

Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu
{lianxiangru, huangyj0, raingomm, ji.liu.uwisc}@gmail.com
Department of Computer Science, University of Rochester

May 20, 2017

Abstract

Asynchronous parallel implementations of stochastic gradient (SG) have been broadly used in solving deep neural network and received many successes in practice recently. However, existing theories cannot explain their convergence and speedup properties, mainly due to the nonconvexity of most deep learning formulations and the asynchronous parallel mechanism. To fill the gaps in theory and provide theoretical supports, this paper studies two asynchronous parallel implementations of SG: one is over a computer network and the other is on a shared memory system. We establish an ergodic convergence rate $O(1/\sqrt{K})$ for both algorithms and prove that the linear speedup is achievable if the number of workers is bounded by \sqrt{K} (K is the total number of iterations). Our results generalize and improve existing analysis for convex minimization.

1 Introduction

The asynchronous parallel optimization recently received many successes and broad attention in machine learning and optimization [Niu et al., 2011, Li et al., 2013, 2014b, Yun et al., 2013, Fercoq and Richtárik, 2013, Zhang and Kwok, 2014, Marecek et al., 2014, Tappenden et al., 2015, Hong, 2014]. It is mainly due to that the asynchronous parallelism largely reduces the system overhead comparing to the synchronous parallelism. The key idea of the asynchronous parallelism is to allow all workers work independently and have no need of synchronization or coordination. The asynchronous parallelism has been successfully applied to speedup many state-of-the-art optimization algorithms including stochastic gradient [Niu et al., 2011, Agarwal and Duchi, 2011, Zhang et al., 2014, Feyzmahdavian et al., 2015, Paine et al., 2013, Mania et al., 2015], stochastic coordinate descent [Avron et al., 2014, Liu et al., 2014a, Sridhar et al., 2013], dual stochastic coordinate ascent [Tran et al., 2015], and randomized Kaczmarz algorithm [Liu et al., 2014b].

In this paper, we are particularly interested in the asynchronous parallel stochastic gradient algorithm (ASYSG) for *nonconvex* optimization mainly due to its recent successes and popularity in deep neural network [Bengio et al., 2003, Dean et al., 2012, Paine et al., 2013, Zhang et al., 2014, Li et al., 2014a] and matrix completion [Niu et al., 2011, Petroni and Querzoni, 2014, Yun et al., 2013]. While some research efforts have been made to study the convergence and speedup properties of ASYSG for *convex* optimization, people still know very little about its properties in *nonconvex* optimization. Existing theories cannot explain its convergence and excellent speedup property in practice, mainly due to the nonconvexity of most deep learning formulations and the asynchronous parallel mechanism. People even have no idea if its convergence is certified for nonconvex optimization, although it has been used widely in solving deep neural network and implemented on different platforms such as computer network and shared memory (for example, multicore and multiGPU) system.

To fill these gaps in theory, this paper tries to make the first attempt to study ASYSG for the following nonconvex optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) := \mathbb{E}_{\xi}[F(x; \xi)] \quad (1)$$

where $\xi \in \Xi$ is a random variable and $f(x)$ is a smooth (but not necessarily convex) function. The most common specification is that Ξ is an index set of all training samples $\Xi = \{1, 2, \dots, N\}$ and $F(x; \xi)$ is the loss function with respect to the training sample indexed by ξ .

We consider two popular asynchronous parallel implementations of SG: one is for the computer network originally proposed in [Agarwal and Duchi, 2011] and the other one is for the shared memory (including multicore/multiGPU) system originally proposed in [Niu et al., 2011]. Note that due to the architecture diversity, it leads to two different algorithms. The key difference lies on that the computer network can naturally (also efficiently) ensure the atomicity of reading and writing the *whole* vector of x , while the shared memory system is unable to do that efficiently and usually only ensures efficiency for atomic reading and writing on a *single* coordinate of parameter x . The implementation on computer cluster is described by the “consistent asynchronous parallel SG” algorithm (ASYSG-CON), because the value of parameter x used for stochastic gradient evaluation is **consistent** – an existing value of parameter x at some time point. Contrarily, we use the “inconsistent asynchronous parallel SG” algorithm (ASYSG-INCON) to describe the implementation on the shared memory platform, because the value of parameter x used is **inconsistent**, that is, it might not be the real state of x at any time point.

This paper studies the theoretical convergence and speedup properties for both algorithms. We establish an asymptotic convergence rate of $O(1/\sqrt{KM})$ for ASYSG-CON where K is the total iteration number and M is the size of minibatch. The linear speedup¹ is proved to be achievable while the number of workers is bounded by $O(\sqrt{K})$. For ASYSG-INCON, we establish an asymptotic convergence and speedup properties similar to ASYSG-CON. The intuition of the linear speedup of asynchronous parallelism for SG can be explained in the following: Recall that the serial

¹The speedup for T workers is defined as the ratio between the total work load using one worker and the average work load using T workers to obtain a solution at the same precision. “The linear speedup is achieved” means that the speedup with T workers greater than cT for any values of T ($c \in (0, 1]$ is a constant independent to T).

SG essentially uses the “stochastic” gradient to surrogate the accurate gradient. ASySG brings additional deviation from the accurate gradient due to using “stale” (or delayed) information. If the additional deviation is relatively minor to the deviation caused by the “stochastic” in SG, the total iteration complexity (or convergence rate) of ASySG would be comparable to the serial SG, which implies a nearly linear speedup. This is the key reason why ASySG works.

The main contributions of this paper are highlighted as follows:

- Our result for ASySG-CON generalizes and improves earlier analysis of ASySG-CON for convex optimization in [Agarwal and Duchi, 2011]. Particularly, we improve the upper bound of the maximal number of workers to ensure the linear speedup from $O(K^{1/4}M^{-3/4})$ to $O(K^{1/2}M^{-1/2})$ by a factor $K^{1/4}M^{1/4}$;
- The proposed ASySG-INCON algorithm provides a more accurate description than HOGWILD! [Niu et al., 2011] for the lock-free implementation of ASySG on the shared memory system. Although our result does not strictly dominate the result for HOGWILD! due to different problem settings, our result can be applied to more scenarios (e.g., nonconvex optimization);
- Our analysis provides theoretical (convergence and speedup) guarantees for many recent successes of ASySG in deep learning. To the best of our knowledge, this is the first work that offers such theoretical support.

Notation

- x^* denotes the global optimal solution to (1).
- $\|x\|_0$ denotes the ℓ_0 norm of vector x , that is, the number of nonzeros in x ;
- $e_i \in \mathbb{R}^n$ denotes the i th natural unit basis vector;
- We use $\mathbb{E}_{\tilde{\zeta}_{k,*}}(\cdot)$ to denote the expectation with respect to a set of variables $\{\tilde{\zeta}_{k,1}, \dots, \tilde{\zeta}_{k,M}\}$;
- $\mathbb{E}(\cdot)$ means taking the expectation in terms of all random variables.
- $G(x; \tilde{\zeta})$ is used to denote $\nabla F(x; \tilde{\zeta})$ for short;
- We use $\nabla_i f(x)$ and $(G(x; \tilde{\zeta}))_i$ to denote the i th element of $\nabla f(x)$ and $G(x; \tilde{\zeta})$ respectively.

Assumption Throughout this paper, we make the following assumption for the objective function. All of them are quite common in the analysis of stochastic gradient algorithms.

Assumption 1. *We assume that the following holds:*

- **(Unbiased Gradient):** *The stochastic gradient $G(x; \tilde{\zeta})$ is unbiased, that is to say,*

$$\nabla f(x) = \mathbb{E}_{\tilde{\zeta}}[G(x; \tilde{\zeta})] \tag{2}$$

- **(Bounded Variance):** The variance of stochastic gradient is bounded:

$$\mathbb{E}_{\xi}(\|G(x; \xi) - \nabla f(x)\|^2) \leq \sigma^2, \quad \forall x. \quad (3)$$

- **(Lipschitzian Gradient):** The gradient function $\nabla f(\cdot)$ is Lipschitzian, that is to say,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, \forall y. \quad (4)$$

Under the Lipschitzian gradient assumption, we can define two more constants L_s and L_{\max} . Let s be any positive integer. Define L_s to be the minimal constant satisfying the following inequality:

$$\left\| \nabla f(x) - \nabla f\left(x + \sum_{i \in S} \alpha_i e_i\right) \right\| \leq L_s \left\| \sum_{i \in S} \alpha_i e_i \right\|, \quad \forall S \subset \{1, 2, \dots, n\} \text{ and } |S| \leq s \quad (5)$$

Define L_{\max} as the minimum constant that satisfies:

$$|\nabla_i f(x) - \nabla_i f(x + \alpha e_i)| \leq L_{\max} |\alpha|, \quad \forall i \in \{1, 2, \dots, n\}. \quad (6)$$

It can be seen that $L_{\max} \leq L_s \leq L$.

2 Related Work

This section reviews stochastic gradient algorithms, synchronous parallel stochastic gradient algorithms, asynchronous parallel gradient algorithms, and asynchronous parallel stochastic gradient algorithms.

Stochastic gradient is a very powerful approach for large scale optimization. The well known convergence rate of stochastic gradient is $O(1/\sqrt{K})$ for convex problems and $O(1/K)$ for strongly convex problems, for example, see [Nemirovski et al., 2009, Moulines and Bach, 2011]. A parallel algorithm of stochastic gradient is the dual averaging approach, which achieves the same convergence rate [Xiao, 2009]. The stochastic gradient is also closely related to online learning algorithms. Please refer to the online learning literature, for example, [Crammer et al., 2006, Shalev-Shwartz, 2011, Yang et al., 2014]. Most studies for stochastic gradient focus on convex optimization. For nonconvex optimization using SG, Ghadimi and Lan [2013] proved an ergodic convergence rate of $O(1/\sqrt{K})$ for stochastic gradient under nonconvex optimization, which is consistent with the rate of stochastic gradient for convex problems.

As for *synchronous parallel stochastic gradient methods*, Dekel et al. [2012] proposed a mini-batch stochastic gradient algorithm – multiple workers compute the stochastic gradient on M data in parallel and need a synchronization step before modifying parameter x in every iteration. The convergence rate is proven to be $O(1/\sqrt{KM})$ for convex optimization. In their follow up work [Dekel et al., 2010], they showed that a variant of their approach is asymptotically robust to asynchrony, as long as most processors remain synchronized for most of time. (Precisely, this variant can be considered as a partially asynchronous parallel stochastic gradient algorithm.)

[Zinkevich et al., 2010] studied a method where each node in the network runs the vanilla stochastic gradient method, using random subsets of the overall data set, and their solutions are only averaged in the final step.

The *asynchronous parallel algorithms* received broad attention in optimization recently, although pioneer studies started from 1980s [Bertsekas and Tsitsiklis, 1989]. Due to the rapid development of hardware resources, the asynchronous parallelism recently received many successes when applied to parallel stochastic gradient [Niu et al., 2011, Agarwal and Duchi, 2011, Zhang et al., 2014, Feyzmahdavian et al., 2015, Paine et al., 2013], stochastic coordinate descent [Avron et al., 2014, Liu et al., 2014a], dual stochastic coordinate ascent [Tran et al., 2015], randomized Kaczmarz algorithm [Liu et al., 2014b], and ADMM [Zhang and Kwok, 2014]. Liu et al. [2014a] and Liu and Wright [2014] studied the asynchronous parallel stochastic coordinate descent algorithm with consistent read and inconsistent read respectively and prove the linear speedup is achievable if $T \leq O(n^{1/2})$ for smooth convex functions and $T \leq O(n^{1/4})$ for functions with “smooth convex loss + nonsmooth convex separable regularization”. Avron et al. [2014] studied this asynchronous parallel stochastic coordinate descent algorithm in solving $Ax = b$ where A is a symmetric positive definite matrix, and showed that the linear speedup is achievable if $T \leq O(n)$ for consistent read and $T \leq O(n^{1/2})$ for inconsistent read. Tran et al. [2015] studied a semi-asynchronous parallel version of Stochastic Dual Coordinate Ascent algorithm which periodically enforces primal-dual synchronization in a separate thread.

