STEM: A Stochastic Two-Sided Momentum Algorithm Achieving Near-Optimal Sample and Communication Complexities for Federated Learning

Prashant Khanduri
University of Minnesota
khand095@umn.edu

Pranay Sharma
Carnegie Mellon University
pranaysh@andrew.cmu.edu

Haibo Yang
The Ohio State University
yang.5952@buckeyemail.osu.edu

Mingyi Hong[*]
University of Minnesota
mhong@umn.edu

Jia Liu
The Ohio State University
liu@ece.osu.edu

Ketan Rajawat
Indian Institute of Technology Kanpur
ketan@iitk.ac.in

Pramod K. Varshney
Syracuse University
varshney@syr.edu

Abstract

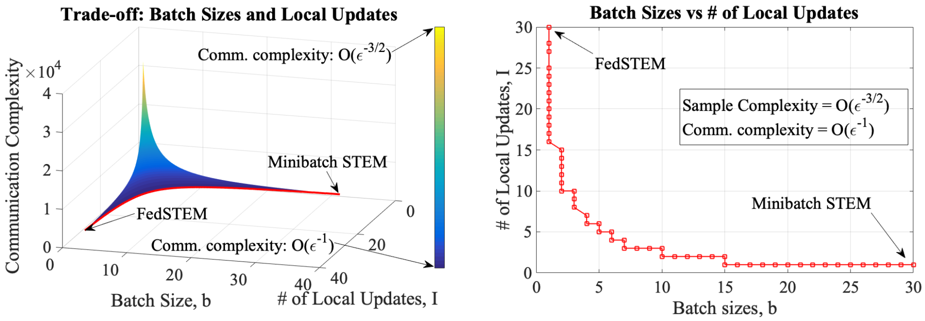
Federated Learning (FL) refers to the paradigm where multiple worker nodes (WNs) build a joint model by using local data. Despite extensive research, for a generic non-convex FL problem, it is not clear, how to choose the WNs' and the server's update directions, the minibatch sizes, and the number of local updates, so that the WNs use the minimum number of samples and communication rounds to achieve the desired solution. This work addresses the above question and considers a class of stochastic algorithms where the WNs perform a few local updates before communication. We show that when both the WN's and the server's directions are chosen based on certain stochastic momentum estimator, the algorithm requires $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ samples and $\tilde{\mathcal{O}}(\epsilon^{-1})$ communication rounds to compute an ϵ-stationary solution. To the best of our knowledge, this is the first FL algorithm that achieves such *near-optimal* sample and communication complexities simultaneously. Further, we show that there is a trade-off curve between the number of local updates and the minibatch sizes, on which the above sample and communication complexities can be maintained. Finally, we show that for the classical FedAvg (a.k.a. Local SGD, which is a momentum-less special case of the STEM), a similar trade-off curve exists, albeit with worse sample and communication complexities. Our insights on this trade-off provides guidelines for choosing the four important design elements for FL algorithms, the number of local updates, WNs' and server's update directions, and minibatch sizes to achieve the best performance.

1 Introduction

In Federated Learning (FL), multiple worker nodes (WNs) collaborate with the goal of learning a joint model, by only using local data. Therefore it has become popular for machine learning problems where datasets are massively distributed [1]. In FL, the data is often collected at or off-loaded to multiple WNs which in collaboration with a server node (SN) jointly aim to learn a centralized model

[*] Corresponding Author: Mingyi Hong.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).



(a) Communication complexity.

(b) Minibatch sizes vs Local Updates.

Figure 1: The 3D surface in (a) plots the communication complexity of the proposed STEM for different minibatch sizes and number of local updates. The surface is generated such that each point represents STEM with a particular choice of (b, I), so that it requires $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ samples to achieve ϵ-stationarity. Plot (b) shows the optimal trade off between the minibatch sizes and the number of local updates at each WN (i.e., achieving the lowest communication and sample complexities). Both plots are generated for an accuracy of $\epsilon = 10^{-3}$ and all the constants dependent on system parameters (variance of stochastic gradients, heterogeneity parameter, optimality gap, Lipschitz constants, etc.) are assumed to be 1. Fed STEM is a special case of STEM where $\mathcal{O}(1)$ minibatch is used; Minibatch STEM is a special case of STEM where $\mathcal{O}(1)$ local updates are used.

[2, 3]. The local WNs share the computational load and since the data is local to each WN, FL also provides some level of data privacy [4]. A classical distributed optimization problem that K WNs aim to solve:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{K} \sum_{k=1}^{K} f^{(k)}(x) := \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\xi^{(k)} \sim \mathcal{D}^{(k)}} [f^{(k)}(x; \xi^{(k)})] \right\}. \tag{1}$$

where $f^{(k)} : \mathbb{R}^d \to \mathbb{R}$ denotes the smooth (possibly non-convex) objective function and $\xi^{(k)} \sim \mathcal{D}^{(k)}$ represents the sample/s drawn from distribution $\mathcal{D}^{(k)}$ at the k^{th} WN with $k \in [K]$. When the distributions $\mathcal{D}^{(k)}$ are different across the WNs, it is referred to as the *heterogeneous* data setting.

The optimization performance of non-convex FL algorithms is typically measured by the total number of samples accessed (cf. Definition 2.2) and the total rounds of communication (cf. Definition 2.3) required by each WN to achieve an ϵ-stationary solution (cf. Definition 2.1). To minimize the sample and the communication complexities, FL algorithms rely on the following *four* key design elements: (i) the WNs' local model update directions, (ii) Minibatch size to compute each local direction, (iii) the number of local updates before WNs share their parameters, and (iv) the SN's update direction. How to find effective FL algorithms by (optimally) designing these parameters has received significant research interest recently.

Contributions. The main contributions of this work are listed below:

1) We propose the Stochastic Two-Sided Momentum (STEM) algorithm, that utilizes certain momentum-assisted stochastic gradient directions for *both* the WNs and SN updates. We show that there exists an *optimal* trade off between the minibatch sizes and number of local updates, such that on the trade-off curve STEM requires $\tilde{\mathcal{O}}(\epsilon^{-3/2})^2$ samples and $\tilde{\mathcal{O}}(\epsilon^{-1})$ communication rounds to reach an ϵ-stationary solution; see Figure 1 for an illustration. These complexity results are the best achievable for first-order stochastic FL algorithms (under certain assumptions, cf. Assumption 1); see [5–8] and [9, 10], as well as Remark 1 of this paper for discussions regarding optimality. To the best of our knowledge, STEM is the first algorithm which – (i) *simultaneously* achieves the optimal sample and communication complexities for FL and (ii) can optimally trade off the minibatch sizes and the number of local updates.

2) A momentum-less special case of our STEM result further reveals some interesting insights of the classical FedAvg algorithm (a.k.a. the Local SGD) [11–13]. Specifically, we show that for FedAvg, there also exists a trade-off between the minibatch sizes and the number of local updates, such that it requires $\mathcal{O}(\epsilon^{-2})$ samples and $\mathcal{O}(\epsilon^{-3/2})$ communication rounds to achieve an ϵ-stationary solution.

[2] The notation $\tilde{\mathcal{O}}(\cdot)$ hides the logarithmic factors.

Algorithm	Work	Sample	Comm.	Minibatch (b)	Local Updates (I) /round
FedAvg$^\diamond$	[12] /[14]	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3/2})$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1/2})$
	[15]/[16]		$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
	this work		$\mathcal{O}(\epsilon^{-3/2})$	$\mathcal{O}\big(\epsilon^{-\frac{2(1-\nu)}{(4-\nu)}}\big)$	$\mathcal{O}\big(\epsilon^{-\frac{3\nu}{2(4-\nu)}}\big)$
SCAFFOLD*	[15]	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
FedPD/FedProx‡	[9]/ [10]	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1})$
MIME†/FedGLOMO	[17]/[18]	$\mathcal{O}(\epsilon^{-3/2})$	$\mathcal{O}(\epsilon^{-3/2})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
STEM$^\diamond$	this work	$\tilde{\mathcal{O}}(\epsilon^{-3/2})$	$\tilde{\mathcal{O}}(\epsilon^{-1})$	$\mathcal{O}\big(\epsilon^{-\frac{3(1-\nu)}{2(3-\nu)}}\big)$	$\mathcal{O}\big(\epsilon^{-\frac{\nu}{(3-\nu)}}\big)$
Fed STEM				$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1/2})$
Minibatch STEM*				$\mathcal{O}(\epsilon^{-1/2})$	$\mathcal{O}(1)$

Table 1: Comparison of FedAvg and STEM with different FL algorithms for various choices of the minibatch sizes (b) and the number of per node local updates between two rounds of communication (I).
$^\diamond\nu \in [0, 1]$ trades off b and I; $\nu = 1$ (resp. $\nu = 0$) uses multiple (resp. $\mathcal{O}(1)$) local updates and $\mathcal{O}(1)$ (resp. multiple) samples. Fed STEM and Minibatch STEM are two variants of the proposed STEM.
‡The data heterogeneity assumption is weaker than Assumption 2 (please see [9] for details).
†Requires bounded Hessian dissimilarity to model data heterogeneity across WNs.
*Guarantees for Minibatch STEM with $I = 1$ and SCAFFOLD are independent of the data heterogeneity.

Collectively, our insights on the trade-offs provide practical guidelines for choosing different design elements for FL algorithms.

Related Works. FL algorithms were first proposed in the form of FedAvg [11], where the local update directions at each WN were chosen to be the SGD updates. Earlier works analyzed these algorithms in the homogeneous data setting [19–25], while many recent studies have focused on designing new algorithms to deal with heterogeneous data settings, as well as problems where the local loss functions are non-convex [9, 10, 12–16, 18, 26–32]. In [12], the authors showed that Parallel Restarted SGD (Local SGD or FedAvg [11]) achieves linear speed up while requiring $\mathcal{O}(\epsilon^{-2})$ samples and $\mathcal{O}(\epsilon^{-3/2})$ rounds of communication to reach an ϵ-stationary solution. In [14], a Momentum SGD was proposed, which achieved the same sample and communication complexities as Parallel Restarted SGD [12], without requiring that the second moments of the gradients be bounded. Further, it was shown that under the homogeneous data setting, the communication complexity can be improved to $\mathcal{O}(\epsilon^{-1})$ while maintaining the same sample complexity. The works in [15, 16] conducted tighter analysis for FedAvg with partial WN participation with $\mathcal{O}(1)$ local updates and batch sizes. Their analysis showed that FedAvg's sample and communication complexities are both $\mathcal{O}(\epsilon^{-2})$. Additionally, SCAFFOLD was proposed in [15], which utilized variance reduction based local update directions [33] to achieve the same sample and communication complexities as FedAvg. Similarly, VRL-SGD proposed in [29] also utilized variance reduction and showed improved communication complexity of $\mathcal{O}(\epsilon^{-1})$, while requiring the same computations as FedAvg. Importantly, both SCAFFOLD and VRL-SGD's guarantees were independent of the data heterogeneity. The FedProx proposed in [10] used a penalty based method to improve the communication complexity of FedAvg (i.e., the Parallel Restarted and Momentum SGD [14, 12]) to $\mathcal{O}(\epsilon^{-1})$. FedProx used a gradient similarity assumption to model data heterogeneity which can be stringent for many practical applications. This assumption was relaxed by FedPD proposed in [9].

Recently, the works [17, 18] proposed to utilize hybrid momentum gradient estimators [7, 8]. The MIME algorithm [17] matched the optimal sample complexity (under certain smoothness assumptions) of $\mathcal{O}(\epsilon^{-3/2})$ of the centralized non-convex stochastic optimization algorithms [5–8]. Similarly, Fed-GLOMO [18] achieved the same sample complexity while employing compression to further reduce communication. Both MIME and Fed-GLOMO required $\mathcal{O}(\epsilon^{-3/2})$ communication rounds to achieve an ϵ-stationary solution. Please see Table 1 for a summary of the above discussion.

The comparison of Local SGD (FedAvg) to Minibatch SGD for convex and strongly convex problems with homogeneous data setting was first conducted in [19] and later extended to heterogeneous setting in [13]. It was shown that Minibatch SGD almost always dominates the Local SGD. In contrast, it was shown in [24] that Local SGD dominates Minibatch SGD in terms of generalization performance. Although existing FL results are rich, but they are somehow ad hoc and there is a lack of principled understanding of the algorithms. We note that the proposed STEM algorithmic framework provides a theoretical framework that unifies all existing FL results on sample and communication complexities.

Notations. The expected value of a random variable X is denoted by $\mathbb{E}[X]$ and its expectation conditioned on an Event A is denoted as $\mathbb{E}[X|\text{Event } A]$. We denote by $\mathbb{R}$ (and $\mathbb{R}^d$) the real line (and the d-dimensional Euclidean space). The set of natural numbers is denoted by $\mathbb{N}$. Given a positive integer $K \in \mathbb{N}$, we denote $[K] \triangleq \{1, 2, \ldots, K\}$. Notation $\|\cdot\|$ denotes the ℓ_2-norm and $\langle \cdot, \cdot \rangle$ the Euclidean inner product. For a discrete set $\mathcal{B}$, $|\mathcal{B}|$ denotes the cardinality of the set. Uniform distribution over a discrete set $\{1, \ldots, T\}$ is denoted as $\mathcal{U}\{1, \ldots, T\}$.

2 Preliminaries

Before we proceed to the algorithms, we make the following assumptions about problem (1).

Assumption 1 (Sample Gradient Lipschitz Smoothness)**.** The stochastic functions $f^{(k)}(\cdot, \xi^{(k)})$ with $\xi^{(k)} \sim \mathcal{D}^{(k)}$ for all $k \in [K]$, satisfy the mean squared smoothness property, i.e, we have

$$\mathbb{E}\|\nabla f^{(k)}(x; \xi^{(k)}) - \nabla f^{(k)}(y; \xi^{(k)})\|^2 \leq L^2 \|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^d.$$

Assumption 2 (Unbiased gradient and Variance Bounds)**.** (i) Unbiased Gradient. The stochastic gradients computed at each WN are unbiased

$$\mathbb{E}[\nabla f^{(k)}(x; \xi^{(k)})] = \nabla f^{(k)}(x), \ \forall\, \xi^{(k)} \sim \mathcal{D}^{(k)}, \ \forall\, k \in [K].$$

(ii) Intra- and inter- node Variance Bound. The following bounds hold:

$$\mathbb{E}\|\nabla f^{(k)}(x; \xi^{(k)}) - \nabla f^{(k)}(x)\|^2 \leq \sigma^2, \|\nabla f^{(k)}(x) - \nabla f^{(\ell)}(x)\|^2 \leq \zeta^2, \ \forall\, \xi^{(k)} \sim \mathcal{D}^{(k)}, \forall k, \ell \in [K].$$

Note that Assumption 1 is stronger than directly assuming $f^{(k)}$'s are Lipschitz smooth (which we will refer to as the *averaged* gradient Lipschitz smooth condition), but it is still a rather standard assumption in SGD analysis. For example it has been used in analyzing centralized SGD algorithms such as SPIDER [5], SNVRG [6], STORM [7] (and many others) as well as in FL algorithms such as MIME [17] and Fed-GLOMO [18]. The second relation in Assumption 2-(ii) quantifies the data heterogeneity, and we call $\zeta > 0$ as the *heterogeneity parameter*. This is a typical assumption required to evaluate the performance of FL algorithms. If data distributions across individual WNs are identical, i.e., $\mathcal{D}^{(k)} = \mathcal{D}^{(\ell)}$ for all $k, \ell \in [K]$ then we have $\zeta = 0$.

Next, we define the ϵ-stationary solution for non-convex optimization problems, as well as quantify the computation and communication complexities to achieve an ϵ-stationary point.

Definition 2.1 (ϵ-Stationary Point)**.** A point x is called ϵ-stationary if $\|\nabla f(x)\|^2 \leq \epsilon$. Moreover, a stochastic algorithm is said to achieve an ϵ-stationary point in t iterations if $\mathbb{E}[\|\nabla f(x_t)\|^2] \leq \epsilon$, where the expectation is over the stochasticity of the algorithm until time instant t.

Definition 2.2 (Sample complexity)**.** We assume an Incremental First-order Oracle (IFO) framework [34], where, given a sample $\xi^{(k)} \sim \mathcal{D}^{(k)}$ at the k^{th} node and iterate x, the oracle returns $(f^{(k)}(x; \xi^{(k)}), \nabla f^{(k)}(x; \xi^{(k)}))$. Each access to the oracle is counted as a single IFO operation. We measure the sample (and computational) complexity in terms of the total number of calls to the IFO by all WNs to achieve an ϵ-stationary point given in Definition 2.1.

Definition 2.3 (Communication complexity)**.** We define a communication round as a one back-and-forth sharing of parameters between the WNs and the SN. Then the communication complexity is defined to be the total number of communication rounds between any WN and the SN required to achieve an ϵ-stationary point given in Definition 2.1.

3 The STEM algorithm and the trade-off analysis

In this section, we discuss the proposed algorithm and present the main results. The key in the algorithm design is to carefully balance *all the four* design elements mentioned in Sec. 1, so that sufficient and useful progress can be made between two rounds of communication.

Let us discuss the key steps of STEM, listed in Algorithm 1. In Step 10, each node locally updates its model parameters using the local direction d_t^k, computed by using b stochastic gradients at two

Algorithm 1 The Stochastic Two-Sided Momemtum (STEM) Algorithm

```
**Input**: Parameters: $c > 0$, the number of local updates $I$, batch size $b$, stepsizes $\{\eta_t\}$.
**Initialize**: Iterate $x_1^{(k)} = \bar{x}_1 = \frac{1}{K}\sum_{k=1}^{K} x_1^{(k)}$, descent direction $d_1^{(k)} = \bar{d}_1 = \frac{1}{K}\sum_{k=1}^{K} d_1^{(k)}$
    with $d_1^{(k)} = \frac{1}{B}\sum_{\xi_1^{(k)} \in \mathcal{B}_1^{(k)}} \nabla f^{(k)}(x_1^{(k)}; \xi_1^{(k)})$ and $|\mathcal{B}_1^{(k)}| = B$ for $k \in [K]$.
Perform: $x_2^{(k)} = x_1^k - \eta_1 d_1^{(k)}, \; \forall\, k \in [K]$
**for** $t = 1$ to $T$ **do**
    **for** $k = 1$ to $K$ **do**                                              #at the WN
        $d_{t+1}^{(k)} = \frac{1}{b}\sum_{\xi_{t+1}^{(k)} \in \mathcal{B}_{t+1}^{(k)}} \nabla f^{(k)}(x_{t+1}^{(k)}; \xi_{t+1}^{(k)}) + (1 - a_{t+1})\Big(d_t^{(k)} - \frac{1}{b}\sum_{\xi_{t+1}^{(k)} \in \mathcal{B}_{t+1}^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_{t+1}^{(k)})\Big)$
    where we choose $|\mathcal{B}_{t+1}^{(k)}| = b$, and $a_{t+1} = c \cdot \eta_t^2$;
        **if** $t$ mod $I = 0$ **then**                                        #at the SN
            $d_{t+1}^{(k)} = \bar{d}_{t+1} := \frac{1}{K}\sum_{k=1}^{K} d_{t+1}^{(k)}$
            $x_{t+2}^{(k)} := \bar{x}_{t+1} - \eta_{t+1}\bar{d}_{t+1} = \frac{1}{K}\sum_{k=1}^{K} x_{t+1}^{(k)} - \eta_{t+1}\bar{d}_{t+1}$    #server-side momentum
        **else** $x_{t+2}^{(k)} = x_{t+1}^{(k)} - \eta_{t+1} d_{t+1}^{(k)}$                    #worker-side momentum
        **end if**
    **end for**
**end for**
**Return:** $\bar{x}_a$ where $a \sim \mathcal{U}\{1, ..., T\}$.
```

consecutive iterates $x_{t+1}^{(k)}$ and $x_t^{(k)}$. After every I local steps, the WNs share their current local models $\{x_{t+1}^{(k)}\}_{k=1}^{K}$ and directions $\{d_{t+1}^{(k)}\}_{k=1}^{K}$ with the SN. The SN aggregates these quantities, and performs a server-side momentum step, before returning $\bar{x}_{t+1}$ and $\bar{d}_{t+1}$ to all the WNs. Because both the WNs and the SN perform momentum based updates, we call the algorithm a stochastic *two-sided* momentum algorithm. The key parameters are: b the minibatch size, I the local update steps between two communication rounds, η_t the stepsizes, and a_t the momentum parameters.

