

# Beyond Scaffold: A Unified Spatio-Temporal Gradient Tracking Method<sup>\*</sup>

Yan Huang<sup>†</sup> Jinming Xu<sup>‡</sup> Jiming Chen<sup>‡</sup>  
Karl Henrik Johansson<sup>†</sup>

<sup>†</sup> Division of Decision and Control Systems, School of EECS, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: [yahuang@kth.se](mailto:yahuang@kth.se), [kallej@kth.se](mailto:kallej@kth.se))

<sup>‡</sup> College of Control Science and Engineering, Zhejiang University, 310027 Hangzhou, China (e-mail: [jimmyxu@zju.edu.cn](mailto:jimmyxu@zju.edu.cn), [cjm@zju.edu.cn](mailto:cjm@zju.edu.cn))

**Abstract:** In distributed and federated learning algorithms, communication overhead is often reduced by performing multiple local updates between communication rounds. However, due to data heterogeneity across nodes and the local gradient noise within each node, this strategy can lead to the drift of local models away from the global optimum. To address this issue, we revisit the well-known federated learning method Scaffold (Karimireddy et al., 2020) under a gradient tracking perspective, and propose a unified spatio-temporal gradient tracking algorithm, termed ST-GT, for distributed stochastic optimization over time-varying graphs. ST-GT tracks the global gradient across neighboring nodes to mitigate data heterogeneity, while maintaining a running average of local gradients to substantially suppress noise, with slightly more storage overhead. Without assuming bounded data heterogeneity, we prove that ST-GT attains a linear convergence rate for strongly convex problems and a sublinear rate for nonconvex cases. Notably, ST-GT achieves the first linear speed-up in communication complexity with respect to the number of local updates per round  $\tau$  for the strongly-convex setting. Compared to traditional gradient tracking methods, ST-GT reduces the topology-dependent noise term from  $\sigma^2$  to  $\sigma^2/\tau$ , where  $\sigma^2$  denotes the noise level, thereby improving communication efficiency.

**Keywords:** Distributed optimization, federated learning, data heterogeneity.

## 1. INTRODUCTION

Data parallelism is a standard paradigm in large-scale machine learning tasks, in which multiple devices collaboratively train a shared model (Dean et al., 2012). Common strategies include employing a centralized server-worker architecture, as in federated learning (Li et al., 2014), where a server aggregates gradients and distributes model updates. Alternatively, a distributed network architecture allows each node to compute locally and exchange information with its neighbors (Yuan et al., 2016), making it well-suited for training over multiple data centers (Lian et al., 2017), wireless sensor networks (Rabbat and Nowak, 2004), and multi-robot systems (Tian et al., 2022). In general, both paradigms aim to solve the following distributed optimization problem with  $n$  nodes:

$$\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}_{\xi_i \sim \mathcal{D}_i} [f_i(x; \xi_i)]}_{:= f_i(x)}, \quad (1)$$

where  $x \in \mathbb{R}^p$  is the global model parameter,  $f_i$  the local objective function, and  $\xi_i$  the random sample drawn from the local data distribution  $\mathcal{D}_i$  accessible only to node  $i$ .

To reduce communication overhead between neighboring nodes or between nodes and the central server, it is common to skip certain communication rounds and perform multiple local updates in federated and distributed learning (McMahan et al., 2017; Nguyen et al., 2023). However, due to data heterogeneity across nodes and the sampling noise inherent in local gradients, reducing communication can substantially increase the drift of local models from the global gradient direction (Karimireddy et al., 2020) and amplify gradient noise (Huang and Xu, 2023), resulting in additional errors or requiring diminishing stepsizes. Addressing these issues while preserving communication efficiency remains an active research problem. Moreover, although distributed learning and federated learning share nearly identical objectives and challenges, existing studies often treat them separately, which hinders a unified understanding of data-parallel learning algorithms.

### 1.1 Related Work

**Federated learning.** Canonical federated learning methods adopt a server-worker architecture, where sampled worker nodes perform local computations in a data-parallel manner and then communicate with the server node to achieve information aggregation. A major issue with this approach is that the server node becomes a communication bottleneck and a potential single point of failure (Zhang et al., 2023). To reduce the communication load per round

\* The computations/data handling/[SIMILAR] were/was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

on the server, McMahan et al. (2017) proposed the FedAvg algorithm, which improves performance by allowing only a subset of nodes to participate in training and by incorporating multiple rounds of local updates. FedAvg has since been widely applied in many real-world scenarios, particularly for the case of identically and independently distributed (i.i.d.) data (Stich, 2019). However, Karimireddy et al. (2020) demonstrated that FedAvg fails to guarantee exact convergence under heterogeneous data distributions. The reason is that each worker may converge toward its own local optimum based solely on its dataset, resulting in client drift. To mitigate this issue, they proposed the Scaffold algorithm, which introduces control variables at both the server and worker sides to correct local gradient directions, thereby counteracting client drift. This method was later extended to various settings, including those with random communication intervals (Mishchenko et al., 2022), finite-sum problems with variance reduction (Jiang et al., 2024), and federated compositional optimization problems (Zhang et al., 2024), to name a few, achieving improved performance in non-i.i.d. scenarios. Another line of research for addressing data heterogeneity is personalized federated learning, which includes approaches such as regularization-based methods (Li et al., 2020) and knowledge distillation (Lee et al., 2022). For a comprehensive overview, we refer readers to recent surveys on this topic (Mora et al., 2024).

**Distributed learning.** In a distributed network without a central server, each node communicates only with its immediate neighbors (Ram et al., 2009). Such a topology is more flexible and robust against single-point failures. However, relying solely on peer-to-peer communication also makes distributed learning algorithms vulnerable to data heterogeneity, and this effect becomes increasingly pronounced as network connectivity weakens (Lian et al., 2017). To address this issue, gradient tracking (GT) algorithms have become a mainstream choice (Xu et al., 2015; Di Lorenzo and Scutari, 2016; Nedic et al., 2017). For instance, Xu et al. (2015) proposed Aug-DGM, which incorporates a global gradient estimator with a dynamic average-consensus protocol. This method progressively eliminates the effect of data heterogeneity and achieves linear convergence with a constant stepsize for strongly-convex and smooth objectives. Aug-DGM was later extended to stochastic gradient settings (Pu and Nedić, 2021) and further applied to various machine learning tasks (Lu and De Sa, 2021). To improve the communication efficiency, Nguyen et al. (2023) integrated GT with multiple local updates and proposed the LU-GT algorithm for deterministic settings. Wu et al. (2025) showed that LU-GT remains communication-efficient under mild data heterogeneity. However, simply skipping communication will, indeed, amplify the impact of gradient noise, thereby increasing computational complexity. Huang and Xu (2023) analyzed the trade-off between communication and computation costs with respect to the number of local updates, and proposed FlexGT, a flexible gradient tracking method for stochastic settings that supports adjustable computation and communication steps. Subsequently, Liu et al. (2024) introduced the K-GT algorithm, which employs a Scaffold-style control variable to correct local gradients and reduce gradient noise by scaling with the number of local updates; however, this improvement hinges on

initializing the control variables with global gradients and exhibits a stronger dependence on the network topology.

## 1.2 Contributions

In this work, we solve the distributed stochastic optimization problem (1) with non-independent and identically distributed (non-i.i.d.) datasets. The main contributions are summarized as follows:

- We propose a unified spatio-temporal gradient tracking algorithm (ST-GT) for time-varying graphs, applicable to both distributed and federated learning. By jointly tracking the global gradient across nodes and the running average of local gradients, ST-GT effectively mitigates data heterogeneity and reduces gradient noise, incurring only a slight increase in storage overhead. Furthermore, by extending Scaffold with a tunable parameter in its global control variable, we show that Scaffold emerges as a special case of ST-GT under a random communication topology, thereby offering a unified framework that bridges distributed and federated learning paradigms.
- Without assuming any bound on data heterogeneity, we prove that ST-GT achieves a linear convergence rate (Theorem 1) and a sublinear rate (Theorem 2) for strongly convex and nonconvex objective functions, respectively. More importantly, ST-GT achieves the first linear speed-up in communication complexity with respect to the number of local updates per round  $\tau$  in the strongly convex setting, improving upon FlexGT by a factor of  $1/\tau$ . It also reduces the network dependence from  $1/(1-\rho)^2$  to  $1/(1-\rho)^{3/2}$  compared with K-GT (see Table 1), where  $\rho$  denotes the graph connectivity. The theoretical results are validated on both synthetic and real-world datasets.

**Paper organization.** The remainder of the paper is organized as follows. Section 2 formulates the distributed stochastic optimization problem and introduces the design of ST-GT. Section 3 studies an extended Scaffold algorithm and establishes its connections with ST-GT. Section 4 presents the convergence results of ST-GT. Section 5 reports numerical experiments that validate the theoretical analysis. Finally, Section 6 concludes the paper, and Appendix A contains several supporting lemmas and the proofs of the main results.

**Notations.** In this work, we use the following notation:  $\|\cdot\|$  denotes the Frobenius norm,  $\langle \cdot, \cdot \rangle$  the inner product,  $|\cdot|$  the cardinality of a set, and  $\mathbb{E}[\cdot]$  the expectation of a vector or matrix. We let  $\mathbf{1}$  be the all-ones vector,  $\mathbf{I}$  the identity matrix, and define the averaging matrix as  $\mathbf{J} = \mathbf{1}\mathbf{1}^\top/n$ . In addition, the asymptotic notation  $\mathcal{O}(\cdot)$  is used to suppress constant factors, while  $\tilde{\mathcal{O}}(\cdot)$  further omits logarithmic factors.