We review the *asynchronous parallel stochastic gradient algorithms* in the last. Agarwal and Duchi [2011] analyzed the ASySG-CON algorithm (on computer cluster) for *convex smooth* optimization and proved a convergence rate of $O\left(\frac{1}{\sqrt{MK}} + \frac{MT^2}{K}\right)$ which implies that linear speedup is achieved when T is bounded by $O(K^{1/4}/M^{3/4})$. In comparison, our analysis for the more general nonconvex smooth optimization improves the upper bound by a factor $K^{1/4}M^{1/4}$. A very recent work [Feyzmahdavian et al., 2015] extended the analysis in Agarwal and Duchi [2011] to minimize functions in the form “smooth convex loss + nonsmooth convex regularization” and obtained similar results. Niu et al. [2011] proposed a lock free asynchronous parallel implementation of SG on the shared memory system and described this implementation as HOGWILD! algorithm. They proved a sublinear convergence rate $O(1/K)$ for *strongly* convex smooth objectives. Another recent work Mania et al. [2015] analyzed asynchronous stochastic optimization algorithms for convex functions by viewing it as a serial algorithm with the input perturbed by bounded noise and proved the convergences rates no worse than using traditional point of view for several algorithms.

3 Asynchronous parallel stochastic gradient for computer network

This section considers the asynchronous parallel implementation of SG on computer network proposed by Agarwal and Duchi [2011]. It has been successfully applied to the distributed neural network [Dean et al., 2012] and the parameter server [Li et al., 2014a] to solve deep neural network.

3.1 Algorithm Description: ASYSG-CON

Algorithm 1 ASYSG-CON

Require: $x_0, K, \{\gamma_k\}_{k=0, \dots, K-1}$

Ensure: x_K

- 1: **for** $k = 0, \dots, K - 1$ **do**
 - 2: Randomly select M training samples indexed by $\xi_{k,1}, \xi_{k,2}, \dots, \xi_{k,M}$;
 - 3: $x_{k+1} = x_k - \gamma_k \sum_{m=1}^M G(x_{k-\tau_{k,m}}, \xi_{k,m})$;
 - 4: **end for**
-

The “star” in the star-shaped network is a master machine² which maintains the parameter x . Other machines in the computer network serve as workers which only communicate with the master. All workers exchange information with the master independently and simultaneously, basically repeating the following steps:

- **(Select):** randomly select a subset of training samples $S \in \Xi$;
- **(Pull):** pull parameter x from the master;
- **(Compute):** compute the stochastic gradient $g \leftarrow \sum_{\xi \in S} G(x; \xi)$;
- **(Push):** push g to the master.

The master basically repeats the following steps:

- **(Aggregate):** aggregate a certain amount of stochastic gradients “ g ” from workers;
- **(Sum):** summarize all “ g ”s into a vector Δ ;
- **(Update):** update parameter x by $x \leftarrow x - \gamma \Delta$.

While the master is aggregating stochastic gradients from workers, it does not care about the sources of the collected stochastic gradients. As long as the total amount achieves the predefined quantity, the master will compute Δ and perform the update on x . The “update” step is performed as an atomic operation – workers cannot read the value of x during this step, which can be efficiently implemented in the network (especially in the parameter server [Li et al., 2014a]). The key difference between this asynchronous parallel implementation of SG and the serial (or synchronous parallel) SG algorithm lies on that in the “update” step, some stochastic gradients “ g ” in “ Δ ” might be computed from some early value of x instead of the current one, while in the serial SG, all g ’s are guaranteed to use the current value of x .

²There could be more than one machines in some networks, but all of them serves the same purpose and can be treated as a single machine.

The asynchronous parallel implementation substantially reduces the system overhead and overcomes the possible large network delay, but the cost is to use the old value of “ x ” in the stochastic gradient evaluation. We will show in Section 3.2 that the negative affect of this cost will vanish asymptotically.

To mathematically characterize this asynchronous parallel implementation, we monitor parameter x in the master. We use the subscript k to indicate the k th iteration on the master. For example, x_k denotes the value of parameter x after k updates, so on and so forth. We introduce a variable $\tau_{k,m}$ to denote how many delays for x used in evaluating the m th stochastic gradient at the k th iteration. This asynchronous parallel implementation of SG on the “star-shaped” network is summarized by the ASYSG-CON algorithm, see Algorithm 1. The suffix “CON” is short for “consistent read”. “Consistent read” means that the value of x used to compute the stochastic gradient is a real state of x no matter at which time point. “Consistent read” is ensured by the atomicity of the “update” step. When the atomicity fails, it leads to “inconsistent read” which will be discussed in Section 4. It is worth noting that on some “non-star” structures the asynchronous implementation can also be described as ASYSG-CON in Algorithm 1, for example, the cyclic delayed architecture and the locally averaged delayed architecture [Agarwal and Duchi, 2011, Figure 2] .

3.2 Analysis for ASYSG-CON

To analyze Algorithm 1, besides Assumption 1 we make the following additional assumptions.

Assumption 2. *We assume that the following holds:*

- **(Independence):** *All random variables in $\{\xi_{k,m}\}_{k=0,1,\dots,K;m=1,\dots,M}$ in Algorithm 1 are independent to each other;*
- **(Bounded Age):** *All delay variables $\tau_{k,m}$'s are bounded: $\max_{k,m} \tau_{k,m} \leq T$.*

The independence assumption strictly holds if all workers select samples with *replacement*. Although it might not be satisfied strictly in practice, it is a common assumption made for the analysis purpose. The bounded delay assumption is much more important. As pointed out before, the asynchronous implementation may use some old value of parameter x to evaluate the stochastic gradient. Intuitively, the age (or “oldness”) should not be too large to ensure the convergence. Therefore, it is a natural and reasonable idea to assume an upper bound for ages. This assumption is commonly used in the analysis for asynchronous algorithms, for example, [Niu et al., 2011, Avron et al., 2014, Liu and Wright, 2014, Liu et al., 2014a, Feyzmahdavian et al., 2015, Liu et al., 2014b]. It is worth noting that the upper bound T is roughly proportional to the number of workers.

Under Assumptions 1 and 2, we have the following convergence rate for nonconvex optimization.

Theorem 1. Assume that Assumptions 1 and 2 hold and the steplength sequence $\{\gamma_k\}_{k=1,\dots,K}$ in Algorithm 1 satisfies

$$LM\gamma_k + 2L^2M^2T\gamma_k \sum_{\kappa=1}^T \gamma_{k+\kappa} \leq 1 \quad \text{for all } k = 1, 2, \dots \quad (7)$$

We have the following ergodic convergence rate for the iteration of Algorithm 1

$$\frac{1}{\sum_{k=1}^K \gamma_k} \sum_{k=1}^K \gamma_k \mathbb{E}(\|\nabla f(x_k)\|^2) \leq \frac{2(f(x_1) - f(x^*)) + \sum_{k=1}^K \left(\gamma_k^2 ML + 2L^2M^2\gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2 \right) \sigma^2}{M \sum_{k=1}^K \gamma_k}. \quad (8)$$

where $\mathbb{E}(\cdot)$ denotes taking expectation in terms of all random variables in Algorithm 1.

To evaluate the convergence rate, the commonly used metrics in convex optimization are not eligible, for example, $f(x_k) - f^*$ and $\|x_k - x^*\|^2$. For nonsmooth optimization, we use the ergodic convergence as the metric, that is, the weighted average of the ℓ_2 norm of all gradients $\|\nabla f(x_k)\|^2$, which is used in the analysis for nonconvex optimization [Ghadimi and Lan, 2013]. Although the metric used in nonconvex optimization is not exactly comparable to $f(x_k) - f^*$ or $\|x_k - x^*\|^2$ used in the analysis for convex optimization, it is not totally unreasonable to think that they are roughly in the same order. The ergodic convergence directly indicates the following convergence: If randomly select an index \tilde{K} from $\{1, 2, \dots, K\}$ with probability $\{\gamma_k / \sum_{k=1}^K \gamma_k\}$, then $\mathbb{E}(\|\nabla f(x_{\tilde{K}})\|^2)$ is bounded by the right hand side of (8) and all bounds we will show in the following.

Taking a close look at Theorem 1, we can properly choose the steplength γ_k as a constant value and obtain the following convergence rate:

Corollary 2. Assume that Assumptions 1 and 2 hold. Set the steplength γ_k to be a constant γ

$$\gamma := \sqrt{\frac{f(x_1) - f(x^*)}{MLK\sigma^2}}. \quad (9)$$

If the delay parameter T is bounded by

$$K \geq \frac{4ML(f(x_1) - f(x^*))}{\sigma^2} (T+1)^2, \quad (10)$$

then the output of Algorithm 1 satisfies the following ergodic convergence rate

$$\min_{k \in \{1, \dots, K\}} \mathbb{E}(\|\nabla f(x_k)\|^2) \leq \frac{1}{K} \sum_{k=1}^K \mathbb{E}(\|\nabla f(x_k)\|^2) \leq 4\sqrt{\frac{(f(x_1) - f(x^*))L}{MK}} \sigma. \quad (11)$$

This corollary basically claims that when the total iteration number K is greater than $O(T^2)$, the convergence rate achieves $O(1/\sqrt{MK})$. Since this rate does not depend on the delay parameter T after sufficient number of iterations, the negative effect of using old values of x for stochastic gradient evaluation vanishes asymptotically. In other words, if the total number of workers is bounded by $O(\sqrt{K/M})$, the linear speedup is achieved.

Note that our convergence rate $O(1/\sqrt{MK})$ is consistent with the serial SG (with $M = 1$) for convex optimization [Nemirovski et al., 2009], the synchronous parallel (or mini-batch) SG for convex optimization [Dekel et al., 2012], and nonconvex smooth optimization [Ghadimi and Lan, 2013]. Therefore, an important observation is that as long as the number of workers (which is proportional to T) is bounded by $O(\sqrt{K/M})$, the iteration complexity to achieve the same accuracy level will be roughly the same. In other words, the average work load for each worker is reduced by the factor T comparing to the serial SG. Therefore, the linear speedup is achievable if $T \leq O(\sqrt{K/M})$. Since our convergence rate meets several special cases, it is tight.

Next we compare with the analysis of AsySG-con for *convex* smooth optimization in Agarwal and Duchi [2011, Corollary 2]. They proved an asymptotic convergence rate $O(1/\sqrt{MK})$, which is consistent with ours. But their results require $T \leq O(K^{1/4}M^{-3/4})$ to guarantee linear speedup. Our result improves it by a factor $O(K^{1/4}M^{1/4})$.

4 Asynchronous parallel stochastic gradient for shared memory architecture

This section considers a widely used lock-free asynchronous implementation of SG on the shared memory system proposed in Niu et al. [2011]. Its advantages have been witnessed in solving SVM, graph cuts [Niu et al., 2011], linear equations [Liu et al., 2014b], and matrix completion [Petroni and Querzoni, 2014]. While the computer network always involves multiple machines, the shared memory platform usually only includes a single machine with multiple cores / GPUs sharing the same memory.

4.1 Algorithm Description: ASYSG-INCON

Algorithm 2 ASYSG-INCON

Require: x_0, K, γ

Ensure: x_K

- 1: **for** $k = 0, \dots, K - 1$ **do**
 - 2: Randomly select M training samples indexed by $\zeta_{k,1}, \zeta_{k,2}, \dots, \zeta_{k,M}$;
 - 3: Randomly select $i_k \in \{1, 2, \dots, n\}$ with uniform distribution;
 - 4: $(x_{k+1})_{i_k} = (x_k)_{i_k} - \gamma \sum_{m=1}^M (G(\hat{x}_{k,m}; \zeta_{k,m}))_{i_k}$;
 - 5: **end for**
-

For the shared memory platform, one can exactly follow AsySG-con on the computer network using software locks, which is expensive³. Therefore, in practice the lock free asynchronous

³The time consumed by locks is roughly equal to the time of 10^4 floating-point computation. The additional cost for

parallel implementation of SG is preferred. This section considers the same implementation as Niu et al. [2011], but provides a more precise algorithm description ASYSG-INCON than HOGWILD! proposed in Niu et al. [2011].

