

Communication-Efficient Federated Optimization over Semi-Decentralized Networks

He Wang*
Carnegie Mellon University
hew2@andrew.cmu.edu

Yuejie Chi*
Carnegie Mellon University
yuejiechi@cmu.edu †

September 2023; Revised January 11, 2024

Abstract

In large-scale federated and decentralized learning, communication efficiency is one of the most challenging bottlenecks. While gossip communication — where agents can exchange information with their connected neighbors — is more cost-effective than communicating with the remote server, it often requires a greater number of communication rounds, especially for large and sparse networks. To tackle the trade-off, we examine the communication efficiency under a *semi-decentralized* communication protocol, in which agents can perform both agent-to-agent and agent-to-server communication *in a probabilistic manner*. We design a tailored communication-efficient algorithm over semi-decentralized networks, referred to as PISCO, which inherits the robustness to data heterogeneity thanks to gradient tracking and allows multiple local updates for saving communication. We establish the convergence rate of PISCO for nonconvex problems and show that PISCO enjoys a linear speedup in terms of the number of agents and local updates. Our numerical results highlight the superior communication efficiency of PISCO and its resilience to data heterogeneity and various network topologies.

Keywords: communication efficiency; semi-decentralized networks; probabilistic communication models; local updates

Contents

1	Introduction	2
1.1	Our contributions	3
1.2	Related works	4
1.3	Notation	5
2	Preliminaries	5
3	Proposed Algorithm	6
4	Theoretical Guarantees	8
4.1	Assumptions	8
4.2	Convergence analysis of PISCO	8
5	Numerical Experiments	11
5.1	Logistic regression with nonconvex regularization	11
5.2	One-hidden-layer neural network training	13
6	Conclusions	14

*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

†A preliminary version of this work will be presented at 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Initial submission: September 2023.

A Proof of Theorem 1	17
A.1 Useful inequalities	18
A.2 Property of gradient tracking	18
A.3 Progress improvement on successive iterates and averages	19
A.4 Bounding consensus errors and tracking errors	21
A.5 Descent lemma	24
A.6 Proof of Theorem 1	26
B Proof of Corollary 1	30

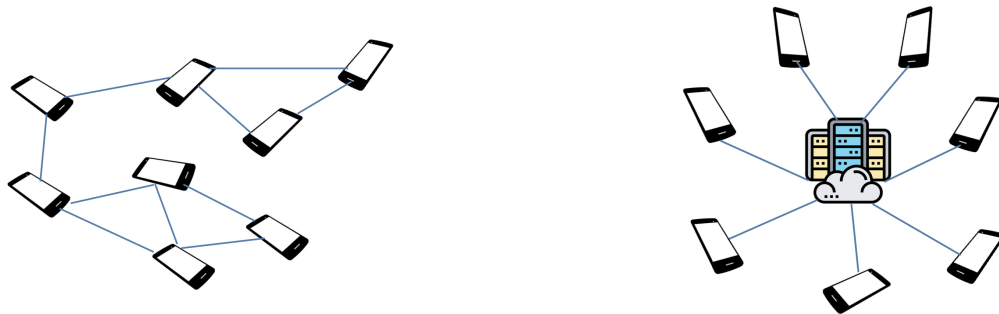
1 Introduction

Consider a networked system that n agents collectively solve the following federated or distributed optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}), \quad \text{where } f_i(\mathbf{x}) := \frac{1}{m} \sum_{\mathbf{z} \in \mathcal{D}_i} \ell(\mathbf{x}; \mathbf{z}). \quad (1)$$

Here, $\mathbf{x} \in \mathbb{R}^d$ denotes the optimization variable, $f_i(\mathbf{x})$ denotes the local and private objective function at agent i , and $f(\mathbf{x})$ denotes the global objective function. In addition, let \mathbf{z} represent one data sample, \mathcal{D}_i stand for the dataset with $|\mathcal{D}_i| = m$ samples at agent i , and $\ell(\mathbf{x}; \mathbf{z})$ denote the empirical loss of \mathbf{x} w.r.t. the data sample \mathbf{z} . Such problems have a wide range of applications, including but not limited to estimation in sensor networks [RN04], resource allocation in smart grids [BNOT14], and coordination in multi-agent systems [CYRC13].

In order to tackle this problem, agents have to communicate with one another for cooperation, since every agent $i \in [n]$ only has access to its own local dataset \mathcal{D}_i . There are two main communication protocols, consisting of agent-to-agent communication model (in decentralized ML) and agent-to-server communication model (in federated ML). Commonly, they are formulated via different network topologies [NOR18], as shown in Figure 1. More specifically, prior works in decentralized ML often use a general graph to capture the local communication, where every agent is only allowed to exchange information with its connected neighbors (cf. Figure 1(a)). In federated ML, the star graph is commonly used to depict the communication between agents and the centralized coordinator (i.e., server) who can both collect information from and broadcast to each agent (cf. Figure 1(b)).



(a) Agent-to-agent communication model with a general connected graph, where each agent only communicates with its adjacent agents.

(b) Agent-to-server communication model with a star graph, where each agent can send messages to and receive messages from the server.

Figure 1: Two communication models for distributed ML.

As the network size increasingly grows, communication efficiency becomes so critical that significantly hinders both decentralized and federated ML from being applied to real-world applications. Compared with agent-to-server communication, agent-to-agent communication is much more affordable and more applicable

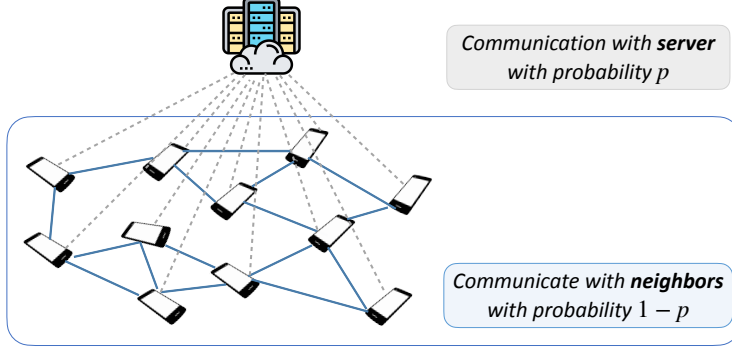


Figure 2: The semi-decentralized communication protocol, where the server can be accessed with probability p and agents can communicate with their neighbors whenever the server is not available. Here, dot lines represents the agent-to-server communication, while the solid ones are for agent-to-agent communication.

to large-scale networks. However, without the coordination of the server, decentralized approaches may need more communication rounds to reach consensus, especially for large and sparse networks.

Given that the communication complexity depends on the trade-off between the communication rounds and the per-round cost, emerging works focus on heterogeneous communication over *semi-decentralized networks*, to gain the best from both agent-to-agent and agent-to-server communication [CYZ⁺21, WFC22, GSHG22]. Such semi-decentralized networks—consisting of a centralized server and a network of agents—widely exist in many applications, such as autonomous vehicles [CRW21], energy systems [NMTA21] and ML systems [MSY⁺23]. It has been observed that heterogeneous communication largely alleviates the heavy network dependence of distributed learning and tackles the communication bottleneck of the server [CYZ⁺21]. However, to the best of our knowledge, all of them rely on the assumption of bounded data dissimilarity across agents and a complete characterization of the convergence behavior with respect to the network heterogeneity is still lacking. More detailed discussions are delegated to Section 1.2.

1.1 Our contributions

To fill the void, we propose a communication-efficient algorithm called PISCO, which incorporates *gradient-tracking techniques* [ZM10] and *multiple local updates* [MMR⁺17] for solving federated nonconvex optimization over semi-decentralized networks modeled by a probabilistic connection model (shown in Figure 2). The highlights of our contributions are as follows.

1. We prove that PISCO converges at a rate of $O(1/\sqrt{nT_oK})$ for sufficiently large K , where K is the number of communication rounds and T_o is the number of local updates. Our result does not impose the strong assumptions on data heterogeneity. Moreover, increasing the number of local updates accelerates the convergence over semi-decentralized networks.
2. We show that the communication heterogeneity offered by a semi-decentralized network largely alleviates the network dependency of communication overheads in decentralized networks via a few agent-to-server communication rounds. For large and sparse networks (i.e., the mixing rate $\lambda_w \rightarrow 0$), with a small probability $p = O(\sqrt{\lambda_w})$ of agent-to-server communication, the network dependency improves from $O(\lambda_w^{-2})$ to $O(\lambda_w^{-1})$.
3. We conduct extensive numerical experiments on real-world datasets over various network topologies to show the superior communication efficiency of PISCO. The results substantiate the convergence speedup brought by multiple local updates and the robustness of PISCO to data heterogeneity and various topologies, even for locally disconnected networks.

1.2 Related works

Over the past few years, distributed optimization has attracted growing attention and has been extensively explored. For the convenience of our readers, we provide a review of the most related works below.

Distributed nonconvex optimization. As the size of the networked system increases, there are considerable algorithms developed for solving distributed nonconvex optimization. Roughly speaking, they can be categorized into two classes — decentralized algorithms [NO09, LS16, HHZ17, TLY+18] where agents are only allowed to exchange information with neighbors, and federated algorithms [LHY+20, KKM+20, WLC22] where agents are able to communicate with the server directly. Early attempts apply (stochastic) gradient descent to distributed optimization, which performs well in practice [NO09, LHY+20]. However, the dissimilarity among local objectives could degenerate the performance under heterogeneous data and thus requires additional assumptions like bounded gradient or diminishing step-sizes. To eliminate such strong assumptions, many following works [LS16, HHZ17, KKM+20] have been developed, including gradient tracking (GT) techniques. The key idea of GT is to utilize dynamic average consensus [ZM10] for global gradient estimation, which has been incorporated with many distributed optimization algorithms to achieve faster convergence rates in nonconvex settings [SS19, SLH20, SMSC22, ZLL+]. Our proposed PISCO also takes advantage of GT to inherit its robustness against data heterogeneity.

Communication-efficient distributed ML. Communication efficiency is one of the most important bottlenecks in distributed ML. In the decentralized setting, the communication complexity largely depends on the network topology, i.e., poor connectivity slows down the information mixing and thus requires more communication rounds to consensus [NOR18]. As for the federated ML, the communication burden of the centralized server may be unaffordable. To overcome such bottlenecks, a number of strategies are proposed for improving communication efficiency [CBD+23], including: 1) *compression methods*: compressing the information for communication [KLSJ20, ZLL+]; 2) *multiple local communication and updates*: executing multiple gossip communication [LCCC20, CYZ+21] or successive local updates within one communication round [NAYU22, GC23, LHY+20, KKM+20], and 3) *heterogeneous communication model over semi-decentralized networks*: allowing both agent-to-agent and agent-to-server communication [CYZ+21, SSM+23, LHA+21]. In this paper, we aim to gain the best communication efficiency from probabilistic heterogeneous communication over semi-decentralized networks and enable multiple local updates for more communication savings.

Semi-decentralized ML. As mentioned above, semi-decentralized ML, resorting to heterogeneous communication, tackles both the communication bottleneck of the centralized server in federated ML and the heavy network dependency of decentralized ML. We mainly summarize the works in nonconvex setting that are mostly related to this paper, while referring readers to [CYZ+21, PHC+23] for the (strongly) convex setting. For nonconvex problems, Gossip-PGA [CYZ+21] first integrates Gossip SGD [NO09] with periodical global averaging and obtains a better scalability. It shows that intermittently communicating with the server can largely alleviate the heavy dependence on the network connectivity, especially for large or sparse networks. However, the theoretical results depend on the assumption of bounded similarity between local objectives. Moreover, HL-SGD [GSHG22] extends Gossip-PGA to the hierarchical networked structure with multiple clusters, while HA-Fed [WFC22] can be viewed as HL-SGD with momentum. Both of them enable intra-cluster gossip averaging and inter-cluster averaging, but also rely on the data heterogeneity assumptions which may be impractical in many real-world applications. Noted that all of them consider deterministic heterogeneous communication, i.e., agents/cluster can only communicate with the server every H communication rounds, but the synchronization largely depends on the availability of the server. To this end, we consider the probabilistic communication model, where agents only exchange information with the server at the probability p . Furthermore, to the best of our knowledge, none of these approaches enables multiple local updates within a single communication round, whereas ours benefits from the linear speedup provided by the local updates. More detailed comparison can be found in Table 1.

Algorithm	Convergence rate ¹		Bounded dissimilarity	Accessibility of server	Multiple local updates
	Mini-batch	Large-batch			
Gossip-PGA [CYZ ⁺ 21]	$O\left(\frac{1}{\sqrt{nK}}\right)$	$O\left(\left(\frac{G}{K}\right)^{\frac{2}{3}}\right)$	✓	every H rounds	✗
HL-SGD [GSHG22]	$O\left(\frac{1}{\sqrt{nK}}\right)$	$O\left(\left(\frac{G_1+G_2}{K}\right)^{\frac{2}{3}}\right)$	✓	every H rounds	✗
This paper	$O\left(\frac{1}{\sqrt{nT_oK}}\right)$	$O\left(\frac{1}{K}\right)$	✗	w.p. p	✓

Table 1: Comparison of ours and other semi-decentralized algorithms in nonconvex optimization, regarding to the convergence rate, algorithm design and data heterogeneity assumptions. Here, K is the number of communication rounds, n is the number of agents, T_o is the number of multiple local updates within a single gossip/global communication round, G is the quantity in bounded dissimilarity between local objective and the global objective, G_1 is the quantity in bounded intra-cluster dissimilarity as [GSHG22, Assumption 4], G_2 is the quantity in bounded inter-cluster dissimilarity as [GSHG22, Assumption 5].