## 2. PROBLEM FORMULATION AND ST-GT ALGORITHM

### 2.1 Distributed Stochastic Optimization

For the implementation purpose, we consider the following equivalent problem with consensus constraints:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times p}} F(X) &= \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}_{\xi_i \sim \mathcal{D}_i} [f_i(x_i; \xi_i)]}_{:= f_i(x_i)}, \\ \text{s.t. } x_i &= x_j, \quad i, j \in [n], \end{aligned} \quad (2)$$

where  $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$  is the collection of the local decision variable  $x_i \in \mathbb{R}^p$  of each node  $i$ . The nodes communicate over a network whose topology is represented by a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V} = 1, 2, \dots, n$  is the set of agents and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  is the set of edges, with each edge  $(i, j)$  indicating a communication link between agents  $i$  and  $j$ . Each agent  $i$  exchanges information only with its neighbors, defined as  $\mathcal{N}_i = j \mid j \neq i, (i, j) \in \mathcal{E}$ , together with itself. Particularly, to accommodate both gossip-based communication in distributed learning and partial client participation in federated learning, we consider a general dynamic graph  $\mathcal{G}_r$  at each communication round  $r$ , satisfying the following assumption.

*Assumption 1.* (Connectivity in expectation). The weight matrix  $W_r$  induced by a dynamic graph  $\mathcal{G}_r$  is doubly stochastic, i.e.,  $W_r \mathbf{1} = \mathbf{1}, \mathbf{1}^\top W_r = \mathbf{1}^\top$  and  $\rho := \mathbb{E} [\|W_r - \mathbf{J}\|_2^2] < 1, \forall r \geq 0$ .

Note that this assumption requires the underlying communication graph to satisfy a contraction property only in expectation, rather than at every round as assumed in Nguyen et al. (2023); Liu et al. (2024). This relaxation allows for more flexible and communication-efficient network topologies (Ying et al., 2021; Nguyen et al., 2025).

## 2.2 The ST-GT Algorithm

In this work, we address Problem (2) with non-i.i.d. local datasets, i.e.,  $\mathcal{D}_i \neq \mathcal{D}_j, i \neq j$ . For brevity, we denote

$$\begin{aligned} X_k &:= [\dots, x_{i,k}, \dots]^\top, Y_k := [\dots, y_{i,k}, \dots]^\top \in \mathbb{R}^{n \times p}, \\ G_k &:= [\dots, \nabla f_i(x_{i,k}; \xi_{i,k}), \dots]^\top \in \mathbb{R}^{n \times p}, \\ \nabla F_k &:= [\dots, \nabla f_i(x_{i,k}), \dots]^\top \in \mathbb{R}^{n \times p}. \end{aligned}$$

as the collections of the local model parameters, gradient tracking variables, stochastic gradient, and full gradient, respectively, at iteration  $k$ .

Recall the GT method with local updates (Nguyen et al., 2023; Huang and Xu, 2023; Wu et al., 2025), which follows the update rules given below:

$$\begin{aligned} X_{(r+1)\tau} &= W \left( X_{r\tau} - \gamma \sum_{t=0}^{\tau-1} Y_{r\tau+t} \right), \\ Y_{(r+1)\tau} &= W Y_{r\tau} + G_{(r+1)\tau} - G_{r\tau}, \end{aligned} \quad (3)$$

where  $W$  indicates a fixed weight matrix induced by the graph, and  $\tau$  is the number of local updates between communications. It can be observed that the update of the model parameter  $X$  relies on the accumulated tracking variables within the  $r$ -th round, whereas  $Y$  uses only the single-step values without aligning with the accumulated quantity in the updates of  $X$ . This mismatch is inconsistent with the principle of gradient tracking.

Motivated by these observations, we propose the spatio-temporal gradient tracking algorithm, ST-GT, whose pseudo-code is given in Algorithm 1. The key idea is illustrated in Fig. 1. By caching the model parameters from

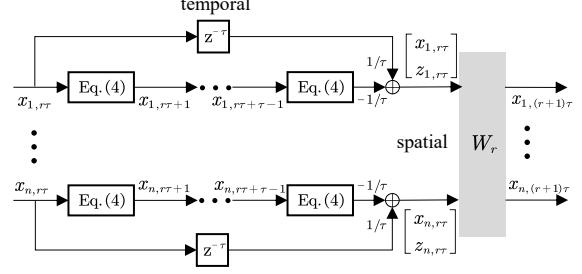


Fig. 1. Illustration of the spatio-temporal gradient tracking. The link from time  $r\tau$  to  $r\tau + \tau$ , passing through a memory element  $z^{-\tau}$ , illustrates how ST-GT tracks local gradient along the temporal dimension.

the previous communication round and passing them to the next communication round for difference calculation (the link from iteration  $r\tau$  to  $r\tau + \tau$ ), the algorithm accumulates the temporal tracking variable  $Z$ , which is then mixed spatially via the weighting matrix  $W_r$ . This spatio-temporal gradient tracking mechanism effectively approximates centralized gradient descent with only a slight increase in storage overhead, improving convergence performance while preserving communication efficiency. In particular, ST-GT can be rewritten in the following compact form: for  $k \in [r\tau, r\tau + \tau - 2]$ ,

$$\begin{aligned} X_{k+1} &= X_k - \gamma Y_k, \\ Y_{k+1} &= Y_k + G_{k+1} - G_k, \end{aligned} \quad (4)$$

and for the  $k = (r+1)\tau$ ,

$$\begin{aligned} Z_{r\tau} &= \frac{1}{\gamma\tau} (X_{r\tau} - X_{r\tau+\tau-1} + \gamma Y_{r\tau+\tau-1}), \\ X_{(r+1)\tau} &= W_r (X_{r\tau} - \tau\gamma Z_{r\tau}), \\ Y_{(r+1)\tau} &= W_r Z_{r\tau} + G_{(r+1)\tau} - \frac{1}{\tau} \sum_{t=0}^{\tau-1} G_{r\tau+t}. \end{aligned} \quad (5)$$

Intuitively, by applying the dynamic consensus protocol with the doubly stochastic matrix  $W_r$ , ST-GT asymptotically tracks the running average of the global gradient within each period, i.e.,

$$\begin{aligned} \bar{z}_{(r+1)\tau} &:= \frac{1}{n} \sum_{i=1}^n z_{i,(r+1)\tau} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{\tau} \sum_{t=0}^{\tau-1} y_{i,r\tau+t} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\tau} \sum_{t=0}^{\tau-1} g_{i,r\tau+t}. \end{aligned} \quad (6)$$

This enhances robustness to the gradient noise with only a slight increase in memory overhead, compared to GT methods with variance reduction (GT-VR) (Xin et al., 2020), which typically incur higher memory or computational costs to approximate the full gradient (see Table 1). Instead, DSGT (Pu and Nedić, 2021) and FlexGT (Huang and Xu, 2023) track the global gradient at a single step, i.e.,

$$\bar{y}_{(r+1)\tau} := \frac{1}{n} \sum_{i=1}^n y_{i,(r+1)\tau} = \frac{1}{n} \sum_{i=1}^n g_{i,(r+1)\tau}. \quad (7)$$

Particularly, ST-GT reduces to DSGT when  $\tau = 1$ , and to FlexGT when communicating  $Y$  instead of  $Z$ . Moreover, unlike K-GT (Liu et al., 2024), ST-GT does not require any extra communication during the initialization phase.

---

**Algorithm 1 ST-GT (distributed)**


---

**Initialization:** Initial points  $x_{i,0} \in \mathbb{R}^p$  and  $\tilde{g}_{i,0} = y_{i,0} = \nabla_x f_i(x_{i,0}; \xi_{i,0})$ , number of local updates  $\tau \geq 1$  and stepsize  $\gamma > 0$ .

- 1: **for** round  $r = 0, 1, \dots, R-1$ , each node  $i \in [n]$ , **do**
- 2:   Re-initial the running-average gradient  $\tilde{g}_{i,r\tau} = 0$
- 3:   **for**  $k = r\tau, r\tau+1, \dots, (r+1)\tau-2$  **do**
- 4:     Sample stochastic gradient  $g_{i,r\tau+t}$ .
- 5:     Perform local update:  

$$x_{i,k+1} = x_{i,k} - \gamma y_{i,k},$$

$$y_{i,k+1} = y_{i,k} + g_{i,k+1} - g_{i,k},$$

$$\tilde{g}_{i,k+1} = \tilde{g}_{i,k} + g_{i,k+1}.$$
- 6:   **end for**
- 7:    $z_{i,(r+1)\tau} = \frac{1}{\gamma\tau} (x_{i,r\tau} - x_{i,r\tau+\tau-1} + \gamma y_{i,r\tau+\tau-1}).$
- 8:   Perform inter-node communication:  

$$x_{i,(r+1)\tau} = \sum_{j \in \mathcal{N}_i} w_{i,j}^r (x_{j,r\tau} - \tau\gamma z_{j,(r+1)\tau}),$$

$$y_{i,(r+1)\tau} = \sum_{j \in \mathcal{N}_i} w_{i,j}^r z_{j,(r+1)\tau} + g_{i,(r+1)\tau} - \frac{1}{\tau} \tilde{g}_{i,(r+1)\tau}.$$
- 9: **end for**

---

**Algorithm 2 Scaffold<sup>+</sup> (server-worker)**


---

- 1: Initialization: server and worker model parameters  $x_0, x_{i,0}$ , global and local control variables  $c_0, c_{i,0}$ , step-sizes  $\gamma_g, \gamma_l$  and  $\gamma_c$ .
- 2: **for**  $r = 0, \dots, R-1$  **do**
- 3:   Sample worker nodes  $S_r \subseteq [1, \dots, n]$ .
- 4:   Send  $x_{r\tau}$  and  $c_{r\tau}$  to sampled nodes.
- 5:   **for**  $i \in S_r$  **do**
- 6:     initial the local mode  $x_{i,r\tau} = x_{r\tau}$ .
- 7:     **for**  $k = r\tau, r\tau+1, \dots, r\tau+\tau-2$  **do**
- 8:       Sample stochastic gradient  $g_{i,k}$ .
- 9:       Local updates  

$$x_{i,k+1} = x_{i,k} - \gamma_l (c_k + g_{i,k} - c_{i,k}),$$

$$x_{k+1} = x_k, c_{i,k+1} = c_{i,k}, c_{k+1} = c_k.$$
- 10:      **end for**
- 11:      For  $k = r\tau + \tau - 1$   

$$x_{i,k+1/2} = x_{i,k} - \gamma_l (c_k + g_{i,k} - c_{i,k}),$$

$$c_{i,k+1} = c_{i,k} - c_k + \frac{1}{\tau\gamma_l} (x_k - x_{i,k+1/2}).$$
- 12:     **end for**
- 13:     Communicate to server and update  

$$x_{k+1} = x_k + \frac{\gamma_g}{|S_r|} \sum_{j \in S_r} (x_{j,k+1/2} - x_k),$$

$$c_{k+1} = c_k + \frac{\gamma_c}{|S_r|} \sum_{j \in S_r} (c_{j,k+1} - c_{j,k}).$$
- 14: **end for**

---

### 3. CONNECTION TO SCAFFOLD

In this section, we investigate an extended version of the Scaffold algorithm (Karimireddy et al., 2020) and build its connection with ST-GT in the context of more general distributed and dynamic network settings.