In this lock free implementation, the shared memory stores the parameter “ x ” and allows all workers reading and modifying parameter x simultaneously without using locks. All workers repeat the following steps independently, concurrently, and simultaneously:

- **(Read):** read the parameter from the shared memory to the local memory *without software locks* (we use \hat{x} to denote its value);
- **(Compute):** sample a training data ξ and use \hat{x} to compute the stochastic gradient $G(\hat{x}; \xi)$ *locally*;
- **(Update):** update parameter x in the shared memory *without software locks* $x \leftarrow x - \gamma G(\hat{x}; \xi)$.

Since we do not use locks in both “read” and “update” steps, it means that multiple workers may manipulate the shared memory simultaneously. It causes the “inconsistent read” at the “read” step, that is, the value of \hat{x} read from the shared memory might not be any state of x in the shared memory at any time point. For example, at time 0, the original value of x in the shared memory is a two dimensional vector $[a, b]$; at time 1, worker W is running the “read” step and first reads a from the shared memory; at time 2, worker W' updates the first component of x in the shared memory from a to a' ; at time 2, worker W' updates the second component of x in the shared memory from b to b' ; at time 3, worker W reads the value of the second component of x in the shared memory as b' . In this case, worker W eventually obtains the value of \hat{x} as $[a, b']$, which is not a real state of x in the shared memory at any time point. Recall that in ASYSG-CON the parameter value obtained by any worker is guaranteed to be some real value of parameter x at some time point.

To precisely characterize this implementation and especially represent \hat{x} , we monitor the value of parameter x in the shared memory. We define one *iteration* as a modification on any *single* component of x in the shared memory since the update on a single component can be considered to be atomic on GPUs and DSPs [Niu et al., 2011]. We use x_k to denote the value of parameter x in the shared memory after k iterations and \hat{x}_k to denote the value read from the shared memory and used for computing stochastic gradient at the k th iteration. \hat{x}_k can be represented by x_k with a few earlier updates missing

$$\hat{x}_k = x_k - \sum_{j \in J(k)} (x_{j+1} - x_j) \quad (12)$$

where $J(k) \subset \{k-1, k, \dots, 0\}$ is a subset of index numbers of previous iterations. This way is also used in analyzing asynchronous parallel coordinate descent algorithms in [Avron et al., 2014, Liu and Wright, 2014]. The k th update happened in the shared memory can be described as

$$(x_{k+1})_{i_k} = (x_k)_{i_k} - \gamma (G(\hat{x}_k; \xi_k))_{i_k}$$

using locks is the waiting time during which multiple worker access the same memory address.

where ζ_k denotes the index of the selected data and i_k denotes the index of the component being updated at k th iteration. In the original analysis for the HOGWILD! implementation [Niu et al., 2011], \hat{x}_k is assumed to be some earlier state of x in the shared memory (that is, the consistent read) for simpler analysis, although it is not true in practice.

One more complication is to apply the mini-batch strategy like before. Since the “update” step needs physical modification in the shared memory, it is usually much more time consuming than both “read” and “compute” steps are. If many workers run the “update” step simultaneously, the memory contention will seriously harm the performance. To reduce the risk of memory contention, a common trick is to ask each worker to gather multiple (say M) stochastic gradients and write the shared memory only once. That is, in each cycle, run both “update” and “compute” steps for M times before you run the “update” step. Thus, the mini-batch updates happen in the shared memory can be written as

$$(x_{k+1})_{i_k} = (x_k)_{i_k} - \gamma \sum_{m=1}^M (G(\hat{x}_{k,m}; \zeta_{k,m}))_{i_k} \quad (13)$$

where i_k denotes the coordinate index updated at the k th iteration, and $G(\hat{x}_{k,m}; \zeta_{k,m})$ is the m th stochastic gradient computed from the data sample indexed by $\zeta_{k,m}$ and the parameter value denoted by $\hat{x}_{k,m}$ at the k th iteration. $\hat{x}_{k,m}$ can be expressed by:

$$\hat{x}_{k,m} = x_k - \sum_{j \in J(k,m)} (x_{j+1} - x_j) \quad (14)$$

where $J(k, m) \subset \{k-1, k, \dots, 0\}$ is a subset of index numbers of previous iterations. The algorithm is summarized in Algorithm 2 from the view of the shared memory.

4.2 Analysis for ASYSG-INCON

To analyze the ASYSG-INCON, we need to make a few assumptions similar to Niu et al. [2011], Liu et al. [2014b], Avron et al. [2014], Liu and Wright [2014].

Assumption 3. *We assume that the following holds for Algorithm 2:*

- **(Independence):** *All groups of variables $\{i_k, \{\zeta_{k,m}\}_{m=1}^M\}$ at different iterations from $k = 1$ to K are independent to each other.*
- **(Bounded Age):** *Let T be the global bound for delay: $J(k, m) \subset \{k-1, \dots, k-T\}$, $\forall k, \forall m$, so $|J(k, m)| \leq T$.*

The independence assumption might not be true in practice, but it is probably the best assumption one can make in order to analyze the asynchronous parallel SG algorithm. This assumption was also used in the analysis for HOGWILD! [Niu et al., 2011] and asynchronous randomized

Kaczmarz algorithm [Liu et al., 2014b]. The bounded delay assumption basically restricts the age of all missing components in $\hat{x}_{k,m}$ ($\forall m, \forall k$). The upper bound “ T ” here serves a similar purpose as in Assumption 2. Thus we abuse this notation in this section. The value of T is proportional to the number of workers and does not depend on the size of mini-batch M . The bounded age assumption is used in the analysis for asynchronous stochastic coordinate descent with “inconsistent read” [Avron et al., 2014, Liu and Wright, 2014]. Under Assumptions 1 and 3, we have the following results:

Theorem 3. *Assume that Assumptions 1 and 3 hold and the constant steplength γ satisfies*

$$\frac{2M^2TL_T^2(\sqrt{n} + T - 1)\gamma^2}{n^{3/2}} + 2ML_{\max}\gamma \leq 1. \quad (15)$$

We have the following ergodic convergence rate for Algorithm 2

$$\frac{1}{K} \sum_{t=1}^K \mathbb{E} (\|\nabla f(x_t)\|^2) \leq \frac{2n}{KM\gamma} (f(x_1) - f(x^*)) + \frac{L_T^2 TM\gamma^2}{2n} \sigma^2 + L_{\max}\gamma\sigma^2. \quad (16)$$

Taking a close look at Theorem 3, we can choose the steplength γ properly and obtain the following error bound:

Corollary 4. *Assume that Assumptions 1 and 3 hold. Set the steplength to be a constant γ*

$$\gamma := \frac{\sqrt{2(f(x_1) - f(x^*))n}}{\sqrt{KL_TM\sigma}}. \quad (17)$$

If the total iterations K is greater than

$$K \geq \frac{16(f(x_1) - f(x^*))L_TM (n^{3/2} + 4T^2)}{\sqrt{n}\sigma^2}, \quad (18)$$

then the output of Algorithm 2 satisfies the following ergodic convergence rate

$$\frac{1}{K} \sum_{k=1}^K \mathbb{E} (\|\nabla f(x_k)\|^2) \leq \sqrt{\frac{72(f(x_1) - f(x^*))L_Tn}{KM}} \sigma. \quad (19)$$

This corollary indicates the asymptotic convergence rate achieves $O(1/\sqrt{MK})$ when the total iteration number K exceeds a threshold in the order of $O(T^2)$ (if n is considered as a constant). We can see that this rate and the threshold are consistent with the result in Corollary 2 for ASYSG-CON. One may argue that why there is an additional factor \sqrt{n} in the numerator of (19). That is due to the way we count iterations – one iteration is defined as updating a single component of x . If we take into account this factor in the comparison to ASYSG-CON, the convergence rates for ASYSG-CON and ASYSG-INCON are essentially consistent. This comparison implies that the “inconsistent read” would not make a big difference from the “consistent read”.

Next we compare our result with the analysis of HOGWILD! by [Niu et al., 2011]. In principle, our analysis and their analysis consider the same implementation of asynchronous parallel SG, but differ in the following aspects: 1) our analysis considers the smooth nonconvex optimization which includes the smooth strongly convex optimization considered in their analysis; 2) our analysis considers the “inconsistent read” model which meets the practice while their analysis assumes the impractical “consistent read” model. Although the two results are not absolutely comparable, it is still interesting to see the difference. Niu et al. [2011] proved that the linear speedup is achievable if the maximal number of nonzeros in stochastic gradients is bounded by $O(1)$ and the number of workers is bounded by $O(n^{1/4})$. Our analysis does not need this prerequisite and guarantees the linear speedup as long as the number of workers is bounded by $O(\sqrt{K})$. Although it is hard to say that our result strictly dominates HOGWILD! in Niu et al. [2011], our asymptotic result is eligible for more scenarios.

We consider an extension of ASYSG-INCON for sparse stochastic gradients. In this scenario, we slightly change Steps 3 and 4 in Algorithm 2:

3: Uniformly select i_k from the support set of $g_k := \sum_{m=1}^M G(\hat{x}_{k,m}; \zeta_{k,m})$;
 4: $(x_{k+1})_{i_k} = (x_k)_{i_k} - \gamma \|g_k\|_0 (g_k)_{i_k}$.

The convergence rate in (19) can be slightly improved by taking the advantage of sparsity:

$$\frac{1}{K} \sum_{k=1}^K \mathbb{E}(\|\nabla f(x_k)\|^2) \leq \frac{6\sqrt{2} (f(x_1) - f(x^*)) L_T \max_k \|g_k\|_0 \sigma}{\sqrt{KM}}. \quad (20)$$

The proof can be simply get by extending the proof for (19).

5 Experiments

The successes of ASYSG-CON and ASYSG-INCON and their advantages over synchronous parallel algorithms have been widely witnessed in many applications such as deep neural network [Dean et al., 2012, Paine et al., 2013, Zhang et al., 2014, Li et al., 2014a], matrix completion [Niu et al., 2011, Petroni and Querzoni, 2014, Yun et al., 2013], SVM [Niu et al., 2011], and solving linear equations [Liu et al., 2014b]. We refer readers to these literatures for more comprehensive comparison and empirical studies. This section mainly provides the empirical study to validate the speedup properties for *completeness*.

We perform experiments for ASYSG-CON and ASYSG-INCON on computer cluster and multicore machine respectively. The main purpose of the following experiments is to validate the speedup property. We are particularly interested in two types of speedup: iteration speedup and running time speedup. The iteration speedup is exactly the speedup we discussed in the whole paper.

Given T workers, it is computed from the ratio

$$\text{iteration speedup of } T \text{ workers} = \frac{\# \text{ of total iterations of the serial SG (or using one worker)}}{\# \text{ of total iterations using } T \text{ workers}} \times T$$

where $\#$ is the iteration count when the same level of precision achieved. This speedup is less affected by the hardware. The running time speedup is the actual speedup. It is defined with respect to the running time:

$$\text{running time speedup of } T \text{ workers} = \frac{\text{running time for the serial SG (or using one worker)}}{\text{running time of using } T \text{ workers}}.$$

The running time speedup is seriously affected by the hardware. It is generally worse than the iteration speedup.

5.1 ASYSG-CON

We implement ASYSG-CON for deep neural network based on the Caffe [Jia et al., 2014] package. Caffe is an open source code base of deep learning algorithms. We evaluate ASYSG-CON on two standard datasets provided in the Caffe package, LENET and CIFAR10-FULL.