One key technical innovation of our algorithm design is to identify the most suitable way to incorporate momentum based directions in FL algorithms. Although the momentum-based gradient estimator itself is not new and has been used in the literature before (see e.g., in [7, 8] and [17, 18] to improve the sample complexities of centralized and decentralized stochastic optimization problems, respectively), it is by no means clear if and how it can contribute to improve the communication complexity of FL algorithms. We show that in the FL setting, the local directions together with the local models have to be aggregated by the SN so to avoid being influenced too much by the local data. More importantly, besides the WNs, the SN also needs to perform updates using the (aggregated) momentum directions. Finally, such *two-sided* momentum updates have to be done carefully with the correct choice of minibatch size b, and the number of local updates I. Overall, it is the judicious choice of all these design elements that results in the optimal sample and communication complexities.

Next, we present the convergence guarantees of the STEM algorithm.

3.1 Main results: convergence guarantees for STEM

In this section, we analyze the performance of STEM. We first present our main result, and then provide discussions about a few parameter choices. In the next subsection, we discuss a special case of STEM related to the classical FedAvg and minibatch SGD algorithms.

Theorem 3.1. *Under the Assumptions 1 and 2, suppose the stepsize sequence is chosen as:*

$$\eta_t = \frac{\bar{\kappa}}{(w_t + \sigma^2 t)^{1/3}}, \tag{2}$$

where we define :

$$\bar{\kappa} = \frac{(bK)^{2/3}\sigma^{2/3}}{L}, \quad w_t = \max\left\{2\sigma^2, 4096L^3I^3\bar{\kappa}^3 - \sigma^2 t, \frac{c^3\bar{\kappa}^3}{4096L^3I^3}\right\}.$$

Further, let us set $c = \frac{64L^2}{bK} + \frac{\sigma^2}{24\bar{\kappa}^3 LI} = L^2\left(\frac{64}{bK} + \frac{1}{24(bK)^2 I}\right)$*, and set the initial batch size as* $B = bI$*; set the local updates* I *and minibatch size* b *as follows:*

$$I = \mathcal{O}\big((T/K^2)^{\nu/3}\big), \quad b = \mathcal{O}\big((T/K^2)^{1/2-\nu/2}\big) \tag{3}$$

where ν *satisfies* $\nu \in [0,1]$*. Then for* STEM *the following holds:*

(i) For $\bar{x}_a$ *chosen according to Algorithm 1, we have:*

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{K^{2\nu/3}T^{1-\nu/3}}\bigg) + \tilde{\mathcal{O}}\bigg(\frac{\sigma^2}{K^{2\nu/3}T^{1-\nu/3}}\bigg) + \tilde{\mathcal{O}}\bigg(\frac{\zeta^2}{K^{2\nu/3}T^{1-\nu/3}}\bigg). \tag{4}$$

(ii) For any $\nu \in [0,1]$*, we have*
Sample Complexity: *The sample complexity of* STEM *is* $\tilde{\mathcal{O}}(\epsilon^{-3/2})$*. This implies that each WN requires at most* $\tilde{\mathcal{O}}(K^{-1}\epsilon^{-3/2})$ *gradient computations, thereby achieving linear speedup with the number of WNs present in the network.*
Communication Complexity: *The communication complexity of* STEM *is* $\tilde{\mathcal{O}}(\epsilon^{-1})$*.*

The proof of this result is relegated to the Supplemental Material. A few remarks are in order.

Remark 1 (Near-Optimal sample and communication complexities)**.** Theorem 3.1 suggests that when I and b are selected appropriately, then STEM achieves $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ and $\tilde{\mathcal{O}}(\epsilon^{-1})$ sample and communication complexities. Taking them separately, these complexity bounds are the best achievable by the existing FL algorithms (upto logarithmic factors regardless of sample or batch Lipschitz smooth assumption) [35]; see Table 1. We note that the $\mathcal{O}(\epsilon^{-3/2})$ complexity is the best possible that can be achieved by centralized SGD with the sample Lipschitz gradient assumption; see [5]. On the other hand, the $\mathcal{O}(\epsilon^{-1})$ complexity bound is also *likely* to be the optimal, since in [9] the authors showed that even when the local steps use a class of (deterministic) first-order algorithms, $\mathcal{O}(\epsilon^{-1})$ is the best achievable communication complexity. The only difference is that [9] does not explicitly assume the inter-node variance bound (i.e., the second relation in Assumption 2-(ii)). We leave the precise characterization of the communication lower bound with inter-node variance as future work. □

Remark 2 (Large Batch Sizes and/or Local Updates)**.** At first glance, it may seem that the requirement of STEM to compute large mini-batches and/or local updates (cf. Table 1) to achieve this (near) optimal performance is a drawback, however, we note that it is in fact an advantage of STEM that it allows the WNs to perform larger number of local updates (or compute large minibatches) without communicating often. This follows from the fact that irrespective of the number of local updates (or batch sizes) STEM achieves near-optimal communication complexity while attaining optimal *overall* sample complexity. Moreover, note that even with $b = I = \mathcal{O}(1)$ (i.e., b and I are chosen as constants), STEM achieves the same (optimal) sample and communication complexities as achieved by FedGLOMO [18] and MIME [17]. We further note that to the best of our knowledge the algorithms that achieve the communication complexity of $\mathcal{O}(\epsilon^{-1})$ either require the number of local updates or the batch-sizes that depend on the solution accuracy ϵ. For example, FedProx [10], FedPD [9], and FedDyn [36] rely on solving the "local problems" to achieve an ϵ-accuracy, which implies that the number of local updates (or the batch sizes) implicitly depends on the desired solution accuracy ϵ, as is the case for STEM. Similarly, as shown in [12] and [14] the communication complexity of FedAvg and its momentum version can be improved from $\mathcal{O}(\epsilon^{-2})$ to $\mathcal{O}(\epsilon^{-3/2})$ when the number of local updates (or batch size) is chosen as $\mathcal{O}(\epsilon^{-1/2})$ (cf. Section 3.2 for a more detailed discussion). □

Remark 3 (The Optimal Batch Sizes and Local Updates Trade-off)**.** The parameter $\nu \in [0,1]$ is used to balance the local minibatch sizes b, and the number of local updates I. Eqs. in (3) suggest that when ν increases from 0 to 1, b decreases and I increases. Specifically, if $\nu = 1$, then b is a constant but $I = \mathcal{O}(T^{1/3}/K^{2/3})$. In this case, each WN chooses a small minibatch while executing multiple local updates, and STEM resembles a FedAvg (a.k.a. Local SGD) algorithm but with double-sided momentum update directions, and is referred to as Fed STEM. In contrast, if $\nu = 0$, then $b = \mathcal{O}(T^{1/2}/K)$ but I is a constant. In this case, each WN chooses a large batch size while executing only a few, or even one, local updates, and STEM resembles the Minibatch SGD, but again with different update directions, and is referred to as Minibatch STEM. Such a trade-off can be seen in Fig. 1b. Due to space limitation, these two special cases will be precisely stated in the supplementary materials as corollaries of Theorem 3.1. □

Algorithm 2 The FedAvg Algorithm

```
Input: $\{\eta_t\}_{t=0}^{T}$; $I$, the # of local updates per communication round; $b$, the minibatch sizes.
for $t = 1$ to $T$ do
    for $k = 1$ to $K$ do
        $d_t^{(k)} = \frac{1}{b}\sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)})$ with $|\mathcal{B}_t^{(k)}| = b$
        $x_{t+1}^{(k)} = x_t^{(k)} - \eta_t d_t^{(k)}$
        if $t$ mod $I = 0$ then
            $x_{t+1}^{(k)} = \bar{x}_{t+1} = \frac{1}{K}\sum_{k=1}^{K} x_{t+1}^{(k)}$
        end if
    end for
end for
Return: $\bar{x}_a$ where $a \sim \mathcal{U}\{1, ..., T\}$.
```

Remark 4 (The Sub-Optimal Batch Sizes and Local Updates Trade-off)**.** From our proof (Theorem C.10 included in the supplemental material), we can see that STEM requires $\tilde{\mathcal{O}}\big(\max\big\{(b \cdot I)\epsilon^{-1}, K^{-1}\epsilon^{-3/2}\big\}\big)$ samples and $\tilde{\mathcal{O}}\big(\max\big\{\epsilon^{-1}, (b \cdot I)^{-1}K^{-1}\epsilon^{-3/2}\big\}\big)$ and communication rounds. According to the above expressions, if $b \cdot I$ increases beyond $\mathcal{O}(K^{-1}\epsilon^{-1/2})$, then the sample complexity will increase from the optimal $\tilde{\mathcal{O}}(\epsilon^{-3/2})$; otherwise, the optimal sample complexity $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ is maintained. On the other hand, if $b \cdot I$ decreases beyond $\mathcal{O}(K^{-1}\epsilon^{-1/2})$, the communication complexity increases from $\tilde{\mathcal{O}}(\epsilon^{-1})$. For instance, if we choose $b = \mathcal{O}(1)$ and $I = \mathcal{O}(1)$ the communication complexity becomes $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ while the optimal sample complexity $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ is maintained. This trade-off is illustrated in Figure 1a, where we maintain the optimal sample complexity, while changing b and I to generate the trade-off surface. □

Remark 5 (Data Heterogeneity)**.** The term $\tilde{\mathcal{O}}\Big(\frac{\zeta^2}{K^{2\nu/3}T^{1-\nu/3}}\Big)$ in the gradient bound (4) captures the effect of the heterogeneity of data across WNs, where ζ is the parameter characterizing the intra-node variance and has been defined in Assumption 2-(ii). Highly heterogeneous data with large ζ^2 can adversely impact the performance of STEM. Note that such a dependency on ζ also appears in other existing FL algorithms, such as [9, 14, 18]. However, there is one special case of STEM that does not depend on the parameter ζ. This is the case where $I = 1$, i.e., the minibatch SGD counterpart of STEM where only a single local iteration is performed between two communication rounds. We have the following corollary. □

Corollary 1 (Minibatch STEM)**.** *Under Assumptions 1 and 2 , and choose the algorithm parameters as in Theorem 3.1. At each WN, choose $I = 1$, $b = (T/K^2)^{1/2}$, and the initial batch size $B = b \cdot I$. Then STEM satisfies:*

(i) For $\bar{x}_a$ chosen according to Algorithm 1, we have

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\Big(\frac{f(\bar{x}_1) - f^*}{T}\Big) + \tilde{\mathcal{O}}\Big(\frac{\sigma^2}{T}\Big).$$

(ii) Minibatch STEM achieves $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ sample and $\tilde{\mathcal{O}}(\epsilon^{-1})$ communication complexity.

Next, we show that FedAvg also exhibits a trade-off similar to that of STEM but with worse sample and communication complexities.

3.2 Special cases: The FedAvg algorithm

We briefly discuss another interesting special case of STEM, where the local momentum update is replaced by the conventional SGD (i.e., $a_t = 1, \ \forall\ t$), while the server does not perform the momentum update (i.e., $\bar{d}_t = 0, \forall\ t$). This is essentially the classical FedAvg algorithm, just that it balances the number of local updates I and the minibatch size b. We show that this algorithm also exhibits a trade-off between b and I and on the trade-off curve it achieves $\mathcal{O}(\epsilon^{-2})$ sample complexity and $\mathcal{O}(\epsilon^{-3/2})$ communication complexity.

Algorithm	Training Acc.	Testing Acc.
FedAvg	78.2	74.1
FedProx	79.2	74.8
FedDyn	68.9	66.0
SCAFFOLD	71.9	74.0
MIME	82.6	76.8
FedGLOMO	76.1	72.8
STEM	80.1	78.8

(a) Mild heterogeneity, $b = 64$, and $I = 7$.

Algorithm	Training Acc.	Testing Acc.
FedAvg	73.6	75.4
FedProx	80.0	75.2
FedDyn	76.1	71.3
SCAFFOLD	72.5	73.7
MIME	61.5	58.6
FedGLOMO	10.0	10.0
STEM	81.1	78.5

(b) Moderate heterogeneity, $b = 8$, and $I = 61$.

Table 2: Training and testing accuracy of different algorithms on CIFAR-10 dataset for different batch-sizes, number of local updates, and heteregeneity settings.

Theorem 3.2 (The FedAvg Algorithm)**.** *Under Assumptions 1 and 2, suppose the stepsize is chosen as:* $\eta = \sqrt{\frac{bK}{T}}$*; Let us set:*

$$I = \mathcal{O}\big((T/K^3)^{\nu/4}\big), \quad b = \mathcal{O}\big((T/K^3)^{1/3-\nu/3}\big) \tag{5}$$

where $\nu \in [0,1]$ *is a constant. Then for FedAvg with* $T \geq 81L^2I^2bK$*, the following holds*

(i) For $\bar{x}_a$ *chosen according to Algorithm 2, we have*

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{K^{\nu/2}T^{2/3-\nu/6}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2}{K^{\nu/2}T^{2/3-\nu/6}}\bigg) + \mathcal{O}\bigg(\frac{\zeta^2}{K^{\nu/2}T^{2/3-\nu/6}}\bigg).$$

(ii) For any choice of $\nu \in [0,1]$ *we have:*
Sample Complexity: *The sample complexity of FedAvg is* $\mathcal{O}(\epsilon^{-2})$*. This implies that each WN requires at most* $\mathcal{O}(K^{-1}\epsilon^{-2})$ *gradient computations, thereby achieving linear speedup with the number of WNs in the network.*
Communication Complexity: *The communication complexity of FedAvg is* $\mathcal{O}(\epsilon^{-3/2})$.

Note that the requirement on T being lower bounded is only relevant for theoretical purposes, a similar requirement was also imposed in [14] to prove convergence. Again, the parameter $\nu \in [0,1]$ in the statement of Theorem 3.2 balances I and b at each WN while maintaining state-of-the-art sample and communication complexities; please see Table 1 for a comparison of those bounds with existing FedAvg bounds. For $\nu = 1$, FedAvg (cf. Theorem 3.2) reduces to FedAvg proposed in [12, 14] and for $\nu = 0$, the algorithm can be viewed as a large batch FedAvg with constant local updates [15, 16]. Note that similar to STEM, it is known that for $I = 1$, the Minibatch SGD's performance is independent of the heterogeneity parameter, ζ [13]. We also point out that if Algorithm 1 uses Nesterov's or Polyak's momentum [14] at local WNs instead of the recursive momentum estimator we get the same guarantees as in Theorem 3.2.

In summary, this section established that once the WN's and the SN's update directions (SGD in FedAvg and momentum based directions in STEM) are fixed, there exists a sequence of optimal choices of the number of local updates I, and the batch sizes b, which guarantees the best possible sample and communication complexities for the particular algorithm. The trade-off analysis presented in this section provides some useful guidelines for how to best select b and I in practice. Our subsequent numerical results will also verify that if b or I are not chosen judiciously, then the practical performance of the algorithms can degrade significantly.

4 Numerical results

In this section, we validate the proposed STEM algorithm and compare its performance with the de facto standard FedAvg [11], and the algorithms stated in Table 1. Note that instead of FedPD we include the performance comparison with FedDyn [36] since they are known to be very closely

Algorithm	Training Acc.	Testing Acc.
FedAvg	57.6	57.1
FedProx	59.1	58.5
FedDyn	51.2	51.3
SCAFFOLD	53.1	54.7
MIME	56.1	55.1
FedGLOMO	56.8	56.1
STEM	58.5	57.4

Table 3: Training and testing accuracy on CIFAR-10 dataset for high heterogeneity, $b = 128$ and $I = 6$.

Algorithm	Training Acc.	Testing Acc.
FedAvg	40.1	39.2
FedProx	43.5	43.2
FedDyn	43.7	43.2
SCAFFOLD	40.3	41.3
MIME	32.1	32.1
FedGLOMO	40.3	40.1
STEM	44.5	43.8

Table 4: Training and testing accuracy on Shakespeare dataset.

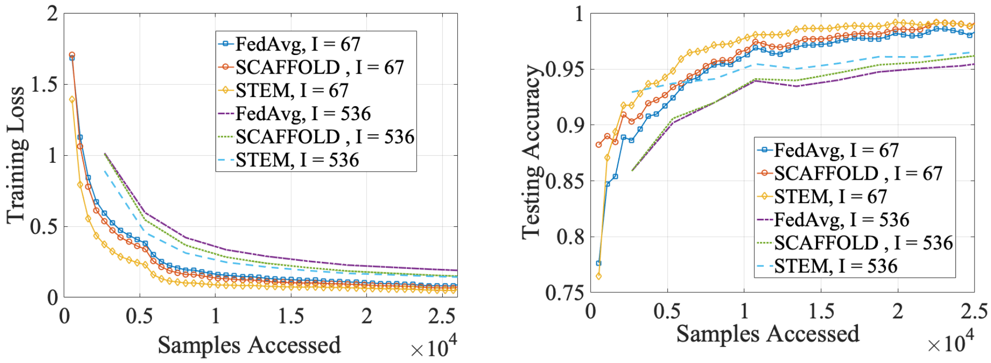


Figure 2: Training loss and the testing accuracy for classification on MNIST data set against the number of samples accessed at each WN for moderate heterogeneity setting with $b = 8$.

related. The goal of our experiments are three-fold: (1) To show that STEM performs on par, if not better, compared to other algorithms in different heterogeneity settings, (2) there are multiple ways to reach the desired solution accuracy, one can either choose a large batch size and perform only a few local updates or select a smaller batch size and perform multiple local updates, and finally, (3) if the local updates and the batch sizes are not chosen appropriately, the WNs might need to perform excessive computations to achieve the desired solution accuracy, thereby slowing down convergence.

Data and Parameter Settings: We compare the algorithms for image classification tasks on CIFAR-10 and MNIST data sets with 100 WNs, and for next character prediction task on Shakespeare dataset [37] with 143 WNs in the network. For both CIFAR-10 and MNIST, each WN implements a two-hidden-layer convolutional neural network (CNN) architecture followed by three linear layers for CIFAR-10 and two for MNIST. For CIFAR-10 (and MNIST) dataset, we consider three settings with mild, moderate and high heterogeneity. For all the three settings, the data is partitioned into disjoint sets among the WNs. In the mild heterogeneity setting, the WNs have access to partitioned data from all the classes. In the moderate (resp. high) heterogeneity setting the data is partitioned such that each WN can access data from only 5 (resp. 2) out of 10 classes. For CIFAR-10 (resp. MNIST), each WN has access to 490 (resp. 540) samples for training and 90 (resp. 80) samples for testing purposes.

We also compare the performance of algorithms on a popular FL benchmarking dataset, Shakespeare dataset [37]. For this task, we adopt the settings from [10] and utilize a 2-Layer LSTM network with 100 hidden units and an 8-D embedding layer at each WN. Each WN has access to 3616 samples on average, and the samples are randomly split into an 80% training set and a 20% testing set. We randomly sample 10 nodes out of 143 for the training purpose. All the experiments are implemented on a single NVIDIA Quadro RTX 5000 GPU. More details are provided in appendix.

For the proposed STEM algorithm, recall that the step-size is $\eta_t = \bar{\kappa}/(w_t + \sigma^2 t)^{1/3}$ with momentum parameter defined as $a_t = c\eta_t^2$. The step-size is used to update the iterates while the momentum parameter is used to construct the stochastic gradient estimate (cf. Algorithm 1 and Theorem 3.1). For the experiments, we set $w_t = \sigma^2 = 1$ and $c = \bar{c}/\bar{\kappa}^2$ and tune for $\bar{\kappa} \in [10^{-1}, 10^{-2}]$ for the CIFAR-10 dataset and for $\bar{\kappa} \in \{10^1, 10^0, 10^{-1}, 10^{-2}\}$ for the Shakespeare dataset. For both the datasets we tune for $\bar{c}$ in the range $[1, 10]$. For FedProx [10] and FedDyn [36] we choose the regularization constant to be 0.1. The momentum parameters for FedGLOMO [18] and MIME [17] are set based on the choices given in the respective papers. Specifically, for FedGLOMO we choose the parameter $\beta_k = 0.2$ and design the momentum gradient using a damping factor given in Appendix A.4 of FedGLOMO [18]. Moreover, for MIME we choose the momentum parameter as 0.9. For the rest of the algorithms (including FedAvg and SCAFFOLD), the step-size is tuned from the set $\{10^1, 10^0, 10^{-1}, 10^{-2}\}$.