#### 3.1 Extended Scaffold with Parameter $\gamma_c$

Scaffold is a popular federated learning algorithm designed to address data heterogeneity in non-i.i.d. settings (Karimireddy et al., 2020). It employs local and global control variables,  $c_i$  for each worker node and  $c$  for the server, to correct the client drift caused by data heterogeneity. Building on Scaffold, we propose an extended version in Algorithm 2, named Scaffold<sup>+</sup>, which introduces a tunable parameter  $\gamma_c \geq 0$  (cf., line 13) in that algorithm. This modification generalizes the original Scaffold (recovered when  $\gamma_c = |S_r|/n$ , where  $S_r \subseteq [1, \dots, n]$  denotes the set of sampled workers at each round  $r$ ) and establishes a connection to ST-GT. In particular, define

$$y_{i,k} := c_k + g_{i,k} - c_{i,k},$$

which serves as the gradient tracking variable in ST-GT. We show that Scaffold<sup>+</sup> can be interpreted as a spatio-temporal gradient tracking method with dynamic networks in the following proposition.

*Proposition 1.* Consider Scaffold<sup>+</sup> in Algorithm 2. For each node  $i \in S_r$  at the  $r$ -th communication round, we have  $x_{i,r\tau} = x_{r\tau}$  and  $c_{i,r\tau} = c_{i,t(i)}$ , where  $t(i) \leq r\tau$  denotes the most recent round at which node  $i$  was sampled. Then, for iteration  $k+1 = r\tau$ ,

$$\begin{aligned} x_{i,k+1} &= (1 - \gamma_g) x_{(r-1)\tau} \\ &\quad + \frac{\gamma_g}{|S_{r-1}|} \sum_{j \in S_{r-1}} \left( x_{j,(r-1)\tau} - \gamma_l \sum_{t=0}^{\tau-1} y_{j,k-t} \right), \\ y_{i,k+1} &= (1 - \gamma_c) c_{(r-1)\tau} \\ &\quad + \frac{\gamma_c}{|S_{r-1}|} \sum_{j \in S_{r-1}} \frac{1}{\tau} \sum_{t=0}^{\tau-1} y_{j,k-t} + g_{i,k+1} - c_{i,t(i)}; \end{aligned} \tag{8}$$

for  $k+1 = r\tau+1, \dots, r\tau+\tau-1$ ,

$$\begin{aligned} x_{i,k+1} &= x_{i,k} - \gamma_l y_{i,k}, \\ y_{i,k+1} &= y_{i,k} + g_{i,k+1} - g_{i,k}. \end{aligned} \tag{9}$$

And, for  $k+1 = r\tau+\tau$ ,

$$c_{i,r\tau+\tau} = \frac{1}{\tau} \sum_{t=0}^{\tau-1} g_{i,r\tau+t} \tag{10}$$

Moreover, for unsampled nodes  $i \notin S_r$ , all associated variables remain unchanged.

*Proof.* At the beginning of the  $r$ -th round with  $k+1 = r\tau$ , each node  $i \in S_r$  receives information from the server, which is updated based on the information uploaded by the nodes sampled in  $S_{r-1}$ , i.e.,

$$\begin{aligned} x_{i,k+1} &= x_{k+1} \\ &= x_k + \frac{\gamma_g}{|S_{r-1}|} \sum_{j \in S_{r-1}} (x_{j,k+1/2} - x_k) \\ &= (1 - \gamma_g) x_{(r-1)\tau} \\ &\quad + \frac{\gamma_g}{|S_{r-1}|} \sum_{j \in S_{r-1}} \left( x_{j,(r-1)\tau} - \gamma_l \sum_{t=0}^{\tau-1} y_{j,k-t} \right) \\ &= \frac{1}{|S_{r-1}|} \sum_{j \in S_{r-1}} \left( x_{j,(r-1)\tau} - \gamma_g \gamma_l \sum_{t=0}^{\tau-1} y_{j,k-t} \right). \end{aligned} \tag{11}$$

Table 1. Relevant algorithms for solving Problem (2) with strongly-convex (SC) and nonconvex (NC) objective functions. We take the size of the model parameters as one unit and compare the overhead of related distributed optimization algorithms in terms of per-node computational (Comp.), communication (Comm.), and memory costs, as well as their communication complexity to achieve an accuracy of  $\epsilon > 0$ . Here,  $m$  denotes the total number of local samples at each node. ‘s/w’ denotes the server–worker architecture, and ‘dist.’ the distributed graph.

Algorithm	Graph	Comp.	Comm.	Memory	Complexity ( $\tilde{\mathcal{O}}(\cdot)$ )	Assum.
Scaffold (Karimireddy et al., 2020) <sup>a</sup>	s/w	1x	2x	6x	$\frac{L}{\mu} + \frac{n}{s} + \frac{\sigma^2}{\mu^2 \tau n \epsilon} \frac{n}{s}$	SC
					$\frac{\sigma^2}{\tau \epsilon^2 n} \frac{n}{s} + \frac{1}{\epsilon} \left( \frac{n}{s} \right)^{2/3}$	NC
FlexGT (Huang and Xu, 2023) <sup>b</sup>	dist.	1x	2x	4x	$\frac{L}{(1-\rho)^2 \mu} + \frac{\sigma^2 / \tau}{\mu^2 n \epsilon} + \frac{\sqrt{L \sigma^2}}{\sqrt{\mu^3 (1-\rho)^3 \epsilon}}$	SC
K-GT (Liu et al., 2024)	dist.	1x	2x	5x	$\frac{L}{(1-\rho)^2 \epsilon} + \frac{L \sigma^2}{\tau n \epsilon^2} + \frac{L \sigma}{(1-\rho)^2 \sqrt{\tau \epsilon^3}}$	NC
GT-VR (Xin et al., 2020) <sup>c</sup>	dist.	{1, m}x	2x	{4+m, 4}x	$\max \left\{ m, \frac{L}{\mu (1-\rho)^2} \right\} \log \frac{1}{\epsilon}$	SC
ST-GT (This work)	dist.	1x	2x	5x	$\frac{L}{(1-\rho)^2 \mu} + \frac{\sigma^2}{\mu^2 n \tau \epsilon} + \sqrt{\frac{L \sigma^2}{\mu^3 (1-\rho)^3 \tau \epsilon}}$ $\frac{L}{(1-\rho)^2 \epsilon} + \frac{\sigma^2 L}{n \tau \epsilon^2} + \frac{L \sqrt{\sigma^2 / \tau}}{\sqrt{(1-\rho)^3 \epsilon^3}} + \frac{C_0}{\epsilon}$	SC NC

<sup>a</sup> Note that  $s/n$  can be interpreted as a measure of connectivity in the server-worker topology. However, it is not directly comparable to  $1 - \rho$ , as obtaining a closed-form expression for  $\mathbb{E} [\|W_r - \mathbf{J}\|^2]$  with  $W_r$  defined in (14) is generally intractable.

<sup>b</sup> This rate corresponds to the case where no multi-round communication is employed, consistent with the setting considered in this paper.

<sup>c</sup> This rate incurs higher memory or computational costs to approximate the full gradient and is obtained under the assumption of sample-wise smoothness of the objective function.

At the end of the  $r$ -th round with  $k+1 = r\tau + \tau$ , the local control variable will be updated, i.e.,

$$\begin{aligned}
 c_{i,k+1} &= c_{i,k} - c_k + \frac{1}{\tau \gamma_l} (x_k - x_{i,k+1/2}) \\
 &= c_{i,k} - c_k + \frac{1}{\tau \gamma_l} (x_k - x_{i,k} + \gamma_l y_{i,k}) \\
 &= c_{i,k} - c_k + x_{r\tau} - x_{i,r\tau} + \frac{1}{\tau \gamma_l} \sum_{t=0}^{\tau-1} \gamma_l y_{i,k-t} \\
 &= \frac{1}{\tau} \sum_{t=0}^{\tau-1} (c_{i,k-t} - c_{k-t} + y_{i,k-t}) = \frac{1}{\tau} \sum_{t=0}^{\tau-1} g_{i,r\tau+t},
 \end{aligned} \tag{12}$$

where we used the facts that  $x_{i,r\tau} = x_{r\tau} = \dots = x_{r\tau+\tau-1}$ ,  $c_{i,k-\tau+1} = \dots = c_{i,k}$  and  $c_{k-\tau+1} = \dots = c_k$ . This equation shows that the local control variable equals the running average of the local stochastic gradient within a period.

Then, for the global control variable, we have

$$\begin{aligned}
 c_{r\tau} &= c_k + \frac{\gamma_c}{|S_{r-1}|} \sum_{j \in S_{r-1}} (c_{j,k+1} - c_{j,k}) \\
 &= (1 - \gamma_c) c_k + \frac{\gamma_c}{|S_{r-1}|} \sum_{j \in S_{r-1}} \frac{1}{\tau} \sum_{t=0}^{\tau-1} y_{j,(r-1)\tau+t}.
 \end{aligned} \tag{13}$$

During the local updates phase, it is easy to get  $y_{i,k+1} = y_{i,k} + g_{i,k+1} - g_{i,k}$ .  $\square$

### 3.2 Connection between Scaffold<sup>+</sup> and ST-GT

Based on Proposition 1, we establish the connections between Scaffold<sup>+</sup> and ST-GT from the perspective of random communication topology and gradient approximation scheme, respectively. To simplify notations, we assume  $s = |S_r| \leq n$  for all  $r \in [R-1]$ .