1.3 Notation

Throughout this paper, we use the lowercase and uppercase boldface letters to represent vectors and matrices, respectively. We use $\|\mathbf{A}\|_F$ for the Frobenius norm of a matrix \mathbf{A} , $\|\mathbf{A}\|_2$ for the largest singular value of a matrix \mathbf{A} , $\|\mathbf{a}\|_2$ for the l_2 norm of a vector \mathbf{a} , and \otimes for the Kronecker product. In addition, we use \mathbf{I}_n for the identity matrix of dimension n , $\mathbf{1}_n$ for the all-one vector of dimension n and $\mathbf{O}_{d \times n}$ for the all-zero matrix of dimension $(d \times n)$. For any two real functions $f(\cdot)$ and $g(\cdot)$ defined on \mathbb{R}^+ , $f(x) = O(g(x))$ if there exists a positive real constant M and x_0 such that $f(x) \leq Mg(x)$ for any $x \geq x_0$. Note that “ \leq ” can be interpreted in an element-wise fashion, if it is applied to vectors or matrices with the same dimension.

2 Preliminaries

Communication graphs. Consider a *semi-decentralized network* that has a centralized server to coordinate all n agents and an undirected *communication graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ denotes the set of n agents and $\mathcal{E} \subseteq \{\{i, j\} | i, j \in \mathcal{V}\}$ represents the local communication between agents. For every agent $i \in \mathcal{V}$, let $\mathcal{N}_i = \{j | \{i, j\} \in \mathcal{E}\}$ denote agent i ’s neighbors whom the agent i can communicate with.

Mixing. For any communication graph \mathcal{G} , the mixing of local communication can be formally characterized by the *mixing matrix* $\mathbf{W} = [w_{ij}]_{1 \leq i, j \leq n}$ defined in Definition 1.

Definition 1 (Mixing matrix and mixing rate). *Given an undirected communication graph \mathcal{G} , a nonnegative matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ is the mixing matrix, whose element $w_{ij} = 0$ if and only if $\{i, j\} \notin \mathcal{E}$ and $i \neq j$ and \mathbf{W} is doubly stochastic, i.e.,*

$$\mathbf{W}\mathbf{1}_n = \mathbf{1}_n \text{ and } \mathbf{1}_n^\top \mathbf{W} = \mathbf{1}_n^\top.$$

The mixing rate of \mathbf{W} is a nonnegative constant, i.e.,

$$\lambda_w := 1 - \left\| \mathbf{W} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \right\|_2^2.$$

Note that the doubly stochasticity implies $\|\mathbf{W}\|_2 \leq 1$ and the mixing rate $\lambda_w = 1 - \lambda^2 \in [0, 1]$, where λ denotes the second largest eigenvalue. The mixing rate can depict the connectivity of the communication graph \mathcal{G} , or to say, the speed of information mixing. Mathematically, for any vector $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{W}\mathbf{x} - \bar{\mathbf{x}}\|_2^2 \leq (1 - \lambda_w) \|\mathbf{x} - \bar{\mathbf{x}}\|_2^2,$$

where $\bar{\mathbf{x}} = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \mathbf{x} \in \mathbb{R}^n$. In other words, a larger mixing rate λ_w indicates a better connectivity as well as a faster process of information mixing, while disconnected graphs have $\lambda_w = 0$. For example, considering a

fully connected graph where every agent can communicate with each other, the mixing matrix can be defined as

$$\mathbf{J} := \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top,$$

and its mixing rate is equal to 1. Specifically, in this paper, we use \mathbf{J} to describe the mixing of the agent-to-server communication.

Stochastic gradient methods. To improve the computational efficiency, one popular approach is to apply stochastic gradient on the mini-batch data samples, instead of using the full-batch gradient. In distributed setting, we define the local stochastic gradient for each agent $i \in [n]$ as:

$$\mathbf{g}_i = \frac{1}{b} \sum_{\mathbf{z}_i \in \mathcal{Z}_i} \nabla \ell(\mathbf{x}_i; \mathbf{z}_i), \quad (2)$$

where $\mathcal{Z}_i \subset \mathcal{D}_i$ denotes the sampled data batch with size $|\mathcal{Z}_i| = b < m$. In this paper, we assume that \mathcal{Z}_i is drawn i.i.d. from \mathcal{D}_i with the same mini-batch size b for every agent $i \in [n]$ and the local stochastic gradient \mathbf{g}_i is an unbiased estimate of $\nabla f_i(\mathbf{x}_i)$, i.e.,

$$\mathbb{E}[\mathbf{g}_i] = \nabla f_i(\mathbf{x}_i).$$

Gradient-tracking techniques. In many real-world applications, the local dataset \mathcal{D}_i on every agent $i \in [n]$ may be quite different from each other, referred to as *data heterogeneity*. Accordingly, there exists some local stationary solution \mathbf{x} satisfying $\nabla f_i(\mathbf{x}) = 0$ for some $i \in [n]$, but not necessarily with $\nabla f(\mathbf{x}) = \sum_{i=1}^n \nabla f_i(\mathbf{x}) = 0$. Under such circumstances, directly incorporating stochastic gradient methods with gossip or global averaging may not converge to the global stationary solution [NO09] without the strong assumption like bounded data dissimilarity.

To address this issue, gradient-tracking (GT) techniques [LS16, NOS17, QL18] have been proposed, which utilizes gossip communication for global gradient estimation leveraging dynamic average consensus [ZM10]. To improve computational efficiency, DSGT [PN21] incorporates GT with stochastic gradient methods. The updates at the k -th iteration are defined as: every agent $i \in [n]$ updates its optimization variable \mathbf{x}_i^k and gradient-tracking variable \mathbf{y}_i^k by

$$\begin{aligned} \mathbf{x}^{k+1} &= \sum_{j=1}^n w_{ij} (\mathbf{x}_j^k - \eta \mathbf{y}_j^k), \\ \mathbf{y}_i^{k+1} &= \sum_{j=1}^n w_{ij} \mathbf{y}_j^k + \mathbf{g}_i^{k+1} - \mathbf{g}_i^k, \end{aligned}$$

where $\eta > 0$ is the step size, the initialization $\mathbf{y}_i^0 = \mathbf{g}_i^0$, and $\mathbf{g}_i^k = \frac{1}{b} \sum_{\mathbf{z}_i^k \in \mathcal{Z}_i^k} \nabla \ell(\mathbf{x}_i^k; \mathbf{z}_i^k)$ represents the local stochastic gradient over the sampled data batch \mathcal{Z}_i^k drawn i.i.d. from \mathcal{D}_i with mini-batch size b . More recent works [LLKS23, GC23] prove that DSGT with multiple local updates is able to converge under high data heterogeneity in nonconvex setting.

3 Proposed Algorithm

In this section, we introduce PISCO, which exploits communication heterogeneity from the probabilistic communication model and inherits the robustness to data heterogeneity from GT. Before the depiction of PISCO, we first introduce some compact notation for convenience. Let $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}^d$ denote the local optimization variable and gradient-tracking variable at every agent $i \in [n]$, respectively. Accordingly, we denote the collection of all the optimization variables and gradient-tracking variables by the matrix form

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n} \quad \text{and} \quad \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbb{R}^{d \times n}.$$

We define the distributed objective as $F(\mathbf{X}) = \sum_{i=1}^n f_i(\mathbf{x}_i)$, where the distributed gradient is denoted by

$$\nabla F(\mathbf{X}) = [\nabla f_1(\mathbf{x}_1), \nabla f_2(\mathbf{x}_2), \dots, \nabla f_n(\mathbf{x}_n)] \in \mathbb{R}^{d \times n}.$$

Algorithm 1 PISCO for semi-decentralized nonconvex optimization

- 1: **input:** $\mathbf{X}^0 = \mathbf{x}^0 \mathbf{1}_n^\top$, local-update and communication step sizes η_l, η_c , number of iterations K , number of local updates T_o , mini-batch size b .
- 2: **initialize:** Draw the mini-batch $\mathcal{Z}^0 = \{\mathcal{Z}_i^0\}_{i=1}^n$ randomly and set $\mathbf{Y}^0 = \mathbf{G}^0 = \frac{1}{b} \nabla \ell(\mathbf{X}^0; \mathcal{Z}^0)$.
- 3: **for** $k = 0, 1, \dots, K - 1$ **do**
- 4: Set $\mathbf{X}^{k+1,0} = \mathbf{X}^k$, $\mathbf{Y}^{k+1,0} = \mathbf{Y}^k$ and $\mathbf{G}^{k+1,0} = \mathbf{G}^k$.
- 5: **for** $t = 1, 2, \dots, T_o$ **do** ▷ T_o -local updates
- 6: Draw the mini-batch $\mathcal{Z}^{k+1,t} = \{\mathcal{Z}_i^{k+1,t}\}_{i=1}^n$ and compute

$$\mathbf{X}^{k+1,t} = \mathbf{X}^{k+1,t-1} - \eta_l \mathbf{Y}^{k+1,t-1}, \quad (3a)$$

$$\mathbf{G}^{k+1,t} = \frac{1}{b} \nabla \ell(\mathbf{X}^{k+1,t}; \mathcal{Z}^{k+1,t}), \quad (3b)$$

$$\mathbf{Y}^{k+1,t} = \mathbf{Y}^{k+1,t-1} + \mathbf{G}^{k+1,t} - \mathbf{G}^{k+1,t-1}. \quad (3c)$$

- 7: **end for**
- 8: Define $\mathbf{W}^k = \begin{cases} \mathbf{J} & \text{with probability } p, \\ \mathbf{W} & \text{otherwise.} \end{cases}$ ▷ probabilistic heterogeneous communication
- 9: Draw the mini-batch $\mathcal{Z}^{k+1} = \{\mathcal{Z}_i^{k+1}\}_{i=1}^n$ and update

$$\mathbf{X}^{k+1} = \left((1 - \eta_c) \mathbf{X}^k + \eta_c (\mathbf{X}^{k+1, T_o} - \eta_l \mathbf{Y}^{k+1, T_o}) \right) \mathbf{W}^k, \quad (4a)$$

$$\mathbf{G}^{k+1} = \frac{1}{b} \nabla \ell(\mathbf{X}^{k+1}; \mathcal{Z}^{k+1}), \quad (4b)$$

$$\mathbf{Y}^{k+1} = \left(\mathbf{Y}^{k+1, T_o} + \mathbf{G}^{k+1} - \mathbf{G}^{k+1, T_o} \right) \mathbf{W}^k. \quad (4c)$$

10: **end for**

Similarly, we denote the gradient of empirical loss given the sampling dataset $\mathcal{Z} = \{\mathcal{Z}_i\}_{i=1}^n$ as

$$\nabla \ell(\mathbf{X}; \mathcal{Z}) = \left[\sum_{\mathbf{z}_1 \in \mathcal{Z}_1} \nabla \ell(\mathbf{x}_1; \mathbf{z}_1), \sum_{\mathbf{z}_2 \in \mathcal{Z}_2} \nabla \ell(\mathbf{x}_2; \mathbf{z}_2), \dots, \sum_{\mathbf{z}_n \in \mathcal{Z}_n} \nabla \ell(\mathbf{x}_n; \mathbf{z}_n) \right] \in \mathbb{R}^{d \times n},$$

where $\{\mathcal{Z}_i\}_{i=1}^n$ is randomly drawn from $\{\mathcal{D}_i\}_{i=1}^n$ with the mini-batch size b . With the local stochastic gradient defined in (2), the distributed stochastic gradient can be represented by

$$\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n] = \frac{1}{b} \nabla \ell(\mathbf{X}; \mathcal{Z}).$$

Then, we are ready to describe PISCO detailed in Algorithm 1, using the above compact notations. At the beginning of the k -th communication round, PISCO maintains the model estimate \mathbf{X}^k , the global gradient estimate \mathbf{Y}^k and the distributed stochastic gradient \mathbf{G}^k . It then boils down to the following two stages for achieving both communication efficiency and exact convergence under data heterogeneity.

- The first stage is to execute T_o local steps without any communication (cf. line 4-7). The key idea is to utilize the local computational resources to facilitate the convergence. At the beginning of the local updates, initialize the local-update variables $\mathbf{X}^{k+1,0} = \mathbf{X}^k$, $\mathbf{Y}^{k+1,0} = \mathbf{Y}^k$ and $\mathbf{G}^{k+1,0} = \mathbf{G}^k$. At the t -th local update, update $\{\mathbf{X}^{k+1,t}, \mathbf{Y}^{k+1,t}, \mathbf{G}^{k+1,t}\}$ via (3), maintaining the fashion of gradient-tracking techniques.
- The second stage is to perform information exchange over the semi-decentralized network via a probabilistic communication model (cf. line 8-10), i.e., there are two possible communication schemes — agent-to-server communication with probability p and agent-to-agent communication otherwise. Different schemes correspond to different mixing matrices (cf. line 8), i.e., if agents implement the global communication, set $\mathbf{W}^k = \mathbf{J}$; otherwise, set $\mathbf{W}^k = \mathbf{W}$. Then, agents update $\{\mathbf{X}^{k+1}, \mathbf{Y}^{k+1}, \mathbf{G}^{k+1}\}$

via (4), using the output of local updates $\{\mathbf{X}^{k+1, T_o}, \mathbf{Y}^{k+1, T_o}, \mathbf{G}^{k+1, T_o}\}$ via the selected communication scheme.

4 Theoretical Guarantees

In this section, we provide the convergence results of our PISCO under different settings: PISCO converges at a rate of $O(1/\sqrt{nT_oK})$ using mini-batch gradients and $O(1/(nK))$ with full-batch gradients.

4.1 Assumptions

Before proceeding to the results, we first impose the following assumptions on the network model, objective functions and data sampling.

Assumption 1 (Semi-decentralized network model). *Given the underlying graph \mathcal{G} and its mixing matrix \mathbf{W} following the Definition 1, then \mathbf{W}^k defined in Algorithm 1 satisfies that for any $\mathbf{x} \in \mathbb{R}^n$*

$$\mathbb{E}[\|\mathbf{W}^k \mathbf{x} - \bar{\mathbf{x}}\|_2^2] \leq (1 - \lambda_p) \|\mathbf{x} - \bar{\mathbf{x}}\|_2^2,$$

where $\bar{\mathbf{x}} = \mathbf{J}\mathbf{x} \in \mathbb{R}^n$ and the expected mixing rate $\lambda_p = \lambda_w + p(1 - \lambda_w) \in (0, 1]$.