The neural network consists of convolution layer, nonlinear layer, max pool layer and fully connected layer [Krizhevsky et al., 2012], and the detailed specification can be found on the Caffe website⁴. LENET is a digit classifier network, training on the MNIST dataset⁵. CIFAR10-full has 10 classes of color images, training on the CIFAR10 dataset [Krizhevsky and Hinton, 2009]. We first initialize a parameter server hosting the parameters, and then spawn up to 8 stochastic gradient workers. The point to point communication between the parameter server and gradient workers are handled by the MPICH library⁶. The parameter server and stochastic gradient workers run on separate machines, and each process uses a single core of a Xeon(R) E5-2430 CPU. The steplength γ for LENET is chosen as the default value. It means that this steplength has been tuned to be the optimal for the serial SG algorithm on LENET. The steplength we used for CIFAR10-FULL is chosen as the default value as well.

We draw the curves of objective loss against iterations and running time in Figures 1 and 2 respectively, and report their speedups in Tables 1 and 2. (More details about datasets and the parameter setting in experiments can be found in Table 3.) We can observe that

- The iteration speedup is always better than the running time speedup, which meets our common sense;
- The speedups for both problems are comparable overall as shown in Tables 1 and 2. The time speedup for CIFAR10-FULL is slightly more stable, while we notice that the time

⁴<https://github.com/BVLC/caffe>

⁵<http://yann.lecun.com/exdb/mnist/>

⁶<https://www.mpich.org/>

speedup for LENET in Table 1 suddenly drops to “2.88” with mpi-8 from “5.29” with mpi-7. That is because it hits the ceiling of communication bandwidth. The number of parameters in LENET is more than CIFAR10-FULL, thus requiring more communication cost. When the number of machines achieves a certain threshold (in this case LENET, it is 8), the performance might become dramatically worse.

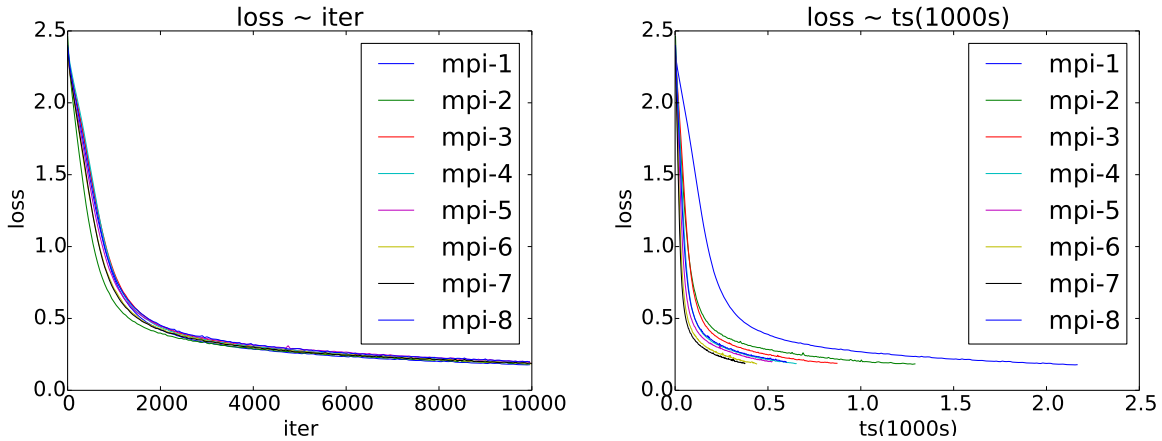


Figure 1: LENET. The AsySG-con algorithm is run on various numbers of machines from 1 to 8 to solve LENET. The curves of the objective loss against the number of iteration and the running time are drawn in the left and the right graphs respectively.

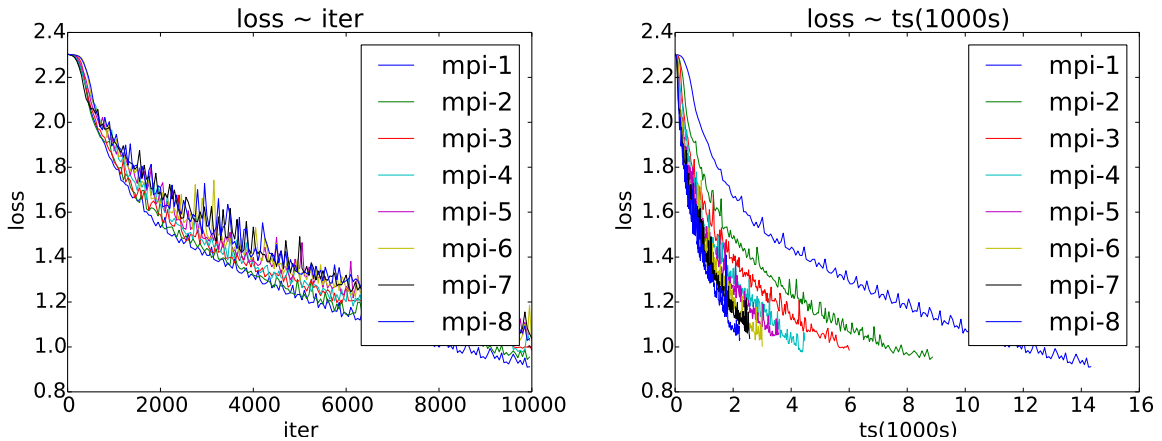


Figure 2: CIFAR10-FULL. The AsySG-con algorithm is run on various numbers of machines from 1 to 8 to solve CIFAR10-FULL. The curves of the objective loss against the number of iteration and the running time are drawn in the left and the right graphs respectively.

Table 1: Iteration speedup and running time speedup of AsySG-CON. (LENET)

	mpi-1	mpi-2	mpi-3	mpi-4	mpi-5	mpi-6	mpi-7	mpi-8
iteration speedup	1.07	2.02	2.77	3.73	4.20	5.82	6.86	6.97
time speedup	0.98	1.69	2.24	2.96	3.30	4.51	5.29	2.88

Table 2: Iteration speedup and running time speedup of AsySG-CON. (CIFAR10-FULL)

	mpi-1	mpi-2	mpi-3	mpi-4	mpi-5	mpi-6	mpi-7	mpi-8
iteration speedup	1.01	1.93	2.65	3.42	4.27	4.92	5.36	5.96
time speedup	1.00	1.73	2.28	2.88	3.56	4.07	4.41	5.00

Table 3: More details about LENET and CIFAR10-FULL.

Datasets	Type	#Images	MiniBatch	#CONV	#FC	#Params
LENET	28x28 grayscale	60K	60	2	2	431,080
CIFAR10-FULL	32x32 RGB	50K	100	3	1	89,578

5.2 ASYSG-INCON

We conduct the empirical study for ASYSG-INCON on the machine (Intel Xeon architecture), which has 4 sockets and 10 cores for each socket. The synthetic data is generated from a full connected neural network with 5 layers ($400 \times 100 \times 50 \times 20 \times 10$) and 46380 parameters totally. The total number of samples is 463800. The data size is about 1.5 GB. The input vector and all parameters are generated from i.i.d. Gaussian distribution. The output vector is constructed by applying the network parameter to the input vector plus some Gaussian random noise.

We run ASYSG-INCON on various numbers of cores from 1 to 32. The size of mini-batch is chosen as $M = 32$ and the steplength is chosen as $\gamma = 1.1 \times 10^{-7}$. Both parameters are chosen based on the best performance of the serial SG to achieve the precision 10^{-1} for the ℓ_2 norm of gradient. Figure 3 draws the curve of the ℓ_2 norm of gradients against the number of iterations and running time respectively. The speedup is reported in Table 4. We observe that the iteration speedup is almost linear while the running time speedup is slightly worse than the iteration speedup. The overall performance of ASYSG-INCON on the shared memory system is better than ASYSG-CON on the computer cluster. The reason is that the computer cluster suffers from serious communication delay and the delay bound T on the computer cluster is usually larger than on the shared memory system.

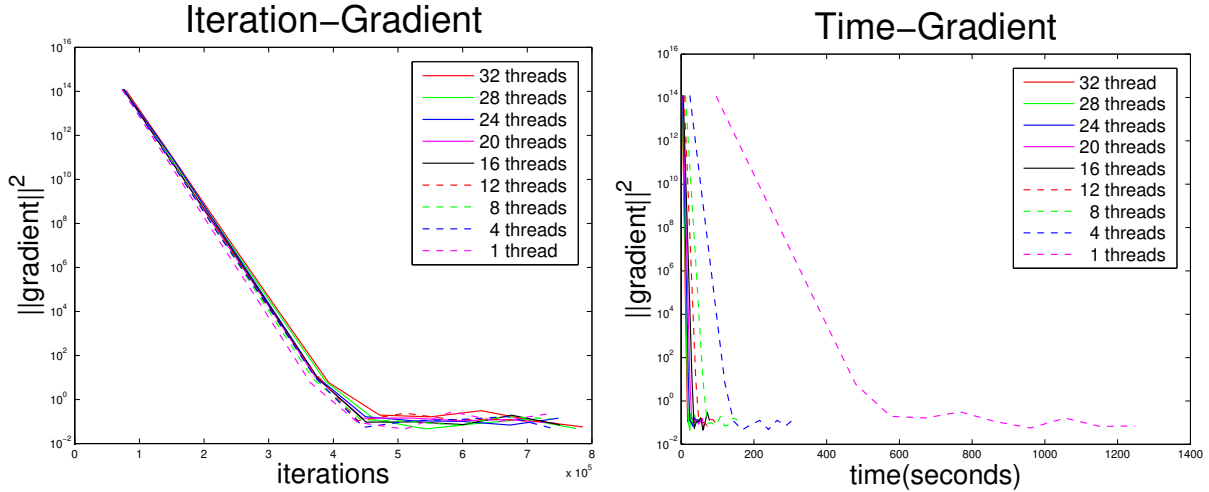


Figure 3: Deep neural network for synthetic data using AsySG-INCON. The AsySG-CON algorithm is run on various numbers of machines from 1 to 32. The curves of the objective loss against the number of iteration and the running time are drawn in the left and the right graphs respectively.

Table 4: Iteration speedup and running time speedup of AsySG-INCON (synthetic data).

	thr-1	thr-4	thr-8	thr-12	thr-16	thr-20	thr-24	thr-28	thr-32
iteration speedup	1	3.9	7.8	11.6	15.4	19.9	24.1	28.7	31.6
time speedup	1	4.0	8.1	11.9	16.3	19.2	22.7	26.1	29.2

6 Conclusion

This paper studied two popular asynchronous parallel implementations for SG on computer cluster and shared memory system respectively. Two algorithms (ASYSG-CON and ASYSG-INCON) are used to describe two implementations. An asymptotic sublinear convergence rate is proven for both algorithms on *nonconvex* smooth optimization. This rate is consistent with the result of SG for convex optimization. The linear speedup is proven to be achievable when the number of workers is bounded by \sqrt{K} , which improves the earlier analysis of ASYSG-CON for convex optimization in [Agarwal and Duchi, 2011]. The proposed ASYSG-INCON algorithm provides a more precise description for lock free implementation on shared memory system than HOGWILD! [Niu et al., 2011]. Our result for ASYSG-INCON can be applied to more scenarios.

Acknowledgements

This project is supported by the NSF grant CNS-1548078, the NEC fellowship, and the startup funding at University of Rochester. We thank Professor Daniel Gildea and Professor Sandhya Dwarkadas at University of Rochester, Professor Stephen J. Wright at University of Wisconsin-Madison, and anonymous (meta-)reviewers for their constructive comments and helpful advices.

References

- A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. *NIPS*, 2011.
- H. Avron, A. Druinsky, and A. Gupta. Revisiting asynchronous linear solvers: Provable convergence rate through randomization. *IPDPS*, 2014.
- Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.
- D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.
- K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.
- J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. Large scale distributed deep networks. *NIPS*, 2012.
- O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Robust distributed online prediction. *arXiv preprint arXiv:1012.1370*, 2010.
- O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1):165–202, 2012.