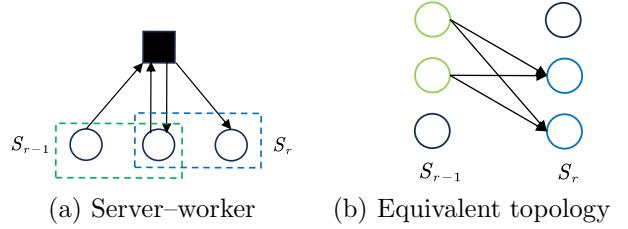


Fig. 2. Illustration of the equivalent topology of Scaffold<sup>+</sup> under the random network perspective. Black solid rectangles represent server nodes, while circles represent worker nodes. Information is transmitted from nodes sampled at round  $r-1$  (green box) to nodes sampled at time  $r$  (blue box).

**Random network perspective.** Intuitively, the server in Scaffold<sup>+</sup> acts as a relay node, transmitting information from the workers in  $S_{r-1}$  to those in  $S_r$ . The communication protocol can be described from the perspective of a bipartite graph, as illustrated in Fig. 2. In particular, according to Proposition 1 with  $\gamma_c = 1$  and noticing that the variables of the unsampled workers are unchanged, we get an equivalent weight matrix at round  $r$  as follows:

$$W_r = \frac{1}{s} \mathbf{e}_{S_r} \mathbf{e}_{S_{r-1}}^\top + \mathbf{I} - \text{diag}(\mathbf{e}_{S_r}), \tag{14}$$

where  $\mathbf{e}_{S_r}$  is an  $n$ -dimensional column vector where the index of the sampled nodes  $S_r$  is 1, and all other elements are 0. Note that  $W_r$  is row-stochastic. Assuming that workers are sampled uniformly and independently at each round, the expectation of  $W_r$  is

$$\mathbb{E}[W_r] = \frac{s}{n} \mathbf{J} + \frac{n-s}{n} \mathbf{I}, \tag{15}$$

which corresponds to a fully connected graph with spectral gap  $1 - s/n$ . This random network perspective provides

a foundation for the unified analysis of algorithms under server-worker architectures with partial participation and distributed topologies.

**Gradient tracking perspective.** As shown in (8), setting  $\gamma_c = 1$ , Scaffold<sup>+</sup> has the same spatio-temporal gradient-tracking scheme as that of ST-GT. Particularly, for  $i \in S_r$ , we have

$$y_{i,r\tau} = \frac{1}{s} \sum_{j \in S_{r-1}} \frac{1}{\tau} \sum_{t=0}^{\tau-1} y_{j,(r-1)\tau+t} + g_{i,r\tau} - c_{i,t(i)}; \quad (16)$$

and for  $i \notin S_r$ ,  $y_{i,r\tau} = y_{i,r\tau-1} = \dots = y_{i,(r-1)\tau}$ . By proper initialization, we have

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}[y_{i,r\tau+t}] = \frac{s}{n} \frac{1}{n} \sum_{i=1}^n \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}[g_{i,r\tau+t}], \quad (17)$$

illustrating the similar gradient tracking property as in ST-GT. Note that the contribution from unsampled nodes is slightly different compared to ST-GT. Instead, by setting  $\gamma_c = s/n$  as in the original Scaffold (cf. line 13 in Algorithm 2) and noting that  $c_k$  stores the average of the most recent  $c_{i,k}$  values and is updated every  $\tau$  iterations, we obtain

$$c_{r\tau} = \frac{1^\top}{n} \left( \begin{bmatrix} c_{1,t(1)} \\ \vdots \\ c_{n,t(n)} \end{bmatrix} + \begin{bmatrix} & & \\ & \ddots & \\ c_{j,r\tau} - c_{j,t(j)} & & \end{bmatrix}_{j \in S_{r-1}} \right), \quad (18)$$

which serves as the same role as the full gradient table in SAGA (Defazio et al., 2014).

These observations demonstrate that Scaffold for federated learning can be interpreted as a spatio-temporal gradient tracking method operating over random networks. In particular, when  $s = n$ , Scaffold becomes exactly equivalent to ST-GT with  $W_r = \mathbf{J}$ . This perspective further offers a unified framework for understanding other federated learning algorithms, such as FedAvg.

## 4. CONVERGENCE RESULTS

In this section, we establish the convergence properties of ST-GT under several standard assumptions on the objective functions and their gradients.

### 4.1 Assumptions

*Assumption 2.* (Convexity). Each  $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex, i.e., for any  $x, x' \in \mathbb{R}^p$ , there exists a constant  $\mu > 0$  such that

$$\langle \nabla f_i(x) - \nabla f_i(x'), x - x' \rangle \geq \mu \|x - x'\|^2. \quad (19)$$

*Assumption 3.* (Smoothness). Each  $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$  is  $L$ -smooth, i.e., for any  $x, x' \in \mathbb{R}^p$ , there exists a constant  $L > 0$  such that

$$\|\nabla f_i(x) - \nabla f_i(x')\| \leq L \|x - x'\|. \quad (20)$$

*Assumption 4.* (Bounded variance). For each node  $i$ , the stochastic gradient is unbiased, i.e.,  $\mathbb{E}_{\xi_i \sim \mathcal{D}_i} [\nabla f_i(x; \xi_i)] = \nabla f_i(x), \forall x \in \mathbb{R}^p$ , and there exists a constant  $\sigma \geq 0$  such that

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i} [\|\nabla f_i(x; \xi_i) - \nabla f_i(x)\|^2] \leq \sigma^2. \quad (21)$$

### 4.2 Strongly convex Case

We are now ready to give the convergence results of the ST-GT algorithm. To this end, we first define a Lyapunov function as follows:

$$V_k = \|\bar{x}_k - x^*\|^2 + c_x \|X_k - \mathbf{1}\bar{x}_k\|^2 + c_y \|\Upsilon_k\|^2, \quad (22)$$

where

$$\Upsilon_k := Y_k - G_k - \nabla f(\mathbf{1}\bar{x}_k) + \nabla F(\mathbf{1}\bar{x}_k), \quad (23)$$

and the coefficients  $c_x$  and  $c_y$  are designed as:

$$c_x = \frac{80\gamma\tau L}{n(1-\rho)}, \quad c_y = \frac{3556\gamma^3\tau^3 L}{n(1-\rho)^3}. \quad (24)$$

Then, for the strongly convex objective functions, we have the following theorem.

*Theorem 1.* (Strongly convex case). Suppose Assumptions 1–4 hold. Let the stepsize  $\gamma = \mathcal{O}\left(\frac{(1-\rho)^2}{\tau L}\right)$ . Then, we have

$$\begin{aligned} \mathbb{E} [\|V_{(r+1)\tau}\|^2] &\leq \left(1 - \min\left\{\frac{\mu\gamma}{2}, \frac{1-\rho}{8}\right\}\right) \mathbb{E} [\|V_{r\tau}\|^2] \\ &\quad + \mathcal{O}\left(\gamma^2\tau\frac{\sigma^2}{n} + \frac{\gamma^3\tau^2L}{(1-\rho)^3}\sigma^2\right). \end{aligned} \quad (25)$$

Further, the ST-GT algorithm achieves an accuracy of  $\epsilon > 0$  after at least the following rounds of communications:

$$R = \tilde{\mathcal{O}}\left(\frac{L}{(1-\rho)^2\mu} + \frac{\sigma^2}{\mu^2 n \tau \epsilon} + \sqrt{\frac{L\sigma^2}{\mu^3(1-\rho)^3\tau\epsilon}}\right). \quad (26)$$

*Proof.* See Appendix A.2. □

*Remark 1.* Theorem 1 shows that ST-GT converges linearly to a neighborhood of the optimal solution for strongly-convex and smooth objectives. The neighborhood size has two components: one matching the centralized SGD algorithm, which achieves linear speedup with respect to  $n$ , and the other induced by the network topology. More importantly, the communication complexity of ST-GT achieves a linear speedup with respect to the number of local updates  $\tau$ , and is significantly lower than that of the FlexGT algorithm. Specifically, the topology-dependent term in (26) is scaled by  $1/\tau$ , thereby reducing the communication cost, particularly in cases of weak network connectivity, i.e., when  $\rho \rightarrow 1$ . Furthermore, in the absence of the gradient noise, ST-GT attains exact linear convergence. A more detailed comparison with existing related algorithms in terms of computation, storage, and communication complexity is summarized in Table 1.

### 4.3 Nonconvex Case

When the objective functions are not convex, the following theorem shows a sublinear convergence rate of ST-GT.

*Theorem 2.* (Nonconvex case). Suppose Assumptions 1, 3 and 4 hold. Let the stepsize satisfy

$$\gamma \leq \min\left\{\frac{1-\rho}{178\tau L}, \frac{(1-\rho)^2}{625\sqrt{\rho}\tau L}\right\}. \quad (27)$$

Then, for the ST-GT algorithm, we have

$$\begin{aligned}
& \frac{1}{nR} \sum_{r=0}^{R-1} \mathbb{E} \left[ \|\nabla F(\mathbf{1}\bar{x}_{r\tau})\|^2 \right] \\
& \leq \frac{16(f(\bar{x}_0) - f(\bar{x}_{R\tau}))}{\gamma\tau R} + \frac{\mathbb{E}[\|\gamma_0\|^2]}{nR} \\
& \quad + \mathcal{O} \left( \gamma L \frac{\sigma^2}{n} + \frac{\tau\gamma^2 L^2 \sigma^2}{(1-\rho)^3} \right).
\end{aligned} \tag{28}$$

Further, the ST-GT algorithm achieves an accuracy of  $\epsilon > 0$  to a stationary point after at least the following rounds of communications

$$R = \mathcal{O} \left( \frac{L}{(1-\rho)^2 \epsilon} + \frac{\sigma^2 L}{n\tau\epsilon^2} + \frac{L\sigma}{\sqrt{\tau(1-\rho)^3 \epsilon^3}} + \frac{C_0}{\epsilon} \right), \tag{29}$$

where  $C_0 := \frac{1}{n} \mathbb{E}[\|\gamma_0\|^2] (f(\bar{x}_0) - f^*)^{-1}$ .