Discussion: We evaluate the training and testing performance of STEM against multiple algorithms for different heterogeneity settings, minibatch sizes, and number of local updates. In Tables 2a, 2b and 3, we compare the training and testing accuracy of STEM to that of other algorithms on the CIFAR-10 dataset. Specific, heterogeneity settings, the choices of minibatches, and number of local updates are stated along with the tables. Note that STEM performs uniformly well under all the conditions. Moreover, note from Table 2b that FedGLOMO diverges once the number of local updates are high. Also, note from Table 3 that FedProx and STEM adapt well to high heterogeneity. Finally, with the next set of experiments we emphasize the importance of choosing b and I carefully. In Figure 2, we compare the training and testing performance of STEM, FedAvg and SCAFFOLD, against the number of samples accessed at each WN for the classification task on MNIST dataset with moderate heterogeneity. We fix $b = 8$ and conduct experiments under two settings, one with $I = 67$, and the other with $I = 536$ local updates at each WN. Note that although a large number of local updates might lead to fewer communication rounds but it can make the sample complexity extremely high as is demonstrated by Figure 2. For example, Figure 2 shows that to reach testing accuracy of $96 - 97\%$ with $I = 67$, STEM requires approximately $5000 - 6000$ samples, in contrast with $I = 536$ it requires more than 25000 samples at each WN. Similar behavior can be observed if we fix $I > 1$ and increase the local batch sizes. This implies not choosing the local updates and the batch sizes judiciously might lead to increased sample complexity. Additional experiments are included in the supplementary material to further evaluate the performance of the proposed algorithms.

Conclusion

In this work, we proposed a novel algorithm STEM, for distributed stochastic non-convex optimization with applications to FL. We showed that STEM reaches an ϵ-stationary point with $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ sample complexity while achieving linear speed-up with the number of WNs. Moreover, the algorithm achieves a communication complexity of $\tilde{\mathcal{O}}(\epsilon^{-1})$. We established a (optimal) trade-off that allows interpolation between varying choices of local updates and the batch sizes at each WN while maintaining (near optimal) sample and communication complexities. We showed that FedAvg (a.k.a LocalSGD) also exhibits a similar trade-off while achieving worse complexities. Our results provide guidelines to carefully choose the number of local updates, update directions, and minibatch sizes to achieve the best performance. The future directions of this work include developing lower bounds on communication complexity that establishes the tightness of the analysis conducted in this work.

Acknowledgement

We thank the anonymous reviewers for their valuable comments and suggestions. The work of Prashant Khanduri and Mingyi Hong was supported by NSF grant CMMI-1727757, AFOSR grant 19RT0424 and ARO grant W911NF-19-1-0247. The work of Mingyi Hong was also supported by an IBM Faculty Research award. The work of Jia Liu has been supported in part by NSF grants CAREER CNS-2110259, CNS-2112471, CNS-2102233, CCF-2110252, ECCS-2140277, and a Google Faculty Research Award.

References

[1] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[2] M. Li, D. G. Andersen, A. J. Smola, and K. Yu, "Communication efficient distributed machine learning with the parameter server," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 19–27.

[3] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang *et al.*, "Large scale distributed deep networks," in *Advances in neural information processing systems*, 2012, pp. 1223–1231.

[4] T. Léauté and B. Faltings, "Protecting privacy through distributed computation in multi-agent decision making," *Journal of Artificial Intelligence Research*, vol. 47, pp. 649–695, 2013.

[5] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Advances in Neural Information Processing Systems*, 2018, pp. 689–699.

[6] D. Zhou, P. Xu, and Q. Gu, "Stochastic nested variance reduction for nonconvex optimization," *arXiv preprint arXiv:1806.07811*, 2018.

[7] A. Cutkosky and F. Orabona, "Momentum-based variance reduction in non-convex SGD," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 15 236–15 245.

[8] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen, "Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization," *arXiv preprint arXiv:1905.05920*, 2019.

[9] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "Fedpd: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, pp. 1–1, 2021.

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[12] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5693–5700.

[13] B. Woodworth, K. K. Patel, and N. Srebro, "Minibatch vs local sgd for heterogeneous distributed learning," *arXiv preprint arXiv:2006.04735*, 2020.

[14] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7184–7193.

[15] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[16] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," *arXiv preprint arXiv:2101.11203*, 2021.

[17] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Mime: Mimicking centralized stochastic algorithms in federated learning," *arXiv preprint arXiv:2008.03606*, 2020.

[18] R. Das, A. Hashemi, S. Sanghavi, and I. S. Dhillon, "Improved convergence rates for non-convex federated learning with compression," *arXiv preprint arXiv:2012.04061*, 2020.

[19] B. Woodworth, K. K. Patel, S. U. Stich, Z. Dai, B. Bullins, H. B. McMahan, O. Shamir, and N. Srebro, “Is local sgd better than minibatch sgd?” *arXiv preprint arXiv:2002.07839*, 2020.

[20] H. Yu and R. Jin, “On the computation and communication complexity of parallel sgd with dynamic batch sizes for stochastic non-convex optimization,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 7174–7183.

[21] J. Wang and G. Joshi, “Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms,” *Journal of Machine Learning Research*, vol. 22, no. 213, pp. 1–50, 2021.

[22] A. Khaled, K. Mishchenko, and P. Richtárik, “Better communication complexity for local sgd,” *arXiv*, 2019.

[23] S. U. Stich, “Local sgd converges fast and communicates little,” *arXiv preprint arXiv:1805.09767*, 2018.

[24] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi, “Don’t use large mini-batches, use local sgd,” in *International Conference on Learning Representations*, 2020.

[25] F. Zhou and G. Cong, “On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 7 2018, pp. 3219–3227.

[26] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, “Robust and communication-efficient federated learning from non-iid data,” *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[27] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, “Federated learning with non-iid data,” *arXiv preprint arXiv:1806.00582*, 2018.

[28] J. Wang, V. Tantia, N. Ballas, and M. Rabbat, “Slowmo: Improving communication-efficient distributed sgd with slow momentum,” *arXiv preprint arXiv:1910.00643*, 2019.

[29] X. Liang, S. Shen, J. Liu, Z. Pan, E. Chen, and Y. Cheng, “Variance reduced local sgd with lower communication complexity,” *arXiv preprint arXiv:1912.12844*, 2019.

[30] P. Sharma, P. Khanduri, S. Bulusu, K. Rajawat, and P. K. Varshney, “Parallel restarted SPIDER – communication efficient distributed nonconvex optimization with optimal computation complexity,” *arXiv preprint arXiv:1912.06036*, 2019.

[31] S. J. Reddi, S. Kale, and S. Kumar, “On the convergence of adam and beyond,” *arXiv preprint arXiv:1904.09237*, 2019.

[32] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, “A unified theory of decentralized sgd with changing topology and local updates,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 5381–5393.

[33] R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 315–323.

[34] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[35] Y. Drori and O. Shamir, “The complexity of finding stationary points with stochastic gradient descent,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 2658–2667.

[36] D. A. E. Acar, Y. Zhao, R. Matas, M. Mattina, P. Whatmough, and V. Saligrama, “Federated learning based on dynamic regularization,” in *International Conference on Learning Representations*, 2020.

[37] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, “Leaf: A benchmark for federated settings,” 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The datasets and the models used for experiments involve large number of parameters. With given computational resources it takes long time to run a single experiment.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the experiments section.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

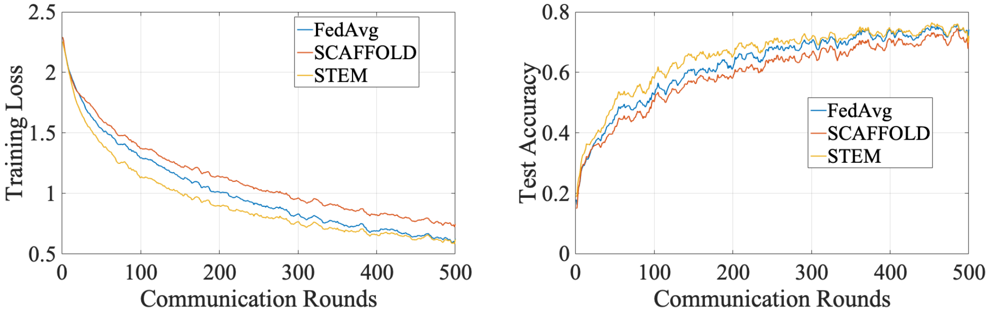


Figure 3: Training loss and testing accuracy for classification on CIFAR-10 dataset against the number of communication rounds for mild heterogeneity setting with $b = 64$ and $I = 7$.

Appendix

The organization of the Appendix is given below. In Appendix A we present the experimental details with along with additional numerical results on CIFAR-10 and MNIST datasets. Then in Appendix B, we present the proof of the convergence guarantees associated with the FedAvg algorithm given in Algorithm 2. Finally in Appendix C, we present the proof of the convergence for STEM given in Algorithm 1. Our proof is further divided into two parts, where in Appendix C.1 we present some useful lemmas, and the main body of the proof is given in Appendix C.2.

A Additional experiments

Shakespeare Dataset. The Shakespeare dataset considers a classification problem of next character prediction with 80 classes in total. We associate with each node a different speaking role (same setting as in [10]). We have a total of 143 nodes with a total of $517,106$ samples that are unevenly split among 143 nodes with each node having 3616 samples on average. We randomly split the data at each node into an 80% training set and a 20% testing set. We randomly sample 10 nodes out of 143 for the training purpose. For this task, we utilize a 2-Layer LSTM network with 100 hidden units and an 8-D embedding layer at each node. For each algorithm, we select a batch size of 128 and tune for the rest of the hyperparameters as discussed in Section 4.

In this section, we present additional numerical results conducted for the classification task on CIFAR-10 and MNIST datasets. Here we focus on mild and moderate heterogeneity settings defined in Section 4. We compare the proposed STEM algorithm to two most popular baselines FedAvg and SCAFFOLD. We show that STEM outperforms both FedAvg and SCAFFOLD. Moreover, we corroborate the theoretical findings by showing that the algorithms converge in both cases, one where large batch sizes with a few local updates are used, and second where small batch sizes with a large number of local updates are employed. We utilize the same experimental settings as discussed in Section 4. Next, we present the results.

Discussion. In Figures 4 and 3, we compare the training and testing performance of STEM with FedAvg and SCAFFOLD for CIFAR-10 dataset under mild heterogeneity setting. For Figure 4, we choose $b = 8$ and $I = 61$, whereas for Figure 3, we choose $b = 64$ and $I = 7$. We first note that for both cases STEM performs better than FedAvg and SCAFFOLD. Moreover, observe that for both settings, small batches with multiple local updates (Figure 4) and large batches with few local updates (Figure 3), the algorithms converge with approximately similar performance, corroborating the theoretical analysis (see Discussion in Section 1). Next, in Figure 5 we evaluate the performance of the proposed algorithms on CIFAR-10 with moderate heterogeneity setting for $b = 8$ and $I = 61$. We note that STEM outperforms FedAvg and SCAFFOLD in this setting as well.

Next, in Figure 6, we compare the training and testing performance of STEM and FedAvg against SCAFFOLD with the number of communication rounds. The figures are generated for local batch size of $b = 64$ while the number of local updates are chosen to be $I = 8$. The initial batch size, B, is chosen the same as b. Note form Figure 6 that STEM performs on par if not better than FedAvg under all settings. Moreover, STEM and FedAvg perform better than SCAFFOLD. In the next set

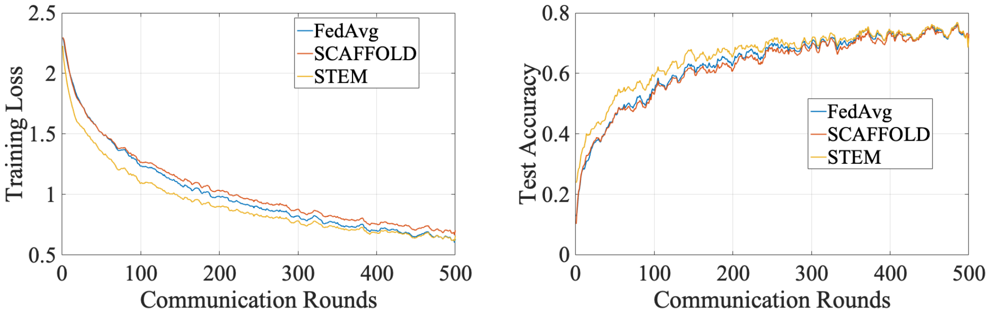


Figure 4: Training loss and testing accuracy for classification on CIFAR-10 dataset against the number of communication rounds for mild heterogeneity setting with $b = 8$ and $I = 61$.

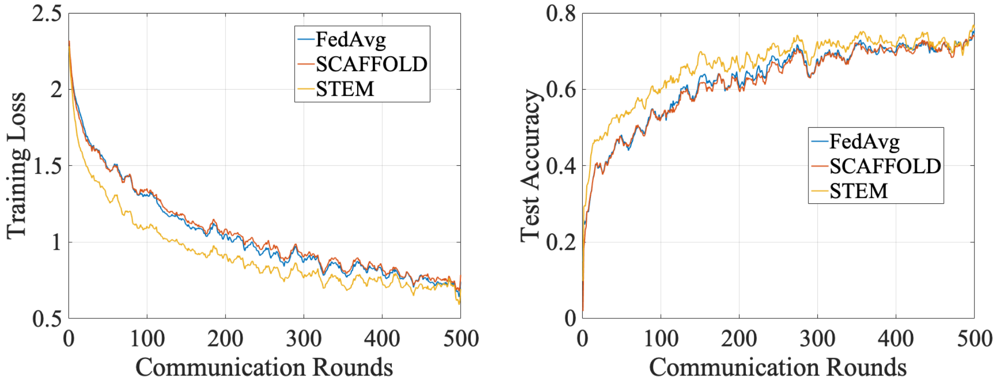


Figure 5: Training loss and testing accuracy for classification on CIFAR-10 dataset against the number of communication rounds for moderate heterogeneity setting with $b = 8$ and $I = 61$.

of simulations we trade the batch sizes for the number of local updates. Specifically, we choose $b = 8$ and $I = 67$, while choosing the same initial batch size, B, as b. The top two figures plot the performance of algorithms with mild heterogeneity setting while the lower two plot the performance for the moderate heterogeneity setting. Again note that STEM performs better than FedAvg and SCAFFOLD in both settings. Importantly, Figures 6 and 7 jointly imply that the algorithms can converge with acceptable performance while employing either "large batch sizes with few local updates" or "smaller batch sizes with multiple local updates".

Next, we present in detail the proofs of the results presented in the paper.

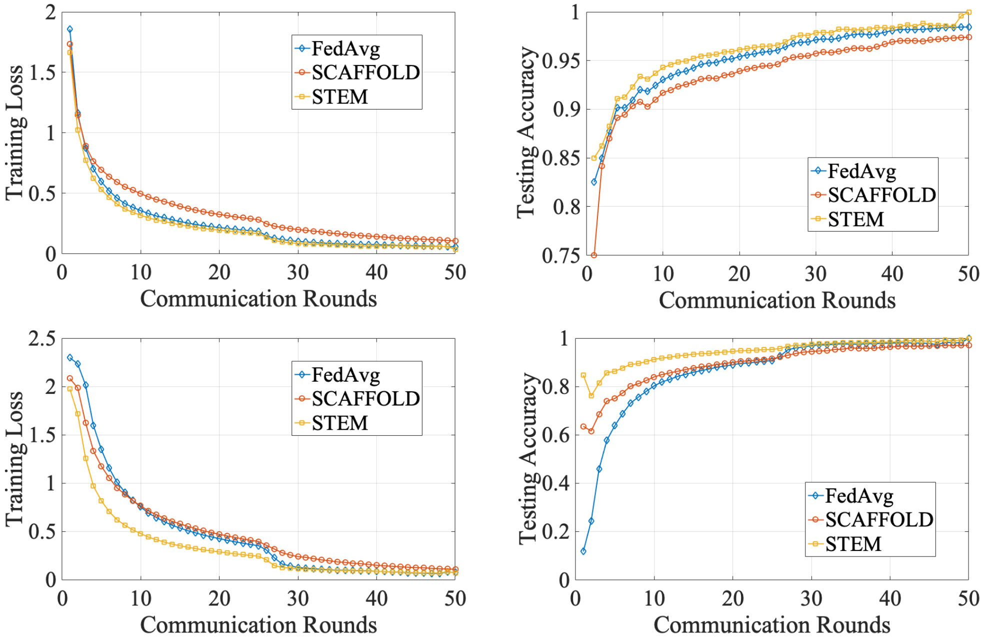


Figure 6: Training loss and the testing accuracy against the number of communication rounds with $b = 64$ and $I = 8$ for MNIST.

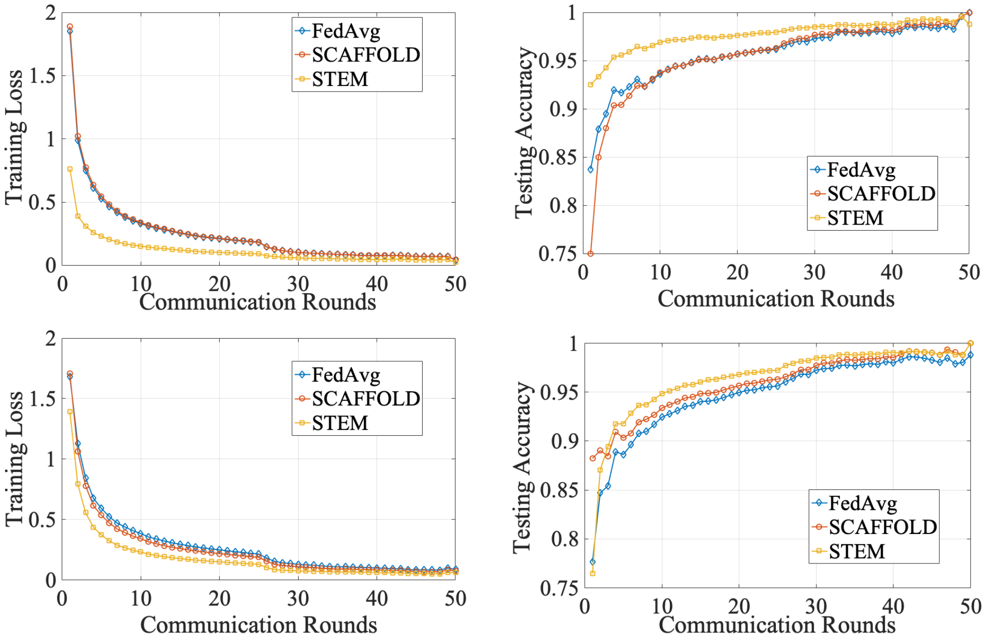


Figure 7: Training loss and the testing accuracy against the number of communication rounds with $b = 8$ and $I = 67$ for MNIST.

B Proofs of Convergence Guarantees for FedAvg

In this section, we present the proofs for the FedAvg algorithm. Before stating the proofs in detail we first present some preliminaries lemmas which shall be used for proving the main results of the paper. We first fix some notations:

We define $\bar{t}_s \coloneqq sI + 1$ with $s \in [S]$. Note from Algorithm 2 that at $(s \times I)^{\text{th}}$ iteration, i.e., when $t \bmod I = 0$, the iterates, $\{x_t^{(k)}\}_{k=1}^K$ corresponding to $t = (\bar{t}_s)^{\text{th}}$ time instant are shared with the SN. We define the filtration $\mathcal{F}_t$ as the sigma algebra generated by iterates $x_1^{(k)}, x_2^{(k)}, \ldots, x_t^{(k)}$ as

$$\mathcal{F}_t = \sigma(x_1^{(k)}, x_2^{(k)}, \ldots, x_t^{(k)}, \text{ for all } k \in [K]).$$

Also, throughout the section we assume Assumptions 1 and 2 to hold.

B.1 Preliminary Lemmas

Lemma B.1. *For $\bar{d}_t = \frac{1}{K}\sum_{k=1}^K d_t^{(k)}$ where $d_t^{(k)}$ for all $k \in [K]$ and $t \in [T]$ is chosen according to Algorithm 2, we have:*

$$\mathbb{E}\Big\|\bar{d}_t - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})\Big\|^2 \leq \frac{\sigma^2}{bK},$$

where the expectation is w.r.t the stochasticity of the the algorithm.