- O. Fercoq and P. Richtárik. Accelerated, parallel and proximal coordinate descent. *arXiv preprint arXiv:1312.5799*, 2013.
- H. R. Feyzmahdavian, A. Aytekin, and M. Johansson. An asynchronous mini-batch algorithm for regularized stochastic optimization. *ArXiv e-prints*, May 18 2015.
- S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- M. Hong. A distributed, asynchronous and incremental algorithm for nonconvex optimization: An ADMM based approach. *arXiv preprint arXiv:1412.6058*, 2014.
- Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 1(4):7, 2009.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, pages 1097–1105, 2012.
- M. Li, L. Zhou, Z. Yang, A. Li, F. Xia, D. G. Andersen, and A. Smola. Parameter server for distributed machine learning. *Big Learning NIPS Workshop*, 2013.
- M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su. Scaling distributed machine learning with the parameter server. *OSDI*, 2014a.
- M. Li, D. G. Andersen, A. J. Smola, and K. Yu. Communication efficient distributed machine learning with the parameter server. *NIPS*, 2014b.
- J. Liu and S. J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *arXiv preprint arXiv:1403.3862*, 2014.
- J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *ICML*, 2014a.
- J. Liu, S. J. Wright, and S. Sridhar. An asynchronous parallel randomized kaczmarz algorithm. *arXiv preprint arXiv:1401.4780*, 2014b.
- H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *arXiv preprint arXiv:1507.06970*, 2015.
- J. Marecek, P. Richtárik, and M. Takác. Distributed block coordinate descent for minimizing partially separable functions. *arXiv preprint arXiv:1406.0238*, 2014.
- E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *NIPS*, 2011.

- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- F. Niu, B. Recht, C. Re, and S. Wright. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. *NIPS*, 2011.
- T. Paine, H. Jin, J. Yang, Z. Lin, and T. Huang. Gpu asynchronous stochastic gradient descent to speed up neural network training. *NIPS*, 2013.
- F. Petroni and L. Querzoni. Gasgd: stochastic gradient descent for distributed asynchronous matrix completion via graph partitioning. *ACM Conference on Recommender systems*, 2014.
- S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.
- S. Sridhar, S. Wright, C. Re, J. Liu, V. Bittorf, and C. Zhang. An approximate, efficient LP solver for lp rounding. *NIPS*, 2013.
- R. Tappenden, M. Takáč, and P. Richtárik. On the complexity of parallel coordinate descent. *arXiv preprint arXiv:1503.03033*, 2015.
- K. Tran, S. Hosseini, L. Xiao, T. Finley, and M. Bilenko. Scaling up stochastic dual coordinate ascent. *ICML*, 2015.
- L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. *NIPS*, 2009.
- T. Yang, M. Mahdavi, R. Jin, and S. Zhu. Regret bounded by gradual variation for online convex optimization. *Machine learning*, 95(2):183–223, 2014.
- H. Yun, H.-F. Yu, C.-J. Hsieh, S. Vishwanathan, and I. Dhillon. Nomad: Non-locking, stochastic multi-machine algorithm for asynchronous and decentralized matrix completion. *arXiv preprint arXiv:1312.0193*, 2013.
- R. Zhang and J. Kwok. Asynchronous distributed ADMM for consensus optimization. *ICML*, 2014.
- S. Zhang, A. Choromanska, and Y. LeCun. Deep learning with elastic averaging SGD. *CoRR*, abs/1412.6651, 2014.
- M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. *NIPS*, 2010.

Appendix: Proofs

Proofs to Theorem 1

Proof. From the Lipschitzian gradient assumption (5), we have

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= - \left\langle \nabla f(x_k), \gamma_k \sum_{m=1}^M G(x_{k-\tau_{k,m}}; \xi_{k,m}) \right\rangle + \frac{\gamma_k^2 L}{2} \left\| \sum_{m=1}^M G(x_{k-\tau_{k,m}}; \xi_{k,m}) \right\|^2. \end{aligned} \quad (21)$$

Taking expectation respect to $\xi_{k,*}$ on both sides of (21), we have

$$\begin{aligned} \mathbb{E}_{\xi_{k,*}}(f(x_{k+1})) - f(x_k) &\leq -M\gamma_k \left\langle \nabla f(x_k), \frac{1}{M} \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\rangle \\ &\quad + \frac{\gamma_k^2 L}{2} \mathbb{E}_{\xi_{k,*}} \left(\left\| \sum_{m=1}^M G(x_{k-\tau_{k,m}}; \xi_{k,m}) \right\|^2 \right) \end{aligned} \quad (22)$$

where we use the unbiased stochastic gradient assumption in (2). From the fact

$$\langle a, b \rangle = \frac{1}{2} (\|a\|^2 + \|b\|^2 - \|a - b\|^2),$$

we have

$$\begin{aligned} &\mathbb{E}_{\xi_{k,*}}(f(x_{k+1})) - f(x_k) \\ &\leq -\frac{M\gamma_k}{2} \left(\left\| \nabla f(x_k) \right\|^2 + \left\| \frac{1}{M} \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 - \underbrace{\left\| \nabla f(x_k) - \frac{1}{M} \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2}_{T_1} \right) \\ &\quad + \underbrace{\frac{\gamma_k^2 L}{2} \mathbb{E}_{\xi_{k,*}} \left(\left\| \sum_{m=1}^M G(x_{k-\tau_{k,m}}; \xi_{k,m}) \right\|^2 \right)}_{T_2}. \end{aligned} \quad (23)$$

Next we estimate the upper bound of T_1 and T_2 . For T_2 we have

$$\begin{aligned} T_2 &= \mathbb{E}_{\xi_{k,*}} \left[\left\| \sum_{m=1}^M G(x_{k-\tau_{k,m}}; \xi_{k,m}) \right\|^2 \right] \\ &= \mathbb{E}_{\xi_{k,*}} \left[\left\| \sum_{m=1}^M \left(G(x_{k-\tau_{k,m}}; \xi_{k,m}) - \nabla f(x_{k-\tau_{k,m}}) \right) + \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{\tilde{\xi}_{k,*}} \left[\left\| \sum_{m=1}^M \left(G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}) \right) \right\|^2 \right. \\
&\quad \left. + \left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 + 2 \left\langle \sum_{m=1}^M \left(G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}) \right), \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\rangle \right] \\
&= \mathbb{E}_{\tilde{\xi}_{k,*}} \left[\left\| \sum_{m=1}^M \left(G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}) \right) \right\|^2 + \left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right] \\
&= \mathbb{E}_{\tilde{\xi}_{k,*}} \left[\sum_{m=1}^M \left\| \left(G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}) \right) \right\|^2 \right. \\
&\quad \left. + 2 \sum_{1 \leq m < m' \leq M} \left\langle G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}), G(x_{k-\tau_{k,m'}}; \tilde{\xi}_{k,m'}) - \nabla f(x_{k-\tau_{k,m'}}) \right\rangle \right. \\
&\quad \left. + \left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right] \\
&\leq M\sigma^2 + \left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \tag{24}
\end{aligned}$$

where the forth equality is due to

$$\begin{aligned}
&\mathbb{E}_{\tilde{\xi}_{k,*}} \left\langle \sum_{m=1}^M \left(G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}) \right), \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\rangle \\
&= \left\langle \sum_{m=1}^M \mathbb{E}_{\tilde{\xi}_{k,*}} \left(G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}) \right), \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\rangle \\
&= 0
\end{aligned}$$

and the last inequality is due to the assumption (3) and

$$\begin{aligned}
&\mathbb{E}_{\tilde{\xi}_{k,*}} \sum_{1 \leq m < m' \leq M} \left\langle G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}), G(x_{k-\tau_{k,m'}}; \tilde{\xi}_{k,m'}) - \nabla f(x_{k-\tau_{k,m'}}) \right\rangle \\
&= \mathbb{E}_{\tilde{\xi}_{k,*}} \sum_{1 \leq m < m' \leq M} \mathbb{E}_{k,m'} \left\langle G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}), G(x_{k-\tau_{k,m'}}; \tilde{\xi}_{k,m'}) - \nabla f(x_{k-\tau_{k,m'}}) \right\rangle \\
&= \mathbb{E}_{\tilde{\xi}_{k,*}} \sum_{1 \leq m < m' \leq M} \left\langle G(x_{k-\tau_{k,m}}; \tilde{\xi}_{k,m}) - \nabla f(x_{k-\tau_{k,m}}), \mathbb{E}_{k,m'} G(x_{k-\tau_{k,m'}}; \tilde{\xi}_{k,m'}) - \nabla f(x_{k-\tau_{k,m'}}) \right\rangle \\
&= 0. \tag{25}
\end{aligned}$$

We next turn to T_1 :

$$\begin{aligned}
T_1 &= \left\| \nabla f(x_k) - \frac{1}{M} \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \\
&= \frac{1}{M^2} \left\| \sum_{m=1}^M \left(\nabla f(x_k) - \nabla f(x_{k-\tau_{k,m}}) \right) \right\|^2
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{M} \sum_{m=1}^M \left\| \left(\nabla f(x_k) - \nabla f(x_{k-\tau_{k,m}}) \right) \right\|^2 \\
&\leq \frac{L^2}{M} \sum_{m=1}^M \|x_k - x_{k-\tau_{k,m}}\|^2 \\
&\leq L^2 \max_{k \in \{1, \dots, M\}} \|x_k - x_{k-\tau_{k,m}}\|^2 \\
&= L^2 \|x_k - x_{k-\tau_{k,\mu}}\|^2. \quad (\text{let } \mu := \arg \max_{m \in \{1, \dots, M\}} \|x_k - x_{k-\tau_{k,m}}\|^2)
\end{aligned}$$

where the second inequality is from the Lipschitzian gradient assumption (5). It follows that

$$\begin{aligned}
T_1 &\leq L^2 \|x_k - x_{k-\tau_{k,\mu}}\|^2 \\
&= L^2 \left\| \sum_{j=k-\tau_{k,\mu}}^{k-1} (x_{j+1} - x_j) \right\|^2 \\
&= L^2 \left\| \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j \sum_{m=1}^M G(x_{j-\tau_{j,m}}; \xi_{j,m}) \right\|^2 \\
&= L^2 \left\| \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j \sum_{m=1}^M \left[G(x_{j-\tau_{j,m}}; \xi_{j,m}) - \nabla f(x_{j-\tau_{j,m}}) \right] + \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2 \\
&\leq 2L^2 \left(\underbrace{\left\| \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j \sum_{m=1}^M \left[G(x_{j-\tau_{j,m}}; \xi_{j,m}) - \nabla f(x_{j-\tau_{j,m}}) \right] \right\|^2}_{T_3} + \underbrace{\left\| \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2}_{T_4} \right) \tag{26}
\end{aligned}$$

where the last inequality uses the fact that $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any real vectors a and b . Taking the expectation in terms of $\{\xi_{j,*} | j \in \{k - \tau_{k,\mu}, \dots, k - 1\}\}$ for T_3 , we have