Noted that Assumption 1 is weaker than the connected assumption in prior semi-decentralized literatures [CYZ⁺21, GSHG22, WFC22], i.e., $\lambda_w > 0$. More specifically, Assumption 1 implies that the underlying graph can be disconnected if and only if $p > 0$. Only in the case that the centralized server is unavailable (i.e., $p = 0$), Assumption 1 presumes the connectivity of \mathcal{G} , i.e., $\lambda_p = \lambda_w > 0$.

Regarding to the objective functions, we assume that the optimal value $f^* := \min_{\mathbf{x}} f(\mathbf{x})$ exists and $f^* > -\infty$. The local objective functions $\{f_i\}_{i=1}^n$ could be nonconvex but satisfy the standard smoothness assumption provided below.

Assumption 2 (L -smooth). *Each local function $f_i(\mathbf{x})$ is differentiable and there exists a constant L such that for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2.$$

In addition, we assume that the local stochastic gradient \mathbf{g}_i is an unbiased estimate with a bounded variance, which is widely used in the literature [CYZ⁺21, WFC22, GSHG22, GC23, KLS21, ZLL⁺].

Assumption 3 (Unbiased gradient and bounded variance). *For every agent $i \in [n]$, the local stochastic gradient \mathbf{g}_i is unbiased, i.e., $\mathbb{E}[\mathbf{g}_i] = \nabla f_i(\mathbf{x}_i)$, and there exists a constant $\sigma \geq 0$ s.t.*

$$\mathbb{E}_{\mathcal{Z}_i \sim \mathcal{D}_i}[\|\mathbf{g}_i - \nabla f_i(\mathbf{x}_i)\|_2^2] \leq \sigma^2/b.$$

or in the compact form

$$\mathbb{E}_{\mathcal{Z} \sim \mathcal{D}}[\|\mathbf{G} - \nabla F(\mathbf{X})\|_F^2] \leq \frac{n\sigma^2}{b}.$$

Note that in the case of the full-batch gradients, i.e., the mini-batch size $b = m$, we can simply set $\sigma = 0$ and thus Assumption 3 still holds.

4.2 Convergence analysis of PISCO

Now, we are ready to present our main results. First, the following theorem demonstrates that our proposed PISCO algorithm is able to converge to the neighborhood of the stationary solution to the problem (1) at a rate of $O(1/K)$ with constant step-sizes.

Theorem 1 (Convergence rate with mini-batch gradients). *Suppose Assumption 1, 2 and 3 hold. Let $\tilde{f} = f(\bar{\mathbf{x}}^0) - f^*$ and $\Phi_y^0 = \mathbf{Y}^0 - \mathbf{Y}^0 \mathbf{J}$. For any $\alpha \geq 1$ s.t. $\eta_c = \alpha \sqrt{(1+p)} \lambda_p$, $\eta_l \leq \frac{\sqrt{(1+p)} \lambda_p}{360 \alpha L (T_o + 1)}$, we have*

$$\begin{aligned} & \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \\ & \leq \underbrace{O\left(\frac{\tilde{f}}{\eta T_o K} + \left(L^2 T_o^2 \eta_l^2 + \frac{L\eta}{n}\right) \frac{\sigma^2}{b}\right)}_{\text{terms due to SGD and local updates}} + \underbrace{O\left(\frac{(1-p)L^2 T_o^2 \eta^2 \sigma^2}{(1+p)^2 \lambda_p^4} \frac{\sigma^2}{b} + \frac{1}{nK} \mathbb{E}[\|\Phi_y^0\|_F^2]\right)}_{\text{terms due to decentralized overhead}}, \end{aligned} \quad (5)$$

where the average model estimate $\bar{\mathbf{x}}^k = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^k \in \mathbb{R}^d$ and $\eta = \eta_c \eta_l$.

Proof. See Appendix A. □

Noted that the above convergence result can hold even under significant data heterogeneity across agents, since we do not assume any bounded similarity between local objectives. Because of the existence of local variance σ^2 , we fine-tune the local-update step-size to obtain the exact convergence rate with the leading term $O(1/\sqrt{nT_o K})$, given by the following corollary.

Corollary 1 (Convergence rate with mini-batch gradients). *Suppose all the conditions in Theorem 1 hold. If the step-sizes satisfy $\eta_c = \alpha \sqrt{1+p} \lambda_p$ and*

$$\eta_l = \frac{1}{\alpha T_o} \min\left\{\left(\frac{nbT_o \tilde{f}}{(1+p)\lambda_p^2 L \sigma^2 K}\right)^{1/2}, \left(\frac{\sqrt{1+p} \lambda_p b \tilde{f}}{(1-p)L^2 \sigma^2 K}\right)^{1/3}, \left(\frac{\alpha^2 b \tilde{f}}{\sqrt{1+p} \lambda_p L^2 \sigma^2 K}\right)^{1/3}\right\}$$

and K is sufficiently large to satisfy (42), PISCO converges at the rate of

$$O\left(\left(\frac{L\sigma^2}{nT_o bK}\right)^{\frac{1}{2}} + \left(\frac{L\sigma}{\alpha \sqrt{1+p} \lambda_p \sqrt{b} K}\right)^{\frac{2}{3}} + \left(\frac{\sqrt{1-p} L \sigma}{(1+p)\lambda_p^2 \sqrt{b} K}\right)^{\frac{2}{3}} + \frac{1}{nK}\right). \quad (6)$$

Proof. See Appendix B. □

From Corollary 1, PISCO can achieve the ϵ -accuracy, i.e., $\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \leq \epsilon^2$ after

$$O\left(\frac{L\sigma^2}{nT_o b \epsilon^4} + \frac{L\sigma}{\alpha \sqrt{1+p} \lambda_p \sqrt{b} \epsilon^3} + \frac{\sqrt{1-p} L \sigma}{(1+p)\lambda_p^2 \sqrt{b} \epsilon^3} + \frac{1}{n\epsilon^2}\right) \quad (7)$$

communication rounds. Notice that if K is sufficiently large and the first term $\left(\frac{L\sigma^2}{nT_o bK}\right)^{\frac{1}{2}}$ correspondingly becomes dominant, increasing the number of agents n or the number of local updates T_o can accelerate the convergence. Such a speedup matches the findings in the special cases of semi-decentralized ML, i.e., decentralized setting [LLKS23, GC23] when $p = 0$ and federated setting [KKM⁺20] when $p = 1$.

In fact, PISCO can be generalized to the decentralized case and federated case by setting $p = 0$ and $p = 1$ respectively, while maintaining comparable convergence guarantees.

Remark 1 (Decentralized cases). When $p = 0$, Algorithm 1 becomes fully decentralized, i.e., agents only perform local communication. Then, the communication complexity to achieve ϵ -accuracy becomes

$$O\left(\frac{L\sigma^2}{nT_o b \epsilon^4} + \frac{L\sigma}{\lambda_w^2 \sqrt{b} \epsilon^3} + \frac{1}{n\epsilon^2}\right),$$

which is better than the rate of Periodical-GT in [LLKS23] and LSGT [GC23], since the network dependency is $O(1/\lambda_w^2)$ and only appears in the second term (see Table 2). The second term is slightly worse than K -GT, the variance-reduction version of Periodical-GT [LLKS23], because it corrects the descent direction as the average of T_o local updates instead of the last local update at communication. However, their theoretical guarantee requires the initial local correction variables to be settled in a centralized way. Combining our analysis with such a variance-reduction method presented in [LLKS23] while avoiding the centralized initialization would be a promising future direction of this paper.

Algorithm	# Agent-to-server communication	# Agent-to-agent communication
SCAFFOLD [KKM ⁺ 20]	$O\left(\frac{\sigma^2}{nT_o\epsilon^4} + \frac{1}{\epsilon^2}\right)$	0
LSGT [GC23]	0	$O\left(\frac{\sigma^4}{nT_o\lambda_w^8\epsilon^4} + \frac{1}{nT_o^{1/3}\lambda_w^{8/3}\epsilon^{4/3}} + \frac{1}{nT_o\epsilon^2}\right)$
Periodical-GT [LLKS23]	0	$O\left(\frac{\sigma^2}{nT_o\epsilon^4} + \frac{\sigma}{\lambda_w^2\epsilon^3} + \frac{1}{\lambda_w^2\epsilon^2}\right)$
K-GT [LLKS23]	0	$O\left(\frac{\sigma^2}{nT_o\epsilon^4} + \frac{\sigma}{\lambda_w^2\sqrt{T_o}\epsilon^3} + \frac{1}{\lambda_w^2\epsilon^2}\right)$
This paper	$O\left(\frac{p\sigma^2}{nT_o\epsilon^4} + \frac{p\sigma}{(\lambda_w + p)^2\epsilon^3} + \frac{p}{n\epsilon^2}\right)$	$O\left(\frac{(1-p)\sigma^2}{nT_o\epsilon^4} + \frac{(1-p)\sigma}{(\lambda_w + p)^2\epsilon^3} + \frac{1-p}{n\epsilon^2}\right)$

Table 2: The number of (expected) server/local communication rounds of existing distributed algorithms with stochastic gradients and local updates, to achieve the ϵ -accuracy. Here, n is the number of agents, T_o is the number of local updates, λ_w is the mixing rate of the underlying graph.

Remark 2 (Federated case). When $p = 1$, every agent can communicate with the server directly and thus PISCO performs like federated algorithms with local updates at any iteration $k \geq 1$. Correspondingly, the communication complexity becomes

$$O\left(\frac{L\sigma^2}{nT_ob\epsilon^4} + \frac{L\sigma}{\sqrt{b}\epsilon^3} + \frac{1}{n\epsilon^2}\right),$$

where the leading term is the same as that of SCAFFOLD in [KKM⁺20] with the linear speedup in terms of the network size n and the number of local updates T_o .

Moreover, the highlight of our work is to fill the void of semi-decentralized ML with the probabilistic communication model and gain the best communication efficiency from both agent-to-agent communication and agent-to-server communication, as shown in Table 2. In addition, PISCO is able to improve the network dependency of the communication overheads from $O(\lambda_w^{-2})$ [LLKS23] to $O(\lambda_p^{-2})$, where the trade-off between the communication probability and the network dependency is illustrated in Figure 3. The flexible heterogeneous communication brings the superior communication efficiency of PISCO in both well-connected and poorly-connected networks.

Implications for well-connected networks. As gossip communication is efficient to mix information for well-connected networks, PISCO is able to achieve a comparable convergence rate with much fewer agent-to-server communication rounds compared with using only agent-to-server communication. Therefore, our PISCO can significantly reduce the communication costs for well-connected networks whenever local agent-to-agent communications are inexpensive.

Implications for poorly-connected networks. When $\lambda_w \rightarrow 0$, performing agent-to-agent only communication often results in a large number of communication rounds and prohibitive communication costs. Surprisingly, as shown in Figure 3, with any probability $p \geq \lambda_w$, the network dependency can be reduced to $O(p^{-2})$. More specifically, even a small agent-to-server probability $p = O(\sqrt{\lambda_w})$ can significantly improve the network dependency from $O(\lambda_w^{-2})$ to $O(\lambda_w^{-1})$. Take the large-scale path graph as an example, where the mixing rate λ_w scales on the order of $O(1/n^2)$ [NOR18]. Our PISCO with $p = O(1/n)$ can improve the network dependency from $O(n^4)$ to $O(n^2)$. Moreover, if $p = O(1)$, the communication complexity can be network-independent like Gossip-PGA [CYZ⁺21], but our theoretical analysis does not require the additional assumption of the bounded dissimilarity between local objectives.

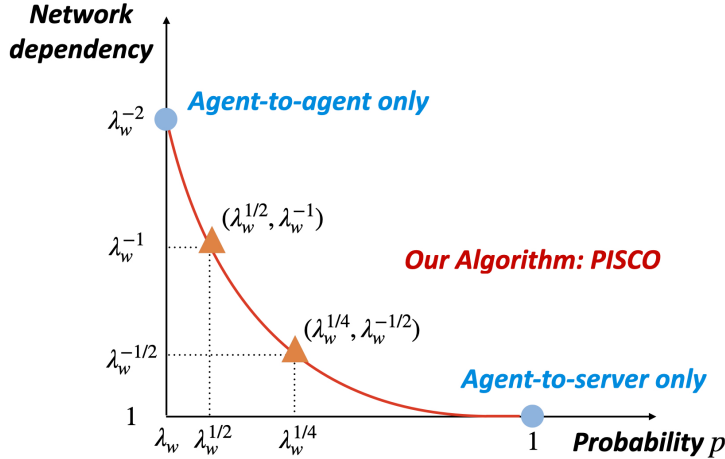


Figure 3: The network dependency of PISCO regarding to agent-to-server communication probability p .

In many real-world scenarios, it is also popular to choose large mini-batch size b to guarantee the exact convergence to the stationary point. As the terms related to local variance on the RHS of (5) scale on the order of $O(\sigma^2/b)$, if we choose a large enough mini-batch size $b = O(\sigma^2/\epsilon^2)$, the following result holds.

Corollary 2 (Communication complexity with large batch size). *Suppose all the conditions in Theorem 1 holds. If the mini-batch size is sufficiently large, i.e., $O(\frac{\sigma^2}{\epsilon^2})$, it holds $\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{x}^k)\|_2^2] \leq \epsilon^2$ after*

$$O\left(\frac{L}{(1+p)\lambda_p^2\epsilon^2} + \frac{1}{\epsilon^2}\right)$$

communication rounds. In addition, if the mini-batch size $b = m$, i.e., we take the full-batch gradient, the communication complexity will be improved to

$$O\left(\frac{L}{(1+p)\lambda_p^2\epsilon^2} + \frac{1}{n\epsilon^2}\right).$$

Note that Corollary 2 matches the result in decentralized setting [NAYU22] and federated setting [KKM⁺20], by setting $p = 0$ and $p = 1$ respectively.

5 Numerical Experiments

In this section, we present the numerical performance of PISCO on real-world datasets, to substantiate its superior performance in terms of communication efficiency and robustness to various topologies and data heterogeneity.