Proof. Using the definition of $\bar{d}_t$ we have:

$$\begin{aligned}
&\mathbb{E}\Big\|\bar{d}_t - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})\Big\|^2 \\
&= \mathbb{E}\Big\|\frac{1}{K}\sum_{k=1}^K \frac{1}{b}\sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})\Big\|^2 \\
&= \mathbb{E}\Big\|\frac{1}{K}\sum_{k=1}^K \frac{1}{b}\sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big)\Big\|^2 \\
&\overset{(a)}{=} \frac{1}{b^2K^2}\sum_{k=1}^K \mathbb{E}\Big\|\sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big)\Big\|^2 \\
&+ \frac{1}{b^2K^2}\sum_{k \neq \ell} \mathbb{E}\,\Big\langle \underbrace{\mathbb{E}\Big[\sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big)\Big|\mathcal{F}_t\Big]}_{=0}, \underbrace{\mathbb{E}\Big[\sum_{\xi_t^{(\ell)} \in \mathcal{B}_t^{(\ell)}} \big(\nabla f^{(\ell)}(x_t^{(\ell)}; \xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_t^{(\ell)})\big)\Big|\mathcal{F}_t\Big]}_{=0} \Big\rangle \\
&\overset{(b)}{=} \frac{1}{b^2K^2}\sum_{k=1}^K \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \mathbb{E}\big\|\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big\|^2 \\
&+ \frac{1}{b^2K^2}\sum_{k=1}^K \sum_{\xi_t^{(k)} \neq \zeta_t^{(k)}} \mathbb{E}\Big\langle \underbrace{\mathbb{E}\big[\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big|\mathcal{F}_t\big]}_{=0}, \underbrace{\mathbb{E}\big[\nabla f^{(k)}(x_t^{(k)}; \zeta_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big|\mathcal{F}_t\big]}_{=0} \Big\rangle \\
&\overset{(c)}{\leq} \frac{\sigma^2}{bK},
\end{aligned}$$

where (a) follows from Assumption 2 that given $\mathcal{F}_t$ we have: $\mathbb{E}\big[\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)})\big] = \nabla f^{(k)}(x_t^{(k)})$, for all $k \in [K]$. Moreover, given $\mathcal{F}_t$ the samples $\xi_t^{(k)}$ and $\xi_t^{(\ell)}$ at the k^{th} and the ℓ^{th} WNs are chosen uniformly randomly, and independent of each other for all $k, \ell \in [K]$ and $k \neq \ell$, therefore we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big), \sum_{\xi_t^{(\ell)} \in \mathcal{B}_t^{(\ell)}} \big(\nabla f^{(\ell)}(x_t^{(\ell)}; \xi_t^{(\ell)}) - \nabla f^{(\ell)}(\bar{x}_t)\big) \Big\rangle\Big] \\
&= \mathbb{E}\Big[\Big\langle \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \underbrace{\mathbb{E}\big[\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big|\mathcal{F}_t\big]}_{=0}, \sum_{\xi_t^{(\ell)} \in \mathcal{B}_t^{(\ell)}} \underbrace{\mathbb{E}\big[\nabla f^{(\ell)}(x_t^{(\ell)}; \xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_t^{(\ell)})\big|\mathcal{F}_t\big]}_{=0} \Big\rangle\Big]
\end{aligned}$$

$= 0.$

The equality (b) follows from the fact that $\xi_1^{(k)}$ and $\zeta_1^{(k)}$ for all $k \in [K]$ are chosen independently of each other. Then we conclude (b) from an argument similar to that of (a). Finally, (c) results from the intra-node variance bound given in Assumption 2(ii).

Hence, the lemma is proved. □

Lemma B.2. *For a finite sequence $x^{(k)} \in \mathbb{R}^d$ for $k \in [K]$ define $\bar{x} := \frac{1}{K}\sum_{k=1}^K x^{(k)}$, we then have*

$$\sum_{k=1}^K \|x^{(k)} - \bar{x}\|^2 \leq \sum_{k=1}^K \|x^{(k)}\|^2.$$

Proof. Using the notation $\mathbf{x} = \left[\left(x^{(1)}\right)^T, \left(x^{(2)}\right)^T, \ldots, \left(x^{(K)}\right)^T\right]^T \in \mathbb{R}^{Kd}$, denoting $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ and $\mathbf{I}_{Kd} \in \mathbb{R}^{Kd\times Kd}$ as identity matrices and representing $\mathbf{1} \in \mathbb{R}^K$ as the vector of all ones. We rewrite the left hand side of the statement as

$$\begin{aligned}
\sum_{k=1}^K \|x^{(k)} - \bar{x}\|^2 &= \left\|\mathbf{x} - \left(\mathbf{I} \otimes \frac{\mathbf{1}\mathbf{1}^T}{K}\right)\mathbf{x}\right\|^2 \\
&= \left\|\left(\mathbf{I}_{Kd} - \left(\mathbf{I}_d \otimes \frac{\mathbf{1}\mathbf{1}^T}{K}\right)\right)\mathbf{x}\right\|^2 \\
&\overset{(a)}{\leq} \|\mathbf{x}\|^2 = \sum_{k=1}^K \|x^{(k)}\|^2,
\end{aligned}$$

where (a) follows from the fact that the induced matrix norm $\left\|\mathbf{I}_{Kd} - \left(\mathbf{I}_d \otimes \frac{\mathbf{1}\mathbf{1}^T}{K}\right)\right\| \leq 1$. □

Lemma B.3 (From [7]). *Let $a_0 > 0$ and $a_1, a_2, \ldots, a_T \geq 0$. We have*

$$\sum_{t=1}^T \frac{a_t}{a_0 + \sum_{i=t}^t a_i} \leq \ln\left(1 + \frac{\sum_{i=1}^t a_i}{a_0}\right).$$

Lemma B.4. *For $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$, we have*

$$\|X_1 + X_2 + \ldots + X_n\|^2 \leq n\|X_1\|^2 + n\|X_2\|^2 + \ldots + n\|X_n\|^2.$$

Next, we present the proof of Theorem 3.2. The proof follows in few steps which are discussed next.

B.2 Proof of Main Results: FedAvg

Lemma B.5 (Error Accumulation from Iterates). *For the choice of stepsize $\eta \leq 1/9 \cdot L \cdot I$, the iterates $x_t^{(k)}$ for each $k \in [K]$ generated from Algorithm 2 satisfy:*

$$\sum_{t=1}^T \frac{1}{K}\sum_{k=1}^K \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 \leq 3\eta^2(I-1)\sigma^2 T + 5\eta^2(I-1)^2\zeta^2 T,$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. Note from Algorithm 2 and the definition of $\bar{t}_s$ that at $t = \bar{t}_{s-1}$ with $s \in [S]$, $x_t^{(k)} = \bar{x}_t$, for all k. This implies

$$\frac{1}{K}\sum_{k=1}^K \|x_{\bar{t}_{s-1}}^{(k)} - \bar{x}_{\bar{t}_{s-1}}\|^2 = 0.$$

Therefore, the statement of the lemma holds trivially. Moreover, for $t \in [\bar{t}_{s-1} + 1, \bar{t}_s - 1]$, with $s \in [S]$, we have from Algorithm 2: $x_t^{(k)} = x_{t-1}^{(k)} - \eta d_{t-1}^{(k)}$, this implies that:

$$x_t^{(k)} = x_{\bar{t}_{s-1}}^{(k)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta d_\ell^{(k)} \quad \text{and} \quad \bar{x}_t = \bar{x}_{\bar{t}_{s-1}} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta \bar{d}_\ell.$$

This implies that for $t \in [\bar{t}_{s-1} + 1, \bar{t}_s - 1]$, with $s \in [S]$ we have

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} \|x_t^{(k)} - \bar{x}_t\|^2 &= \frac{1}{K}\sum_{k=1}^{K} \Big\| x_{\bar{t}_{s-1}}^{(k)} - \bar{x}_{\bar{t}_{s-1}} - \Big(\sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta d_\ell^{(k)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta \bar{d}_\ell \Big) \Big\|^2 \\
&\overset{(a)}{=} \frac{\eta^2}{K} \sum_{k=1}^{K} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \big(d_\ell^{(k)} - \bar{d}_\ell \big) \Big\|^2 \\
&\overset{(b)}{=} \frac{\eta^2}{K} \sum_{k=1}^{K} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big(\frac{1}{b} \sum_{\xi_\ell^{(k)} \in \mathcal{B}_\ell^{(k)}} \nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_\ell^{(j)} \in \mathcal{B}_\ell^{(j)}} \nabla f^{(j)}(x_\ell^{(j)}; \xi_\ell^{(j)}) \Big) \Big\|^2 \\
&\overset{(c)}{\leq} \frac{2\eta^2}{K} \sum_{k=1}^{K} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big[\Big(\frac{1}{b} \sum_{\xi_\ell^{(k)} \in \mathcal{B}_\ell^{(k)}} \nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \Big) \\
&\qquad\qquad - \frac{1}{K} \sum_{j=1}^{K} \Big(\frac{1}{b} \sum_{\xi_\ell^{(j)} \in \mathcal{B}_\ell^{(j)}} \nabla f^{(j)}(x_\ell^{(j)}; \xi_\ell^{(j)}) - \nabla f^{(j)}(x_\ell^{(j)}) \Big) \Big] \Big\|^2 \\
&\qquad\qquad\qquad + \frac{2\eta^2}{K} \sum_{k=1}^{K} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big(\nabla f^{(k)}(x_\ell^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} \nabla f^{(j)}(x_\ell^{(j)}) \Big) \Big\|^2 \\
&\overset{(d)}{\leq} \frac{2\eta^2}{K} \sum_{k=1}^{K} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big(\frac{1}{b} \sum_{\xi_\ell^{(k)} \in \mathcal{B}_\ell^{(k)}} \nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \Big) \Big\|^2 \\
&\qquad + \frac{2\eta^2}{K} \sum_{k=1}^{K} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big(\nabla f^{(k)}(x_\ell^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} \nabla f^{(j)}(x_\ell^{(j)}) \Big) \Big\|^2, \qquad (6)
\end{aligned}$$

where the equality (a) follows from the fact that $x_{\bar{t}_{s-1}}^{(k)} = \bar{x}_{\bar{t}_{s-1}}$ for $t = \bar{t}_{s-1}$; (b) results from the definition of the stochastic gradient employed by FedAvg in Algorithm 2; (c) uses Lemma B.4 and (d) follows from the application of Lemma B.2.

Taking expectation on both sides and let us next consider each term of (6) above separately, we have for any $k \in [K]$ from the first term of (6) above

$$\begin{aligned}
\mathbb{E} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big(\frac{1}{b} \sum_{\xi_\ell^{(k)} \in \mathcal{B}_\ell^{(k)}} \nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \Big) \Big\|^2 &\overset{(a)}{=} \sum_{\ell=\bar{t}_{s-1}}^{t-1} \mathbb{E} \Big\| \frac{1}{b} \sum_{\xi_\ell^{(k)} \in \mathcal{B}_\ell^{(k)}} \nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \Big\|^2 \\
&\overset{(b)}{=} \sum_{\ell=\bar{t}_{s-1}}^{t-1} \frac{1}{b^2} \sum_{\xi_\ell^{(k)} \in \mathcal{B}_\ell^{(k)}} \mathbb{E} \big\| \nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \big\|^2 \\
&\overset{(c)}{\leq} \frac{(I-1)}{b} \sigma^2 \\
&\overset{(d)}{\leq} (I-1)\sigma^2, \qquad (7)
\end{aligned}$$

where (a) results from the fact that $\mathbb{E}\Big[\frac{1}{b} \sum_{\xi_\ell^{(k)} \in \mathcal{B}_\ell^{(k)}} \nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \Big| \mathcal{F}_{\bar{\ell}} \Big] = 0$ for any $\bar{\ell} < \ell$; (b) uses the fact that $\mathbb{E}\big[\nabla f^{(k)}(x_\ell^{(k)}; \xi_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \big| \nabla f^{(k)}(x_\ell^{(k)}; \zeta_\ell^{(k)}) - \nabla f^{(k)}(x_\ell^{(k)}) \big] = 0$ for samples $\xi_\ell^{(k)}, \zeta_\ell^{(k)} \sim \mathcal{D}^{(k)}$ chosen independent; (c) utilizes intra-node variance bound in

Assumption 2(ii) and the fact that $(t-1)-\bar{t}_{s-1} \leq I-1$ for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$; and finally, (d) uses the fact that $b \geq 1$.

Next, we consider the second term of (6) for any $k \in [K]$, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big(\nabla f^{(k)}(x_\ell^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \nabla f^{(j)}(x_\ell^{(j)}) \Big) \Big\|^2 \\
&\overset{(a)}{\leq} (I-1) \sum_{\ell=\bar{t}_{s-1}}^{t-1} \sum_{k=1}^{K} \mathbb{E}\Big\| \nabla f^{(k)}(x_\ell^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \nabla f^{(j)}(x_\ell^{(j)}) \Big\|^2 \\
&\overset{(b)}{\leq} (I-1) \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big[4\sum_{k=1}^{K} \mathbb{E}\big\| \nabla f^{(k)}(x_\ell^{(k)}) - \nabla f^{(k)}(\bar{x}_\ell) \big\|^2 + 4\sum_{k=1}^{K} \mathbb{E}\Big\| \nabla f(\bar{x}_\ell) - \frac{1}{K}\sum_{j=1}^{K} \nabla f(x_\ell^{(j)}) \Big\|^2 \\
&\qquad\qquad + 2\sum_{k=1}^{K} \mathbb{E}\big\| \nabla f^{(k)}(\bar{x}_\ell) - \nabla f(\bar{x}_\ell) \big\|^2 \Big] \\
&\overset{(c)}{\leq} (I-1) \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big[8L^2 \sum_{k=1}^{K} \mathbb{E}\big\| x_\ell^{(k)} - \bar{x}_\ell \big\|^2 + 2\sum_{k=1}^{K} \mathbb{E}\Big\| \nabla f^{(k)}(\bar{x}_\ell) - \frac{1}{K}\sum_{j=1}^{K} \nabla f^{(j)}(\bar{x}_\ell) \Big\|^2 \Big] \\
&\overset{(d)}{\leq} 8L^2(I-1) \sum_{\ell=\bar{t}_{s-1}}^{t-1} \sum_{k=1}^{K} \mathbb{E}\big\| x_\ell^{(k)} - \bar{x}_\ell \big\|^2 + 2K(I-1)^2\zeta^2, \qquad (8)
\end{aligned}$$

where (a) utilizes the fact that $(t-1)-\bar{t}_{s-1} \leq I-1$ for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$; (b) results from the application of Lemma B.4; (c) follows from Assumption 1; and (d) utilizes the inter-node variance Assumption 2 and the fact that $(t-1)-\bar{t}_{s-1} \leq I-1$ for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$.

Substituting (7) and (8) in (6) and taking expectation on both sides we get

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 &\leq 2\eta^2(I-1)\sigma^2 + 4\eta^2(I-1)^2\zeta^2 \\
&\qquad + 16L^2(I-1)\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \frac{1}{K}\sum_{k=1}^{k} \mathbb{E}\|x_\ell^{(k)} - \bar{x}_\ell\|^2.
\end{aligned}$$

Summing both sides from $t = \bar{t}_{s-1}$ to $\bar{t}_s - 1$, we get

$$\begin{aligned}
&\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 \\
&\qquad \leq 2\eta^2(I-1)\sigma^2 I + 4\eta^2(I-1)^2\zeta^2 I + 16L^2(I-1)\eta^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \sum_{\ell=\bar{t}_{s-1}}^{t-1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|x_\ell^{(k)} - \bar{x}_\ell\|^2 \\
&\qquad \overset{(a)}{\leq} 2\eta^2(I-1)\sigma^2 I + 4\eta^2(I-1)^2\zeta^2 I + 16L^2(I-1)\eta^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s - 1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|x_\ell^{(k)} - \bar{x}_\ell\|^2 \\
&\qquad \overset{(b)}{\leq} 2\eta^2(I-1)\sigma^2 I + 4\eta^2(I-1)^2\zeta^2 I + 16L^2(I-1)\eta^2 I \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2,
\end{aligned}$$

where (a) uses that fact that $t \leq \bar{t}_s - 1$; (b) results from $t_s - t_{s-1} \leq I$ for all $s \in [S]$. Finally, summing over $s \in [S]$ and using $T = SI$ we get

$$\sum_{t=1}^{T} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 \leq 2\eta^2(I-1)\sigma^2 T + 4\eta^2(I-1)^2\zeta^2 T + 16L^2 I^2 \eta^2 \sum_{t=1}^{T} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2.$$

Rearranging the terms, we get

$$(1 - 16L^2 I^2 \eta^2) \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 \leq 2\eta^2 (I-1)\sigma^2 T + 4\eta^2 (I-1)^2 \zeta^2 T.$$

Finally, using the fact that $\eta \leq 1/9 \cdot L \cdot I$ we have $1 - 16L^2 I^2 \eta^2 \geq 4/5$. Multiplying, both sides by $5/4$ we get

$$\sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 \leq 3\eta^2 (I-1)\sigma^2 T + 5\eta^2 (I-1)^2 \zeta^2 T.$$

Therefore, the lemma is proved. □

Lemma B.6 (Descent Lemma). *For all $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ and $s \in [S]$, with the choice of stepsizes $\eta \leq 1/9 \cdot L \cdot I$, the iterates generated by Algorithm 2 satisfy:*

$$\mathbb{E} f(\bar{x}_{t+1}) \leq \mathbb{E} f(\bar{x}_t) - \frac{\eta}{2} \mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{\eta L^2}{2K} \sum_{k=1}^{K} \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 + \frac{\eta^2 L}{bK}\sigma^2,$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. Using the smoothness of f (Assumption 1) we have:

$$\begin{aligned}
&\mathbb{E}[f(\bar{x}_{t+1})] \\
&\leq \mathbb{E}\Big[f(\bar{x}_t) + \langle \nabla f(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[f(\bar{x}_t) - \eta \langle \nabla f(\bar{x}_t), \bar{d}_t \rangle + \frac{\eta^2 L}{2}\|\bar{d}_t\|^2\Big] \\
&\overset{(b)}{=} \mathbb{E}\Big[f(\bar{x}_t) - \eta \Big\langle \nabla f(\bar{x}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla f^{(k)}(x_t^{(k)}) \Big\rangle + \frac{\eta^2 L}{2}\|\bar{d}_t\|^2\Big] \\
&\overset{(c)}{=} \mathbb{E}\Bigg[f(\bar{x}_t) - \frac{\eta}{2}\Big\|\frac{1}{K}\sum_{k=1}^{K} \nabla f^{(k)}(x_t^{(k)})\Big\|^2 - \frac{\eta}{2}\|\nabla f(\bar{x}_t)\|^2 + \frac{\eta}{2}\Big\|\nabla f(\bar{x}_t) - \frac{1}{K}\sum_{k=1}^{K} \nabla f^{(k)}(x_t^{(k)})\Big\|^2 \\
&\qquad\qquad + \eta^2 L \Big\|\bar{d}_t - \frac{1}{K}\sum_{k=1}^{K} \nabla f^{(k)}(x_t^{(k)})\Big\|^2 + \eta^2 L \Big\|\frac{1}{K}\sum_{k=1}^{K} \nabla f^{(k)}(x_t^{(k)})\Big\|^2\Bigg] \\
&\overset{(d)}{\leq} \mathbb{E}\Bigg[f(\bar{x}_t) - \Big(\frac{\eta}{2} - \eta^2 L\Big)\Big\|\frac{1}{K}\sum_{k=1}^{K} \nabla f^{(k)}(x_t^{(k)})\Big\|^2 - \frac{\eta}{2}\|\nabla f(\bar{x}_t)\|^2 + \frac{\eta L^2}{2K}\sum_{k=1}^{K}\|x_t^{(k)} - \bar{x}_t\|^2 + \frac{\eta^2 L}{bK}\sigma^2\Bigg] \\
&\overset{(e)}{\leq} \mathbb{E}\Bigg[f(\bar{x}_t) - \frac{\eta}{2}\|\nabla f(\bar{x}_t)\|^2 + \frac{\eta L^2}{2K}\sum_{k=1}^{K}\|x_t^{(k)} - \bar{x}_t\|^2 + \frac{\eta^2 L}{bK}\sigma^2\Bigg],
\end{aligned}$$

where equality (a) follows from the iterate update given in Step 5 of Algorithm 2; (b) results from $\mathbb{E}[\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) | \mathcal{F}_t] = \nabla f^{(k)}(x_t^{(k)})$; (c) uses $\langle a, b \rangle = \frac{1}{2}[\|a\|^2 + \|b\|^2 - \|a - b\|^2]$ and Lemma B.4; (d) results from (9) below and Lemma B.1; and (e) results from the stepsize choice of $\eta \leq 1/9LI$.

$$\begin{aligned}
\mathbb{E}\Big\|\frac{1}{K}\sum_{k=1}^{K}\big(\nabla f^{(k)}(x_t^{(k)}) - \nabla f^{(k)}(\bar{x}_t)\big)\Big\|^2 &\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big\|\nabla f^{(k)}(x_t^{(k)}) - \nabla f^{(k)}(\bar{x}_t)\big\|^2 \\
&\leq \frac{L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2, \qquad (9)
\end{aligned}$$

where the first inequality follows from Lemma B.4, and the second follows from the L-Smoothness of $f^{(k)}(\cdot)$ (Assumption 1).