$$\begin{aligned}
&\mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}}(T_3) \\
&= \mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} \left(\left\| \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j \sum_{m=1}^M \left[G(x_{j-\tau_{j,m}}; \xi_{j,m}) - \nabla f(x_{j-\tau_{j,m}}) \right] \right\|^2 \right) \\
&= \mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} \left(\sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j^2 \left\| \sum_{m=1}^M \left[G(x_{j-\tau_{j,m}}; \xi_{j,m}) - \nabla f(x_{j-\tau_{j,m}}) \right] \right\|^2 \right) \\
&\quad + 2 \mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} \left(\sum_{k-1 \geq j'' > j' \geq k-\tau_{k,\mu}} \gamma_{j'} \gamma_{j''} \left\langle \sum_{m=1}^M \left[G(x_{j''-\tau_{j'',m}}; \xi_{j'',m}) - \nabla f(x_{j''-\tau_{j'',m}}) \right], \right. \right. \\
&\quad \left. \left. \sum_{m=1}^M \left[G(x_{j'-\tau_{j',m}}; \xi_{j',m}) - \nabla_{i_{j'}} f(x_{j'-\tau_{j',m}}) \right] \right\rangle \right)
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} \left(\sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j^2 \left\| \sum_{m=1}^M \left[G(x_{j-\tau_{j,m}}; \xi_{j,m}) - \nabla f(x_{j-\tau_{j,m}}) \right] \right\|^2 \right) \\
&= \mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} \left(\sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j^2 \sum_{m=1}^M \left\| \left[G(x_{j-\tau_{j,m}}; \xi_{j,m}) - \nabla f(x_{j-\tau_{j,m}}) \right] \right\|^2 \right) \\
&\leq M \sum_{j=k-T}^{k-1} \gamma_j^2 \sigma^2
\end{aligned} \tag{27}$$

where the second last equality is due to (25) and the third equality is due to

$$\begin{aligned}
&\mathbb{E}_{\xi_j, k-1 \geq j \geq k-\tau_{k,\mu}} \left(\sum_{k-1 \geq j'' > j' \geq k-\tau_{k,\mu}} \gamma_{j'} \gamma_{j''} \left\langle \sum_{m=1}^M \left[G(x_{j''-\tau_{j'',m}}; \xi_{j'',m}) - \nabla f(x_{j''-\tau_{j'',m}}) \right], \right. \right. \\
&\quad \left. \left. \sum_{m=1}^M \left[G(x_{j'-\tau_{j',m}}; \xi_{j',m}) - \nabla f(x_{j'-\tau_{j',m}}) \right] \right\rangle \right) \\
&= \mathbb{E}_{\xi_j, k-1 \geq j \geq k-\tau_{k,\mu}} \left(\sum_{k-1 \geq j'' > j' \geq k-\tau_{k,\mu}} \gamma_{j'} \gamma_{j''} \left\langle \sum_{m=1}^M \left[\mathbb{E}_{j'',*} G(x_{j''-\tau_{j'',m}}; \xi_{j'',m}) - \nabla f(x_{j''-\tau_{j'',m}}) \right], \right. \right. \\
&\quad \left. \left. \sum_{m=1}^M \left[G(x_{j'-\tau_{j',m}}; \xi_{j',m}) - \nabla f(x_{j'-\tau_{j',m}}) \right] \right\rangle \right) \\
&= 0.
\end{aligned}$$

Taking the expectation in terms of $\xi_{j,*}$ for T_4 , we have

$$\begin{aligned}
&\mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} (T_4) \\
&= \mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} \left(\left\| \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2 \right) \\
&\leq T \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j^2 \mathbb{E}_{\xi_{j,*}, j \in \{k-\tau_{k,\mu}, \dots, k-1\}} \left(\left\| \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2 \right)
\end{aligned} \tag{28}$$

where the last inequality uses the upper bound of the delay age: $\tau_{k,\mu} \leq T$.

We take full expectation on both sides of (26) and substitute $\mathbb{E}(T_3)$ and $\mathbb{E}(T_4)$ by their upper bounds in (27) and (28) respectively:

$$\mathbb{E}(T_1) \leq 2L^2 \left(M \sum_{j=k-T}^{k-1} \gamma_j^2 \sigma^2 + T \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j^2 \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2 \right) \right). \tag{29}$$

Applying the upper bounds for $\mathbb{E}(T_1)$ in (29) and $\mathbb{E}(T_2)$ in (24) to (23), and take full expectation on both sides, we obtain

$$\begin{aligned}
& \mathbb{E}(f(x_{k+1})) - f(x_k) \\
& \leq -\frac{M\gamma_k}{2} \left[\mathbb{E} \left(\|\nabla f(x_k)\|^2 \right) + \mathbb{E} \left(\left\| \frac{1}{M} \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right) \right. \\
& \quad \left. - 2L^2 \left(M \sum_{j=k-T}^{k-1} \gamma_j^2 \sigma^2 + T \sum_{j=k-\tau_{k,\mu}}^{k-1} \gamma_j^2 \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2 \right) \right) \right] \\
& \quad + \frac{\gamma_k^2 L}{2} \left(M\sigma^2 + \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right) \right) \\
& \leq -\frac{M\gamma_k}{2} \mathbb{E} \|\nabla f(x_k)\|^2 + \left(\frac{\gamma_k^2 L}{2} - \frac{\gamma_k}{2M} \right) \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right) \\
& \quad + \left(\frac{\gamma_k^2 ML}{2} + L^2 M^2 \gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2 \right) \sigma^2 \\
& \quad + L^2 MT \gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2 \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2 \right). \tag{30}
\end{aligned}$$

Summarizing the inequality (30) from $k = 1$ to $k = K$, we have

$$\begin{aligned}
& \mathbb{E}(f(x_{K+1})) - f(x_1) \\
& \leq -\frac{M}{2} \sum_{k=1}^K \gamma_k \mathbb{E} \left(\|\nabla f(x_k)\|^2 \right) + \sum_{k=1}^K \left(\frac{\gamma_k^2 L}{2} - \frac{\gamma_k}{2M} \right) \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right) \\
& \quad + \sum_{k=1}^K \left(\frac{\gamma_k^2 ML}{2} + L^2 M^2 \gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2 \right) \sigma^2 \\
& \quad + L^2 MT \sum_{k=1}^K \left(\gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2 \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{j-\tau_{j,m}}) \right\|^2 \right) \right) \\
& = -\frac{M}{2} \sum_{k=1}^K \gamma_k \mathbb{E} \left(\|\nabla f(x_k)\|^2 \right) \\
& \quad + \sum_{k=1}^K \left(\gamma_k^2 \left(\frac{L}{2} + L^2 MT \sum_{\kappa=1}^T \gamma_{k+\kappa} \right) - \frac{\gamma_k}{2M} \right) \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(x_{k-\tau_{k,m}}) \right\|^2 \right) \\
& \quad + \sum_{k=1}^K \left(\frac{\gamma_k^2 ML}{2} + L^2 M^2 \gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2 \right) \sigma^2 \\
& \leq -\frac{M}{2} \sum_{k=1}^K \gamma_k \mathbb{E} \left(\|\nabla f(x_k)\|^2 \right) + \sum_{k=1}^K \left(\frac{\gamma_k^2 ML}{2} + L^2 M^2 \gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2 \right) \sigma^2
\end{aligned}$$

where the last inequality is due to (7). Note that x^* is the global optimization point. Thus we have

$$\frac{1}{\sum_{k=1}^K \gamma_k} \sum_{k=1}^K \gamma_k \mathbb{E}(\|\nabla f(x_k)\|^2) \leq \frac{2(f(x_1) - f(x^*)) + \sum_{k=1}^K (\gamma_k^2 ML + 2L^2 M^2 \gamma_k \sum_{j=k-T}^{k-1} \gamma_j^2) \sigma^2}{M \sum_{k=1}^K \gamma_k}.$$

It completes the proof. \square

Proofs to Corollary 2

Proof. From (9) and (10), we have

$$\gamma \leq \frac{1}{2ML(T+2)}. \quad (31)$$

It follows that

$$\frac{1}{2} \gamma L + L^2 M T^2 \gamma^2 \leq \frac{1}{4M(T+2)} + \frac{T^2}{4M(T+2)^2} \leq \frac{1}{2M'}$$

which implies that the condition (7) in Theorem 1 is satisfied globally. Then we can safely apply (8) in Theorem 1:

$$\begin{aligned} \frac{1}{K} \sum_{i=1}^K \mathbb{E}(\|\nabla f(x_i)\|^2) &\leq \frac{2(f(x_1) - f(x^*)) + K (\gamma^2 ML + 2L^2 M^2 T \gamma^3) \sigma^2}{MK\gamma} \\ &= \frac{2(f(x_1) - f(x^*))}{MK\gamma} + L\sigma^2\gamma + 2L^2 MT\sigma^2\gamma^2 \\ &\leq \frac{2(f(x_1) - f(x^*))}{MK\gamma} + 2L\sigma^2\gamma \\ &= 4\sqrt{\frac{(f(x_1) - f(x^*))L}{MK}}\sigma, \end{aligned}$$

where the second last inequality is due to (31) and the last equality uses (9). It completes the proof. \square

Proofs to Theorem 3

Proof. From the Lipschitzian gradient assumption (5), we have

$$f(x_{k+1}) \leq f(x_k) - \gamma \left\langle \nabla_{i_k} f(x_k), \sum_{m=1}^M (G(\hat{x}_{k,m}; \xi_{k,m}))_{i_k} \right\rangle + \frac{L_{\max} \gamma^2}{2} \left(\sum_{m=1}^M (G(\hat{x}_{k,m}; \xi_{k,m}))_{i_k} \right)^2. \quad (32)$$

Taking the expectation of i_k on both sides of (32), we have

$$\mathbb{E}_{i_k} f(x_{k+1}) \leq f(x_k) - \frac{\gamma}{n} \left\langle \nabla f(x_k), \sum_{m=1}^M G(\hat{x}_{k,m}; \xi_{k,m}) \right\rangle + \frac{L_{\max} \gamma^2}{2n} \left\| \sum_{m=1}^M G(\hat{x}_{k,m}; \xi_{k,m}) \right\|^2. \quad (33)$$

Taking the expectation of $\zeta_{k,*}$ on both sides of (33), we obtain

$$\begin{aligned}
& \mathbb{E}_{\zeta_{k,*}}(f(x_{k+1})) \\
& \leq f(x_k) - \frac{\gamma}{n} \mathbb{E}_{\zeta_{k,*}} \left[\left\langle \nabla f(x_k), \sum_{m=1}^M G(\hat{x}_{k,m}; \zeta_{k,m}) \right\rangle \right] + \frac{L_{\max} \gamma^2}{2n} \mathbb{E}_{\zeta_{k,*}} \left[\left\| \sum_{m=1}^M G(\hat{x}_{k,m}; \zeta_{k,m}) \right\|^2 \right] \\
& = f(x_k) - \frac{\gamma}{n} \left\langle \nabla f(x_k), \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\rangle + \frac{L_{\max} \gamma^2}{2n} \mathbb{E}_{\zeta_{k,*}} \left[\left\| \sum_{m=1}^M G(\hat{x}_{k,m}; \zeta_{k,m}) \right\|^2 \right] \\
& = f(x_k) - \frac{M\gamma}{2n} \left(\left\| \nabla f(x_k) \right\|^2 + \left\| \frac{1}{M} \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 - \underbrace{\left\| \nabla f(x_k) - \frac{1}{M} \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2}_{T_1} \right) \\
& \quad + \frac{L_{\max} \gamma^2}{2n} \underbrace{\mathbb{E}_{\zeta_{k,*}} \left[\left\| \sum_{m=1}^M G(\hat{x}_{k,m}; \zeta_{k,m}) \right\|^2 \right]}_{T_2} \tag{34}
\end{aligned}$$

where the last equation is deduced by the fact $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$. We next consider T_1 and T_2 respectively.