5.1 Logistic regression with nonconvex regularization

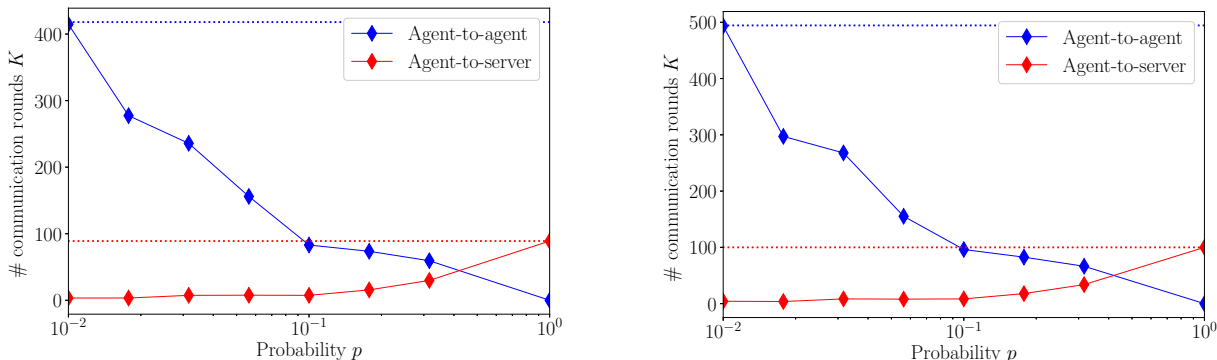
To investigate communication efficiency of PISCO, we conduct experiments on logistic regression with a nonconvex regularization term [WJZ⁺19] using the a9a dataset [CL11]. Given the model parameter \mathbf{x} and data sample $\mathbf{z} = (\mathbf{a}, y)$, the empirical loss $\ell(\mathbf{x}; \mathbf{z})$ is defined as:

$$\ell(\mathbf{x}; \mathbf{z}) = \log(1 + \exp(-y\mathbf{a}^\top \mathbf{x})) + \rho \sum_{l=1}^d \frac{\mathbf{x}(l)^2}{1 + \mathbf{x}(l)^2},$$

where $\mathbf{a} \in \mathbb{R}^d$ is the feature vector, $y \in \{-1, 1\}$ is the corresponding label, the regularizer coefficient ρ is set as 0.01, and $\mathbf{x}(l)$ denotes the l -th coordinate of \mathbf{x} .

In this subsection, we consider a ring topology with $n = 10$ agents and evenly partition the sorted **a9a** dataset to 10 agents to augment the data heterogeneity. Roughly speaking, every agent will receive $m = 3256$ training samples of dimension $d = 124$, where 5 agents will receive data with label 1 and the others will receive data with label 0. Regarding the mixing matrix, we follow the symmetric FDLA matrix [XB04] to aggregate information among neighbors. In addition, we set the batch size $b = 256$ for the following experiments. To reduce the impact of randomness, we run every experiment with 5 different seeds and show the average results.

The impact of different agent-to-server probabilities. First, we study the influence of the agent-to-server communication probability p on the training and test performance. To this end, we vary the probability p from $\{1, 1/10^{0.5}, 1/10^{0.75}, 1/10, 1/10^{1.25}, 1/10^{1.5}, 1/10^{1.75}, 1/10^2, 0\}$ and presents the number of communication rounds of PISCO with different p to achieve 0.05 training accuracy (i.e., $\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(\bar{\mathbf{x}}^k)\|_2^2 \leq 0.05$) and 80% test accuracy ($\geq 95\%$ of the peak accuracy within 1000 communication rounds), in Figure 4.



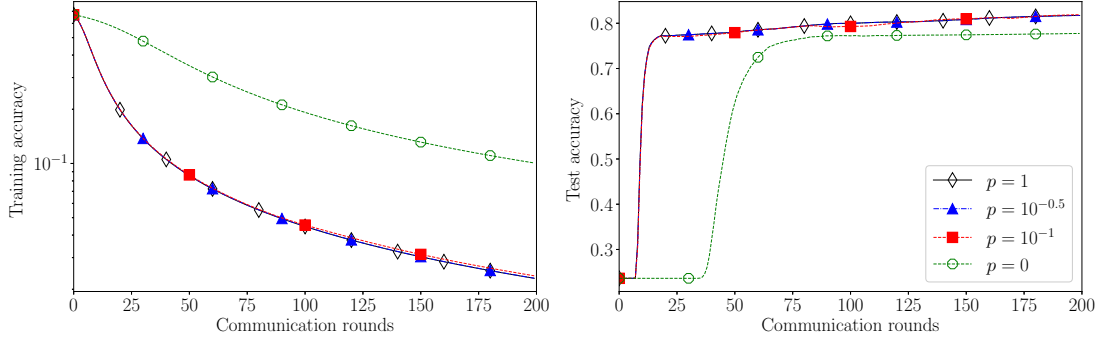
(a) Training accuracy $\frac{1}{K+1} \sum_{k=0}^K \|\nabla f(\bar{\mathbf{x}}^k)\|_2^2 \leq 0.05$.

(b) Test accuracy $\leq 80\%$.

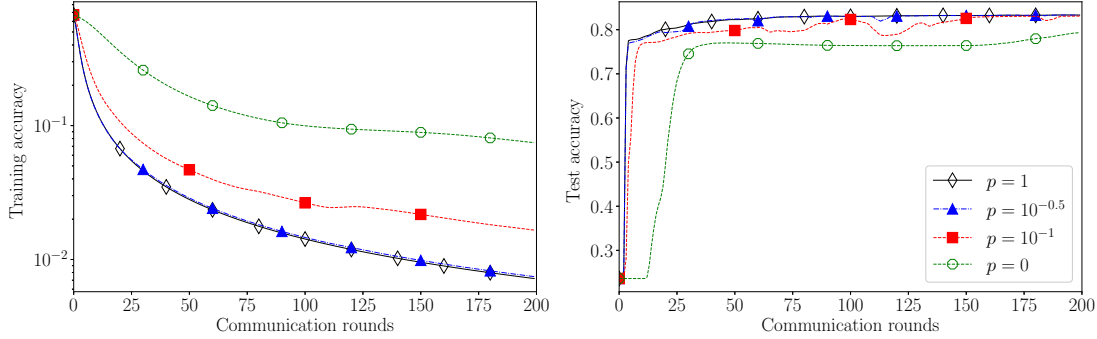
Figure 4: The number of agent-to-agent and agent-to-server communication rounds required to achieve 5% training accuracy (the left panel) and 80% test accuracy (the right panel) for PISCO with $T_o = 1$ and different $p \in \{0, 10^{-2}, 10^{-1.75}, 10^{-1.5}, 10^{-1.25}, 10^{-1}, 10^{-0.75}, 10^{-0.5}, 1\}$. Here, the blue (red) dotted line represents the number of agent-to-agent (agent-to-server) communication rounds that PISCO with $p = 0$ (with $p = 1$) requires.

From Figure 4, we observe that just a small agent-to-server probability (e.g., $p \leq 0.1$) can considerably reduce the number of communication rounds required to attain a specific accuracy during both training and testing phases. For instance, PISCO with $p = 10^{-1.25} \approx 0.06$ can reduce agent-to-agent communication rounds by 60%, with several agent-to-server communication rounds. Moreover, even if the server is more accessible (e.g., $p \geq 0.1$), increasing the agent-to-server communication probability p might not further save the total communication rounds. This indicates that not all costly communications between agents and the server are crucial for accelerating the convergence compared with decentralized methods. Therefore, by leveraging heterogeneous communication, we can reduce the average per-round communication expense while preserving a comparable rate of convergence.

The speedup of multiple local updates. To verify the speedup of multiple local updates, we plot the training accuracy and test accuracy of PISCO with different numbers of local updates $T_o = 1$ (cf. Figure 5(a)) and $T_o = 10$ (cf. Figure 5(b)). In both cases, we vary the probability $p \in \{1, 10^{-0.5}, 10^{-1}, 0\}$. It is worth noting that with only $p = 0.1$ or $p = 10^{-0.5}$, PISCO already achieves almost the same performance as PISCO with $p = 1$. Comparing Figure 5(b) with Figure 5(a), we can clearly observe the speedup brought by multiple local updates for different probabilities. For example, for PISCO with $p = 0.1$, the number of communication rounds required to attain 0.05 training accuracy or 80% testing accuracy decreases roughly by 50% if we increase T_o from 1 to 10.



(a) The number of local updates $T_o = 1$.



(b) The number of local updates $T_o = 10$.

Figure 5: The training accuracy (left two panels) and testing accuracy (right two panels) against communication rounds with different probabilities $p = 1, 10^{-0.5}, 10^{-1}, 0$ and different number of local updates $T_o = 1, 10$, over a ring topology for logistic regression with a nonconvex regularizer on the sorted **a9a** dataset.

5.2 One-hidden-layer neural network training

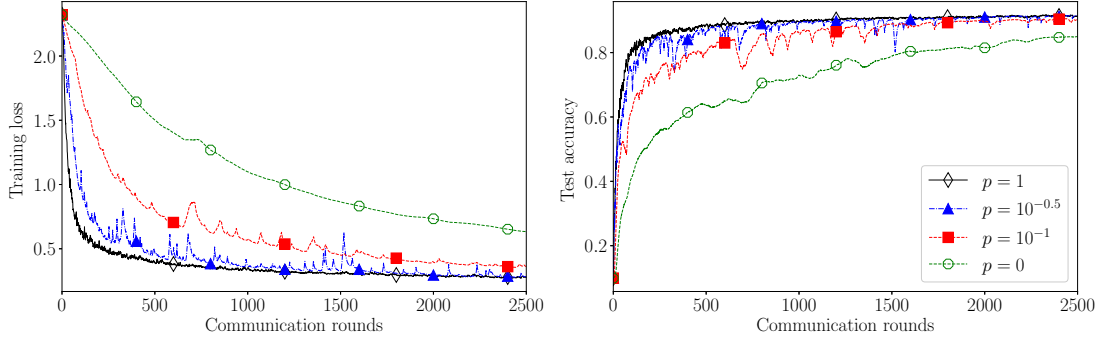
Further, we run the single hidden-layer neural network training with 32 hidden neurons on the MNIST dataset [Den12]. More specifically, we use the sigmoid and softmax function as the activation function, where the empirical loss w.r.t. the training parameter $\mathbf{x} = \text{vec}(\mathbf{W}_1, \mathbf{c}_1, \mathbf{W}_2, \mathbf{c}_2)$ and the sample $\mathbf{z} = (\mathbf{a}, y)$ is defined using the cross entropy loss as:

$$\text{CrossEntropy}(\text{softmax}(\mathbf{W}_2 \text{sigmoid}(\mathbf{W}_1 \mathbf{a} + \mathbf{c}_1) + \mathbf{c}_2), y),$$

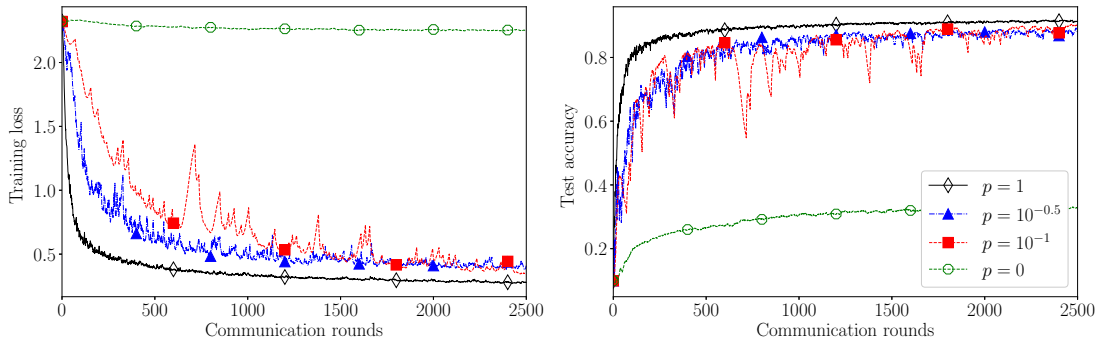
where the training weights $\mathbf{W}_1 \in \mathbb{R}^{32 \times 784}$, $\mathbf{W}_2 \in \mathbb{R}^{10 \times 32}$, $\mathbf{c}_1 \in \mathbb{R}^{32}$, and $\mathbf{c}_2 \in \mathbb{R}^{10}$.

To verify the robustness of PISCO to diverse topologies, we consider a well-connected Erdős-Rényi topology with a connectivity probability of 0.3 (corresponding to $\lambda_w = 0.38$) and a disconnected Erdős-Rényi topology with a connectivity probability of 0.1 (corresponding to $\lambda_w = 0$). To simulate the highly data-heterogeneous scenario, we evenly split the sorted MNIST dataset to $n = 10$ agents, where agent i will receive the training data associated with label i , for $i \in [n]$. Moreover, we set the batch size $b = 100$, the number of local updates $T_o = 10$ and the agent-to-server communication probability $p \in \{1, 1/\sqrt{n}, 1/n, 0\} = \{1, 10^{-0.5}, 10^{-1}, 0\}$. To reduce the impact of randomness, we run every experiment with 3 different seeds and show the average results in Figure 6.

In Figure 6, our PISCO shows impressive robustness to high data heterogeneity and different topologies, including the well-connected network (cf. Fig 6(a)) and the disconnected network (cf. Figure 6(b)). By comparing Figure 6(a) with Figure 6(b), we observe that better connectivity makes gossip communication sufficiently efficient to mix information. As a result, heterogeneous communication with a smaller p can attain a comparable performance to that of PISCO with $p = 1$ in Figure 6(a). Notice that the performance



(a) A connected Erdős-Rényi graph with 0.3 connectivity probability.



(b) A disconnected Erdős-Rényi graph with the 0.1 connectivity probability.