Hence, the lemma is proved. □

B.2.1 Proof of Theorem 3.2

The proof of Theorem 3.2 follows by replacing the choices of b and I given in (5) in the following result.

Theorem B.7. *Under Assumptions 1 and 2, with stepsize $\eta = \sqrt{\frac{bk}{T}}$. Then for $T \geq 81L^2I^2bK$ with any choice of minibatch sizes, $b \geq 1$, and number of local updates, $I \geq 1$, the iterates generated from Algorithm 2 satisfy*

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 \leq \frac{2(f(\bar{x}_t)) - f^*)}{(bk)^{1/2}T^{1/2}} + \frac{2L}{(bk)^{1/2}T^{1/2}}\sigma^2 + \frac{3L^2bK(I-1)}{T}\sigma^2 + \frac{5L^2bK(I-1)^2}{T}\zeta^2.$$

Proof. Summing the result of Lemma B.6 for $t = [T]$ and multiplying both sides by $2/\eta T$ we get

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 &\leq \frac{2(f(\bar{x}_t) - f(\bar{x}_{t+1}))}{\eta T} + \frac{2\eta L}{bK}\sigma^2 + \frac{L^2}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 \\ &\leq \frac{2(f(\bar{x}_t) - f^*)}{\eta T} + \frac{2\eta L}{bK}\sigma^2 + \frac{L^2}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2\end{aligned}$$

where the second inequality uses $f(\bar{x}_{t-1}) \geq f^*$. Next, using Lemma B.5 we get

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 \leq \frac{2(f(\bar{x}_t) - f^*)}{\eta T} + \frac{2\eta L}{bK}\sigma^2 + 3L^2\eta^2(I-1)\sigma^2 + 5L^2\eta^2(I-1)^2\zeta^2.$$

Finally, using the definition of $\bar{x}_a$ from Algorithm 2 and the choice of $\eta = \sqrt{\frac{bK}{T}}$, we get

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 \leq \frac{2(f(\bar{x}_t) - f^*)}{(bK)^{1/2}T^{1/2}} + \frac{2L}{(bK)^{1/2}T^{1/2}}\sigma^2 + \frac{3L^2bK(I-1)}{T}\sigma^2 + \frac{5L^2bK(I-1)^2}{T}\zeta^2.$$

Therefore, we have the theorem. □

Finally, substituting the choice of I and b given in (5) we get the statement of Theorem 3.2. Next two remarks characterize the behavior of FedAvg for two extreme choices of I and b.

Remark 6 (FedAvg: multiple local updates)**.** Choosing $\nu = 1$ in Theorem 3.2 implies $I = (T/b^3K^3)^{1/4}$ and $b = \mathcal{O}(1)$, we have

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{K^{1/2}T^{1/2}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2}{K^{1/2}T^{1/2}}\bigg) + \mathcal{O}\bigg(\frac{\zeta^2}{K^{1/2}T^{1/2}}\bigg),$$

while the sample and communication complexities are still $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-3/2})$, respectively. Note that these are the same guarantees for FedAvg analyzed in [14, 20]. □

Remark 7 (FedAvg: large batch)**.** Choosing $\nu = 0$ in Theorem 3.2 implies $I = \mathcal{O}(1) > 1$ (we allow multiple local updates, i.e. $I > 1$) and $b = (T/I^4K^3)^{1/3}$, then we have

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{T^{2/3}}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2}{T^{2/3}}\bigg) + \mathcal{O}\bigg(\frac{\zeta^2}{T^{2/3}}\bigg).$$

while the sample and communication complexities are again $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-3/2})$, respectively. □

Minibatch SGD: When the parameters are shared after each local update, for such case we have $I = 1$ and for the choice of $b = \mathcal{O}(T/K)$ we have:

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{T}\bigg) + \mathcal{O}\bigg(\frac{\sigma^2}{T}\bigg).$$

This implies that the sample and communication complexitiess are $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-1})$. Again, this result is independent of the heterogeniety parameter ζ (cf. Assumption 2) as the algorithm for $I = 1$ is essentially a centralized algorithm.

Next, we present the main result of the work presented in Theorem 3.1.

Algorithm 3 The Stochastic Two-Sided Momemtum (STEM) Algorithm

```
**Input**: Parameters: $c > 0$, the number of local updates $I$, batch size $b$, stepsizes $\{\eta_t\}$.
**Initialize**: Iterate $x_1^{(k)} = \bar{x}_1 = \frac{1}{K}\sum_{k=1}^K x_1^{(k)}$, descent direction $d_1^{(k)} = \bar{d}_1 = \frac{1}{K}\sum_{k=1}^K d_1^{(k)}$
    with $d_1^{(k)} = \frac{1}{B}\sum_{\xi_1^{(k)}\in\mathcal{B}_1^{(k)}} \nabla f^{(k)}(x_1^{(k)};\xi_1^{(k)})$ and $|\mathcal{B}_1^{(k)}| = B$ for $k \in [K]$.
Perform: $x_2^{(k)} = x_1^k - \eta_1 d_1^{(k)}, \ \forall\, k \in [K]$
**for** $t = 1$ to $T$ **do**
    **for** $k = 1$ to $K$ **do**                                                      #at the WN
        $d_{t+1}^{(k)} = \frac{1}{b}\sum_{\xi_{t+1}^{(k)}\in\mathcal{B}_{t+1}^{(k)}} \nabla f^{(k)}(x_{t+1}^{(k)};\xi_{t+1}^{(k)}) + (1-a_{t+1})\Big(d_t^{(k)} - \frac{1}{b}\sum_{\xi_{t+1}^{(k)}\in\mathcal{B}_{t+1}^{(k)}} \nabla f^{(k)}(x_t^{(k)};\xi_{t+1}^{(k)})\Big)$
    where we choose $|\mathcal{B}_{t+1}^{(k)}| = b$, and $a_{t+1} = c\cdot\eta_t^2$;
        **if** $t \bmod I = 0$ **then**                                               #at the SN
            $d_{t+1}^{(k)} = \bar{d}_{t+1} := \frac{1}{K}\sum_{k=1}^K d_{t+1}^{(k)}$
            $x_{t+2}^{(k)} := \bar{x}_{t+1} - \eta_{t+1}\bar{d}_{t+1} = \frac{1}{K}\sum_{k=1}^K x_{t+1}^{(k)} - \eta_{t+1}\bar{d}_{t+1}$   #server-side momentum
        **else** $x_{t+2}^{(k)} = x_{t+1}^{(k)} - \eta_{t+1}d_{t+1}^{(k)}$                  #worker-side momentum
        **end if**
    **end for**
**end for**
**Return:** $\bar{x}_a$ where $a \sim \mathcal{U}\{1,...,T\}$.
```

C Proofs of Convergence Guarantees for STEM

In this section we present the proofs of the convergence of STEM. First, we present some preliminary lemmas to be utilized throughout the proof. For reader's convenience here we restate the steps of the Algorithm 1 in Algorithm 3.

C.1 Preliminary Lemmas

Lemma C.1. *Define* $\bar{e}_t := \bar{d}_t - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})$*, then the iterates generated according to Algorithm 3 satisfy*

$$\mathbb{E}\Bigg[\Bigg\langle (1-a_t)\bar{e}_{t-1}, \frac{1}{K}\sum_{k=1}^K \frac{1}{b}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}} \bigg[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) \\ - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\bigg]\Bigg\rangle\Bigg] = 0,$$

where the expectation is w.r.t. the stochasticity of the algorithm.

Proof. Note that, given the filtration

$$\mathcal{F}_t = \sigma(x_1^{(k)}, x_2^{(k)}, \dots, x_t^{(k)}, d_1^{(k)}, d_2^{(k)}, \dots, d_{t-1}^{(k)} \text{ for all } k \in [K]),$$

the gradient error term, $\bar{e}_{t-1}$, is fixed. The only randomness in the left hand side of the statement of the Lemma is with respect to $\xi_t^{(k)}$, for all $k \in [K]$. This implies that we can write it as

$$\mathbb{E}\Bigg[\Bigg\langle (1-a_t)\bar{e}_{t-1}, \frac{1}{K}\sum_{k=1}^K \frac{1}{b}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}} \bigg[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) \\ - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\bigg]\Bigg\rangle\Bigg]$$

$$= \mathbb{E}\Bigg[\Bigg\langle (1-a_t)\bar{e}_{t-1}, \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Bigg[\frac{1}{b}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\Bigg[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Bigg]\Bigg|\mathcal{F}_t\Bigg]\Bigg\rangle\Bigg].$$

The result then follows from the fact that $\xi_t^{(k)}$ is chosen uniformly randomly at each $k \in [K]$, and we have from (Assumption 2) that: $\mathbb{E}\big[\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)})\big] = \nabla f^{(k)}(x_t^{(k)})$. This implies we have

$$\mathbb{E}\Bigg[\frac{1}{b}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\Big[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Big]\Bigg|\mathcal{F}_t\Bigg] = 0$$

for all $k \in [K]$.

Therefore the lemma is proved. □

Lemma C.2. *For $k, \ell \in [K]$ with $k \neq \ell$, the iterates generated according to Algorithm 3 satisfy*

$$\mathbb{E}\Bigg[\Bigg\langle \sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\Big[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Big],$$
$$\sum_{\xi_t^{(\ell)}\in\mathcal{B}_t^{(\ell)}}\Big[\Big(\nabla f^{(\ell)}(x_t^{(\ell)};\xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_t^{(\ell)})\Big) - (1-a_t)\Big(\nabla f^{(\ell)}(x_{t-1}^{(\ell)};\xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_{t-1}^{(\ell)})\Big)\Big]\Bigg\rangle\Bigg] = 0$$

Proof. Again note from the fact that conditioned on $\mathcal{F}_t$ the batches $\mathcal{B}_t^{(k)}$ and $\mathcal{B}_t^{(\ell)}$ for all $k, \ell \in [K]$ with $k \neq \ell$ across WNs are chosen independently of each other. Therefore, we have

$$\mathbb{E}\Bigg[\Bigg\langle \sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\Big[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Big],$$
$$\sum_{\xi_t^{(\ell)}\in\mathcal{B}_t^{(\ell)}}\Big[\Big(\nabla f^{(\ell)}(x_t^{(\ell)};\xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_t^{(\ell)})\Big) - (1-a_t)\Big(\nabla f^{(\ell)}(x_{t-1}^{(\ell)};\xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_{t-1}^{(\ell)})\Big)\Big]\Bigg\rangle\Bigg]$$
$$= \mathbb{E}\Bigg[\Bigg\langle \mathbb{E}\Bigg[\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\Big[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Big]\Bigg|\mathcal{F}_t\Bigg],$$
$$\mathbb{E}\Bigg[\sum_{\xi_t^{(\ell)}\in\mathcal{B}_t^{(\ell)}}\Big[\Big(\nabla f^{(\ell)}(x_t^{(\ell)};\xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_t^{(\ell)})\Big) - (1-a_t)\Big(\nabla f^{(\ell)}(x_{t-1}^{(\ell)};\xi_t^{(\ell)}) - \nabla f^{(\ell)}(x_{t-1}^{(\ell)})\Big)\Big]\Bigg|\mathcal{F}_t\Bigg]\Bigg\rangle\Bigg].$$

The result then follows from the fact that $\xi_t^{(k)}$ is chosen uniformly randomly across $k \in [K]$ and we have from the unbiased gradient Assumption 2 that: $\mathbb{E}\big[\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)})\big] = \nabla f^{(k)}(x_t^{(k)})$. This implies we have

$$\mathbb{E}\Bigg[\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\Big[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Big]\Bigg|\mathcal{F}_t\Bigg] = 0$$

for all $k \in [K]$.

Therefore, the lemma is proved. □

Lemma C.3. *For $\bar{e}_1 := \bar{d}_1 - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_1^{(k)})$ where $\bar{d}_1$ chosen according to Algorithm 3, we have:*

$$\mathbb{E}\|\bar{e}_1\|^2 \leq \frac{\sigma^2}{KB}.$$

Proof. The proof follows from an argument similar to that of Lemma B.1

□

Next, using the preliminary lemmas developed in this section we prove the main results of the work.

C.2 Proof of Main Results: STEM

In this section, we utilize the results developed in earlier sections to derive the main result of the paper presented in Section 3.1. Throughout the section we assume Assumptions 1 and 2 to hold. Before proceeding, we first define some notations.

We define $\bar{t}_s := sI + 1$ with $s \in [S]$. Note from Algorithm 3 that at $(s \times I)^{\text{th}}$ iteration, i.e., when $t \bmod I = 0$, the descent directions, $\{d_t^{(k)}\}_{k=1}^K$, corresponding to $t = (\bar{t}_s)^{\text{th}}$ time instant are shared with the SN. At the same time instant, the iterates, $\{x_t^{(k)}\}_{k=1}^K$ are also shared and the SN performs the "server side momentum step" (cf. Step 9 of Algorithm 3).

C.2.1 Proof of Descent Lemma

In the first step, we bound the error accumulation via the iterates generated by Algorithm 3.

Lemma C.4 (Error Accumulation from Iterates)**.** *For each $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ and $s \in [S]$, the iterates $x_t^{(k)}$ for each $k \in [K]$ generated from Algorithm 3 satisfy:*

$$\sum_{k=1}^K \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2 \leq (I-1) \sum_{\ell=\bar{t}_{s-1}}^{t} \eta_\ell^2 \sum_{k=1}^K \mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2,$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. Note from Algorithm 3 and the definition of $\bar{t}_s$ that at $t = \bar{t}_{s-1}$ with $s \in [S]$, $x_t^{(k)} = \bar{x}_t$, for all k. This implies

$$\sum_{k=1}^K \|x_{\bar{t}_{s-1}}^{(k)} - \bar{x}_{\bar{t}_{s-1}}\|^2 = 0.$$

Therefore, the statement of the lemma holds trivially. Moreover, for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$, with $s \in [S]$, we have from Algorithm 3: $x_t^{(k)} = x_{t-1}^{(k)} - \eta_{t-1} d_{t-1}^{(k)}$, this implies that:

$$x_t^{(k)} = x_{\bar{t}_{s-1}}^{(k)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta_\ell d_\ell^{(k)} \quad \text{and} \quad \bar{x}_t = \bar{x}_{\bar{t}_{s-1}} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta_\ell \bar{d}_\ell.$$

This implies that for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$, with $s \in [S]$ we have

$$\begin{aligned}
\sum_{k=1}^K \|x_t^{(k)} - \bar{x}_t\|^2 &= \sum_{k=1}^K \bigg\| x_{\bar{t}_{s-1}}^{(k)} - \bar{x}_{\bar{t}_{s-1}} - \Big(\sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta_\ell d_\ell^{(k)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta_\ell \bar{d}_\ell \Big) \bigg\|^2 \\
&\overset{(a)}{=} \sum_{k=1}^K \bigg\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \big(\eta_\ell d_\ell^{(k)} - \eta_\ell \bar{d}_\ell \big) \bigg\|^2 \\
&\overset{(b)}{\leq} (I-1) \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta_\ell^2 \sum_{k=1}^K \|d_\ell^{(k)} - \bar{d}_\ell\|^2 \\
&\leq (I-1) \sum_{\ell=\bar{t}_{s-1}}^{t} \eta_\ell^2 \sum_{k=1}^K \|d_\ell^{(k)} - \bar{d}_\ell\|^2,
\end{aligned}$$

where the equality (a) follows from the fact that $x_{\bar{t}_{s-1}}^{(k)} = \bar{x}_{\bar{t}_{s-1}}$ and inequality (b) uses the Lemma B.4 along with the fact that we have $d_t^{(k)} = \bar{d}_t$ for $t = \bar{t}_{s-1}$.

Taking expectation on both sides yields the statement of the lemma. □

Next, we utilize Lemma C.4 along with the smoothness of the function $f(\cdot)$ (Assumption 1) to show descent in the objective function value at consecutive iterates.

Lemma C.5 (Descent Lemma)**.** *With $\bar{e}_t := \bar{d}_t - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})$, for all $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ and $s \in [S]$, then the iterates generated by Algorithm 3 satisfy:*

$$\mathbb{E}f(\bar{x}_{t+1}) \leq \mathbb{E}f(\bar{x}_t) - \left(\frac{\eta_t}{2} - \frac{\eta_t^2 L}{2}\right)\mathbb{E}\|\bar{d}_t\|^2 - \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \eta_t \mathbb{E}\|\bar{e}_t\|^2 + \frac{\eta_t L^2(I-1)}{K}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^2 \sum_{k=1}^K \mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2,$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. Using the smoothness of f (Assumption 1) we have:

$$\begin{aligned}
f(\bar{x}_{t+1}) &\leq f(\bar{x}_t) + \langle \nabla f(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&\overset{(a)}{=} f(\bar{x}_t) - \eta_t\langle \nabla f(\bar{x}_t), \bar{d}_t\rangle + \frac{\eta_t^2 L}{2}\|\bar{d}_t\|^2 \\
&\overset{(b)}{=} f(\bar{x}_t) - \eta_t\|\bar{d}_t\|^2 + \eta_t\langle \bar{d}_t - \nabla f(\bar{x}_t), \bar{d}_t\rangle + \frac{\eta_t^2 L}{2}\|\bar{d}_t\|^2 \\
&\overset{(c)}{=} f(\bar{x}_t) - \left(\frac{\eta_t}{2} - \frac{\eta_t^2 L}{2}\right)\|\bar{d}_t\|^2 - \frac{\eta_t}{2}\|\nabla f(\bar{x}_t)\|^2 + \frac{\eta_t}{2}\|\bar{d}_t - \nabla f(\bar{x}_t)\|^2 \\
&\overset{(d)}{\leq} f(\bar{x}_t) - \left(\frac{\eta_t}{2} - \frac{\eta_t^2 L}{2}\right)\|\bar{d}_t\|^2 - \frac{\eta_t}{2}\|\nabla f(\bar{x}_t)\|^2 + \eta_t\left\|\bar{d}_t - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})\right\|^2 \\
&\qquad + \eta_t\left\|\frac{1}{K}\sum_{k=1}^K \left(\nabla f^{(k)}(x_t^{(k)}) - \nabla f^{(k)}(\bar{x}_t)\right)\right\|^2, \qquad (10)
\end{aligned}$$

where equality (a) follows from the iterate update given in Step 10 of Algorithm 3, (b) results by adding and subtracting $\bar{d}_t$ to $\nabla f(\bar{x}_t)$ in the inner product term and using the linearity of the inner product, (c) follows from the relation $\langle x, y\rangle = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x-y\|^2$, finally inequality (d) results from adding and subtracting $\frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})$ in the last term of (c) and using Lemma B.4.

Taking expectation on both sides and considering the last term of (10), we have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^K \left(\nabla f^{(k)}(x_t^{(k)}) - \nabla f^{(k)}(\bar{x}_t)\right)\right\|^2 &\leq \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left\|\nabla f^{(k)}(x_t^{(k)}) - \nabla f^{(k)}(\bar{x}_t)\right\|^2 \\
&\leq \frac{L^2}{K}\sum_{k=1}^K \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2, \qquad (11)
\end{aligned}$$

where the first inequality follows from Lemma B.4, and the second follows from the L-smoothness of $f^{(k)}(\cdot)$ (Assumption 1).