*Proof.* See Appendix A.3.  $\square$

*Remark 2.* Theorem 2 shows that ST-GT converges to a neighborhood of a local optimum at a sublinear rate, where the neighborhood size depends on the gradient noise level, objective properties, and network connectivity. Compared with K-GT (Liu et al., 2024), ST-GT reduces the network dependence from  $1/(1-\rho)^2$  to  $1/(1-\rho)^{3/2}$  and achieves scale invariance to gradient noise by a factor of  $\tau$ .

## 5. NUMERICAL RESULTS

### 5.1 Synthetic Example

To validate our theoretical findings and illustrate the effectiveness of ST-GT, we consider the following distributed ridge regression problem over a network of  $n = 32$  nodes:

$$\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{n} \sum_{i=1}^n \underbrace{\left( \mathbb{E}_{d_i} \left[ (\theta_i^\top x - d_i)^2 + \frac{\mu}{2} \|x\|^2 \right] \right)}_{=: f_i}, \tag{30}$$

where  $\mu > 0$  is the regularization parameter,  $\theta_i \in [0, 1]^p$  denotes the feature parameters of node  $i$  with dimension  $p = 10$ , and  $d_i \sim \mathcal{N}(\bar{d}_i, \sigma^2)$  with  $\bar{d}_i \in [0, 1]$ . The algorithms can obtain an unbiased noisy gradient  $\nabla f_i(x_{i,k}) + \delta_{i,k}$  with  $\delta_{i,k} \sim \mathcal{N}(0, \sigma^2)$  at each iteration  $k$ .

We compare the convergence performance of FlexGT, Scaffold, and ST-GT in terms of the residual  $\|\bar{x}_k - x^*\|^2$  as shown in Fig. 3. For the communication topology, Scaffold samples  $s = \{4, 16\}$  nodes at each round, while in FlexGT and ST-GT, each node in the exponential graph has  $s - 1$  neighbors. The number of local updates is set to  $\tau = 50$ , the stepsize to  $\gamma = 0.4$ , and the noise variance to  $\sigma^2 = 0.1$  for all algorithms. The results show that ST-GT achieves the lowest steady-state error. While Scaffold attains a smaller error than FlexGT, its performance remains inferior to ST-GT due to the uncertainty introduced by its random topology. Moreover, as the number of communicating nodes increases from 4 to 16, the performance of all algorithms approaches that of the centralized setting, and the gap among the three algorithms becomes smaller.

To illustrate the effect of local update frequency on ST-GT, Fig. 4 shows the convergence behavior measured in communication rounds under different network topologies for

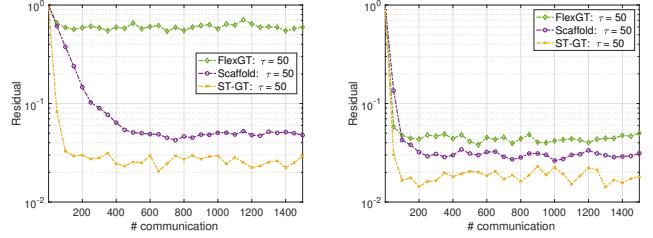


Fig. 3. Comparison of the convergence between Scaffold, FlexGT, and ST-GT. The number of nodes is  $n = 32$ . For Scaffold with partial node participation, we set  $s = 4$  (left) and  $s = 16$  (right). For the exponential graph used in the other two algorithms, each node is connected to 3 and 15 neighbors, respectively.

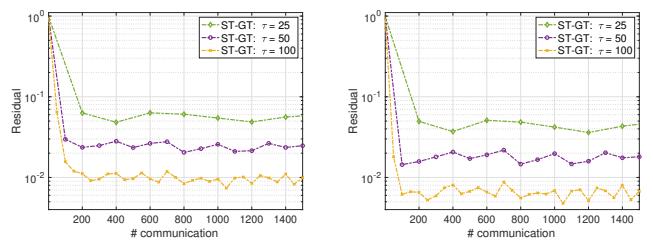


Fig. 4. Impact of the number of local updates  $\tau$ . We use an exponential graph with  $n = 32$  nodes, where each node is connected to 3 (left) or 15 (right) neighbors.

$\tau = 25, 50$ , and  $100$ . The stepsizes follow the proportional relationship recommended in Theorem 1, with  $\gamma = 0.8, 0.4$ , and  $0.2$ . The results confirm that ST-GT achieves an almost linear speed-up with respect to  $\tau$ .

### 5.2 Training ResNet-18 on CIFAR-10

We further evaluate the performance of ST-GT on the real-world dataset CIFAR-10 (Krizhevsky, 2009). Specifically, we perform distributed training of ResNet-18 using multiple processes to emulate nodes on a single A40 GPU, with the Gloo backend handling inter-node communication. The training data are unevenly partitioned across eight nodes, where each node contains samples from only eight of the ten classes, leading to heterogeneous local datasets. The learning rate is set to 1, the batch size is 200, and the number of local updates is set to  $\tau = 25$ . We compare the training loss and testing accuracy of FlexGT, Scaffold, and ST-GT across communication rounds. Figure 5 shows that ST-GT achieves the best performance in both training and testing. Scaffold, which also incorporates spatio-temporal gradient tracking, attains the second-best results, whereas FlexGT, relying solely on local updates, performs the worst. These results further corroborate our theoretical analysis and confirm the effectiveness of the proposed algorithm.

## 6. CONCLUSION

We have proposed a unified spatio-temporal gradient tracking algorithm, ST-GT, for distributed stochastic optimization with non-i.i.d. datasets. By simultaneously tracking the global gradient across nodes and the time-averaged local stochastic gradients at each node, ST-GT has improved

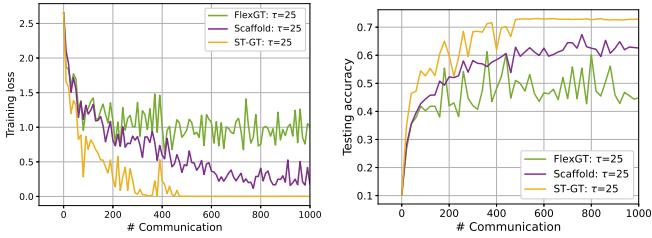


Fig. 5. Comparison among ST-GT, FlexGT, and Scaffold algorithms for distributed training of ResNet-18 on the CIFAR-10 dataset with  $n = 8$  nodes. The plots illustrate the training loss (left) and test accuracy (right) as functions of the communication rounds.

robustness against data heterogeneity and mitigated the effect of the gradient noise. By investigating an extended version of Scaffold, we further revealed that it could also be interpreted as a spatio-temporal gradient tracking method. Without assuming any data similarity, we proved that ST-GT achieved a linear and a sublinear convergence rate for strongly convex and nonconvex objective functions, respectively, while significantly reducing communication complexity compared with the FlexGT and K-GT algorithms, especially under poor network connectivity. Simulation results corroborated the theoretical analysis and demonstrated the effectiveness of ST-GT.

## REFERENCES

- Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., et al. (2012). Large scale distributed deep networks. *Advances in neural information processing systems*, 25.
- Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.
- Di Lorenzo, P. and Scutari, G. (2016). Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2), 120–136.
- Huang, Y. and Xu, J. (2023). On the computation-communication trade-off with a flexible gradient tracking approach. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, 284–289. IEEE.
- Jiang, X., Rodomanov, A., and Stich, S.U. (2024). Federated optimization with doubly regularized drift correction. *arXiv preprint arXiv:2404.08447*.
- Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A.T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 5132–5143. PMLR.
- Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.
- Lee, G., Jeong, M., Shin, Y., Bae, S., and Yun, S.Y. (2022). Preservation of the global knowledge by not-true distillation in federated learning. *Advances in Neural Information Processing Systems*, 35, 38461–38474.
- Li, M., Andersen, D.G., Park, J.W., Smola, A.J., Ahmed, A., Josifovski, V., Long, J., Shekita, E.J., and Su, B.Y. (2014). Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation*, OSDI’14, 583–598. USENIX Association, USA.
- Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2, 429–450.
- Lian, X., Zhang, C., Zhang, H., Hsieh, C.J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30.
- Liu, Y., Lin, T., Koloskova, A., and Stich, S.U. (2024). Decentralized gradient tracking with local steps. *Optimization Methods and Software*, 1–28.
- Lu, Y. and De Sa, C. (2021). Optimal complexity in decentralized training. In *International conference on machine learning*, 7111–7123. PMLR.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B.A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.
- Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. (2022). Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, 15750–15769. PMLR.
- Mora, A., Bujari, A., and Bellavista, P. (2024). Enhancing generalization in federated learning with heterogeneous data: A comparative literature review. *Future Generation Computer Systems*, 157, 1–15.
- Nedic, A., Olshevsky, A., and Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4), 2597–2633.
- Nguyen, E.D.H., Alghunaim, S.A., Yuan, K., and Uribe, C.A. (2023). On the performance of gradient tracking with local updates. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, 4309–4313. IEEE.
- Nguyen, E.D.H., Jiang, X., Ying, B., and Uribe, C.A. (2025). On graphs with finite-time consensus and their use in gradient tracking. *SIAM Journal on Optimization*, 35(2), 872–898.
- Pu, S. and Nedić, A. (2021). Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1), 409–457.
- Rabbat, M. and Nowak, R. (2004). Distributed optimization in sensor networks. In *Proceedings of The 3rd International Symposium on Information Processing in Sensor Networks*, 20–27.
- Ram, S.S., Nedić, A., and Veeravalli, V.V. (2009). Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, 3581–3586. IEEE.
- Stich, S.U. (2019). Local SGD converges fast and communicates little. In *International Conference on Learning Representations*.
- Tian, Y., Bedi, A.S., Koppel, A., Calvo-Fullana, M., Rosen, D.M., and How, J.P. (2022). Distributed riemannian optimization with lazy communication for collaborative geometric estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4391–4398. IEEE.