For T_2 , we have

$$\begin{aligned}
\mathbb{E}_{\zeta_{k,*}}(T_2) & = \mathbb{E}_{\zeta_{k,*}} \left[\left\| \sum_{m=1}^M G(\hat{x}_{k,m}; \zeta_{k,m}) \right\|^2 \right] \\
& = \mathbb{E}_{\zeta_{k,*}} \left[\left\| \sum_{m=1}^M (G(\hat{x}_{k,m}; \zeta_{k,m}) - \nabla f(\hat{x}_{k,m})) + \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right] \\
& = \mathbb{E}_{\zeta_{k,*}} \left[\left\| \sum_{m=1}^M (G(\hat{x}_{k,m}; \zeta_{k,m}) - \nabla f(\hat{x}_{k,m})) \right\|^2 \right. \\
& \quad \left. + \left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 + 2 \left\langle \sum_{m=1}^M (G(\hat{x}_{k,m}; \zeta_{k,m}) - \nabla f(\hat{x}_{k,m})), \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\rangle \right] \\
& = \mathbb{E}_{\zeta_{k,*}} \left[\left\| \sum_{m=1}^M (G(\hat{x}_{k,m}; \zeta_{k,m}) - \nabla f(\hat{x}_{k,m})) \right\|^2 + \left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right] \\
& = \mathbb{E}_{\zeta_{k,*}} \left[\sum_{m=1}^M \|G(\hat{x}_{k,m}; \zeta_{k,m}) - \nabla f(\hat{x}_{k,m})\|^2 \right. \\
& \quad \left. + 2 \sum_{1 \leq m < m' \leq M} \langle G(\hat{x}_{k,m}; \zeta_{k,m}) - \nabla f(\hat{x}_{k,m}), G(\hat{x}_{k,m'}; \zeta_{k,m'}) - \nabla f(\hat{x}_{k,m'}) \rangle \right. \\
& \quad \left. + \left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right]
\end{aligned}$$

$$\leq M\sigma^2 + \left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \quad (35)$$

where the fourth equality is due to

$$\begin{aligned} & \mathbb{E}_{\tilde{\zeta}_{k,*}} \left\langle \sum_{m=1}^M (G(\hat{x}_{k,m}; \tilde{\zeta}_{k,m}) - \nabla f(\hat{x}_{k,m})), \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\rangle \\ &= \left\langle \sum_{m=1}^M \mathbb{E}_{\tilde{\zeta}_{k,*}} (G(\hat{x}_{k,m}; \tilde{\zeta}_{k,m}) - \nabla f(\hat{x}_{k,m})), \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\rangle \\ &= 0 \end{aligned}$$

and the last inequality is due to the assumption (3) and

$$\begin{aligned} & \mathbb{E}_{\tilde{\zeta}_{k,*}} \sum_{1 \leq m < m' \leq M} \langle G(\hat{x}_{k,m}; \tilde{\zeta}_{k,m}) - \nabla f(\hat{x}_{k,m}), G(\hat{x}_{k,m'}; \tilde{\zeta}_{k,m'}) - \nabla f(\hat{x}_{k,m'}) \rangle \\ &= \mathbb{E}_{\tilde{\zeta}_{k,*}} \sum_{1 \leq m < m' \leq M} \mathbb{E}_{k,m'} \langle G(\hat{x}_{k,m}; \tilde{\zeta}_{k,m}) - \nabla f(\hat{x}_{k,m}), G(\hat{x}_{k,m'}; \tilde{\zeta}_{k,m'}) - \nabla f(\hat{x}_{k,m'}) \rangle \\ &= \mathbb{E}_{\tilde{\zeta}_{k,*}} \sum_{1 \leq m < m' \leq M} \langle G(\hat{x}_{k,m}; \tilde{\zeta}_{k,m}) - \nabla f(\hat{x}_{k,m}), \mathbb{E}_{k,m'} G(\hat{x}_{k,m'}; \tilde{\zeta}_{k,m'}) - \nabla f(\hat{x}_{k,m'}) \rangle \\ &= 0. \end{aligned} \quad (36)$$

As for T_1 , we have:

$$\begin{aligned} T_1 &= \left\| \nabla f(x_k) - \frac{1}{M} \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \\ &= \frac{1}{M^2} \left\| \sum_{m=1}^M (\nabla f(x_k) - \nabla f(\hat{x}_{k,m})) \right\|^2 \\ &\leq \frac{1}{M} \sum_{m=1}^M \|\nabla f(x_k) - \nabla f(\hat{x}_{k,m})\|^2 \\ &\leq \frac{L_T^2}{M} \sum_{m=1}^M \|x_k - \hat{x}_{k,m}\|^2 \\ &\leq L_T^2 \max_{k \in \{1, \dots, M\}} \|x_k - \hat{x}_{k,m}\|^2 \\ &= L_T^2 \|x_k - \hat{x}_{k,\mu}\|^2. \quad (\text{let } \mu := \arg \max_{m \in \{1, \dots, M\}} \|x_k - \hat{x}_{k,m}\|^2) \end{aligned}$$

It follows that

$$\begin{aligned} T_1 &\leq L_T^2 \|x_k - \hat{x}_{k,\mu}\|^2 \\ &= L_T^2 \left\| \sum_{j \in J(k,\mu)} (x_{j+1} - x_j) \right\|^2 \end{aligned}$$

$$\begin{aligned}
&= L_T^2 \gamma^2 \left\| \sum_{j \in J(k, \mu)} \sum_{m=1}^M (G(\hat{x}_{j,m}; \xi_{j,m}))_{i_j} e_{i_j} \right\|^2 \\
&= L_T^2 \gamma^2 \left\| \sum_{j \in J(k, \mu)} \sum_{m=1}^M \left[(G(\hat{x}_{j,m}; \xi_{j,m}))_{i_j} - \nabla_{i_j} f(\hat{x}_{j,m}) \right] e_{i_j} + \sum_{j \in J(k, \mu)} \sum_{m=1}^M \nabla_{i_j} f(\hat{x}_{j,m}) e_{i_j} \right\|^2 \\
&\leq 2L_T^2 \gamma^2 \left(\underbrace{\left\| \sum_{j \in J(k, \mu)} \sum_{m=1}^M \left[(G(\hat{x}_{j,m}; \xi_{j,m}))_{i_j} - \nabla_{i_j} f(\hat{x}_{j,m}) \right] e_{i_j} \right\|^2}_{T_3} + \underbrace{\left\| \sum_{j \in J(k, \mu)} \sum_{m=1}^M \nabla_{i_j} f(\hat{x}_{j,m}) e_{i_j} \right\|^2}_{T_4} \right) \quad (37)
\end{aligned}$$

where the last inequality uses the fact that $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any real vector a and b . Taking the expectation in terms of i_j and $\xi_{j,*}$ with all j 's in $J(k, \mu)$ for T_3 , we have

$$\begin{aligned}
&\mathbb{E}_{\xi_{j,*}, i_j, j \in J(k, \mu)}(T_3) \\
&= \mathbb{E}_{\xi_{j,*}, i_j, j \in J(k, \mu)} \left(\left\| \sum_{j \in J(k, \mu)} \sum_{m=1}^M \left[(G(\hat{x}_{j,m}; \xi_{j,m}))_{i_j} - \nabla_{i_j} f(\hat{x}_{j,m}) \right] e_{i_j} \right\|^2 \right) \\
&= \mathbb{E}_{\xi_{j,*}, i_j, j \in J(k, \mu)} \left(\sum_{j \in J(k, \mu)} \left\| \sum_{m=1}^M \left[(G(\hat{x}_{j,m}; \xi_{j,m}))_{i_j} - \nabla_{i_j} f(\hat{x}_{j,m}) \right] e_{i_j} \right\|^2 \right) \\
&\quad + \mathbb{E}_{\xi_{j,*}, i_j, j \in J(k, \mu)} \left(\sum_{j'' \neq j', j'', j' \in J(k, \mu)} \left\langle \sum_{m=1}^M \left[(G(\hat{x}_{j'',m}; \xi_{j'',m}))_{i_{j''}} - \nabla_{i_{j''}} f(\hat{x}_{j'',m}) \right] e_{i_{j''}}, \right. \right. \\
&\quad \left. \left. \sum_{m=1}^M \left[(G(\hat{x}_{j',m}; \xi_{j',m}))_{i_{j'}} - \nabla_{i_{j'}} f(\hat{x}_{j',m}) \right] e_{i_{j'}} \right\rangle \right) \\
&= \mathbb{E}_{\xi_{j,*}, i_j, j \in J(k, \mu)} \left(\sum_{j \in J(k, \mu)} \left\| \sum_{m=1}^M \left[(G(\hat{x}_{j,m}; \xi_{j,m}))_{i_j} - \nabla_{i_j} f(\hat{x}_{j,m}) \right] e_{i_j} \right\|^2 \right) \\
&= \frac{1}{n} \mathbb{E}_{\xi_{j,*}, j \in J(k, \mu)} \left(\sum_{j \in J(k, \mu)} \left\| \sum_{m=1}^M \left[G(\hat{x}_{j,m}; \xi_{j,m}) - \nabla f(\hat{x}_{j,m}) \right] \right\|^2 \right) \\
&= \frac{1}{n} \mathbb{E}_{\xi_{j,*}, j \in J(k, \mu)} \left(\sum_{j \in J(k, \mu)} \sum_{m=1}^M \left\| \left[G(\hat{x}_{j,m}; \xi_{j,m}) - \nabla f(\hat{x}_{j,m}) \right] \right\|^2 \right) \\
&\leq \frac{TM\sigma^2}{n}. \quad (38)
\end{aligned}$$

where the second last equality is due to (36) and the third equality is due to

$$\begin{aligned}
&\mathbb{E}_{\xi_{j,*}, i_j, j \in J(k, \mu)} \left(\sum_{j'' \neq j', j'', j' \in J(k, \mu)} \left\langle \sum_{m=1}^M \left[(G(\hat{x}_{j'',m}; \xi_{j'',m}))_{i_{j''}} - \nabla_{i_{j''}} f(\hat{x}_{j'',m}) \right] e_{i_{j''}}, \right. \right. \\
&\quad \left. \left. \sum_{m=1}^M \left[(G(\hat{x}_{j',m}; \xi_{j',m}))_{i_{j'}} - \nabla_{i_{j'}} f(\hat{x}_{j',m}) \right] e_{i_{j'}} \right\rangle \right)
\end{aligned}$$

$$\begin{aligned}
&= 2\mathbb{E}_{\xi_{j^*}, i_j, j \in J(k, \mu)} \left(\sum_{j'' > j', j'', j' \in J(k, \mu)} \left\langle \sum_{m=1}^M \left[(G(\hat{x}_{j'', m}; \xi_{j'', m}))_{i_{j''}} - \nabla_{i_{j''}} f(\hat{x}_{j'', m}) \right] e_{i_{j''}}, \right. \right. \\
&\quad \left. \left. \sum_{m=1}^M \left[(G(\hat{x}_{j', m}; \xi_{j', m}))_{i_{j'}} - \nabla_{i_{j'}} f(\hat{x}_{j', m}) \right] e_{i_{j'}} \right\rangle \right) \\
&= 2\mathbb{E}_{\xi_{j^*}, i_j, j \in J(k, \mu)} \left(\sum_{j'' > j', j'', j' \in J(k, \mu)} \left\langle \sum_{m=1}^M \left[(\mathbb{E}_{\xi_{j'', m}} G(\hat{x}_{j'', m}; \xi_{j'', m}))_{i_{j''}} - \nabla_{i_{j''}} f(\hat{x}_{j'', m}) \right] e_{i_{j''}}, \right. \right. \\
&\quad \left. \left. \sum_{m=1}^M \left[(G(\hat{x}_{j', m}; \xi_{j', m}))_{i_{j'}} - \nabla_{i_{j'}} f(\hat{x}_{j', m}) \right] e_{i_{j'}} \right\rangle \right) \\
&= 0.
\end{aligned}$$