Substituting (11) in (10) and using the definition $\bar{e}_t := \bar{d}_t - \dfrac{1}{K}\displaystyle\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})$ we get:

$$\mathbb{E}f(\bar{x}_{t+1}) \leq \mathbb{E}f(\bar{x}_t) - \left(\frac{\eta_t}{2} - \frac{\eta_t^2 L}{2}\right)\mathbb{E}\|\bar{d}_t\|^2 - \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \eta_t\mathbb{E}\|\bar{e}_t\|^2 + \frac{\eta_t L^2}{K}\sum_{k=1}^K \mathbb{E}\|x_t^{(k)} - \bar{x}_t\|^2. \qquad (12)$$

Finally, using Lemma C.4 to bound the last term of (12), we get:

$$\mathbb{E}f(\bar{x}_{t+1}) \leq \mathbb{E}f(\bar{x}_t) - \left(\frac{\eta_t}{2} - \frac{\eta_t^2 L}{2}\right)\mathbb{E}\|\bar{d}_t\|^2 - \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \eta_t\mathbb{E}\|\bar{e}_t\|^2 + \frac{\eta_t L^2(I-1)}{K}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^2\sum_{k=1}^{K}\mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2.$$

Hence, the lemma is proved. □

Lemma C.5 shows that the expected descent in the function f depends on the magnitude of the expected gradient error term $\bar{e}_t$, and the expected gradient drift across WNs, i.e., $\mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2$. This implies that to ensure sufficient descent we need to control the gradient error, and the gradient drift across WNs. We achieve this by carefully designing the number of local updates, I, at each WN, and the batch-sizes b (and initial batch size B), that each WN uses to compute the descent direction.

Next, we present the error contraction lemma which analyzes how the term $\mathbb{E}\|\bar{e}_t\|^2$ contracts across time.

C.2.2 Proof of Gradient Error Contraction

Lemma C.6 (Gradient Error Contraction)**.** *Define $\bar{e}_t := \bar{d}_t - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})$, then for every $t \in [T]$ the iterates generated by Algorithm 3 satisfy*

$$\mathbb{E}\|\bar{e}_{t+1}\|^2 \leq (1-a_{t+1})^2\mathbb{E}\|\bar{e}_t\|^2 + \frac{8(1-a_{t+1})^2L^2}{bK^2}\frac{(I-1)}{I}\eta_t^2\sum_{k=1}^{K}\mathbb{E}\big\|d_t^{(k)} - \bar{d}_t\big\|^2 + \frac{4(1-a_{t+1})^2L^2\eta_t^2}{bK}\mathbb{E}\|\bar{d}_t\|^2 + \frac{2a_{t+1}^2\sigma^2}{bK},$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. Consider the error term $\|\bar{e}_t\|^2$ as

$$\mathbb{E}\|\bar{e}_t\|^2 = \mathbb{E}\bigg\|\bar{d}_t - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_t^{(k)})\bigg\|^2$$

$$\overset{(a)}{=} \mathbb{E}\bigg\|\frac{1}{K}\sum_{k=1}^{K}\frac{1}{b}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) + (1-a_t)\bigg(\bar{d}_{t-1} - \frac{1}{K}\sum_{k=1}^{K}\frac{1}{b}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)})\bigg) - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_t^{(k)})\bigg\|^2$$

$$\overset{(b)}{=} \mathbb{E}\bigg\|\frac{1}{K}\sum_{k=1}^{K}\frac{1}{b}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\bigg[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\bigg] + (1-a_t)\bar{e}_{t-1}\bigg\|^2,$$

where (a) follows from the definition of descent direction given in Step 6 of Algorithm 3; (b) follows by adding and subtracting $(1-a_t)\frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_{t-1}^{(k)})$ and using the definition of $\bar{e}_{t-1}$. Further simplifying the above expression, we get

$$\mathbb{E}\|\bar{e}_t\|^2 \overset{(c)}{=} (1-a_t)^2\mathbb{E}\|\bar{e}_{t-1}\|^2 + \frac{1}{b^2K^2}\mathbb{E}\bigg\|\sum_{k=1}^{K}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\bigg[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) - (1-a_t)\left(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\right)\bigg]\bigg\|^2$$

$$\begin{aligned}
&\overset{(d)}{=} (1-a_t)^2\mathbb{E}\|\bar{e}_{t-1}\|^2 + \frac{1}{b^2K^2}\sum_{k=1}^{K}\mathbb{E}\Big\|\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\Big[\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) \\
&\qquad\qquad\qquad\qquad - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Big]\Big\|^2, \\
&\overset{(e)}{=} (1-a_t)^2\mathbb{E}\|\bar{e}_{t-1}\|^2 + \frac{1}{b^2K^2}\sum_{k=1}^{K}\sum_{\xi_t^{(k)}\in\mathcal{B}_t^{(k)}}\mathbb{E}\Big\|\Big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\Big) \\
&\qquad\qquad\qquad\qquad - (1-a_t)\Big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\Big)\Big\|^2,
\end{aligned} \tag{13}$$

where (c) results from expanding the norm using inner product and noting that the cross terms are zero in expectation from Lemma C.1; (d) follows from expanding the norm using the inner products across $k \in [K]$ and noting that the cross term is zero in expectation from Lemma C.2; finally, (e) results from expanding the norm using the inner product across samples used to compute the minibatch gradients and the inner product is zero since at each node $k \in [K]$, the samples in the minibatch $\mathcal{B}_t^{(k)}$ are sampled independently of each other.

Now considering the 2nd term of (13) above, we have

$$\begin{aligned}
&\mathbb{E}\big\|\big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big) - (1-a_t)\big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\big)\big\|^2 \\
&\quad = \mathbb{E}\big\|(1-a_t)\big[\big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big) - \big(\nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\big)\big] \\
&\qquad\qquad\qquad + a_t\big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big)\big\|^2 \\
&\quad \overset{(a)}{\leq} 2(1-a_t)^2\mathbb{E}\big\|\big(\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)})\big) - \big(\nabla f^{(k)}(x_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)})\big)\big\|^2 \\
&\qquad\qquad\qquad + 2a_t^2\mathbb{E}\big\|\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_t^{(k)})\big\|^2 \\
&\quad \overset{(b)}{\leq} 2(1-a_t)^2\mathbb{E}\big\|\nabla f^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)};\xi_t^{(k)})\big\|^2 + 2a_t^2\sigma^2 \\
&\quad \overset{(c)}{\leq} 2(1-a_t)^2L^2\mathbb{E}\|x_t^{(k)} - x_{t-1}^{(k)}\|^2 + 2a_t^2\sigma^2 \\
&\quad \overset{(d)}{\leq} 2(1-a_t)^2L^2\eta_{t-1}^2\mathbb{E}\|d_{t-1}^{(k)}\|^2 + 2a_t^2\sigma^2 \\
&\quad \overset{(e)}{\leq} 8(1-a_t)^2L^2\frac{(I-1)}{I}\eta_{t-1}^2\mathbb{E}\|d_{t-1}^{(k)} - \bar{d}_{t-1}\|^2 + 4(1-a_t)^2L^2\eta_{t-1}^2\mathbb{E}\|\bar{d}_{t-1}\|^2 + 2a_t^2\sigma^2,
\end{aligned} \tag{14}$$

where (a) follows from Lemma B.4; (b) results from use of Assumption 2 and mean variance inequality: For a random variable Z we have $\mathbb{E}\|Z - \mathbb{E}[Z]\|^2 \leq \mathbb{E}\|Z\|^2$; (c) follows from the Lipschitz continuity of the gradient given in Assumption 1; (d) results from the iterate update equation given in Step 10 of Algorithm 3; finally, (e) uses the fact that: (i) for $I = 1$ we have $d_t^{(k)} = \bar{d}_t$ for all $t \in [T]$ and (ii) for $I \geq 2$ we use Lemma B.4 and the fact that $(I-1)/I \geq 1/2$.

Substituting (14) in (13) we get:

$$\begin{aligned}
\mathbb{E}\|\bar{e}_t\|^2 \leq (1-a_t)^2\mathbb{E}\|\bar{e}_{t-1}\|^2 &+ \frac{8(1-a_t)^2L^2}{bK^2}\frac{(I-1)}{I}\eta_{t-1}^2\sum_{k=1}^{K}\mathbb{E}\|d_{t-1}^{(k)} - \bar{d}_{t-1}\|^2 \\
&+ \frac{4(1-a_t)^2L^2\eta_{t-1}^2}{bK}\mathbb{E}\|\bar{d}_{t-1}\|^2 + \frac{2a_t^2\sigma^2}{bK}.
\end{aligned}$$

Finally, the lemma is proved by replacing t by $t+1$. □

Lemma C.6 shows that the gradient error contracts in each iteration. Next, we first define a potential function and then utilize Lemmas C.5 and C.6 to show descent in the potential function.

C.2.3 Descent in Potential Function

We define the potential function as a linear combination of the objective function and the gradient estimation error: $\bar{e}_t := \bar{d}_t - \frac{1}{K}\sum_{k=1}^K \nabla f^{(k)}(x_t^{(k)})$

$$\Phi_t := f(\bar{x}_t) + \frac{bK}{64L^2}\frac{\|\bar{e}_t\|^2}{\eta_{t-1}}. \tag{15}$$

Next, we characterize the descent in the potential function.

Lemma C.7 (Potential Function Descent). *For $\bar{t} \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ and for $\eta_t \leq \frac{1}{16LI}$ we have*

$$\mathbb{E}[\Phi_{\bar{t}+1} - \Phi_{\bar{t}_{s-1}}] \leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \left(\frac{7\eta_t}{16} - \frac{\eta_t^2 L}{2}\right)\mathbb{E}\|\bar{d}_t\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{\sigma^2 c^2}{32L^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3$$
$$+ \frac{33}{256K}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{k=1}^K \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. To get the descent on the the potential function, i.e. $\mathbb{E}[\Phi_{t+1} - \Phi_t]$, we first consider the term: $\dfrac{\mathbb{E}\|\bar{e}_{t+1}\|^2}{\eta_t} - \dfrac{\mathbb{E}\|\bar{e}_t\|^2}{\eta_{t-1}}$.

Using Lemma C.6 we get

$$\begin{aligned}
\frac{\mathbb{E}\|\bar{e}_{t+1}\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{e}_t\|^2}{\eta_{t-1}} &\leq \left[\frac{(1-a_{t+1})^2}{\eta_t} - \frac{1}{\eta_{t-1}}\right]\mathbb{E}\|\bar{e}_t\|^2 + \frac{8(1-a_{t+1})^2L^2}{bK^2}\frac{(I-1)}{I}\eta_t\sum_{k=1}^K \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \\
&\quad + \frac{4(1-a_{t+1})^2L^2\eta_t}{bK}\mathbb{E}\|\bar{d}_t\|^2 + \frac{2a_{t+1}^2\sigma^2}{\eta_t bK} \\
&\overset{(a)}{\leq} \left(\eta_t^{-1} - \eta_{t-1}^{-1} - c\eta_t\right)\mathbb{E}\|\bar{e}_t\|^2 + \frac{8L^2}{bK^2}\frac{(I-1)}{I}\eta_t\sum_{k=1}^K \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \\
&\quad + \frac{4L^2\eta_t}{bK}\mathbb{E}\|\bar{d}_t\|^2 + \frac{2\sigma^2c^2\eta_t^3}{bK},
\end{aligned} \tag{16}$$

where inequality (a) utilizes the fact that $(1-a_t)^2 \leq 1 - a_t \leq 1$ for all $t \in [T]$.

Let us consider $\eta_t^{-1} - \eta_{t-1}^{-1}$ in the first term of the inequality in (16) and using the definition of the stepsize η_t from Theorem 3.1, we have

$$\begin{aligned}
\eta_t^{-1} - \eta_{t-1}^{-1} &= \frac{(w_t + \sigma^2 t)^{1/3}}{\bar{\kappa}} - \frac{(w_{t-1} + \sigma^2(t-1))^{1/3}}{\bar{\kappa}} \\
&\overset{(a)}{\leq} \frac{(w_t + \sigma^2 t)^{1/3}}{\bar{\kappa}} - \frac{(w_t + \sigma^2(t-1))^{1/3}}{\bar{\kappa}} \\
&\overset{(b)}{\leq} \frac{\sigma^2}{3\bar{\kappa}(w_t + \sigma^2(t-1))^{2/3}} \\
&\overset{(c)}{\leq} \frac{2^{2/3}\sigma^2\bar{\kappa}^2}{3\bar{\kappa}^3(w_t + \sigma^2 t)^{2/3}} \\
&\overset{(d)}{=} \frac{2^{2/3}\sigma^2}{3\bar{\kappa}^3}\eta_t^2 \\
&\overset{(e)}{\leq} \frac{\sigma^2}{24\bar{\kappa}^3 LI}\eta_t,
\end{aligned} \tag{17}$$

where inequality (a) follows from the fact that we choose $w_t \leq w_{t-1}$ (see definition of w_t in Theorem 3.1), (b) results from the concavity of $x^{1/3}$ as:

$$(x+y)^{1/3} - x^{1/3} \leq \frac{y}{3x^{2/3}}.$$

In inequality (c), we have used the fact that $w_t \geq 2\sigma^2$, finally, (d) and (e) utilize the definition of η_t and the fact that $\eta_t \leq 1/16LI$ for all $t \in [T]$, respectively.

Now combining the first term of inequality in (16) with (17) and choosing $c = \frac{64L^2}{bK} + \frac{\sigma^2}{24\bar{\kappa}^3 LI}$ we get:

$$\eta_t^{-1} - \eta_{t-1}^{-1} - c\eta_t \leq -\frac{64L^2}{bK}\eta_t.$$

Therefore, we have from (16):

$$\begin{aligned}\frac{\mathbb{E}\|\bar{e}_{t+1}\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{e}_t\|^2}{\eta_{t-1}} &\leq -\frac{64L^2\eta_t}{bK}\mathbb{E}\|\bar{e}_t\|^2 + \frac{8L^2}{bK^2}\frac{(I-1)}{I}\eta_t\sum_{k=1}^K \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \\ &\qquad + \frac{4L^2\eta_t}{bK}\mathbb{E}\|\bar{d}_t\|^2 + \frac{2\sigma^2c^2\eta_t^3}{bK} \\ \frac{bK}{64L^2}\bigg(\frac{\mathbb{E}\|\bar{e}_{t+1}\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{e}_t\|^2}{\eta_{t-1}}\bigg) &\leq -\eta_t\mathbb{E}\|\bar{e}_t\|^2 + \frac{1}{8K}\frac{(I-1)}{I}\eta_t\sum_{k=1}^K \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 + \frac{\eta_t}{16}\mathbb{E}\|\bar{d}_t\|^2 + \frac{\sigma^2c^2\eta_t^3}{32L^2}.\end{aligned}$$

Finally, using Lemma C.5 and the definition of potential function given in (15), using the above we get the descent in the potential function for any $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ with $s \in [S]$ as:

$$\begin{aligned}\mathbb{E}[\Phi_{t+1} - \Phi_t] \leq &-\bigg(\frac{7\eta_t}{16} - \frac{\eta_t^2 L}{2}\bigg)\mathbb{E}\|\bar{d}_t\|^2 - \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{\eta_t L^2(I-1)}{K}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^2\sum_{k=1}^K\mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2 \\ &+ \frac{1}{8K}\frac{(I-1)}{I}\eta_t\sum_{k=1}^K\mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 + \frac{\sigma^2c^2\eta_t^3}{32L^2}.\end{aligned}$$

Summing the above over $t = \bar{t}_{s-1}$ to $\bar{t}$ for $\bar{t} \in [\bar{t}_{s-1}, \bar{t}_s - 1]$, we get:

$$\begin{aligned}\mathbb{E}[\Phi_{\bar{t}+1} - \Phi_{\bar{t}_{s-1}}] \leq &-\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\bigg(\frac{7\eta_t}{16} - \frac{\eta_t^2 L}{2}\bigg)\mathbb{E}\|\bar{d}_t\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}}\frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{\sigma^2c^2}{32L^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t^3 \\ &+ \frac{L^2(I-1)}{K}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^2\sum_{k=1}^K\mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2 + \frac{1}{8K}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t\sum_{k=1}^K\mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \\ \leq &-\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\bigg(\frac{7\eta_t}{16} - \frac{\eta_t^2 L}{2}\bigg)\mathbb{E}\|\bar{d}_t\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}}\frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{\sigma^2c^2}{32L^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t^3 \\ &+ \frac{L^2(I-1)}{K}\bigg(\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t\bigg)\bigg(\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}}\eta_\ell^2\sum_{k=1}^K\mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2\bigg) \\ &+ \frac{1}{8K}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t\sum_{k=1}^K\mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2.\end{aligned}$$

Finally, using the fact that we have: $\eta_t \leq 1/16LI$ for all $t \in [T]$, we get:

$$\begin{aligned}\mathbb{E}[\Phi_{\bar{t}+1} - \Phi_{\bar{t}_{s-1}}] \leq &-\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\bigg(\frac{7\eta_t}{16} - \frac{\eta_t^2 L}{2}\bigg)\mathbb{E}\|\bar{d}_t\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}}\frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{\sigma^2c^2}{32L^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t^3 \\ &+ \frac{L^2(I-1)}{K}\bigg(I \times \frac{1}{16LI} \times \frac{1}{16LI}\bigg)\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t\sum_{k=1}^K\mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \\ &+ \frac{1}{8K}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t\sum_{k=1}^K\mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2\end{aligned}$$

$$= - \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \left(\frac{7\eta_t}{16} - \frac{\eta_t^2 L}{2}\right) \mathbb{E}\|\bar{d}_t\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \frac{\eta_t}{2} \mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{\sigma^2 c^2}{32L^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3$$
$$+ \frac{33}{256K} \frac{(I-1)}{I} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2.$$

Therefore, the lemma is proved. □

Multiple local updates at each WN on heterogeneous data can cause the local descent directions to drift away from each other. Next, we bound this error accumulated via gradient drift across WNs.

C.2.4 Accumulated Gradient Consensus Error

We first upper bound the gradient consensus error given by term $\sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2$.

Lemma C.8 (Gradient Consensus Error)**.** *For every $t \in [T]$ and some $\beta > 0$ we have*

$$\sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \leq \left[(1-a_t)^2(1+\beta) + 4L^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2\right] \sum_{k=1}^{K} \mathbb{E}\|d_{t-1}^{(k)} - \bar{d}_{t-1}\|^2$$
$$+ 4KL^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 \mathbb{E}\|\bar{d}_{t-1}\|^2 + \frac{4K\sigma^2}{b}\left(1+\frac{1}{\beta}\right)a_t^2 + 8K\zeta^2\left(1+\frac{1}{\beta}\right)a_t^2$$
$$+32L^2\left(1+\frac{1}{\beta}\right)(I-1)a_t^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{t-1} \eta_{\bar{\ell}}^2 \sum_{k=1}^{K} \mathbb{E}\|d_{\bar{\ell}}^{(k)} - \bar{d}_{\bar{\ell}}\|^2.$$

where the expectation is w.r.t. the stochasticity of the algorithm.

Proof. Using the definition of the descent direction $d_t^{(k)}$ from Algorithm 3 we have

$$\sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \tag{18}$$
$$= \sum_{k=1}^{K} \mathbb{E}\Bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) + (1-a_t)\big(d_{t-1}^{(k)} - \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)})\big)$$
$$- \left(\frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_t^{(j)}; \xi_t^{(j)}) + (1-a_t)\big(\bar{d}_{t-1} - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)})\big)\right)\Bigg\|^2$$
$$= \sum_{k=1}^{K} \mathbb{E}\Bigg\| (1-a_t)\big(d_{t-1}^{(k)} - \bar{d}_{t-1}\big) + \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_t^{(j)}; \xi_t^{(j)})$$
$$- (1-a_t)\left(\frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)})\right)\Bigg\|^2$$
$$\overset{(a)}{\leq} (1+\beta)(1-a_t)^2 \sum_{k=1}^{K} \mathbb{E}\|d_{t-1}^{(k)} - \bar{d}_{t-1}\|^2$$
$$+ \left(1+\frac{1}{\beta}\right) \sum_{k=1}^{K} \mathbb{E}\Bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_t^{(j)}; \xi_t^{(j)})$$
$$- (1-a_t)\left(\frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)})\right)\Bigg\|^2 \tag{19}$$

where inequality (a) follows from the Young's inequality for some $\beta > 0$. Now considering the second term in (19), we get

$$
\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_t^{(j)}; \xi_t^{(j)}) \\
&\qquad\qquad - (1-a_t)\bigg(\frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg) \bigg\|^2 \\
&= \sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_t^{(j)}; \xi_t^{(j)}) \\
&\qquad\qquad - \bigg(\frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg) \\
&\qquad\qquad\qquad + a_t\bigg(\frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg) \bigg\|^2 \\
&\overset{(a)}{\leq} 2\sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_t^{(j)}; \xi_t^{(j)}) \\
&\qquad - \bigg(\frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg) \bigg\|^2 \\
&\qquad\qquad\qquad + 2a_t^2 \sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg\|^2 \\
&\overset{(b)}{\leq} 2\sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) \big) \bigg\|^2 \\
&\qquad\qquad\qquad + 2a_t^2 \sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg\|^2 \\
&\overset{(c)}{\leq} 2\sum_{k=1}^{K} \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \mathbb{E}\big\| \nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) \big\|^2 \\
&\qquad\qquad\qquad + 2a_t^2 \sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg\|^2 \\
&\overset{(d)}{\leq} 2L^2 \sum_{k=1}^{K} \mathbb{E}\| x_t^{(k)} - x_{t-1}^{(k)} \|^2 + 2a_t^2 \sum_{k=1}^{K} \mathbb{E}\bigg\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K}\sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \bigg\|^2,
\end{aligned} \tag{20}
$$

where inequality (a) above follows from Lemma B.4, (b) follows from Lemma B.2, inequality (c) again uses Lemma B.4 and (d) follows from the Lipschitz-smoothness of the individual functions $f^{(k)}$ (Assumption 1).