- Wu, T., Li, Z., and Sun, Y. (2025). The effectiveness of local updates for decentralized learning under data heterogeneity. *IEEE Transactions on Signal Processing*.
- Xin, R., Khan, U.A., and Kar, S. (2020). Variance-reduced decentralized stochastic optimization with accelerated convergence. *IEEE Transactions on Signal Processing*, 68, 6255–6271.
- Xu, J., Zhu, S., Soh, Y.C., and Xie, L. (2015). Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, 2055–2060. IEEE.
- Ying, B., Yuan, K., Chen, Y., Hu, H., Pan, P., and Yin, W. (2021). Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34, 13975–13987.
- Yuan, K., Ling, Q., and Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3), 1835–1854.
- Zhang, J., Hu, J., and Johansson, M. (2024). Composite federated learning with heterogeneous data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8946–8950. IEEE.
- Zhang, T., Feng, T., Alam, S., Lee, S., Zhang, M., Narayanan, S.S., and Avestimehr, S. (2023). Fedaudio: A federated learning benchmark for audio tasks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

## Appendix A. PROOF OF THE MAIN RESULTS

In this section, we provide the detailed convergence analysis for the ST-GT algorithm.

### A.1 Supporting Lemmas

We first bound the extent to which the local model parameters diverge from the averaged model between two communication rounds.

*Lemma 1.* (Client divergence within a period). Suppose Assumptions 1-4 hold. Let the stepsize satisfy  $\gamma \leq \frac{1}{8\tau L}$ . We have for all  $t \in [0, \tau - 1]$ ,

$$\begin{aligned} & \frac{1}{n} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \bar{x}_{r\tau}\|^2] \\ & \leq 3\tau \frac{1}{n} \mathbb{E} [\|X_{r\tau} - \bar{x}_{r\tau}\|^2] + \frac{9}{2} \gamma^2 \tau^3 \frac{1}{n} \mathbb{E} [\|\gamma_{r\tau}\|^2] \quad (\text{A.1}) \\ & \quad + \frac{9}{2} \gamma^2 \tau^3 \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] + \frac{3\gamma^2 \tau^2}{2} \sigma^2. \end{aligned}$$

*Proof.* By the update rule of ST-GT in (5), we have

$$\begin{aligned} & \mathbb{E} [\|x_{i,r\tau+t} - \bar{x}_{r\tau}\|^2] \\ & = \mathbb{E} [\|x_{i,r\tau+t-1} - \gamma(y_{i,r\tau} + g_{i,r\tau+t} - g_{i,r\tau}) - \bar{x}_{r\tau}\|^2]. \quad (\text{A.2}) \end{aligned}$$

Then, by Assumptions 2 and 4, and using Young's inequality with parameter  $\beta > 0$ , we get

$$\begin{aligned} & \mathbb{E} [\|X_{r\tau+t} - \bar{x}_{r\tau}\|^2] \\ & \leq (1 + \beta) \mathbb{E} [\|X_{r\tau+t-1} - \bar{x}_{r\tau}\|^2] + \gamma^2 n \sigma^2 \\ & \quad + (1 + \beta^{-1}) \gamma^2 \mathbb{E} [\|Y_{r\tau} - G_{r\tau} + \nabla F_{r\tau+t-1}\|^2] \\ & \leq (1 + \beta + 3(1 + \beta^{-1}) \gamma^2 L^2) \mathbb{E} [\|X_{r\tau+t-1} - \bar{x}_{r\tau}\|^2] \\ & \quad + 3(1 + \beta^{-1}) \gamma^2 \mathbb{E} [\|\gamma_{r\tau}\|^2] \\ & \quad + 3(1 + \beta^{-1}) \gamma^2 \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] + \gamma^2 n \sigma^2. \end{aligned} \quad (\text{A.3})$$

Letting  $\beta = \frac{1}{\tau}$ ,  $\gamma \leq \frac{1}{8\tau L}$  and noticing that

$$(1 + \beta + 3(1 + \beta^{-1}) \gamma^2 L^2)^t < 3,$$

we obtain the result by iteratively applying (A.3).  $\square$

Building on the above lemma, we can establish the contraction property of the consensus error.

*Lemma 2.* (Consensus error). Suppose Assumptions 1-4 hold. Let the stepsize satisfy  $\gamma \leq \min \left\{ \frac{1-\rho}{11\tau L\sqrt{\rho}}, \frac{1}{8\tau L} \right\}$ . We have

$$\begin{aligned} & \mathbb{E} [\|X_{(r+1)\tau} - \bar{x}_{(r+1)\tau}\|^2] \\ & \leq \frac{3+\rho}{4} \mathbb{E} [\|X_{r\tau} - \bar{x}_{r\tau}\|^2] + \frac{5\rho}{1-\rho} \gamma^2 \tau^2 \mathbb{E} [\|\gamma_{r\tau}\|^2] \\ & \quad + 3\rho \gamma^2 \tau n \sigma^2 + \frac{9\rho}{1-\rho} \gamma^2 \tau^2 n \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2]. \end{aligned} \quad (\text{A.4})$$

*Proof.* By the update rule of ST-GT (5) and under Assumptions 1 and 2, we have

$$\begin{aligned} & \mathbb{E} [\|X_{(r+1)\tau} - \bar{x}_{(r+1)\tau}\|^2] \\ & \leq \frac{1+\rho}{2} \mathbb{E} [\|X_{r\tau} - \bar{x}_{r\tau}\|^2] \\ & \quad + \frac{2(1+\rho)\rho}{1-\rho} \gamma^2 \tau^2 \mathbb{E} [\|\gamma_{r\tau}\|^2] + \rho \gamma^2 \tau n \sigma^2 \\ & \quad + \frac{4(1+\rho)\rho}{1-\rho} \gamma^2 \tau L^2 \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \bar{x}_{r\tau}\|^2] \\ & \quad + \frac{4(1+\rho)\rho}{1-\rho} \gamma^2 \tau^2 n \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] \\ & \quad + 2\mathbb{E} \left[ \left\langle \sum_{t=0}^{\tau-1} (G(X_{r\tau+t}) - \nabla F_{r\tau+t}), \sum_{t=0}^{\tau-1} \nabla F_{r\tau+t} \right\rangle \right]. \end{aligned} \quad (\text{A.5})$$

With the help of Lemma 1, we get

$$\begin{aligned} & \mathbb{E} [\|X_{(r+1)\tau} - \bar{x}_{(r+1)\tau}\|^2] \\ & \leq \left( \frac{1+\rho}{2} + \frac{30\rho}{1-\rho} \gamma^2 \tau^2 L^2 \right) \mathbb{E} [\|X_{r\tau} - \bar{x}_{r\tau}\|^2] \\ & \quad + \left( \frac{4\rho}{1-\rho} \gamma^2 \tau^2 + \frac{45\rho}{1-\rho} \gamma^4 \tau^4 L^2 \right) \mathbb{E} [\|\gamma_{r\tau}\|^2] \\ & \quad + \left( \frac{8\rho}{1-\rho} \gamma^2 \tau^2 + \frac{45\rho}{1-\rho} \gamma^4 \tau^4 L^2 \right) n \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] \\ & \quad + 2\rho \gamma^2 \tau n \sigma^2 + \frac{15\rho}{1-\rho} \gamma^4 \tau^3 L^2 n \sigma^2. \end{aligned} \quad (\text{A.6})$$

Letting the stepsize satisfy  $\gamma \leq \min\left\{\frac{1-\rho}{11L\sqrt{\rho}}, \frac{1}{8\tau L}\right\}$ , we obtain the result.  $\square$

Similarly, we prove the contraction property of the gradient tracking error.

*Lemma 3.* (Gradient-tracking error). Suppose Assumptions 1-4 hold. Let the stepsize  $\gamma \leq \frac{1-\rho}{12\tau L}$ . Then, we have

$$\begin{aligned} & \mathbb{E} [\|\mathcal{Y}_{(r+1)\tau}\|^2] \\ & \leq \frac{3+\rho}{4} \mathbb{E} [\|\mathcal{Y}_{r\tau}\|^2] + \frac{27(1+\rho)L^2}{1-\rho} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\ & \quad + \frac{8(1+\rho)}{1-\rho} n \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] + 7n \frac{\sigma^2}{\tau}. \end{aligned} \quad (\text{A.7})$$

*Proof.* By the update rule of ST-GT in (5), we have

$$\begin{aligned} \mathcal{Y}_{(r+1)\tau} &= (W - \mathbf{J}) \mathcal{Y}_{r\tau} - W (\nabla F(\mathbf{1}\bar{x}_{r\tau}) - \nabla f(\mathbf{1}\bar{x}_{r\tau})) \\ &\quad + (W - \mathbf{I}) \frac{1}{\tau} \sum_{t=0}^{\tau-1} G_{r\tau+t} \\ &\quad + (\nabla F(\mathbf{1}\bar{x}_{(r+1)\tau}) - \nabla f(\mathbf{1}\bar{x}_{(r+1)\tau})). \end{aligned} \quad (\text{A.8})$$

Then, by Assumptions 1-4 and using Young's inequality, we get

$$\begin{aligned} & \mathbb{E} [\|\mathcal{Y}_{(r+1)\tau}\|^2] \\ & \leq \frac{1+\rho}{2} \mathbb{E} [\|\mathcal{Y}_{r\tau}\|^2] + 6n \frac{\sigma^2}{\tau} \\ & \quad + \left(3n \frac{(1+\rho)L^2}{1-\rho} + 1\right) \mathbb{E} [\|\bar{x}_{(r+1)\tau} - \bar{x}_{rK}\|^2] \\ & \quad + 2 \left(\frac{3(1+\rho)L^2}{1-\rho} + 1\right) \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\ & \quad + \frac{6(1+\rho)}{1-\rho} n \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2]. \end{aligned} \quad (\text{A.9})$$

With the help of Lemma 1, and noticing that

$$\begin{aligned} & \mathbb{E} [\|\bar{x}_{(r+1)\tau} - \bar{x}_{r\tau}\|^2] \\ & = \gamma^2 \mathbb{E} \left[ \left\| \sum_{t=0}^{\tau-1} \frac{1}{n} \nabla G_{r\tau+t} \right\|^2 \right] \\ & \leq \gamma^2 \tau^2 \mathbb{E} \left[ \left\| \frac{1}{\tau} \sum_{t=0}^{\tau-1} \frac{1}{n} \nabla F_{r\tau+t} \right\|^2 \right] + \gamma^2 \frac{\tau\sigma^2}{n} \\ & \leq 2\gamma^2 \tau^2 \frac{L^2}{n\tau} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\ & \quad + 2\gamma^2 \tau^2 \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] + \gamma^2 \frac{\tau\sigma^2}{n}, \end{aligned} \quad (\text{A.10})$$

we obtain the result with stepsize  $\gamma \leq \frac{1-\rho}{18\tau L}$ .  $\square$

## A.2 Proof of Theorem 1

With these supporting lemmas in hand, we can prove Theorem 1 under Assumptions 1-4, and the stepsize condition.