Taking the expectation in terms of i_j with all j 's in $J(k, \mu)$ for T_4 , we have

$$\begin{aligned}
&\mathbb{E}_{i_j, j \in J(k, \mu)}(T_4) \\
&= \mathbb{E}_{i_j, j \in J(k, \mu)} \left(\left\| \sum_{j \in J(k, \mu)} \sum_{m=1}^M \nabla_{i_j} f(\hat{x}_{j, m}) e_{i_j} \right\|^2 \right) \\
&= \mathbb{E}_{i_j, j \in J(k, \mu)} \left[\sum_{j \in J(k, \mu)} \left\| \sum_{m=1}^M \nabla_{i_j} f(\hat{x}_{j, m}) e_{i_j} \right\|^2 \right. \\
&\quad \left. + 2 \sum_{j'' > j', j'', j' \in J(k, \mu)} \left\langle \sum_{m=1}^M \nabla_{i_{j''}} f(\hat{x}_{j'', m}) e_{i_{j''}}, \sum_{m=1}^M \nabla_{i_{j'}} f(\hat{x}_{j', m}) e_{i_{j'}} \right\rangle \right] \\
&= \mathbb{E}_{i_j, j \in J(k, \mu)} \left[\frac{1}{n} \sum_{j \in J(k, \mu)} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{j, m}) \right\|^2 \right. \\
&\quad \left. + 2 \sum_{j'' > j', j'', j' \in J(k, \mu)} \left\langle \sum_{m=1}^M \nabla_{i_{j''}} f(\hat{x}_{j'', m}) e_{i_{j''}}, \sum_{m=1}^M \nabla_{i_{j'}} f(\hat{x}_{j', m}) e_{i_{j'}} \right\rangle \right] \\
&\leq \mathbb{E}_{i_j, j \in J(k, \mu)} \left[\frac{1}{n} \sum_{j \in J(k, \mu)} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{j, m}) \right\|^2 \right. \\
&\quad \left. + \frac{1}{n} \sum_{j'' > j', j'', j' \in J(k, \mu)} \left(\frac{1}{\alpha} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{j'', m}) \right\|^2 + \frac{\alpha}{n} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{j', m}) \right\|^2 \right) \right] \quad (\text{Let } \alpha = \sqrt{n}) \\
&\leq \left(\frac{\sqrt{n} + T - 1}{n^{3/2}} \right) \sum_{j \in J(k, \mu)} \mathbb{E}_{i_j, j \in J(k, \mu)} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{j, m}) \right\|^2 \right), \tag{39}
\end{aligned}$$

where the second last inequality is due to that for any $j'' > j'$:

$$\mathbb{E}_{i_{j''}, i_{j'}} \left\langle \sum_{m=1}^M \nabla_{i_{j''}} f(\hat{x}_{j'', m}) e_{i_{j''}}, \sum_{m=1}^M \nabla_{i_{j'}} f(\hat{x}_{j', m}) e_{i_{j'}} \right\rangle$$

$$\begin{aligned}
&= \frac{1}{n} \mathbb{E}_{i_{j'}} \left\langle \sum_{m=1}^M \nabla f(\hat{x}_{j'',m}), \sum_{m=1}^M \nabla_{i_{j'}} f(\hat{x}_{j',m}) e_{i_{j'}} \right\rangle \\
&\leq \frac{1}{n} \mathbb{E}_{i_{j'}} \left(\frac{1}{2\alpha} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{j'',m}) \right\|^2 + \frac{\alpha}{2} \left\| \sum_{m=1}^M \nabla_{i_{j'}} f(\hat{x}_{j',m}) e_{i_{j'}} \right\|^2 \right) \\
&= \frac{1}{n} \mathbb{E}_{i_{j'}} \left(\frac{1}{2\alpha} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{j'',m}) \right\|^2 + \frac{\alpha}{2n} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{j',m}) \right\|^2 \right).
\end{aligned}$$

Take full expectation on both sides of (39), (38) and (37). Then substituting the upper bound of $\mathbb{E}(T_3)$ and $\mathbb{E}(T_4)$ into $\mathbb{E}(T_1)$, we have

$$\mathbb{E}(T_1) \leq 2L_T^2 \gamma^2 \left(\frac{TM\sigma^2}{n} + \left(\frac{\sqrt{n} + T - 1}{n^{3/2}} \right) \sum_{j \in J(k,\mu)} \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{j,m}) \right\|^2 \right) \right) \quad (40)$$

Take full expectation on both sides of (35) and (34). Substituting $\mathbb{E}(T_1)$ and $\mathbb{E}(T_2)$ into (34), we have

$$\begin{aligned}
\mathbb{E}(f(x_{k+1})) &\leq \mathbb{E}(f(x_k)) - \frac{M\gamma}{2n} \left[\mathbb{E}(\|\nabla f(x_k)\|^2) + \mathbb{E} \left(\left\| \frac{1}{M} \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \right. \\
&\quad \left. - 2L_T^2 \gamma^2 \left(\frac{TM\sigma^2}{n} + \left(\frac{\sqrt{n} + T - 1}{n^{3/2}} \right) \sum_{j \in J(k,\mu)} \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{j,m}) \right\|^2 \right) \right) \right] \\
&\quad + \frac{L_{\max} \gamma^2}{2n} \left(M\sigma^2 + \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \right) \\
&= \mathbb{E}(f(x_k)) - \frac{M\gamma}{2n} \mathbb{E}(\|\nabla f(x_k)\|^2) - \frac{\gamma}{2Mn} \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \\
&\quad + \frac{M\gamma^3}{n} L_T^2 \left(\left(\frac{\sqrt{n} + T - 1}{n^{3/2}} \right) \sum_{j \in J(k,\mu)} \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{j,m}) \right\|^2 \right) \right) \\
&\quad + \frac{L_{\max} \gamma^2}{2n} \left(\mathbb{E} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) + \frac{L_T^2 TM^2 \gamma^3}{n^2} \sigma^2 + \frac{L_{\max} M \gamma^2}{2n} \sigma^2. \quad (41)
\end{aligned}$$

Summarizing (41) from $k = 1$ to K , we have

$$\begin{aligned}
\mathbb{E}(f(x_{K+1})) &\leq f(x_1) - \sum_{k=1}^K \frac{M\gamma}{2n} \mathbb{E}(\|\nabla f(x_k)\|^2) - \frac{\gamma}{2Mn} \sum_{k=1}^K \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \\
&\quad + \frac{M\gamma^3}{n} L_T^2 \left(\left(\frac{\sqrt{n} + T - 1}{n^{3/2}} \right) \sum_{k=1}^K \sum_{j \in J(k,\mu)} \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{j,m}) \right\|^2 \right) \right)
\end{aligned}$$

$$\begin{aligned}
& + \frac{L_{\max}\gamma^2}{2n} \sum_{k=1}^K \left(\mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \right) + \frac{KL_T^2 TM^2 \gamma^3}{n^2} \sigma^2 + \frac{KL_{\max} M \gamma^2}{2n} \sigma^2 \\
& \leq f(x_1) - \sum_{k=1}^K \frac{M\gamma}{2n} \mathbb{E} \left(\|\nabla f(x_k)\|^2 \right) - \frac{\gamma}{2Mn} \sum_{k=1}^K \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \\
& \quad + \frac{MT\gamma^3}{n} L_T^2 \left(\left(\frac{\sqrt{n} + T - 1}{n^{3/2}} \right) \sum_{k=1}^K \mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \right) \\
& \quad + \frac{L_{\max}\gamma^2}{2n} \sum_{k=1}^K \left(\mathbb{E} \left(\left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \right) + \frac{KL_T^2 TM^2 \gamma^3}{n^2} \sigma^2 + \frac{KL_{\max} M \gamma^2}{2n} \sigma^2 \\
& \leq f(x_1) - \sum_{k=1}^K \frac{M\gamma}{2n} \mathbb{E} \left(\|\nabla f(x_k)\|^2 \right) + \frac{KL_T^2 TM^2 \gamma^3}{n^2} \sigma^2 + \frac{KL_{\max} M \gamma^2}{2n} \sigma^2 \\
& \quad + \left(\frac{MTL_T^2 (\sqrt{n} + T - 1) \gamma^3}{n^{5/2}} - \frac{\gamma}{2Mn} + \frac{L_{\max}\gamma^2}{2n} \right) \sum_{k=1}^K \left(\mathbb{E} \left\| \sum_{m=1}^M \nabla f(\hat{x}_{k,m}) \right\|^2 \right) \\
& \leq f(x_1) - \sum_{k=1}^K \frac{M\gamma}{2n} \mathbb{E} \left(\|\nabla f(x_k)\|^2 \right) + \frac{KL_T^2 TM^2 \gamma^3}{n^2} \sigma^2 + \frac{KL_{\max} M \gamma^2}{2n} \sigma^2
\end{aligned}$$

where the last inequality comes from (15). Together with $\mathbb{E}(f(x_{k+1})) \geq f(x^*)$, we have

$$\frac{1}{K} \sum_{t=1}^K \mathbb{E} \|\nabla f(x_t)\|^2 \leq \frac{2n}{KM\gamma} (f(x_1) - f(x^*)) + \frac{2L_T^2 TM\gamma^2}{n} \sigma^2 + L_{\max}\gamma\sigma^2. \quad (42)$$

It completes the proof. \square

Proofs to Corollary 4

Proof. From the definition of the steplength (17) and the lower bound of K (18), we have

$$\begin{aligned}
\gamma & = \sqrt{\frac{2n(f(x_1) - f(x^*))}{L_T M \sigma^2}} \frac{1}{\sqrt{K}} \\
& \leq \sqrt{\frac{2n(f(x_1) - f(x^*))}{L_T M \sigma^2}} \sqrt{\frac{\sqrt{n}\sigma^2}{16(f(x_1) - f(x^*))L_T M (n^{3/2} + 4T^2)}} \\
& = \frac{1}{2} \sqrt{\frac{1}{2n^{3/2} + 8T^2}} \frac{n^{3/4}}{L_T M}
\end{aligned} \quad (43)$$

which gives an upper bound for γ . We can further relax this upper bound by

$$\gamma \leq \frac{1}{2} \sqrt{\frac{1}{2n^{3/2} + 8T^2}} \frac{n^{3/4}}{L_T M} \leq \frac{1}{2} \sqrt{\frac{1}{8T^2}} \frac{n^{3/4}}{L_T M} < \frac{2n}{L_T M T}. \quad (44)$$

Next we will show that the steplength satisfies the condition in (15):

$$\begin{aligned}
\text{LHS} &= \frac{2M^2TL_T^2(\sqrt{n}+T-1)\gamma^2}{n^{3/2}} + 2ML_{\max}\gamma \\
&= \frac{2M^2TL_T^2(\sqrt{n}+T-1)}{n^{3/2}} \frac{1}{4} \frac{1}{2n^{3/2}+8T^2} \frac{n^{3/2}}{L_T^2M^2} + \frac{L_{\max}M}{2} \sqrt{\frac{1}{2n^{3/2}+8T^2}} \frac{n^{3/4}}{L_TM} \\
&\leq \frac{1}{2} \frac{T(\sqrt{n}+T)}{n^{3/2}+4T^2} + \frac{1}{2} \\
&\leq \frac{1}{2} \frac{T\sqrt{n}+T^2}{n^{3/2}+4T^2} + \frac{1}{2} \\
&\leq \frac{1}{2} \frac{3T^2}{n^{3/2}+4T^2} + \frac{1}{2} \frac{2n}{n^{3/2}+4T^2} + \frac{1}{2} \\
&\leq \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{2}{5} + \frac{1}{2} \\
&\leq 1 = \text{RHS},
\end{aligned}$$

where the first inequality uses the fact $L_{\max} \leq L_T$ and the first inequality uses the upper bound of γ in (43). The third last inequality comes from $\sqrt{n}T \leq 2n + 2T^2$. It means that the condition (15) in Theorem 3 is satisfied globally. Then we can safely apply (16) in Theorem 3:

$$\begin{aligned}
\frac{1}{K} \sum_{t=1}^K \mathbb{E} \|\nabla f(x_t)\|^2 &\leq \frac{2n}{KM\gamma} (f(x_1) - f(x^*)) + \frac{2L_T^2TM\gamma^2}{n} \sigma^2 + L_{\max}\gamma\sigma^2 \\
&\leq \frac{2n}{KM\gamma} (f(x_1) - f(x^*)) + 5L_T\gamma\sigma^2 \\
&= \frac{6\sqrt{2}(f(x_1) - f(x^*))L_Tn\sigma}{\sqrt{KM}}
\end{aligned}$$

where the second inequality is due to the upper bound of γ in (44) and the last equality is acquired by substituting γ by its definition in (17). It completes the proof. \square