Now considering the second term in (20) above, we have

$$
\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\Big\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) \Big\|^2 \\
&\overset{(a)}{=} \sum_{k=1}^{K} \mathbb{E}\Big\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)}) \big) \\
&\qquad - \frac{1}{K} \sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \big(\nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) - \nabla f^{(j)}(x_{t-1}^{(j)}) \big) + \nabla f^{(k)}(x_{t-1}^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} \nabla f^{(j)}(x_{t-1}^{(j)}) \Big\|^2 \\
&\overset{(b)}{\leq} 2 \sum_{k=1}^{K} \mathbb{E}\Big\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)}) \big) \\
&\qquad\qquad - \frac{1}{K} \sum_{j=1}^{K} \frac{1}{b} \sum_{\xi_t^{(j)} \in \mathcal{B}_t^{(j)}} \big(\nabla f^{(j)}(x_{t-1}^{(j)}; \xi_t^{(j)}) - \nabla f^{(j)}(x_{t-1}^{(j)}) \big) \Big\|^2 \\
&\qquad\qquad\qquad + 2 \sum_{k=1}^{K} \mathbb{E}\Big\| \nabla f^{(k)}(x_{t-1}^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} \nabla f^{(j)}(x_{t-1}^{(j)}) \Big\|^2 \\
&\overset{(c)}{\leq} 2 \sum_{k=1}^{K} \mathbb{E}\Big\| \frac{1}{b} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \big(\nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)}) \big) \Big\|^2 \\
&\qquad\qquad + 2 \sum_{k=1}^{K} \mathbb{E}\Big\| \nabla f^{(k)}(x_{t-1}^{(k)}) - \frac{1}{K} \sum_{j=1}^{K} \nabla f^{(j)}(x_{t-1}^{(j)}) \Big\|^2 \\
&\overset{(d)}{\leq} 2 \sum_{k=1}^{K} \frac{1}{b^2} \sum_{\xi_t^{(k)} \in \mathcal{B}_t^{(k)}} \mathbb{E}\big\| \big(\nabla f^{(k)}(x_{t-1}^{(k)}; \xi_t^{(k)}) - \nabla f^{(k)}(x_{t-1}^{(k)}) \big) \big\|^2 + 4 \sum_{k=1}^{K} \mathbb{E}\big\| \nabla f^{(k)}(\bar{x}_{t-1}) - \nabla f(\bar{x}_{t-1}) \big\|^2 \\
&\qquad + 8 \sum_{k=1}^{K} \mathbb{E}\big\| \nabla f^{(k)}(x_{t-1}^{(k)}) - \nabla f^{(k)}(\bar{x}_{t-1}) \big\|^2 + 8 \sum_{k=1}^{K} \mathbb{E}\Big\| \nabla f(\bar{x}_{t-1}) - \frac{1}{K} \sum_{j=1}^{K} \nabla f^{(j)}(x_{t-1}^{(j)}) \Big\|^2 \\
&\overset{(e)}{\leq} \frac{2K\sigma^2}{b} + 4 \sum_{k=1}^{K} \frac{1}{K} \sum_{j=1}^{K} \mathbb{E}\| \nabla f^{(k)}(\bar{x}_{t-1}) - \nabla f^{(j)}(\bar{x}_{t-1}) \|^2 + 16L^2 \sum_{k=1}^{K} \mathbb{E}\| x_{t-1}^{(k)} - \bar{x}_{t-1} \|^2 \\
&\overset{(g)}{\leq} \frac{2K\sigma^2}{b} + 4K\zeta^2 + 16L^2 \sum_{k=1}^{K} \mathbb{E}\| x_{t-1}^{(k)} - \bar{x}_{t-1} \|^2, \qquad (21)
\end{aligned}
$$

where equality (a) follows from adding and subtracting $\nabla f^{(k)}(x_{t-1}^{(k)})$ and $\frac{1}{K}\sum_{j=1}^{K} \nabla f^{(j)}(x_{t-1}^{(j)})$ inside the norm; inequality (b) uses Lemma B.4; inequality (c) results from the use of Lemma B.2; inequality (d) expands the sum of the first term using inner products and utilizes the fact that the cross product terms are zero in expectation. This follows from the fact that conditioned on $\mathcal{F}_t$ we have $\mathbb{E}[\nabla f^{(k)}(x_t^{(k)}; \xi_t^{(k)})] = \nabla f^{(k)}(x_t^{(k)})$ for all $k \in [K]$ and $t \in [T]$; inequality (e) utilizes Intra-Node Variance Bound (Assumption 2), and Lemma B.4; finally, (g) follows from Inter-Node Variance Bound (Assumption 2).

Finally, substituting (21) and (20) in (19), we get

$$
\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}\| d_t^{(k)} - \bar{d}_t \|^2 \leq (1 - a_t)^2 (1 + \beta) \sum_{k=1}^{K} \mathbb{E}\big\| d_{t-1}^{(k)} - \bar{d}_{t-1} \big\|^2 + 2L^2 \Big(1 + \frac{1}{\beta} \Big) \sum_{k=1}^{K} \mathbb{E}\| x_t^{(k)} - x_{t-1}^{(k)} \|^2 \\
+ \frac{4K\sigma^2}{b} \Big(1 + \frac{1}{\beta} \Big) a_t^2 + 8K\zeta^2 \Big(1 + \frac{1}{\beta} \Big) a_t^2 + 32L^2 \Big(1 + \frac{1}{\beta} \Big) a_t^2 \sum_{k=1}^{K} \mathbb{E}\| x_{t-1}^{(k)} - \bar{x}_{t-1} \|^2
\end{aligned}
$$

$$\begin{aligned}
&\overset{(a)}{\leq} (1-a_t)^2(1+\beta)\sum_{k=1}^{K}\mathbb{E}\big\|d_{t-1}^{(k)}-\bar{d}_{t-1}\big\|^2 + 2L^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2\sum_{k=1}^{K}\mathbb{E}\|d_{t-1}^{(k)}\|^2 \\
&\qquad + \frac{4K\sigma^2}{b}\left(1+\frac{1}{\beta}\right)a_t^2 + 8K\zeta^2\left(1+\frac{1}{\beta}\right)a_t^2 \\
&\qquad\qquad + 32L^2\left(1+\frac{1}{\beta}\right)(I-1)a_t^2\sum_{\bar{\ell}=\bar{t}_{s-1}}^{t-1}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2 \\
&\overset{(b)}{\leq} (1-a_t)^2(1+\beta)\sum_{k=1}^{K}\mathbb{E}\big\|d_{t-1}^{(k)}-\bar{d}_{t-1}\big\|^2 + 4L^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2\sum_{k=1}^{K}\mathbb{E}\|d_{t-1}^{(k)}-\bar{d}_{t-1}\|^2 \\
&\qquad + 4L^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2\sum_{k=1}^{K}\mathbb{E}\|\bar{d}_{t-1}\|^2 + \frac{4K\sigma^2}{b}\left(1+\frac{1}{\beta}\right)a_t^2 + 8K\zeta^2\left(1+\frac{1}{\beta}\right)a_t^2 \\
&\qquad\qquad + 32L^2\left(1+\frac{1}{\beta}\right)(I-1)a_t^2\sum_{\bar{\ell}=\bar{t}_{s-1}}^{t-1}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2 \\
&= \left[(1-a_t)^2(1+\beta) + 4L^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2\right]\sum_{k=1}^{K}\mathbb{E}\|d_{t-1}^{(k)}-\bar{d}_{t-1}\|^2 \\
&\qquad + 4KL^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2\mathbb{E}\|\bar{d}_{t-1}\|^2 + \frac{4K\sigma^2}{b}\left(1+\frac{1}{\beta}\right)a_t^2 + 8K\zeta^2\left(1+\frac{1}{\beta}\right)a_t^2 \\
&\qquad\qquad + 32L^2\left(1+\frac{1}{\beta}\right)(I-1)a_t^2\sum_{\bar{\ell}=\bar{t}_{s-1}}^{t-1}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2,
\end{aligned}$$

where inequality (a) follows from the iterate update given in Step 10 of Algorithm 3 and inequality (b) utilizes Lemma B.4. □

Using the above Lemma C.8, we bound the accumulated gradient consensus error in the potential function's descent derived in Lemma C.7.

Lemma C.9 (Accumulated Gradient Consensus Error)**.** *For $\bar{t}\in[\bar{t}_{s-1},\bar{t}_s-1]$ with $s\in[S]$ we have*

$$\frac{33}{256K}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t\sum_{k=1}^{K}\mathbb{E}\|d_t^{(k)}-\bar{d}_t\|^2 \leq \sum_{t=\bar{t}_{s-1}}^{\bar{t}}\frac{\eta_t}{64}\mathbb{E}\|\bar{d}_t\|^2 + \frac{\sigma^2c^2}{64bL^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t^3 + \frac{\zeta^2c^2}{32L^2}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}}\eta_t^3.$$

Proof. First, from the statement of Lemma C.8, considering the coefficient of first term on the right hand side of the expression, we have:

$$\begin{aligned}
(1-a_t)^2(1+\beta) + 4L^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 &\overset{(a)}{\leq} 1+\beta+4L^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 \\
&\overset{(b)}{\leq} 1+\frac{1}{I}+4L^2(I+1)\eta_{t-1}^2 \\
&\overset{(c)}{\leq} 1+\frac{1}{I}+\frac{I+1}{64I^2} \\
&\overset{(d)}{\leq} 1+\frac{33}{32I},
\end{aligned}$$

where inequality (a) uses the fact that $(1-a_t)^2\leq 1$; the second inequality (b) follows from taking $\beta=1/I$, inequality (c) uses the bound $\eta_t\leq 1/16LI$ for all $t\in[T]$. Finally, the last inequality (d) results by using the fact that we have $I+1\leq 2I$. Substituting in the statement of Lemma C.8 above, we get

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\|d_t^{(k)}-\bar{d}_t\|^2 \le \left(1+\frac{33}{32I}\right)\sum_{k=1}^{K}\mathbb{E}\|d_{t-1}^{(k)}-\bar{d}_{t-1}\|^2 + 4KL^2\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2\mathbb{E}\|\bar{d}_{t-1}\|^2 + \frac{4K\sigma^2}{b}\left(1+\frac{1}{\beta}\right)a_t^2 \\ + 8K\zeta^2\left(1+\frac{1}{\beta}\right)a_t^2 + 32L^2\left(1+\frac{1}{\beta}\right)(I-1)a_t^2\sum_{\bar{\ell}=\bar{t}_{s-1}}^{t-1}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2. \\ \overset{(a)}{\le}\left(1+\frac{33}{32I}\right)\sum_{k=1}^{K}\mathbb{E}\|d_{t-1}^{(k)}-\bar{d}_{t-1}\|^2 + 8KL^2I\eta_{t-1}^2\mathbb{E}\|\bar{d}_{t-1}\|^2 + \frac{8KI\sigma^2}{b}c^2\eta_{t-1}^4 \\ + 16KI\zeta^2c^2\eta_{t-1}^4 + 64L^2I^2c^2\eta_{t-1}^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{t-1}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2 \\ \overset{(b)}{\le}\left(1+\frac{33}{32I}\right)\sum_{k=1}^{K}\mathbb{E}\|d_{t-1}^{(k)}-\bar{d}_{t-1}\|^2 + \frac{KL}{2}\eta_{t-1}\mathbb{E}\|\bar{d}_{t-1}\|^2 + \frac{K\sigma^2c^2}{2bL}\eta_{t-1}^3 \\ + \frac{K\zeta^2c^2}{L}\eta_{t-1}^3 + 64L^2I^2c^2\eta_{t-1}^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{t-1}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2\end{aligned} \tag{22}$$

where (a) follows from using $\beta = 1/I$, the fact that $I+1 \le 2I$ and the definition of a_t from Algorithm 3.

Note form Algorithm 3 that we have $d_t^{(k)} = \bar{d}_t$ for $t = \bar{t}_{s-1}$ with $s \in [S]$. This implies that for $t = \bar{t}_{s-1}$ with $s \in [S]$, we have, $\sum_{k=1}^{K}\|d_t^{(k)} - \bar{d}_t\|^2 = 0$. Applying (22) above recursively for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$ we get:

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\|d_t^{(k)}-\bar{d}_t\|^2 \le \frac{KL}{2}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\left(1+\frac{33}{32I}\right)^{t-1-\ell}\eta_\ell\mathbb{E}\|\bar{d}_\ell\|^2 + \frac{K\sigma^2c^2}{2bL}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\left(1+\frac{33}{32I}\right)^{t-1-\ell}\eta_\ell^3 \\ + \frac{K\zeta^2c^2}{L}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\left(1+\frac{3}{2I}\right)^{t-1-\ell}\eta_\ell^3 + 64L^2I^2c^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\left(1+\frac{33}{32I}\right)^{t-1-\ell}\eta_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell}\eta_{\bar{\ell}}^2\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2 \\ \overset{(a)}{\le}\frac{KL}{2}\left(1+\frac{33}{32I}\right)^{I}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell\mathbb{E}\|\bar{d}_\ell\|^2 + \frac{K\sigma^2c^2}{2bL}\left(1+\frac{33}{32I}\right)^{I}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^3 \\ + \frac{K\zeta^2c^2}{L}\left(1+\frac{33}{32I}\right)^{I}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^3 + 64L^2I^3c^2\left(\frac{1}{16LI}\right)^5\left(1+\frac{33}{32I}\right)^{I}\sum_{\bar{\ell}=\bar{t}_{s-1}}^{t}\eta_{\bar{\ell}}\sum_{k=1}^{K}\mathbb{E}\|d_{\bar{\ell}}^{(k)}-\bar{d}_{\bar{\ell}}\|^2 \\ \overset{(b)}{\le}\frac{3KL}{2}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell\mathbb{E}\|\bar{d}_\ell\|^2 + \frac{3K\sigma^2c^2}{2bL}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^3 + \frac{3K\zeta^2c^2}{L}\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell^3 \\ + 192L^2I^3c^2\left(\frac{1}{16LI}\right)^5\sum_{\ell=\bar{t}_{s-1}}^{t}\eta_\ell\sum_{k=1}^{K}\mathbb{E}\|d_\ell^{(k)}-\bar{d}_\ell\|^2,\end{aligned} \tag{23}$$

where inequality (a) follows from the fact that $1+33/32I > 1$ and $t-1-\ell \le I$ for $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ and $\ell \in [\bar{t}_{s-1}, t]$ and inequality (b) follows from the fact that $(1+33/32I)^I \le \mathrm{e}^{33/32} < 3$ and $\eta_t \le 1/16LI$ for all $t \in [T]$.

Multiplying (23) by η_t and summing over $t = \bar{t}_{s-1}$ to $\bar{t}$ for $\bar{t} \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ with $s \in [S]$

$$
\begin{aligned}
\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 &\leq \frac{3KL}{2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{\ell=\bar{t}_{s-1}}^{t} \eta_\ell \mathbb{E}\|\bar{d}_\ell\|^2 + \frac{3K\sigma^2c^2}{2bL} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{\ell=\bar{t}_{s-1}}^{t} \eta_\ell^3 \\
&\quad + \frac{3K\zeta^2c^2}{L} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{\ell=\bar{t}_{s-1}}^{t} \eta_\ell^3 + 192L^2I^3c^2\bigg(\frac{1}{16LI}\bigg)^5 \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{\ell=\bar{t}_{s-1}}^{t} \eta_\ell \sum_{k=1}^{K} \mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2 \\
&\overset{(a)}{\leq} \frac{3KL}{2}\bigg(\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t\bigg) \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}} \eta_\ell \mathbb{E}\|\bar{d}_\ell\|^2 + \frac{3K\sigma^2c^2}{2bL}\bigg(\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t\bigg) \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}} \eta_\ell^3 \\
&\quad + \frac{3K\zeta^2c^2}{L}\bigg(\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t\bigg) \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}} \eta_\ell^3 + 192L^2I^3c^2\bigg(\frac{1}{16LI}\bigg)^5\bigg(\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t\bigg) \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}} \eta_\ell \sum_{k=1}^{K} \mathbb{E}\|d_\ell^{(k)} - \bar{d}_\ell\|^2 \\
&\overset{(b)}{\leq} \frac{3K}{32} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \mathbb{E}\|\bar{d}_t\|^2 + \frac{3K\sigma^2c^2}{32bL^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 + \frac{3K\zeta^2c^2}{16L^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 \\
&\quad + 192L^2I^4c^2\bigg(\frac{1}{16LI}\bigg)^6 \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2
\end{aligned}
$$

where inequality (a) uses the fact that $t \in [\bar{t}_{s-1}, \bar{t}]$ and (b) follows from the fact that we have $\eta_t \leq 1/16LI$ for all $t \in [T]$. Rearranging the terms we get

$$
\begin{aligned}
\bigg[1 - 192L^2I^4c^2\bigg(\frac{1}{16LI}\bigg)^6\bigg] \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 &\leq \frac{3K}{32} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \mathbb{E}\|\bar{d}_t\|^2 \\
&\quad + \frac{3K\sigma^2c^2}{32bL^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 + \frac{3K\zeta^2c^2}{16L^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3
\end{aligned}
$$

using the fact that $c \leq 128L^2/bK$, $b \geq 1$, $K \geq 1$ and $I \geq 1$, we have $\Big[1 - 192L^2I^4c^2\Big(\frac{1}{16LI}\Big)^6\Big] \geq \frac{4}{5}$, therefore, we get

$$
\frac{33}{256K}\frac{(I-1)}{I} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t \sum_{k=1}^{K} \mathbb{E}\|d_t^{(k)} - \bar{d}_t\|^2 \leq \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \frac{\eta_t}{64} \mathbb{E}\|\bar{d}_t\|^2 + \frac{\sigma^2c^2}{64bL^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 + \frac{\zeta^2c^2}{32L^2}\frac{(I-1)}{I} \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3.
$$

Hence, the lemma is proved. □

C.2.5 Proof of Theorem 3.1

Next, to prove Theorem 3.1 we first prove an intermediate theorem by utilizing Lemmas C.9 and C.7 derived above.