To this end, we first prove the contraction of the optimality gap in the following lemma for the strongly convex and smooth objective function.

*Lemma 4.* (Optimality gap). Suppose Assumptions 1-4 hold. Let the stepsize satisfy  $\gamma \leq \frac{1}{11\tau L}$ . We have

$$\begin{aligned} & \mathbb{E} [\|\bar{x}_{(r+1)\tau} - x^*\|^2] \\ & \leq \left(1 - \frac{\gamma\mu\tau}{2}\right) \mathbb{E} [\|\bar{x}_{r\tau} - x^*\|^2] \\ & \quad + 9\gamma\tau L \frac{1}{n} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] + 14\gamma^3\tau^3 L \frac{1}{n} \mathbb{E} [\|\mathcal{Y}_{r\tau}\|^2] \\ & \quad + \gamma^2 \frac{\tau\sigma^2}{n} + 5\gamma^3\tau^2 L\sigma^2 - \gamma\tau \mathbb{E} [f(\bar{x}_{rK}) - \nabla f(x^*)]. \end{aligned} \quad (\text{A.11})$$

*Proof.* By the update rules, we have

$$\begin{aligned} & \mathbb{E} [\|\bar{x}_{(r+1)\tau} - x^*\|^2] \\ & = \mathbb{E} [\|\bar{x}_{r\tau} - x^*\|^2] + \gamma^2 \mathbb{E} \left[ \left\| \sum_{t=0}^{\tau-1} \frac{1}{n} G_{r\tau+t} \right\|^2 \right] \\ & \quad - 2\gamma \mathbb{E} \left[ \left\langle \bar{x}_{r\tau} - x^*, \sum_{t=0}^{\tau-1} \frac{1}{n} \nabla F_{r\tau+t} \right\rangle \right]. \end{aligned} \quad (\text{A.12})$$

For the second term on the RHS, with the help of Assumption 4, we have

$$\begin{aligned} & \mathbb{E} \left[ \left\| \sum_{t=0}^{K-1} \frac{1}{n} G_{rK+t} \right\|^2 \right] \\ & = \mathbb{E} \left[ \left\| \sum_{t=0}^{K-1} \frac{1}{n} (G_{rK+t} - \nabla F_{rK+t} + \nabla F_{rK+t}) \right\|^2 \right] \\ & \leq 2K \sum_{t=0}^{K-1} \mathbb{E} \left[ \left\| \frac{1}{n} \nabla F_{rK+t} - \nabla f(\bar{x}_{rK}) \right\|^2 \right] \\ & \quad + \frac{K\sigma^2}{n} + 2K^2 \mathbb{E} [\|\nabla f(\bar{x}_{rK})\|^2]. \end{aligned} \quad (\text{A.13})$$

For the third term, we have

$$\begin{aligned} & \mathbb{E} \left[ \left\langle \bar{x}_{r\tau} - x^*, \sum_{t=0}^{\tau-1} \frac{1}{n} \nabla F_{r\tau+t} \right\rangle \right] \\ & = \sum_{t=0}^{\tau-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\langle \bar{x}_{r\tau} - x^*, \nabla f_i(x_{i,r\tau+t}) \rangle] \\ & = \sum_{t=0}^{\tau-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\langle x_{i,r\tau+t} - x^*, \nabla f_i(x_{i,r\tau+t}) \rangle] \\ & \quad - \sum_{t=0}^{\tau-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\langle x_{i,r\tau+t} - \bar{x}_{r\tau}, \nabla f_i(x_{i,r\tau+t}) \rangle] \\ & \geq \tau \mathbb{E} [f(\bar{x}_{r\tau}) - \nabla f(x^*)] \\ & \quad + \frac{\mu}{2} \sum_{t=0}^{\tau-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|x_{i,r\tau+t} - x^*\|] \\ & \quad - \frac{L}{2} \sum_{t=0}^{\tau-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|x_{i,r\tau+t} - \bar{x}_{r\tau}\|], \end{aligned} \quad (\text{A.14})$$

wherein the last inequality we used the convexity and smoothness of  $f_i$  assumed in Assumption 2. Then, we get

$$\begin{aligned}
& \mathbb{E} [\|\bar{x}_{(r+1)\tau} - x^*\|^2] \\
& \leq \mathbb{E} [\|\bar{x}_{r\tau} - x^*\|^2] + \gamma^2 \frac{\tau\sigma^2}{n} \\
& \quad + 2\gamma^2\tau L^2 \frac{1}{n} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\
& \quad - 2\gamma\tau(1-2\gamma\tau L) \mathbb{E} [f(\bar{x}_{r\tau}) - \nabla f(x^*)] \\
& \quad - \gamma\mu \sum_{t=0}^{\tau-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|x_{i,r\tau+t} - x^*\|] \\
& \quad + \gamma L \sum_{t=0}^{K-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|x_{i,r\tau+t} - \bar{x}_{rK}\|].
\end{aligned} \tag{A.15}$$

Noticing that

$$\begin{aligned}
& -\frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|x_{i,r\tau+t} - x^*\|] \\
& = -\frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|x_{i,r\tau+t} - \bar{x}_{r\tau}\|] - \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|\bar{x}_{r\tau} - x^*\|] \\
& \quad - 2\frac{1}{n} \sum_{i=1}^n \mathbb{E} \langle x_{i,r\tau+t} - \bar{x}_{r\tau}, \bar{x}_{r\tau} - x^* \rangle \\
& \leq -\frac{1}{2n} \sum_{i=1}^n \mathbb{E} [\|\bar{x}_{r\tau} - x^*\|] + \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|x_{i,r\tau+t} - \bar{x}_{r\tau}\|].
\end{aligned} \tag{A.16}$$

We get

$$\begin{aligned}
& \mathbb{E} [\|\bar{x}_{(r+1)\tau} - x^*\|^2] \\
& \leq \left(1 - \frac{\gamma\mu\tau}{2}\right) \mathbb{E} [\|\bar{x}_{r\tau} - x^*\|^2] + \gamma^2 \frac{\tau\sigma^2}{n} \\
& \quad - 2\gamma\tau(1-2\gamma\tau L) \mathbb{E} [f(\bar{x}_{r\tau}) - \nabla f(x^*)] \\
& \quad + \gamma(L+\mu+2\gamma\tau L^2) \frac{1}{n} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \mathbf{1}\bar{x}_{r\tau}\|^2].
\end{aligned} \tag{A.17}$$

Then, using Lemma 1 and letting the stepsize satisfy  $\gamma \leq \frac{1}{11\tau L}$ , we complete the proof.  $\square$

Recalling the design of the Lyapunov function in (22), we can obtain

$$\begin{aligned}
& \mathbb{E} [\|V_{(r+1)\tau}\|^2] \\
& \leq \left(1 - \min \left\{ \frac{\gamma\mu}{2}, \frac{1-\rho}{8} \right\} \right) \mathbb{E} [\|V_{r\tau}\|^2] \\
& \quad + e_1 \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] + e_2 \mathbb{E} [\|\Upsilon_{r\tau}\|^2] \\
& \quad + e_3 \mathbb{E} [f(\bar{x}_{r\tau}) - f(x^*)] + \gamma^2 \tau \frac{\sigma^2}{n} + \frac{9}{2} \gamma^2 \gamma^3 L \sigma^2 \\
& \quad + \frac{240\gamma^3\tau^2 L \rho}{1-\rho} \sigma^2 + \frac{24892\gamma^3\tau^2 L}{(1-\rho)^3},
\end{aligned} \tag{A.18}$$

where

$$\begin{aligned}
e_1 &:= 9\gamma\tau L \frac{1}{n} + \frac{54L^2}{1-\rho} c_y - \frac{1-\rho}{8} c_x, \\
e_2 &:= 14\gamma^3\tau^3 L \frac{1}{n} + \frac{5\rho}{1-\rho} \gamma^2 \tau^2 c_x - \frac{1-\rho}{8} c_y, \\
e_3 &:= -\gamma\tau + \frac{18\rho}{1-\rho} \gamma^2 \tau^2 L n c_x + \frac{36}{1-\rho} n L c_y.
\end{aligned} \tag{A.19}$$

Noticing the coefficients of the Lyapunov function are designed as (24) and letting the stepsize satisfy  $\gamma = \mathcal{O}\left(\frac{(1-\rho)^2}{L\tau}\right)$ , we have  $e_1, e_2, e_3 < 0$ , we obtain the convergence rate in (25).

