (37) 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

$$\eta_{t-\tau}^2 \leq \frac{4\alpha^2}{\mu^2}\frac{1}{(t+E)^2}. \tag{40}$$

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

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

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

$$\mathbb{E}[\|w_{t+1} - w_*\|^2] \leq \left(\sum_{i=0}^{t}\left[\frac{(i+E)^2}{(t+E-1)^2}\right]\gamma_i\right) + \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0 - w_*\|^2]. \tag{41}$$

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 (41) 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

$$\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] \leq 2\mathbb{E}[\|\hat{w}_{t+1} - w_{t+1}\|^2] + 2\mathbb{E}[\|w_{t+1} - w_*\|^2], \tag{42}$$

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

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

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

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

where the last inequality follows from (40), and

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

From these inequalities we infer

$$\begin{aligned}\mathbb{E}[\|\hat{w}_{t+1}-w_*\|^2] &\leq 8(1+\sqrt{\Delta}\tau)\tau D\frac{\alpha^2}{\mu^2}\frac{1}{(t+E-1)^2}(2L^2Z_{t+1-\tau}+N)+\\ &\quad 2\frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left(\sum_{i=1}^{t}\left[4a_i(1+\sqrt{\Delta}\tau)[N\tau+2L^2\tau Z_{i-\tau}]+2N\right]\right)+\\ &\quad \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0-w_*\|^2].\end{aligned}\tag{43}$$

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

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

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

$$\begin{aligned}\sum_{i=1}^{t}a_i &= (L+\mu)\sum_{i=1}^{t}\eta_i+2L^2D\sum_{i=1}^{t}\eta_i^2\\ &\leq (L+\mu)\sum_{i=1}^{t}\frac{\alpha}{\mu(i+E)}+2L^2D\sum_{i=1}^{t}\frac{\alpha^2}{\mu^2(i+E)^2}\\ &\leq \frac{(L+\mu)\alpha}{\mu}\sum_{i=1}^{t}\frac{1}{i}+\frac{2L^2\alpha^2D}{\mu^2}\sum_{i=1}^{t}\frac{1}{i^2}\\ &\leq \frac{(L+\mu)\alpha}{\mu}(1+\ln t)+\frac{L^2\alpha^2D\pi^2}{3\mu^2},\end{aligned}\tag{45}$$

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 (45) in (44) and collecting terms yields

$$\begin{aligned} & \mathbb{E}[\|\hat{w}_{t+1}-w_*\|^2] \\ \leq\ & 2\frac{\alpha^2 D}{\mu^2}\frac{1}{(t+E-1)^2}\left(2Nt+4(1+\sqrt{\Delta}\tau)\tau[N+2L^2Z]\left\{\frac{(L+\mu)\alpha}{\mu}(1+\ln t)+\frac{L^2\alpha^2 D\pi^2}{3\mu^2+1}\right\}\right)+ \\ & \frac{(E+1)^2}{(t+E-1)^2}\mathbb{E}[\|w_0-w_*\|^2]. \end{aligned} \tag{46}$$

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 (46) 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 5. *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 (32) 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 (36) 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 (46) with a similar change finally leading to Theorem 6, i.e., Theorem 5 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 (41)

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

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 5 and conclude that there exists a constant Z such that, see (44),

$$\begin{aligned}
\mathbb{E}[\|\hat{w}_{t+1}-w_*\|^2] \quad \leq \quad & 8(1+\sqrt{\Delta}\tau(t))\tau(t)D\frac{\alpha^2}{\mu^2}\frac{1}{t^2}(2L^2Z+N) + \\
& 2\frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left(\sum_{i=1}^{t}\left[4a_i(1+\sqrt{\Delta}\tau(i))[N\tau(i)+2L^2\tau(i)Z]+2N\right]\right) + \\
& \frac{4}{t^2}\mathbb{E}[\|w_0-w_*\|^2]. \qquad (47)
\end{aligned}$$

Let us assume

$$\tau(t) \leq \sqrt{t\cdot L(t)}, \qquad (48)$$

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 (47) gives

$$\begin{aligned}\mathbb{E}[\|\hat{w}_{t+1} - w_*\|^2] &\leq 8(1+\sqrt{\Delta}\tau(t))\tau(t)D\frac{\alpha^2}{\mu^2}\frac{1}{t^2}(2L^2Z+N) +\\ &\quad 2\frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left(2Nt + 4\sqrt{\Delta}[\frac{(L+\mu)\alpha}{\mu}\frac{t}{\ln t} + O(\ln t)][N + 2L^2 Z] + O(\sqrt{t})\right) +\\ &\quad \frac{4}{t^2}\mathbb{E}[\|w_0 - w_*\|^2]\\ &\leq 2\frac{\alpha^2 D}{\mu^2}\frac{1}{t^2}\left(2Nt + 4\sqrt{\Delta}[(1+\frac{(L+\mu)\alpha}{\mu})\frac{t}{\ln t} + O(\ln t)][N + 2L^2 Z] + O(\sqrt{t})\right) +\\ &\quad \frac{4}{t^2}\mathbb{E}[\|w_0 - w_*\|^2] \qquad (49)\end{aligned}$$

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*

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

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 (43) where we may bound $Z_{i-\tau(i)} = O(1/(i-\tau(i))$ which replaces Z in the second line of (44). On careful examination this leads to a new upper bound (49) where the $2L^2Z$ 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 (49) 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.