Theorem C.10. *Choosing the parameters as*

(i) $\bar{\kappa} = \dfrac{(bK)^{2/3}\sigma^{2/3}}{L}$,

(ii) $c = \dfrac{64L^2}{bK} + \dfrac{\sigma^2}{24\bar{\kappa}^3LI} \overset{(i)}{=} L^2\bigg(\dfrac{64}{bK} + \dfrac{1}{24(bK)^2I}\bigg) \leq \dfrac{128L^2}{bK}$,

(iii) We choose $\{w_t\}_{t=0}^{T}$ *as*

$$
w_t = \max\bigg\{2\sigma^2, 4096L^3I^3\bar{\kappa}^3 - \sigma^2 t, \frac{c^3\bar{\kappa}^3}{4096L^3I^3}\bigg\} \overset{(i)(ii)}{\leq} \sigma^2 \max\bigg\{2,\ 4096I^3(bK)^2 - t,\ \frac{512}{bKI^3}\bigg\}.
$$

Moreover, for any number of local updates, $I \geq 1$, batch sizes, $b \geq 1$, and initial batch size, $B \geq 1$, computed at individual WNs, STEM satisfies:

$$\begin{aligned}\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 \leq &\left[\frac{32LI}{T} + \frac{2L}{(bK)^{2/3}T^{2/3}}\right](f(\bar{x}_1) - f^*) + \left[\frac{8bI^2}{BT} + \frac{bI}{2(bK)^{2/3}BT^{2/3}}\right]\sigma^2 \\ &+ \left[\frac{256^2 I}{T} + \frac{64^2}{(bK)^{2/3}T^{2/3}}\right]\sigma^2 \log(T+1) + \left[\frac{256^2 I}{T} + \frac{64^2}{(bK)^{2/3}T^{2/3}}\right]\zeta^2 \frac{(I-1)}{I}\log(T+1).\end{aligned}$$

Proof. Substituting the gradient consensus error derived in Lemma C.9 into the Potential function descent derived in Lemma C.7, we can write the descent of potential function for $\bar{t} \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ with $s \in [S]$ as:

$$\begin{aligned}\mathbb{E}[\Phi_{\bar{t}+1} - \Phi_{\bar{t}_{s-1}}] \leq & - \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \left(\frac{27\eta_t}{64} - \frac{\eta_t^2 L}{2}\right)\mathbb{E}\|\bar{d}_t\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 \\ & + \frac{c^2\sigma^2}{32L^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 + \frac{c^2\sigma^2}{64bL^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 + \frac{c^2\zeta^2}{32L^2}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 \\ & \overset{(a)}{\leq} - \sum_{t=\bar{t}_{s-1}}^{\bar{t}} \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{3c^2\sigma^2}{64L^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3 + \frac{c^2\zeta^2}{32L^2}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}} \eta_t^3.\end{aligned}$$

where (a) follows from the fact that $\eta_t \leq 1/16LI$ for all $t \in [T]$ and $b \geq 1$. Taking $\bar{t} = \bar{t}_s - 1 = sI$, the above expression can be written as:

$$\mathbb{E}[\Phi_{\bar{t}_s} - \Phi_{\bar{t}_{s-1}}] \leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{3c^2\sigma^2}{64L^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \eta_t^3 + \frac{c^2\zeta^2}{32L^2}\frac{(I-1)}{I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \eta_t^3.$$

Summing over all the restarts, i.e, $s \in [S]$, we get:

$$\mathbb{E}[\Phi_{\bar{t}_S} - \Phi_{\bar{t}_0}] \leq - \sum_{t=\bar{t}_0}^{\bar{t}_S - 1} \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{3c^2\sigma^2}{64L^2}\sum_{t=\bar{t}_0}^{\bar{t}_S - 1} \eta_t^3 + \frac{c^2\zeta^2}{32L^2}\frac{(I-1)}{I}\sum_{t=\bar{t}_0}^{\bar{t}_S - 1} \eta_t^3.$$

Assuming that $T = SI$, then from the definition of $\bar{t}_s$ that $\bar{t}_0 = 1$ and $\bar{t}_S = SI + 1 = T + 1$, we get

$$\begin{aligned}\sum_{t=1}^{T} \frac{\eta_t}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 & \leq \mathbb{E}[\Phi_1 - \Phi_{T+1}] + \frac{3c^2\sigma^2}{64L^2}\sum_{t=1}^{T}\eta_t^3 + \frac{c^2\zeta^2}{32L^2}\frac{(I-1)}{I}\sum_{t=1}^{T}\eta_t^3 \\ & \overset{(a)}{\leq} f(\bar{x}_1) - f^* + \frac{bK}{64L^2}\frac{\mathbb{E}\|\bar{e}_1\|^2}{\eta_0} + \frac{3c^2\sigma^2}{64L^2}\sum_{t=1}^{T}\eta_t^3 + \frac{c^2\zeta^2}{32L^2}\frac{(I-1)}{I}\sum_{t=1}^{T}\eta_t^3 \\ & \overset{(b)}{\leq} f(\bar{x}_1) - f^* + \frac{\sigma^2}{64L^2}\frac{b}{B\eta_0} + \frac{3c^2\sigma^2}{64L^2}\sum_{t=1}^{T}\eta_t^3 + \frac{c^2\zeta^2}{32L^2}\frac{(I-1)}{I}\sum_{t=1}^{T}\eta_t^3. \quad (24)\end{aligned}$$

where (a) follows from the fact that $f^* \leq \Phi_{T+1}$ and (b) results from application of Lemma C.3.

First, let us consider the last term of the (24) above, we have from the definition of the stepsize η_t

$$\begin{aligned}\sum_{t=1}^{T}\eta_t^3 & = \sum_{t=1}^{T}\frac{\bar{\kappa}^3}{w_t + \sigma^2 t} \\ & \overset{(a)}{\leq} \sum_{t=1}^{T}\frac{\bar{\kappa}^3}{\sigma^2 + \sigma^2 t} \\ & = \frac{\bar{\kappa}^3}{\sigma^2}\sum_{t=1}^{T}\frac{1}{1+t}\end{aligned}$$

$$\overset{(b)}{\leq} \frac{\bar{\kappa}^3}{\sigma^2} \ln(T+1). \tag{25}$$

where inequality (a) above follows from the fact that we have $w_t \geq 2\sigma^2 > \sigma^2$ and inequality (b) follows from the application of Lemma B.3.

Substituting (25) in (24), dividing both sides by T and using the fact that η_t is non-increasing in t we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(\bar{x}_t)\|^2 &\leq \frac{2(f(\bar{x}_1) - f^*)}{\eta_T T} + \frac{1}{\eta_T T}\frac{\sigma^2}{32L^2}\frac{b}{B\eta_0} + \frac{1}{\eta_T T}\frac{3c^2\bar{\kappa}^3}{32L^2}\log(T+1) \\ &\qquad + \frac{1}{\eta_T T}\frac{c^2\bar{\kappa}^3}{16L^2}\frac{\zeta^2}{\sigma^2}\frac{(I-1)}{I}\log(T+1) \\ &\overset{(a)}{\leq} \frac{2(f(\bar{x}_1) - f^*)}{\eta_T T} + \frac{1}{\eta_T T}\frac{\sigma^2}{32L^2}\frac{b}{B\eta_0} + \frac{1}{\eta_T T}\frac{c^2\bar{\kappa}^3}{4L^2}\log(T+1) \\ &\qquad + \frac{1}{\eta_T T}\frac{c^2\bar{\kappa}^3}{4L^2}\frac{\zeta^2}{\sigma^2}\frac{(I-1)}{I}\log(T+1). \end{aligned} \tag{26}$$

where (a) above utilizes the fact that $1/16 < 3/32 < 1/4$.

Now considering each term of (26) above separately and using the definition of $\eta_t = \dfrac{\bar{\kappa}}{(w_t + \sigma^2 t)^{1/3}}$ we get from the coefficient of the first term:

$$\frac{1}{\eta_T T} = \frac{(w_T + \sigma^2 T)^{1/3}}{\bar{\kappa} T} \overset{(a)}{\leq} \frac{w_T^{1/3}}{\bar{\kappa} T} + \frac{\sigma^{2/3}}{\bar{\kappa} T^{2/3}} \overset{(b)}{\leq} \frac{16LI}{T} + \frac{L}{(bK)^{2/3}T^{2/3}}. \tag{27}$$

where inequality (a) follows from identity $(x+y)^{1/3} \leq x^{1/3} + y^{1/3}$ and inequality (b) follows from the definition of $\bar{\kappa}$ and w_T

$$w_T = \max\left\{2\sigma^2, 4096L^3I^3\bar{\kappa}^3 - \sigma^2 T, \frac{c^3\bar{\kappa}^3}{4096L^3I^3}\right\} \leq \sigma^2 \max\left\{2,\ 4096I^3(bK)^2 - T,\ \frac{512}{bKI^3}\right\},$$

where we used $4096L^3I^3\bar{\kappa}^3 > 4096L^3I^3\bar{\kappa}^3 - \sigma^2 T \geq \max\left\{2\sigma^2, \dfrac{c^3\bar{\kappa}^3}{4096L^3I^3}\right\}$. Note that this choice of w_T captures the worst case guarantees for STEM.

Now, let us consider the second term of (26), we have from the definition of η_0 and η_T

$$\begin{aligned}\frac{1}{\eta_T T}\frac{\sigma^2}{32L^2}\frac{b}{B\eta_0} &\leq \left(\frac{16LI}{T} + \frac{L}{(bK)^{2/3}T^{2/3}}\right) \times \frac{\sigma^2}{32L^2} \times \frac{bw_0^{1/3}}{B\bar{\kappa}} \\ &\overset{(a)}{\leq} \left(\frac{16LI}{T} + \frac{L}{(bK)^{2/3}T^{2/3}}\right) \times \frac{\sigma^2}{32L^2} \times \frac{16LIb}{B} \\ &\overset{(b)}{\leq} \frac{8bI^2}{BT}\sigma^2 + \frac{bI}{(bK)^{2/3}BT^{2/3}}\frac{\sigma^2}{2}. \end{aligned} \tag{28}$$

where inequality (a) follows from the identity $(x+y)^{1/3} \leq x^{1/3} + y^{1/3}$ and (b) follows from the definition of $\bar{\kappa}$ and using $w_0 \leq 4096L^3I^3\bar{\kappa}^3$ and $w_T \leq 4096L^3I^3\bar{\kappa}^3$ (Similar to the approach in (27) this choice of w_0 and w_T capture the worst case convergence guarantees for STEM.)

Finally, considering the term $\frac{1}{\eta_T T}\frac{c^2\bar{\kappa}^3}{4L^2}$ common to the last two terms in (26) above, we have from the definition of the stepsize, η_t,

$$\begin{aligned}\frac{1}{\eta_T T}\frac{c^2\bar{\kappa}^3}{4L^2} &\leq \left(\frac{16LI}{T} + \frac{L}{(bK)^{2/3}T^{2/3}}\right) \times \left(\frac{128L^2}{bK}\right)^2 \times \frac{(bK)^2\sigma^2}{L^3} \times \frac{1}{4L^2} \\ &\overset{(a)}{\leq} 256^2\sigma^2\frac{I}{T} + 64^2\sigma^2\frac{1}{(bK)^{2/3}T^{2/3}}. \end{aligned} \tag{29}$$

where inequality (a) follows from the identity $(x+y)^{1/3} \leq x^{1/3} + y^{1/3}$ and (b) again uses $w_T \leq 4096L^3I^3\bar{\kappa}^3$ along with the definition of $\bar{\kappa}$ and c.

Finally, substituting the bounds obtained in (27), (28) and (29) into (26), we get

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 \leq \bigg[\frac{32LI}{T} + \frac{2L}{(bK)^{2/3}T^{2/3}}\bigg](f(\bar{x}_1) - f^*) + \bigg[\frac{8bI^2}{BT} + \frac{bI}{2(bK)^{2/3}BT^{2/3}}\bigg]\sigma^2$$
$$+ \bigg[\frac{256^2 I}{T} + \frac{64^2}{(bK)^{2/3}T^{2/3}}\bigg]\sigma^2 \log(T+1) + \bigg[\frac{256^2 I}{T} + \frac{64^2}{(bK)^{2/3}T^{2/3}}\bigg]\zeta^2 \frac{(I-1)}{I}\log(T+1).$$

Hence, the theorem is proved. □

Next, using Theorem C.10 we prove Theorem 3.1.

Theorem C.11 (Theorem 3.1: Trade-off: Local Updates vs Batch Sizes)**.** *With the parameters chosen according to Theorem C.10 and for any $\nu \in [0, 1]$ at each WN we set the total number of local updates as $I = \mathcal{O}\big((T/K^2)^{\nu/3}\big)$, batch size, $b = \mathcal{O}\big((T/K^2)^{1/2-\nu/2}\big)$, and the initial batch size, $B = bI$. Then STEM satisfies:*

(i) We have:

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{K^{2\nu/3}T^{1-\nu/3}}\bigg) + \tilde{\mathcal{O}}\bigg(\frac{\sigma^2}{K^{2\nu/3}T^{1-\nu/3}}\bigg) + \tilde{\mathcal{O}}\bigg(\frac{(I-1)}{I} \times \frac{\zeta^2}{K^{2\nu/3}T^{1-\nu/3}}\bigg).$$

(ii) Sample Complexity: To achieve an ϵ-stationary point STEM requires at most $\mathcal{O}(\epsilon^{-3/2})$ gradient computations. This implies that each WN requires at most $\mathcal{O}(K^{-1}\epsilon^{-3/2})$ gradient computations, thereby achieving linear speedup with the number of WNs present in the network.

(iii) Communication Complexity: To achieve an ϵ-stationary point STEM requires at most $\mathcal{O}(\epsilon^{-1})$ communication rounds.

Proof. The proof of statement (i) follows from the statement of Theorem C.10 and substituting the values of parameters B, I and b in the expression. First, replacing $B = bI$ in the statement of Theorem C.10 yields

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 \leq \bigg[\frac{32LI}{T} + \frac{2L}{(bK)^{2/3}T^{2/3}}\bigg](f(\bar{x}_1) - f^*) + \bigg[\frac{8I}{T} + \frac{1}{2(bK)^{2/3}T^{2/3}}\bigg]\sigma^2$$
$$+ \bigg[\frac{256^2 I}{T} + \frac{64^2}{(bK)^{2/3}T^{2/3}}\bigg]\sigma^2 \log(T+1) + \bigg[\frac{256^2 I}{T} + \frac{64^2}{(bK)^{2/3}T^{2/3}}\bigg]\zeta^2 \frac{(I-1)}{I}\log(T+1).$$

Then using the fact that $I = \mathcal{O}\big((T/K^2)^{\nu/3}\big)$ and $b = \mathcal{O}\big((T/K^2)^{1/2-\nu/2}\big)$ yields the expression of statement (i).

Next, we compute the computation and communication complexity of the algorithm.

- *Sample Complexity* [Theorem C.11(ii)]: From the statement of Theorem C.11(i), total iterations required to achieve an ϵ-stationary point are:

$$\tilde{\mathcal{O}}\bigg(\frac{1}{K^{2\nu/3}T^{1-\nu/3}}\bigg) = \epsilon \quad \Rightarrow \quad T = \tilde{\mathcal{O}}\bigg(\frac{1}{K^{2\nu/(3-\nu)}\epsilon^{3/(3-\nu)}}\bigg). \tag{30}$$

  In each iteration, each WN computes $2b$ stochastic gradients, therefore, the total gradient computations at each WN are $2bT$. Using $b = \mathcal{O}\big((T/K^2)^{1/2-\nu/2}\big)$, we get the total gradient computations required at each WN as:

$$bT = \tilde{\mathcal{O}}\bigg(\frac{T^{3/2-\nu/2}}{K^{1-\nu}}\bigg) \overset{(30)}{=} \tilde{\mathcal{O}}\bigg(\frac{1}{K\epsilon^{3/2}}\bigg)$$

  This implies that the sample complexity is $\tilde{\mathcal{O}}(\epsilon^{-3/2})$.

- *Communication Complexity* [Theorem C.11(iii)]: The total rounds of communication to achieve an ϵ-stationary point are T/I, with $I = \mathcal{O}\big((T/K^2)^{\nu/3}\big)$ and T given in (30), therefore, we have the communication complexity as:

$$\frac{T}{I} = \tilde{\mathcal{O}}\big(T^{1-\nu/3}K^{2\nu/3}\big) \overset{(30)}{=} \tilde{\mathcal{O}}\bigg(\frac{1}{\epsilon}\bigg).$$

Hence, the theorem is proved. □

Corollary 2 (Fed$\textsf{STEM}$: Local Updates)**.** *With the choice of parameters given in Theorem C.10. At each WN, setting constant batch size, $b \geq 1$, number of local updates, $I = (T/b^2K^2)^{1/3}$, and the initial batch size, $B = bI$. Then $\textsf{STEM}$ satisfies the following:*

(i) We have:

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{(bK)^{2/3}T^{2/3}}\bigg) + \tilde{\mathcal{O}}\bigg(\frac{\sigma^2}{(bK)^{2/3}T^{2/3}}\bigg) + \tilde{\mathcal{O}}\bigg(\frac{\zeta^2}{(bK)^{2/3}T^{2/3}}\bigg).$$

(ii) Sample Complexity: To achieve an ϵ-stationary point Fed$\textsf{STEM}$ requires at most $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ gradient computations while achieving linear speedup with the number of WNs.

(iii) Communication Complexity: To achieve an ϵ-stationary point Fed$\textsf{STEM}$ requires at most $\tilde{\mathcal{O}}(\epsilon^{-1})$ communication rounds.

Proof. The proof of statement (i) follows from substituting the values of the parameters b, I and B as defined in the statement of the Corollary in the statement of Theorem C.10.

Next, we compute the sample and communication complexity of the algorithm.

- *Sample Complexity:* From the statement of Corollary 2(i), total iterations, T, required to achieve an ϵ-stationary point are:

$$\tilde{\mathcal{O}}\bigg(\frac{1}{(bK)^{2/3}T^{2/3}}\bigg) = \epsilon \qquad \Rightarrow \qquad T = \tilde{\mathcal{O}}\bigg(\frac{1}{bK\epsilon^{3/2}}\bigg). \tag{31}$$

  At each iteration the algorithm computes $2b$ stochastic gradients. Therefore, the total number of gradient computations required at each WN are of the order of $2bT$, which is $\tilde{\mathcal{O}}(K^{-1}\epsilon^{-3/2})$. Therefore, the sample complexity of the algorithm is $\tilde{\mathcal{O}}(\epsilon^{-3/2})$.

- *Communication Complexity:* Total rounds of communication to achieve an ϵ-stationary point is T/I, therefore we have from the choice of I that

$$\frac{T}{I} = \tilde{\mathcal{O}}\big((bK)^{2/3}T^{2/3}\big) \overset{(31)}{=} \tilde{\mathcal{O}}\bigg(\frac{1}{\epsilon}\bigg).$$

Hence, the corollary is proved. □

An alternate design choice for the algorithm is to design large batch-size gradients and communicate more often. The next corollary captures this idea.

Corollary 3 (Corollary 1: Minibatch $\textsf{STEM}$)**.** *With the choice of parameters given in Theorem C.10. At each WN, choosing the number of local updates, $I = 1$, the batch size, $b = T^{1/2}/K$, and the initial batch size, $B = bI$. Then $\textsf{STEM}$ satisfies:*

(i) We have:

$$\mathbb{E}\|\nabla f(\bar{x}_a)\|^2 = \mathcal{O}\bigg(\frac{f(\bar{x}_1) - f^*}{T}\bigg) + \tilde{\mathcal{O}}\bigg(\frac{\sigma^2}{T}\bigg).$$

(ii) Sample Complexity: To achieve an ϵ-stationary point Minibatch $\textsf{STEM}$ requires at most $\tilde{\mathcal{O}}(\epsilon^{-3/2})$ gradient computations while achieving linear speedup with the number of WNs.

(iii) Communication Complexity: To achieve an ϵ-stationary point Minibatch $\textsf{STEM}$ requires at most $\tilde{\mathcal{O}}(\epsilon^{-1})$ communication rounds.

Proof. The proof of statement (i) follows from substituting the values of the parameters b, I and B given in the statement of the Corollary in the statement of Theorem C.10.

Next, we compute the sample and communication complexity of the algorithm.

- *Sample Complexity:* From the statement of Corollary 3(i), total iterations, T, required to achieve an ϵ-stationary point are:

  $$\tilde{\mathcal{O}}\bigg(\frac{I}{T}\bigg) = \epsilon \qquad \Rightarrow \qquad T = \tilde{\mathcal{O}}\bigg(\frac{I}{\epsilon}\bigg). \tag{32}$$

  In each iteration, each WN computes $2b$ stochastic gradients, therefore, the total gradient computations at each WN are $2bT$. Using the fact that $b = \mathcal{O}\bigg(\frac{T^{1/2}}{I^{3/2}K}\bigg)$. The total gradients computed at each WN to reach an ϵ-stationary point are:

  $$\tilde{\mathcal{O}}\bigg(\frac{I}{\epsilon} \times \frac{I^{1/2}}{\epsilon^{1/2}I^{3/2}K}\bigg) = \tilde{\mathcal{O}}\bigg(\frac{1}{K\epsilon^{3/2}}\bigg).$$

  Therefore, the communication complexity if $\tilde{\mathcal{O}}(\epsilon^{-3/2})$.

- *Communication Complexity:* The total rounds of communication required to reach an ϵ-stationary point are T/I, therefore we have

  $$\frac{T}{I} \overset{(32)}{=} \tilde{\mathcal{O}}\bigg(\frac{1}{\epsilon}\bigg).$$

Hence, the corollary is proved. □