Figure 6: The training loss (the left two panels) and testing accuracy (the right two panels) against communication rounds with different probabilities $p = 1, 10^{-0.5}, 10^{-1}, 0$ and the number of local updates $T_o = 10$ over both well-connected and disconnected Erdős-Rényi graphs for 1-hidden-layer network training on the sorted MNIST dataset.

of PISCO with no agent-to-server communication degenerates remarkably when the network is disconnected. In contrast, semi-decentralized PISCO (i.e., $0 < p < 1$) maintains performance levels similar to PISCO with $p = 1$. It illustrates that a few number of agent-to-server communication rounds can largely mitigate the impact of the network connectivity, even for disconnected graphs.

6 Conclusions

In this paper, we develop a communication-efficient algorithm called PISCO for solving federated nonconvex optimization over semi-decentralized networks, which enjoys the linear speedup of local updates and addresses data dissimilarity without any additional assumptions. By leveraging the heterogeneous communication model, PISCO largely reduces communication overheads in terms of the network dependency with a few agent-to-server communication rounds, particularly evident in poorly-connected networks. Both theoretical guarantees and empirical experiments underscore PISCO’s outstanding communication efficiency and robustness to data heterogeneity and various network topologies. In the future, it will be of interest to incorporate variance reduction [LLC22] and communication compression [ZLL⁺] into the algorithm design.

Acknowledgement

This work is supported in part by the grants ONR N00014-19-1-2404, NSF CIF-2007911, CNS-2148212, ECCS-2318441, and AFRL FA8750-20-2-0504. He Wang is also gratefully supported by the Bob Lee Gregory

Fellowship at Carnegie Mellon University.

References

- [BNOT14] A. Beck, A. Nedić, A. Ozdaglar, and M. Teboulle. An $o(1/k)$ gradient method for network resource allocation problems. *IEEE Transactions on Control of Network Systems*, 1(1):64–73, 2014.
- [CBD⁺23] X. Cao, T. Başar, S. Diggavi, Y. C. Eldar, K. B. Letaief, H. V. Poor, and J. Zhang. Communication-efficient distributed learning: An overview. *IEEE Journal on Selected Areas in Communications*, 41(4):851–873, 2023.
- [CL11] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.
- [CRW21] I. E. Carvajal-Roca and J. Wang. A semi-decentralized security framework for connected and autonomous vehicles. In *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pages 1–6, 2021.
- [CYRC13] Y. Cao, W. Yu, W. Ren, and G. Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9(1):427–438, 2013.
- [CYZ⁺21] Y. Chen, K. Yuan, Y. Zhang, P. Pan, Y. Xu, and W. Yin. Accelerating gossip sgd with periodic global averaging. In *International Conference on Machine Learning*, pages 1791–1802, 2021.
- [Den12] L. Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- [GC23] S. Ge and T.-H. Chang. Gradient and variable tracking with multiple local sgd for decentralized non-convex learning. *arXiv preprint arXiv:2302.01537*, 2023.
- [GSHG22] Y. Guo, Y. Sun, R. Hu, and Y. Gong. Hybrid local SGD for federated learning with heterogeneous communications. In *International Conference on Learning Representations*, 2022.
- [HHZ17] M. Hong, D. Hajinezhad, and M.-M. Zhao. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1529–1538, 2017.
- [HJ12] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [KKM⁺20] 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*, pages 5132–5143. PMLR, 2020.
- [KLB⁺20] 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*, pages 5381–5393. PMLR, 2020.
- [KLS21] A. Koloskova, T. Lin, and S. U. Stich. An improved analysis of gradient tracking for decentralized machine learning. In *Advances in Neural Information Processing Systems*, 2021.
- [KLSJ20] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*, 2020.
- [LCCC20] B. Li, S. Cen, Y. Chen, and Y. Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *The Journal of Machine Learning Research*, 21(1):7331–7381, 2020.

- [LHA⁺21] F. P.-C. Lin, S. Hosseinalipour, S. S. Azam, C. G. Brinton, and N. Michelusi. Semi-decentralized federated learning with cooperative D2D local model aggregations. *IEEE Journal on Selected Areas in Communications*, 39(12):3851–3869, 2021.
- [LHY⁺20] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*, 2020.
- [LLC22] B. Li, Z. Li, and Y. Chi. DESTRESS: Computation-optimal and communication-efficient decentralized nonconvex finite-sum optimization. *SIAM Journal on Mathematics of Data Science*, 4(3):1031–1051, 2022.
- [LLKS23] Y. Liu, T. Lin, A. Koloskova, and S. U. Stich. Decentralized gradient tracking with local steps. *arXiv preprint arXiv:2301.01313*, 2023.
- [LS16] P. D. Lorenzo and G. Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.
- [MMR⁺17] 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*, pages 1273–1282. PMLR, 2017.
- [MSY⁺23] X. Miao, Y. Shi, Z. Yang, B. Cui, and Z. Jia. SDPipe: A semi-decentralized framework for heterogeneity-aware pipeline-parallel training. *Proc. VLDB Endow.*, 16(9):2354–2363, jul 2023.
- [NAYU22] E. D. H. Nguyen, S. A. Alghunaim, K. Yuan, and C. A. Uribe. On the performance of gradient tracking with local updates. *arXiv preprint arXiv:2210.04757*, 2022.
- [NMTA21] M. Navidi, S. M. Moghaddas-Tafreshi, and A. M. Alishvandi. A semi-decentralized framework for simultaneous expansion planning of privately owned multi-regional energy systems and sub-transmission grid. *International Journal of Electrical Power & Energy Systems*, 128:106795, 2021.
- [NO09] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- [NOR18] A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.
- [NOS17] A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.
- [PHC⁺23] R. Parasnis, S. Hosseinalipour, Y.-W. Chu, C. G. Brinton, and M. Chiang. Connectivity-aware semi-decentralized federated learning over time-varying D2D networks. *arXiv preprint arXiv:2303.08988*, 2023.
- [PN21] S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.
- [QL18] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.
- [RN04] M. Rabbat and R. Nowak. Distributed optimization in sensor networks. In *Third International Symposium on Information Processing in Sensor Networks, 2004. IPSN 2004*, pages 20–27, 2004.
- [SLH20] H. Sun, S. Lu, and M. Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International Conference on Machine Learning*, pages 9217–9228. PMLR, 2020.
- [SMSC22] Y. Sun, M. Maros, G. Scutari, and G. Cheng. High-dimensional inference over networks: Linear convergence and statistical guarantees. *arXiv preprint arXiv:2201.08507*, 2022.

- [SS19] G. Scutari and Y. Sun. Distributed nonconvex constrained optimization over time-varying di-graphs. *Mathematical Programming*, 176(1-2):497–544, 2019.
- [SSM⁺23] Y. Sun, J. Shao, Y. Mao, J. H. Wang, and J. Zhang. Semi-decentralized federated edge learning with data and device heterogeneity. *IEEE Transactions on Network and Service Management*, 20(2):1487–1501, 2023.
- [TLY⁺18] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu. D^2 : Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018.
- [WFC22] Y. Wang, P. Fang, and J. Chen. Accelerating adaptive federated optimization with local gossip communications. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022.
- [WJZ⁺19] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pages 2406–2416, 2019.
- [WLC22] Y. Wang, L. Lin, and J. Chen. Communication-efficient adaptive federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 22802–22838, 2022.
- [XB04] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.
- [ZLL⁺] H. Zhao, B. Li, Z. Li, P. Richtárik, and Y. Chi. BEER: Fast $O(1/T)$ rate for decentralized nonconvex optimization with communication compression. In *Advances in Neural Information Processing Systems*.
- [ZM10] M. Zhu and S. Martínez. Discrete-time dynamic average consensus. *Automatica*, 46(2):322–329, 2010.

A Proof of Theorem 1

In this section, we present the convergence analysis of PISCO. For the ease of the analysis, we begin with rewriting the updates of PISCO in a more compact form. First, by unrolling the local updates (3) in PISCO, we have

$$\mathbf{X}^{k+1, T_o} = \mathbf{X}^k - \eta_l \mathbf{Y}^k - \eta_l \sum_{t=1}^{T_o-1} \mathbf{Y}^{k+1, t}, \quad (8)$$

$$\mathbf{Y}^{k+1, T_o} = \mathbf{Y}^k + \sum_{t=1}^{T_o} \left(\mathbf{G}^{k+1, t} - \mathbf{G}^{k+1, t-1} \right).$$

Then, summing (3c) from $t = 1$ to $T_o - 1$ and telescoping results in

$$\begin{aligned} \sum_{t=1}^{T_o-1} \mathbf{Y}^{k+1, t} &= \sum_{t=1}^{T_o-1} \mathbf{Y}^{k+1, t-1} + \sum_{t=1}^{T_o-1} \left(\mathbf{G}^{k+1, t} - \mathbf{G}^{k+1, t-1} \right) \\ &= \mathbf{Y}^k + \sum_{t=1}^{T_o-2} \mathbf{Y}^{k+1, t-1} + \sum_{t=1}^{T_o-2} \left(\mathbf{G}^{k+1, t} - \mathbf{G}^{k+1, t-1} \right) + \sum_{t=1}^{T_o-1} \left(\mathbf{G}^{k+1, t} - \mathbf{G}^{k+1, t-1} \right) \\ &\quad \vdots \\ &= (T_o - 1) \mathbf{Y}^k + \sum_{l=1}^{T_o-1} \sum_{t=1}^l \left(\mathbf{G}^{k+1, t} - \mathbf{G}^{k+1, t-1} \right). \end{aligned} \quad (9)$$

Substituting the above equation (9) into (8) and combining with (4), the updates of PISCO can be rewritten as the following compact form:

$$\begin{aligned}\mathbf{X}^{k+1} &= \left(\mathbf{X}^k - \eta(T_o + 1)\mathbf{Y}^k - \eta \sum_{l=1}^{T_o} \sum_{t=1}^l (\mathbf{G}^{k+1,t} - \mathbf{G}^{k+1,t-1}) \right) \mathbf{W}^k \\ &= \left(\mathbf{X}^k - \eta(T_o + 1)\mathbf{Y}^k - \eta \sum_{t=1}^{T_o} \mathbf{G}^{k+1,t} + \eta T_o \mathbf{G}^k \right) \mathbf{W}^k,\end{aligned}\tag{10a}$$

$$\begin{aligned}\mathbf{Y}^{k+1} &= \left(\mathbf{Y}^k + \mathbf{G}^{k+1} - \mathbf{G}^{k+1,T_o} + \sum_{t=1}^{T_o} (\mathbf{G}^{k+1,t} - \mathbf{G}^{k+1,t-1}) \right) \mathbf{W}^k \\ &= \left(\mathbf{Y}^k + \mathbf{G}^{k+1} - \mathbf{G}^k \right) \mathbf{W}^k,\end{aligned}\tag{10b}$$

where $\eta = \eta_c \eta_l$.

A.1 Useful inequalities

Before proceeding, we first introduce several propositions that will be used in the following analysis.

Proposition 1. *Let $\{\mathbf{v}_i\}_{i=1}^M$ be a set of M vectors in \mathbb{R}^d . Then, for any $0 < \delta \leq 1$,*

$$\left\| \sum_{i=1}^M \mathbf{v}_i \right\|_2^2 \leq M \sum_{i=1}^M \|\mathbf{v}_i\|_2^2,\tag{11}$$

$$\|\mathbf{v}_i + \mathbf{v}_j\|_2^2 \leq (1 + \delta)\|\mathbf{v}_i\|_2^2 + \left(1 + \frac{1}{\delta}\right)\|\mathbf{v}_j\|_2^2.\tag{12}$$

Proposition 2. *For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times n}$, $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$.*

Proposition 3 (Lemma 4 in [KKM⁺20]). *Let $\{\Xi_i\}_{i=1}^n$ be n random variables in \mathbb{R}^d , which may not be independent of each other. Suppose that $\mathbb{E}[\Xi_i | \Xi_{i-1}, \dots, \Xi_1] = \xi_i$ and $\mathbb{E}[\|\Xi_i - \xi_i\|_2^2] \leq \sigma^2$. Then, we have*

$$\mathbb{E} \left[\left\| \sum_{i=1}^n (\Xi_i - \xi_i) \right\|_2^2 \right] \leq n\sigma^2.$$

A.2 Property of gradient tracking

In this section, we first show that $\bar{\mathbf{Y}}^k = \mathbf{Y}^k \mathbf{J}$ is able to track the average of local stochastic gradients, i.e., $\bar{\mathbf{G}}^k = \mathbf{G}^k \mathbf{J} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{g}_i^k\right) \mathbf{1}_n^\top$. Define the average model estimate as $\bar{\mathbf{x}}^k = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^k \in \mathbb{R}^d$ and $\bar{\mathbf{X}}^k = \mathbf{X}^k \mathbf{J}$.

Lemma 1 (Gradient tracking property). *Suppose Assumption 1 holds. Then for any $k \in \mathbb{N}$,*

$$\bar{\mathbf{Y}}^k = \bar{\mathbf{G}}^k,\tag{13}$$

In addition, if Assumption 2 and 3 hold, we have

$$\mathbb{E}[\|\bar{\mathbf{Y}}^k\|_F^2] \leq \frac{3\sigma^2}{b} + 3L^2 \mathbb{E}[\|\Phi_x^k\|_F^2] + 3n \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2].\tag{14}$$

where $\Phi_x^k = \mathbf{X}^k - \bar{\mathbf{X}}^k$ denotes the consensus error at iteration k .