To further get the communication complexity, we tune the stepsize between

$$\left\{ \gamma, \frac{2 \ln (\max \{2, \tau^2 \mu^2 R/H_1, \tau^3 \mu^3 R/H_2\})}{\tau \mu R} \right\} \tag{A.20}$$

to ensure that gradient noise-related errors match the linear part, where

$$H_1 = \frac{\tau\sigma^2}{n \mathbb{E} [\|V_0\|^2]}, H_2 = \frac{\tau^2 L \sigma^2}{(1-\rho)^3 \mathbb{E} [\|V_0\|^2]}.$$

Then, we get the required number of communication rounds to achieve an accuracy of  $\epsilon$  in (26).

### A.3 Proof of Theorem 2

For the nonconvex case, we first derive the following descent lemma using the smoothness of  $f_i$ .

*Lemma 5.* Suppose Assumptions 1, 3 and 4 hold. Let the stepsize satisfy  $\gamma \leq \frac{1}{2\tau L}$ . We get

$$\begin{aligned}
& \mathbb{E} [f(\bar{x}_{(r+1)\tau})] \\
& \leq \mathbb{E} [f(\bar{x}_{r\tau})] - \frac{\gamma\tau}{4} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] \\
& \quad + 3\gamma L^2 \frac{1}{n} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \mathbf{1}\bar{x}_{r\tau}\|^2] + \frac{\gamma^2 \tau L \sigma^2}{2n}.
\end{aligned} \tag{A.21}$$

*Proof.* By the smoothness of  $f_i$  assumed in Assumption 3, we have

$$\begin{aligned}
& f(\bar{x}_{(r+1)\tau}) \\
& \leq f(\bar{x}_{r\tau}) + \langle \nabla f(\bar{x}_{r\tau}), \bar{x}_{(r+1)\tau} - \bar{x}_{r\tau} \rangle \\
& \quad + \frac{L}{2} \|\bar{x}_{(r+1)\tau} - \bar{x}_{r\tau}\|^2.
\end{aligned} \tag{A.22}$$

Then, for the last term on the RHS of (A.22), we have

$$\begin{aligned}
& \mathbb{E} [\|\bar{x}_{(r+1)\tau} - \bar{x}_{r\tau}\|^2] \\
& \leq \gamma^2 \tau^2 \mathbb{E} \left[ \left\| \frac{1}{\tau} \sum_{t=0}^{\tau-1} \frac{1}{n} \nabla F_{r\tau+t} \right\|^2 \right] + \gamma^2 \frac{\tau\sigma^2}{n} \\
& \leq 2\gamma^2 \tau^2 \frac{L^2}{n\tau} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\
& \quad + 2\gamma^2 \tau^2 \mathbb{E} [\nabla f(\bar{x}_{r\tau})^2] + \gamma^2 \frac{\tau\sigma^2}{n}.
\end{aligned} \tag{A.23}$$

For the inner-product term above, we have

$$\begin{aligned}
& \mathbb{E} [\langle \nabla f(\bar{x}_{r\tau}), \bar{x}_{(r+1)\tau} - \bar{x}_{r\tau} \rangle] \\
&= \mathbb{E} \left[ \left\langle \nabla f(\bar{x}_{r\tau}), -\gamma \sum_{t=0}^{\tau-1} \frac{\mathbf{1}^\top}{n} \nabla F_{r\tau+t} \right\rangle \right] \\
&= -\gamma \tau \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] \\
&\quad + \mathbb{E} \left[ \left\langle \nabla f(\bar{x}_{r\tau}), -\gamma \sum_{t=0}^{\tau-1} \left( \frac{\mathbf{1}^\top}{n} \nabla F_{r\tau+t} - \nabla f(\bar{x}_{r\tau}) \right) \right\rangle \right] \\
&\leq -\frac{\gamma \tau}{2} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] \\
&\quad + 2\gamma L^2 \frac{1}{n} \sum_{t=0}^{\tau-1} \mathbb{E} [\|X_{r\tau+t} - \mathbf{1}\bar{x}_{r\tau}\|^2]. \tag{A.24}
\end{aligned}$$

Applying (A.24) and (A.23) into (A.22) and letting the stepsize satisfy  $\gamma \leq \frac{1}{\tau L}$ , we complete the proof.  $\square$

Then, using Lemma 1 to bound the client divergence during the local updates and doing accumulation, we have

$$\begin{aligned}
& \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] \\
&\leq \frac{4(f(\bar{x}_0) - f(\bar{x}_{r\tau}))}{\gamma \tau R} + 2\gamma L \frac{\sigma^2}{n} + 18\gamma^2 \tau L^2 \sigma^2 \\
&\quad + 36L^2 \frac{1}{nR} \sum_{r=0}^{R-1} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\
&\quad + 54\gamma^2 \tau^2 L^2 \frac{1}{nR} \sum_{r=0}^{R-1} \mathbb{E} [\|\gamma_{r\tau}\|^2]. \tag{A.25}
\end{aligned}$$

For the accumulated consensus error, invoking Lemma 2, we have

$$\begin{aligned}
& \frac{1}{R} \sum_{r=1}^R \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\
&\leq \frac{3+\rho}{4} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\
&\quad + \frac{5\rho\gamma^2\tau^2}{1-\rho} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\gamma_{r\tau}\|^2] + 3\rho\gamma^2\tau n\sigma^2 \\
&\quad + \frac{9\rho\gamma^2\tau^2 n}{1-\rho} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2]. \tag{A.26}
\end{aligned}$$

Adding  $\|X_0 - \mathbf{1}\bar{x}_0\|^2$  on the both side and noticing that

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] \leq \frac{1}{R} \sum_{r=0}^R \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2],$$

we obtain

$$\begin{aligned}
& \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\
&\leq \frac{4\mathbb{E} [\|X_0 - \mathbf{1}\bar{x}_0\|^2]}{(1-\rho)R} + \frac{20\rho\gamma^2\tau^2}{(1-\rho)^2} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\gamma_{r\tau}\|^2] \\
&\quad + \frac{36\gamma^2\tau^2\rho}{(1-\rho)^2} n \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] + \frac{12\gamma^2\tau\rho}{1-\rho} n\sigma^2. \tag{A.27}
\end{aligned}$$

Similarly, for the accumulated gradient tracking error, we have

$$\begin{aligned}
& \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\gamma_{r\tau}\|^2] \\
&\leq \frac{4\mathbb{E} [\|\gamma_0\|^2]}{(1-\rho)R} + \frac{28n\sigma^2}{\tau(1-\rho)} \\
&\quad + \frac{72}{(1-\rho)^2} n \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2] \\
&\quad + \frac{216L^2}{(1-\rho)^2} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2]. \tag{A.28}
\end{aligned}$$

Next, we decouple the accumulated consensus error and gradient tracking error in the following two lemmas.

*Lemma 6.* Suppose Assumptions 1, 3 and 4 hold. Let the stepsize satisfy  $\gamma \leq \min \left\{ \frac{1-\rho}{12\tau L}, \frac{(1-\rho)^2}{93\tau L\sqrt{\rho}} \right\}$ . We get

$$\begin{aligned}
& \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|X_{r\tau} - \mathbf{1}\bar{x}_{r\tau}\|^2] \\
&\leq C_{0,x} + \frac{24\gamma^2\tau\rho}{1-\rho} n\sigma^2 + \frac{1120\rho\gamma^2\tau}{(1-\rho)^3} n\sigma^2 \\
&\quad + \frac{2952\rho\gamma^2\tau^2 n}{(1-\rho)^4} \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2], \tag{A.29}
\end{aligned}$$

where

$$C_{0,x} := \frac{8\mathbb{E} [\|X_0 - \mathbf{1}\bar{x}_0\|^2]}{(1-\rho)R} + \frac{160\rho\gamma^2\tau^2}{(1-\rho)^3} \frac{\mathbb{E} [\|\gamma_0\|^2]}{R}. \tag{A.30}$$

*Proof.* Applying (A.28) to (A.27), we complete the proof by combining like terms and letting the stepsize satisfy the condition.  $\square$

*Lemma 7.* Suppose Assumptions 1, 3 and 4 hold. Let the stepsize  $\gamma \leq \min \left\{ \frac{1-\rho}{12\tau L}, \frac{(1-\rho)^2}{62\tau L\sqrt{\rho}} \right\}$ . Then, we have

$$\begin{aligned}
& \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\gamma_{r\tau}\|^2] \\
&\leq C_{0,y} + \frac{54n\sigma^2}{\tau(1-\rho)} + \frac{2304\gamma^2\tau L^2\rho}{(1-\rho)^3} n\sigma^2 \\
&\quad + \frac{146}{(1-\rho)^2} n \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\|\nabla f(\bar{x}_{r\tau})\|^2], \tag{A.31}
\end{aligned}$$

where

$$C_{0,y} := \frac{8\mathbb{E} [\|\gamma_0\|^2]}{(1-\rho)R} + \frac{768L^2}{(1-\rho)^3} \frac{\mathbb{E} [\|X_0 - \mathbf{1}\bar{x}_0\|^2]}{R}. \tag{A.32}$$

*Proof.* Applying (A.27) into (A.28), we complete the proof by combining like terms and letting the stepsize satisfy the condition.  $\square$

Applying Lemmas 6 and 7 to (A.25) and letting the stepsize satisfy

$$\gamma \leq \min \left\{ \frac{1-\rho}{178\tau L}, \frac{(1-\rho)^2}{625\sqrt{\rho}\tau L} \right\},$$

we obtain the sublinear convergence rate (28) in Theorem 2.

To further derive the communication complexity, we tune the stepsize between

$$\left\{ \gamma, \frac{1}{\sqrt{H_1 R}}, \frac{1}{\sqrt[3]{H_2 R}} \right\}, \quad (\text{A.33})$$

where

$$H_1 = \frac{\tau \sigma^2 L}{n(f(\bar{x}_0) - f(\bar{x}_{R\tau}))},$$

$$H_2 = \frac{\tau^3 L^2}{(1-\rho)^3(f(\bar{x}_0) - f(\bar{x}_{R\tau}))} \frac{\sigma^2}{\tau},$$

and we obtain the required communication rounds to achieve an accuracy of  $\epsilon > 0$  in (29).