Proof. From the equation (10b),

$$\mathbf{Y}^{k+1} \mathbf{J} = \left(\mathbf{Y}^k + \mathbf{G}^{k+1} - \mathbf{G}^k \right) \mathbf{W}^k \mathbf{J}.$$

Because of Assumption 1, i.e., $\mathbf{W}^k \mathbf{1}_n = \mathbf{1}_n$, and the fact that $\bar{\mathbf{Y}}^{k+1} = \mathbf{Y}^{k+1} \mathbf{J}$,

$$\bar{\mathbf{Y}}^{k+1} = \bar{\mathbf{Y}}^k + \bar{\mathbf{G}}^{k+1} - \bar{\mathbf{G}}^k.$$

Summing from $k = 0$ to $t - 1$ gives

$$\bar{\mathbf{Y}}^t = \bar{\mathbf{Y}}^0 + \bar{\mathbf{G}}^t - \bar{\mathbf{G}}^0 = \bar{\mathbf{G}}^t,$$

due to the initialization of Algorithm 1 s.t. $\mathbf{Y}^0 = \mathbf{G}^0$. Moreover,

$$\begin{aligned} \mathbb{E} \left[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2 \right] &= \mathbb{E} \left[\|\bar{\mathbf{G}}^k\|_{\mathbb{F}}^2 \right] = \mathbb{E} \left[\|\mathbf{G}^k \mathbf{J}\|_{\mathbb{F}}^2 \right] \\ &= \mathbb{E} \left[\left\| \left(\bar{\mathbf{G}}^k - \nabla F(\mathbf{X}^k) + \nabla F(\mathbf{X}^k) - \nabla F(\bar{\mathbf{X}}^k) + \nabla F(\bar{\mathbf{X}}^k) \right) \mathbf{J} \right\|_{\mathbb{F}}^2 \right] \\ &\stackrel{(a)}{=} n \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n (\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k) + \nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\bar{\mathbf{x}}^k) + \nabla f_i(\bar{\mathbf{x}}^k)) \right\|_2^2 \right] \\ &\stackrel{(b)}{=} \frac{3}{n} \mathbb{E} \left[\left\| \sum_{i=1}^n (\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k)) \right\|_2^2 \right] + 3 \sum_{i=1}^n \mathbb{E} \left[\left\| \nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\bar{\mathbf{x}}^k) \right\|_2^2 \right] + 3n \mathbb{E} \left[\left\| \nabla f(\bar{\mathbf{x}}^k) \right\|_2^2 \right] \\ &\stackrel{(c)}{\leq} 3\sigma^2/b + 3L^2 \sum_{i=1}^n \mathbb{E} \left[\left\| \mathbf{x}_i^k - \bar{\mathbf{x}}^k \right\|_2^2 \right] + 3n \mathbb{E} \left[\left\| \nabla f(\bar{\mathbf{x}}^k) \right\|_2^2 \right] \\ &= 3\sigma^2/b + 3L^2 \mathbb{E} \left[\left\| \Phi_x^k \right\|_{\mathbb{F}}^2 \right] + 3n \mathbb{E} \left[\left\| \nabla f(\bar{\mathbf{x}}^k) \right\|_2^2 \right], \end{aligned} \quad (15)$$

where (a) is based on $\|\mathbf{x} \mathbf{1}_n^\top\|_{\mathbb{F}}^2 = n \|\mathbf{x}\|_2^2$, (b) uses (11) in Proposition 1, (c) follows Assumption 2 and 3 and applies Proposition 3. \square

A.3 Progress improvement on successive iterates and averages

Before continuing, we first show two auxiliary lemmas to bound the progress improvement between the successive iterates and their averages. Similar to $\Phi_x^k = \mathbf{X}^k - \bar{\mathbf{X}}^k$, we also use $\Phi_y^k = \mathbf{Y}^k - \bar{\mathbf{Y}}^k$ to represent the tracking error at the k -th iteration. As for the local updates, we define t -th local-update consensus error as $\Phi_x^{k,t} = \mathbf{X}^{k+1,t} - \bar{\mathbf{X}}^k$ and tracking error as $\Phi_y^{k,t} = \mathbf{Y}^{k+1,t} - \bar{\mathbf{Y}}^k$.

Lemma 2 (Progress improvement between successive iterates). *Suppose Assumption 1, 2 and 3 hold. Then, we have*

$$\begin{aligned} \mathbb{E} \left[\left\| \mathbf{X}^k - \mathbf{X}^{k-1} \right\|_{\mathbb{F}}^2 \right] &\leq 12 (1 + 2T_o^2 L^2 \eta^2) \mathbb{E} \left[\left\| \Phi_x^{k-1} \right\|_{\mathbb{F}}^2 \right] + 6(1-p) \lambda^2 (T_o + 1)^2 \eta^2 \mathbb{E} \left[\left\| \Phi_y^{k-1} \right\|_{\mathbb{F}}^2 \right] \\ &\quad + \frac{48nT_o^2 \eta^2 \sigma^2}{b} + 24T_o L^2 \eta^2 \sum_{t=1}^{T_o} \mathbb{E} \left[\left\| \Phi_x^{k-1,t} \right\|_{\mathbb{F}}^2 \right] + 3(T_o + 1)^2 \eta^2 \mathbb{E} \left[\left\| \bar{\mathbf{Y}}^{k-1} \right\|_{\mathbb{F}}^2 \right]. \end{aligned}$$

Proof. Define $\Delta_x^{k-1} = \mathbf{X}^k - \mathbf{X}^{k-1} \mathbf{W}^{k-1}$. Then,

$$\begin{aligned} \mathbb{E} \left[\left\| \mathbf{X}^k - \mathbf{X}^{k-1} \right\|_{\mathbb{F}}^2 \right] &= \mathbb{E} \left[\left\| \mathbf{X}^{k-1} (\mathbf{W}^{k-1} - \mathbf{I}) + \Delta_x^{k-1} \right\|_{\mathbb{F}}^2 \right] \\ &= \mathbb{E} \left[\left\| \left(\mathbf{X}^{k-1} - \bar{\mathbf{X}}^{k-1} \right) (\mathbf{W}^{k-1} - \mathbf{I}) + \Delta_x^{k-1} + \eta(T_o + 1) \bar{\mathbf{G}}^{k-1} - \eta(T_o + 1) \bar{\mathbf{G}}^{k-1} \right\|_{\mathbb{F}}^2 \right] \\ &\stackrel{(a)}{\leq} 3 \mathbb{E} \left[\left\| \Phi_x^{k-1} \right\|_{\mathbb{F}}^2 \left\| \mathbf{W}^{k-1} - \mathbf{I} \right\|_2^2 + \left\| \Delta_x^{k-1} + \eta(T_o + 1) \bar{\mathbf{G}}^{k-1} \right\|_{\mathbb{F}}^2 + \eta^2 (T_o + 1)^2 \left\| \bar{\mathbf{G}}^{k-1} \right\|_{\mathbb{F}}^2 \right] \\ &\stackrel{(b)}{\leq} 12 \mathbb{E} \left[\left\| \Phi_x^{k-1} \right\|_{\mathbb{F}}^2 \right] + 3 \mathbb{E} \left[\left\| \Delta_x^{k-1} + \eta(T_o + 1) \bar{\mathbf{G}}^{k-1} \right\|_{\mathbb{F}}^2 \right] + 3\eta^2 (T_o + 1)^2 \mathbb{E} \left[\left\| \bar{\mathbf{G}}^{k-1} \right\|_{\mathbb{F}}^2 \right], \end{aligned} \quad (16)$$

where (a) is from Proposition 1 and (b) is due to $\|\mathbf{W}^{k-1} - \mathbf{I}\|_2 \leq 2$, for any $k \geq 1$.

Next, we are going to bound the second term $\mathbb{E}[\|\Delta_x^{k-1} + \eta(T_o + 1)\overline{\mathbf{G}}^{k-1}\|_{\mathbb{F}}^2]$. Because of (10a), we have

$$\Delta_x^{k-1} = -\eta \left((T_o + 1)\mathbf{Y}^{k-1} + \sum_{t=1}^{T_o} \mathbf{G}^{k,t} - T_o \mathbf{G}^{k-1} \right) \mathbf{W}^{k-1}.$$

Then,

$$\begin{aligned} & \mathbb{E}[\|\Delta_x^{k-1} + \eta(T_o + 1)\overline{\mathbf{G}}^{k-1}\|_{\mathbb{F}}^2] \\ &= \eta^2 \mathbb{E} \left[\left\| (T_o + 1) \left(\mathbf{Y}^{k-1} - \overline{\mathbf{Y}}^{k-1} \right) \left(\mathbf{W}^{k-1} - \mathbf{J} \right) + \sum_{t=1}^{T_o} \left(\mathbf{G}^{k,t} - \mathbf{G}^{k-1} \right) \mathbf{W}^{k-1} \right\|_{\mathbb{F}}^2 \right] \\ &\leq 2\eta^2 \left((T_o + 1)^2 \mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] \|\mathbf{W}^{k-1} - \mathbf{J}\|_2^2 + \mathbb{E} \left[\left\| \sum_{t=1}^{T_o} \left(\mathbf{G}^{k,t} - \mathbf{G}^{k-1} \right) \right\|_{\mathbb{F}}^2 \|\mathbf{W}^{k-1}\|_2^2 \right] \right) \\ &\leq 2\eta^2 \left((1-p)\lambda^2 (T_o + 1)^2 \mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + T_o \sum_{t=1}^{T_o} \mathbb{E} \left[\|\mathbf{G}^{k,t} - \mathbf{G}^{k-1}\|_{\mathbb{F}}^2 \right] \right), \end{aligned} \quad (17)$$

where the first equality is based on Lemma 1, the penultimate inequality follows the Proposition 1 and 2 and the last one is from Proposition 1 and Assumption 1, i.e., $\mathbb{E}[\|\mathbf{W}^{k-1} - \mathbf{J}\|_2^2] = (1-p)\lambda^2$. Here,

$$\begin{aligned} & \mathbb{E} \left[\|\mathbf{G}^{k,t} - \mathbf{G}^{k-1}\|_{\mathbb{F}}^2 \right] \\ &= \mathbb{E} \left[\left\| \mathbf{G}^{k,t} - \nabla F(\mathbf{X}^{k,t}) + \nabla F(\mathbf{X}^{k,t}) - \nabla F(\overline{\mathbf{X}}^{k-1}) - \mathbf{G}^{k-1} + \nabla F(\mathbf{X}^{k-1}) - \nabla F(\mathbf{X}^{k-1}) + \nabla F(\overline{\mathbf{X}}^{k-1}) \right\|_{\mathbb{F}}^2 \right] \\ &\stackrel{(a)}{\leq} 4 \left(\mathbb{E}[\|\mathbf{G}^{k,t} - \nabla F(\mathbf{X}^{k,t})\|_{\mathbb{F}}^2] + \mathbb{E}[\|\mathbf{G}^{k-1} - \nabla F(\mathbf{X}^{k-1})\|_{\mathbb{F}}^2] + L^2 \mathbb{E}[\|\Phi_x^{k-1,t}\|_{\mathbb{F}}^2] + L^2 \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] \right) \\ &\stackrel{(b)}{\leq} \frac{8n\sigma^2}{b} + 4L^2 \mathbb{E}[\|\Phi_x^{k-1,t}\|_{\mathbb{F}}^2] + 4L^2 \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2], \end{aligned} \quad (18)$$

where (a) is based on Proposition 1 and Assumption 2 and (b) is from Assumption 3. Then,

$$\begin{aligned} \mathbb{E}[\|\Delta_x^{k-1} + \eta(T_o + 1)\overline{\mathbf{G}}^{k-1}\|_{\mathbb{F}}^2] &\leq 2(1-p)\lambda^2 (T_o + 1)^2 \eta^2 \mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + \frac{16nT_o^2 \eta^2 \sigma^2}{b} \\ &\quad + 8T_o L^2 \eta^2 \sum_{t=1}^{T_o} \mathbb{E}[\|\Phi_x^{k-1,t}\|_{\mathbb{F}}^2] + 8T_o^2 L^2 \eta^2 \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2]. \end{aligned} \quad (19)$$

Finally, substituting (19) into (16) and applying Lemma 1 result in

$$\begin{aligned} \mathbb{E}[\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_{\mathbb{F}}^2] &\leq 12 (1 + 2T_o^2 L^2 \eta^2) \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] + 6(1-p)\lambda^2 (T_o + 1)^2 \eta^2 \mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] \\ &\quad + \frac{48nT_o^2 \eta^2 \sigma^2}{b} + 24T_o L^2 \eta^2 \sum_{t=1}^{T_o} \mathbb{E}[\|\Phi_x^{k-1,t}\|_{\mathbb{F}}^2] + 3(T_o + 1)^2 \eta^2 \mathbb{E}[\|\overline{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2]. \end{aligned}$$

□

Lemma 3 (Progress improvement between the averages). *Suppose Assumption 1, 2 and 3 hold. Then, we have*

$$\mathbb{E} \left[\|\overline{\mathbf{X}}^{k+1} - \overline{\mathbf{X}}^k\|_{\mathbb{F}}^2 \right] \leq \frac{3T_o \eta^2 \sigma^2}{b} + 3T_o L^2 \eta^2 \sum_{t=0}^{T_o} \mathbb{E} \left[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2 \right] + 3nT_o^2 \eta^2 \mathbb{E} \left[\|\nabla f(\overline{\mathbf{x}}^k)\|_2^2 \right]. \quad (20)$$

Proof. From Lemma 1 and the (10a), we have

$$\overline{\mathbf{X}}^{k+1} = \overline{\mathbf{X}}^k - \eta \sum_{t=0}^{T_o} \overline{\mathbf{G}}^{k+1,t}. \quad (21)$$

Similar to (15), we have

$$\begin{aligned}
& \frac{1}{\eta^2} \mathbb{E} \left[\|\bar{\mathbf{X}}^{k+1} - \bar{\mathbf{X}}^k\|_{\mathbb{F}}^2 \right] \\
&= \mathbb{E} \left[\left\| \sum_{t=0}^{T_o} \bar{\mathbf{G}}^{k+1,t} \right\|_{\mathbb{F}}^2 \right] \\
&= n \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \sum_{t=0}^{T_o} \left(\mathbf{g}_i^{k+1,t} - \nabla f_i(\mathbf{x}_i^{k+1,t}) + \nabla f_i(\mathbf{x}_i^{k+1,t}) - \nabla f_i(\bar{\mathbf{x}}^k) + \nabla f_i(\bar{\mathbf{x}}^k) \right) \right\|_2^2 \right] \\
&\leq \frac{3}{n} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{t=0}^{T_o} \left(\mathbf{g}_i^{k+1,t} - \nabla f_i(\mathbf{x}_i^{k+1,t}) \right) \right\|_2^2 \right] + 3T_o \sum_{i=1}^n \sum_{t=0}^{T_o} \mathbb{E} \left[\|\nabla f_i(\mathbf{x}_i^{k+1,t}) - \nabla f_i(\bar{\mathbf{x}}^k)\|_2^2 \right] + 3nT_o^2 \mathbb{E} \left[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2 \right] \\
&\leq \frac{3T_o\sigma^2}{b} + 3T_oL^2 \sum_{i=1}^n \sum_{t=0}^{T_o} \mathbb{E} \left[\|\mathbf{x}_i^{k+1,t} - \bar{\mathbf{x}}^k\|_2^2 \right] + 3nT_o^2 \mathbb{E} \left[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2 \right] \\
&\leq \frac{3T_o\sigma^2}{b} + 3T_oL^2 \sum_{t=0}^{T_o} \mathbb{E} \left[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2 \right] + 3nT_o^2 \mathbb{E} \left[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2 \right].
\end{aligned}$$

□

A.4 Bounding consensus errors and tracking errors

In this subsection, we present three key lemmas for controlling consensus and tracking errors, which will be used to establish the convergence results of PISCO.

We begin with bounding the accumulated consensus errors for local updates in the following lemma.

Lemma 4 (Accumulated consensus drift for local updates). *Suppose Assumption 2 and 3 hold. If $\eta_l \leq \frac{1}{8L(T_o+1)}$, we have*

$$\sum_{t=1}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2] \leq 9T_o \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + 8\eta_l^2(T_o+1)^3 \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + \frac{64n\eta_l^2T_o^3\sigma^2}{b} + 3\eta_l^2T_o(T_o+1)^2 \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2].$$

Proof. For the t -th local update at the k -th iteration, it follows (3a) that

$$\begin{aligned}
\Phi_x^{k,t} &= \mathbf{X}^{k+1,t} - \bar{\mathbf{X}}^k = \mathbf{X}^k - \eta_l \mathbf{Y}^k - \eta_l \sum_{l=1}^{t-1} \mathbf{Y}^{k+1,l} - \bar{\mathbf{X}}^k + t\eta_l \bar{\mathbf{Y}}^k - t\eta_l \bar{\mathbf{Y}}^k \\
&= \mathbf{X}^k - \bar{\mathbf{X}}^k - \eta_l (\mathbf{Y}^k - \bar{\mathbf{Y}}^k) - \eta_l \sum_{l=1}^{t-1} (\mathbf{Y}^{k+1,l} - \bar{\mathbf{Y}}^k) - t\eta_l \bar{\mathbf{Y}}^k \\
&= \Phi_x^k - \eta_l \Phi_y^k - \eta_l \sum_{l=1}^{t-1} \Phi_y^{k,l} - t\eta_l \bar{\mathbf{Y}}^k.
\end{aligned} \tag{22}$$

To bound the accumulated consensus errors for T_o local updates, we sum the the expectation of the squared norm of (22) over t from $t = 1$ to T_o , leading to

$$\begin{aligned}
\sum_{t=1}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2] &\stackrel{(11)}{\leq} 4T_o \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + 4\eta_l^2 T_o \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + 4\eta_l^2 \sum_{t=2}^{T_o} (t-1) \sum_{l=1}^{t-1} \mathbb{E}[\|\Phi_y^{k,l}\|_{\mathbb{F}}^2] + \sum_{t=1}^{T_o} 4\eta_l^2 t^2 \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] \\
&\leq 4T_o \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + 4\eta_l^2 T_o \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + 4\eta_l^2 \sum_{t=1}^{T_o-1} \frac{(T_o+t-1)(T_o-t)}{2} \mathbb{E}[\|\Phi_y^{k,t}\|_{\mathbb{F}}^2] \\
&\quad + \frac{2\eta_l^2 T_o(T_o+1)(2T_o+1)}{3} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2],
\end{aligned} \tag{23}$$

where the last inequality uses $\sum_{t=1}^{T_o} t^2 = \frac{T_o(T_o+1)(2T_o+1)}{6}$.

From (3c), we have

$$\mathbf{Y}^{k+1,t} = \mathbf{Y}^k + \mathbf{G}^{k+1,t} - \mathbf{G}^k,$$

such that

$$\begin{aligned} \mathbb{E}[\|\Phi_y^{k,t}\|_{\mathbb{F}}^2] &= \mathbb{E}[\|\mathbf{Y}^k + \mathbf{G}^{k+1,t} - \mathbf{G}^k - \bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] \leq 2\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + 2\mathbb{E}[\|\mathbf{G}^{k+1,t} - \mathbf{G}^k\|_{\mathbb{F}}^2] \\ &\stackrel{(18)}{\leq} 2\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + 16n\sigma^2/b + 8L^2\mathbb{E}[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2] + 8L^2\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2]. \end{aligned} \quad (24)$$

Due to the fact $\frac{(T_o+t-1)(T_o-t)}{2} \leq \frac{T_o^2}{2}$ for $1 \leq t \leq T_o$,

$$\sum_{t=1}^{T_o-1} \frac{(T_o+t-1)(T_o-t)}{2} \mathbb{E}[\|\Phi_y^{k,t}\|_{\mathbb{F}}^2] \leq T_o^3 \left(\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + 8n\sigma^2/b + 4L^2\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] \right) + 4L^2 \sum_{t=1}^{T_o} T_o^2 \mathbb{E}[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2].$$

Combining with (23), we have

$$\begin{aligned} \sum_{t=1}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2] &\leq (4T_o + 16\eta_l^2 T_o^3 L^2) \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + (4\eta_l^2 T_o + 4\eta_l^2 T_o^3) \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + 32n\eta_l^2 T_o^3 \sigma^2/b \\ &\quad + 16\eta_l^2 L^2 \sum_{t=1}^{T_o} T_o^2 \mathbb{E}[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2] + \frac{2\eta_l^2 T_o (T_o+1)(2T_o+1)}{3} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2]. \end{aligned}$$

Therefore, if $\eta_l \leq \frac{1}{8L(T_o+1)}$, we have

$$\begin{aligned} \sum_{t=1}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_{\mathbb{F}}^2] \\ \leq 8T_o(1 + 4\eta_l^2 T_o^2 L^2) \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + 8\eta_l^2 T_o (T_o+1) \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] + 64n\eta_l^2 T_o^3 \sigma^2/b + 3\eta_l^2 T_o (T_o+1)^2 \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2]. \end{aligned}$$

□

With Lemma 4 in hand, we are ready to bound the consensus error $\mathbb{E}[\|\Phi_x\|_{\mathbb{F}}^2]$ and the tracking error $\mathbb{E}[\|\Phi_y\|_{\mathbb{F}}^2]$, respectively.

Lemma 5 (Consensus error for communication updates). *Suppose Assumption 1,2 and 3 hold. If $\eta_l \leq \frac{1}{8L(T_o+1)}$ and $\eta \leq \frac{\lambda_p}{80L(T_o+1)}$, we have*

$$\begin{aligned} \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] &< (1-p) \left[\frac{1 + (1+p)\lambda^2}{2} \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] + \frac{40\lambda^2}{\lambda_p} (T_o+1)^2 \eta^2 \mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] \right. \\ &\quad \left. + \frac{240\lambda^2 L^2 (T_o+1)^4 \eta^2 \eta_l^2}{\lambda_p} \mathbb{E}[\|\bar{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2] + \frac{320\lambda^2 n (T_o+1)^2 \eta^2 \sigma^2}{\lambda_p b} \right], \end{aligned}$$

Proof. Let $\Delta_x^{k-1} = \mathbf{X}^k - \mathbf{X}^{k-1} \mathbf{W}^{k-1}$ and $\bar{\Delta}_x^{k-1} = \bar{\mathbf{X}}^k - \bar{\mathbf{X}}^{k-1}$. Then,

$$\begin{aligned} \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] &\leq \mathbb{E}[\|(\mathbf{X}^{k-1} - \bar{\mathbf{X}}^{k-1})(\mathbf{W}^{k-1} - \mathbf{J}) + \Delta_x^{k-1} - \bar{\Delta}_x^{k-1}\|_{\mathbb{F}}^2] \\ &\stackrel{(a)}{\leq} (1 + \delta_1) \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2 \|\mathbf{W}^{k-1} - \mathbf{J}\|_2^2] + (1 + \frac{1}{\delta_1}) \mathbb{E}[\|\Delta_x^{k-1} - \bar{\Delta}_x^{k-1}\|_{\mathbb{F}}^2] \\ &\stackrel{(b)}{\leq} (1-p)(1 + \delta_1) \lambda^2 \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] + (1 + \frac{1}{\delta_1}) \mathbb{E}[\|\Delta_x^{k-1} - \bar{\Delta}_x^{k-1}\|_{\mathbb{F}}^2], \end{aligned} \quad (25)$$

where (a) is based on (12) and Proposition 2, and (b) is from the Assumption 1, i.e., $\mathbb{E}[\|\mathbf{W}^{k-1} - \mathbf{J}\|_2^2] = 1 - \lambda_p$. Further from (10a), we have

$$\mathbb{E}[\|\Delta_x^{k-1} - \bar{\Delta}_x^{k-1}\|_{\mathbb{F}}^2] = \eta^2 \mathbb{E} \left[\left\| \left((T_o+1) \Phi_y^{k-1} - \sum_{t=1}^{T_o} (\mathbf{G}^{k,t} - \mathbf{G}^{k-1}) \right) (\mathbf{W}^{k-1} - \mathbf{J}) \right\|_{\mathbb{F}}^2 \right]$$

$$\begin{aligned}
&\leq (1-p)\lambda^2\eta^2\mathbb{E}[\|(T_o+1)\Phi_y^{k-1} - \sum_{t=1}^{T_o}(\mathbf{G}^{k,t} - \mathbf{G}^{k-1})\|_{\mathbb{F}}^2] \\
&\leq 2(1-p)\lambda^2\eta^2\left(\mathbb{E}[\|(T_o+1)\Phi_y^{k-1}\|_{\mathbb{F}}^2] + T_o\sum_{t=1}^{T_o}\mathbb{E}[\|(\mathbf{G}^{k,t} - \mathbf{G}^{k-1})\|_{\mathbb{F}}^2]\right). \quad (26)
\end{aligned}$$

Substituting (26) and (18) into (25),

$$\begin{aligned}
&\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] \\
&\leq (1-p)\lambda^2\left[\left(1+\delta_1\right)\mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] + 2\left(1+\frac{1}{\delta_1}\right)\eta^2\left(\left(T_o+1\right)^2\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + T_o\sum_{t=1}^{T_o}\|(\mathbf{G}^{k,t} - \mathbf{G}^{k-1})\|_{\mathbb{F}}^2\right)\right] \\
&\leq (1-p)\lambda^2\left[\left(1+\delta_1+8\left(1+\frac{1}{\delta_1}\right)T_o^2L^2\eta^2\right)\mathbb{E}[\|\Phi_x\|_{\mathbb{F}}^2] + 2\left(1+\frac{1}{\delta_1}\right)(T_o+1)^2\eta^2\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2]\right. \\
&\quad \left.+ 8\left(1+\frac{1}{\delta_1}\right)T_oL^2\eta^2\sum_{t=1}^{T_o}\mathbb{E}[\|\Phi_x^{k-1,t}\|_{\mathbb{F}}^2] + 16\left(1+\frac{1}{\delta_1}\right)nT_o^2\eta^2\sigma^2/b\right].
\end{aligned}$$

By Lemma 4, the consensus error $\mathbb{E}[\|\Phi_x\|_{\mathbb{F}}^2]$ can be upper bounded by

$$\begin{aligned}
\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] &\leq (1-p)\lambda^2\left[c_0\mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] + 2\left(1+\frac{1}{\delta_1}\right)(T_o+1)^2\eta^2(32(T_o+1)^2L^2\eta_l^2+1)\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2]\right. \\
&\quad \left.+ 24\left(1+\frac{1}{\delta_1}\right)L^2T_o^2(T_o+1)^2\eta^2\eta_l^2\mathbb{E}[\|\bar{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2] + 16\left(1+\frac{1}{\delta_1}\right)nT_o^2\eta^2(32T_o^2L^2\eta_l^2+1)\sigma^2/b\right],
\end{aligned}$$

where $c_0 = 1 + \delta_1 + 80\left(1 + \frac{1}{\delta_1}\right)T_o^2L^2\eta^2$. Let $\delta_1 = \frac{\lambda_p}{8} \in (0, 1]$. If $\eta \leq \frac{\lambda_p}{80(T_o+1)L}$, we have

$$c_0 \leq 1 + \frac{\lambda_p}{4} \quad \text{and} \quad 1 + \frac{1}{\delta_1} < \frac{10}{\lambda_p}.$$

Note that $\lambda^2\left(1 + \frac{1-(1-p)\lambda^2}{4}\right) < \frac{1+(1+p)\lambda^2}{2}$. Then,

$$\begin{aligned}
\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] &< (1-p)\left[\frac{1+(1+p)\lambda^2}{2}\mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] + \frac{40\lambda^2}{\lambda_p}(T_o+1)^2\eta^2\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2]\right. \\
&\quad \left.+ \frac{240\lambda^2L^2(T_o+1)^4\eta^2\eta_l^2}{\lambda_p}\mathbb{E}[\|\bar{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2] + \frac{320\lambda^2n(T_o+1)^2\eta^2\sigma^2}{\lambda_p b}\right],
\end{aligned}$$

since $\eta_l \leq \frac{1}{8L(T_o+1)}$. □

Lemma 6 (Tracking error for communication updates). *Suppose Assumption 1, 2 and 3 hold. If $\eta_l \leq \frac{1}{8L(T_o+1)}$ and $\eta \leq \frac{\lambda_p}{80L(T_o+1)}$,*

$$\begin{aligned}
\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] &\leq (1-p)\left[\frac{1+(1+p)\lambda^2}{2}\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + \frac{400\lambda^2}{\lambda_p}L^2\mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2]\right] \\
&\quad + \frac{(1-p)\lambda^2}{\lambda_p}\left[125L^2\eta^2(T_o+1)^2\mathbb{E}[\|\bar{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2] + 180n\sigma^2/b\right].
\end{aligned}$$

Proof. Let $\Delta_y^{k-1} = \mathbf{Y}^k - \mathbf{Y}^{k-1}\mathbf{W}^{k-1}$ and $\bar{\Delta}_y^{k-1} = \bar{\mathbf{Y}}^k - \bar{\mathbf{Y}}^{k-1}$. From (10b) and Lemma 1, we have

$$\begin{aligned}
\mathbb{E}[\|\Delta_y^{k-1} - \bar{\Delta}_y^{k-1}\|_{\mathbb{F}}^2] &= \mathbb{E}[\|(\mathbf{G}^k - \mathbf{G}^{k-1})(\mathbf{W}^{k-1} - \mathbf{J})\|_{\mathbb{F}}^2] \\
&\leq (1-p)\lambda^2\mathbb{E}[\|\mathbf{G}^k - \mathbf{G}^{k-1}\|_{\mathbb{F}}^2]
\end{aligned}$$

$$\begin{aligned}
&= (1-p)\lambda^2 \mathbb{E}[\|\mathbf{G}^k - \nabla F(\mathbf{X}^k) + \nabla F(\mathbf{X}^k) - \nabla F(\mathbf{X}^{k-1}) - (\mathbf{G}^{k-1} - \nabla F(\mathbf{X}^{k-1}))\|_{\mathbb{F}}^2] \\
&\leq (1-p)\lambda^2 \left(\frac{6n\sigma^2}{b} + 3\mathbb{E}[\|\nabla F(\mathbf{X}^k) - \nabla F(\mathbf{X}^{k-1})\|_{\mathbb{F}}^2] \right) \\
&\leq (1-p)\lambda^2 \left(\frac{6n\sigma^2}{b} + 3L^2\mathbb{E}[\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_{\mathbb{F}}^2] \right).
\end{aligned}$$

where the first inequality is based on Assumption 1, the penultimate inequality is due to Assumption 3 and Proposition 1, and the last one is from Assumption 2. Together with Lemma 1, we have

$$\begin{aligned}
\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] &= \mathbb{E}[\|(\mathbf{Y}^{k-1} - \bar{\mathbf{Y}}^{k-1})(\mathbf{W}^{k-1} - \mathbf{J}) + \Delta_y^{k-1} - \bar{\Delta}_y^{k-1}\|_{\mathbb{F}}^2] \\
&\leq (1-p)\lambda^2(1+\delta_1)\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + (1+\frac{1}{\delta_1})\mathbb{E}[\|\Delta_y^{k-1} - \bar{\Delta}_y^{k-1}\|_{\mathbb{F}}^2] \\
&\leq (1-p)\lambda^2 \left[(1+\delta_1)\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + (1+\frac{1}{\delta_1}) \left(\frac{6n\sigma^2}{b} + 3L^2\mathbb{E}[\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_{\mathbb{F}}^2] \right) \right]. \quad (27)
\end{aligned}$$

Moreover, from Lemma 2 and Lemma 4, if $\eta_l \leq \frac{1}{8L(T_o+1)}$,

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_{\mathbb{F}}^2] \\
&\leq 6\eta^2(T_o+1)^2 \left((1-p)\lambda^2 + 32T_o(T_o+1)L^2\eta_l^2 \right) \mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + 12(1+20T_o^2L^2\eta^2) \mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] \\
&\quad + 48nT_o^2\eta^2(1+32T_o^2L^2\eta_l^2)\sigma^2/b + 3\eta^2(T_o+1)^2(1+24T_o^2L^2\eta_l^2) \mathbb{E}[\|\bar{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2]
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] &\leq (1-p)\lambda^2 \left[c_1\mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + 36(1+\frac{1}{\delta_1})(1+20T_o^2L^2\eta^2)L^2\mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] + 6n(1+\frac{1}{\delta_1}) \right. \\
&\quad \left. (1+48T_o^2L^2\eta^2(1+32T_o^2L^2\eta_l^2))\frac{\sigma^2}{b} + 9(1+\frac{1}{\delta_1})L^2(T_o+1)^2(1+24T_o^2L^2\eta_l^2)\eta^2\mathbb{E}[\|\bar{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2] \right]
\end{aligned}$$

where $c_1 = 1 + \delta_1 + 36(1 + \frac{1}{\delta_1})(T_o + 1)^2L^2\eta^2$. If $\eta \leq \frac{\lambda_p}{80L(T_o+1)}$ and $\delta_1 = \frac{\lambda_p}{8} \in (0, 1)$,

$$c_1 \leq 1 + \frac{\lambda_p}{4} \quad \text{and} \quad 1 + \frac{1}{\delta_1} \leq \frac{10}{\lambda_p}.$$

Note that $\lambda^2(1 + \frac{1-(1-p)\lambda^2}{4}) \leq \frac{1+(1+p)\lambda^2}{2}$. Thus,

$$\begin{aligned}
\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] &\leq (1-p) \left[\frac{1+(1+p)\lambda^2}{2} \mathbb{E}[\|\Phi_y^{k-1}\|_{\mathbb{F}}^2] + \frac{360\lambda^2}{\lambda_p} (1+20T_o^2L^2\eta^2) L^2\mathbb{E}[\|\Phi_x^{k-1}\|_{\mathbb{F}}^2] \right. \\
&\quad \left. + \frac{90\lambda^2}{\lambda_p} L^2\eta^2(T_o+1)^2(1+24T_o^2L^2\eta_l^2) \mathbb{E}[\|\bar{\mathbf{Y}}^{k-1}\|_{\mathbb{F}}^2] + \frac{180n\lambda^2\sigma^2}{\lambda_p b} \right].
\end{aligned}$$

Since $\eta_l \leq \frac{1}{8L(T_o+1)}$ and $\eta \leq \frac{\lambda_p}{80L(T_o+1)}$, we arrive at Lemma 6. \square

A.5 Descent lemma

Lemma 7 (Descent lemma). *Suppose Assumption 1, 2 and 3 hold. If $\eta_l \leq \frac{1}{8L(T_o+1)}$ and $\eta \leq \frac{1}{6L(T_o+1)}$, we have*

$$\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}^{k+1})] - \mathbb{E}[f(\bar{\mathbf{x}}^k)] &\leq -\frac{\eta(T_o+1)}{4} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_{\mathbb{F}}^2] + \frac{10(T_o+1)L^2\eta}{n} \left(\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + (T_o+1)^2\eta_l^2\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] \right) \\
&\quad + \left(64L^2T_o^3\eta_l^2 + \frac{3LT_o\eta^2}{2n} \right) \frac{\sigma^2}{b} + \frac{3L^2(T_o+1)^3\eta\eta_l^2}{n} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2]. \quad (28)
\end{aligned}$$

Proof. Because of the L -smooth of the objective f ,

$$\mathbb{E}[f(\bar{\mathbf{x}}^{k+1})] - \mathbb{E}[f(\bar{\mathbf{x}}^k)] \leq \mathbb{E}[\nabla f(\bar{\mathbf{x}}^k)^\top (\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k)] + \frac{L}{2n} \mathbb{E}[\|\bar{\mathbf{X}}^{k+1} - \bar{\mathbf{X}}^k\|_F^2]. \quad (29)$$

We first handle the first term on the right-hand side. Let $\mathbf{g}^{k+1,t} = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i^{k+1,t}$. With (21), we have

$$\begin{aligned} & \mathbb{E}[\nabla f(\bar{\mathbf{x}}^k)^\top (\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k)] \\ &= -\eta \sum_{t=0}^{T_o} \mathbb{E}[\nabla f(\bar{\mathbf{x}}^k)^\top \bar{\mathbf{g}}^{k+1,t}] \\ &= -\eta \sum_{t=0}^{T_o} \mathbb{E}[\nabla f(\bar{\mathbf{x}}^k)^\top (\bar{\mathbf{g}}^{k+1,t} + \nabla f(\bar{\mathbf{x}}^k) - \nabla f(\bar{\mathbf{x}}^k))] \\ &\stackrel{(a)}{=} -\eta(T_o + 1) \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] - \eta \sum_{t=0}^{T_o} \mathbb{E} \left[\nabla f(\bar{\mathbf{x}}^k)^\top \left(\frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{k+1,t}) - \nabla f(\bar{\mathbf{x}}^k) \right) \right] \\ &\stackrel{(b)}{\leq} -\eta(T_o + 1) \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{\eta}{2} \sum_{t=0}^{T_o} \mathbb{E} \left[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2 + \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{k+1,t}) - \nabla f(\bar{\mathbf{x}}^k) \right\|_2^2 \right] \\ &\leq -\frac{\eta(T_o + 1)}{2} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{\eta}{2} \sum_{t=0}^{T_o} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{k+1,t}) - \nabla f(\bar{\mathbf{x}}^k) \right\|_2^2 \right] \\ &\stackrel{(c)}{\leq} -\frac{\eta(T_o + 1)}{2} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{\eta}{2n} \sum_{t=0}^{T_o} \sum_{i=1}^n \mathbb{E}[\|\nabla f_i(\mathbf{x}_i^{k+1,t}) - \nabla f_i(\bar{\mathbf{x}}^k)\|_2^2] \\ &\stackrel{(d)}{\leq} -\frac{\eta(T_o + 1)}{2} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{\eta L^2}{2n} \sum_{t=0}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_F^2]. \end{aligned} \quad (30)$$

Here, (a) is from Assumption 3, i.e., $\mathbb{E}[\mathbf{g}_i^{k+1,t}] = \nabla f_i(\mathbf{x}_i^{k+1,t})$; (b) is based on $2\mathbf{a}^\top \mathbf{b} \leq \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2$ for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$; (c) uses the Proposition 1; and (d) follows Assumption 2.

Next, we will bound the second term on the right-hand side of (29). From Lemma 3, we can control the term $\mathbb{E}[\|\bar{\mathbf{X}}^{k+1} - \bar{\mathbf{X}}^k\|_F^2]$ by

$$\mathbb{E}[\|\bar{\mathbf{X}}^{k+1} - \bar{\mathbf{X}}^k\|_F^2] \leq 3\eta^2 \left[\frac{T_o \sigma^2}{b} + T_o L^2 \sum_{t=0}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_F^2] + n T_o^2 \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \right]. \quad (31)$$

Substituting (30) and (31) into (29), we have

$$\begin{aligned} & \mathbb{E}[f(\bar{\mathbf{x}}^{k+1})] - \mathbb{E}[f(\bar{\mathbf{x}}^k)] \\ &\leq \left(-\frac{\eta(T_o + 1)}{2} + \frac{3LT_o^2\eta^2}{2} \right) \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{\eta L^2}{2n} (1 + 3T_o L \eta) \sum_{t=0}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_F^2] + \frac{3LT_o\eta^2\sigma^2}{2nb} \\ &\leq -\frac{\eta(T_o + 1)}{4} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{\eta L^2}{2n} (1 + 3T_o L \eta) \sum_{t=0}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_F^2] + \frac{3LT_o\eta^2\sigma^2}{2nb}. \end{aligned}$$

where the last inequality holds if $\eta \leq \frac{1}{6L(T_o+1)}$. Recall Lemma 4 that if $\eta_l \leq \frac{1}{8L(T_o+1)}$, we have

$$\sum_{t=0}^{T_o} \mathbb{E}[\|\Phi_x^{k,t}\|_F^2] \leq 10T_o \mathbb{E}[\|\Phi_x^k\|_F^2] + 8\eta_l^2 (T_o + 1)^3 \mathbb{E}[\|\Phi_y^k\|_F^2] + \frac{64n\eta_l^2 T_o^3 \sigma^2}{b} + 3\eta_l^2 (T_o + 1)^3 \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_F^2].$$

Therefore,

$$\mathbb{E}[f(\bar{\mathbf{x}}^{k+1})] - \mathbb{E}[f(\bar{\mathbf{x}}^k)]$$

$$\begin{aligned}
&\leq -\frac{\eta(T_o+1)}{4}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{5T_oL^2\eta}{n}(1+3T_oL\eta)\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + \frac{4(T_o+1)^3L^2\eta\eta_l^2}{n}(1+3T_oL\eta)\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] \\
&\quad + \frac{64n(1+3T_oL\eta)L^2T_o^3\eta\eta_l^2 + 3LT_o\eta^2\sigma^2}{2nb} + \frac{3L^2(T_o+1)^3\eta\eta_l^2}{2n}(1+3T_oL\eta)\mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] \\
&\leq -\frac{\eta(T_o+1)}{4}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{10(T_o+1)L^2\eta}{n}\left(\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + (T_o+1)^2\eta_l^2\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2]\right) \\
&\quad + \left(64L^2T_o^3\eta\eta_l^2 + \frac{3LT_o\eta^2}{2n}\right)\frac{\sigma^2}{b} + \frac{3L^2(T_o+1)^3\eta\eta_l^2}{n}\mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2].
\end{aligned}$$

□

A.6 Proof of Theorem 1

From the decent lemma, i.e., Lemma 7, summing (28) from $k = 0$ to $k = K - 1$ gives

$$\begin{aligned}
\frac{\eta(T_o+1)}{4}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] &\leq \tilde{f} + \frac{10(T_o+1)L^2\eta}{n}\left(\sum_{k=0}^{K-1}\mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + (T_o+1)^2\eta_l^2\sum_{k=0}^{K-1}\mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2]\right) \\
&\quad + \left(64L^2T_o^3\eta\eta_l^2 + \frac{3LT_o\eta^2}{2n}\right)\frac{K\sigma^2}{b} + \frac{3L^2(T_o+1)^3\eta\eta_l^2}{n}\sum_{k=0}^{K-1}\mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2], \quad (32)
\end{aligned}$$

where $\tilde{f} = f(\bar{\mathbf{x}}^0) - f^*$.

To show the convergence, we need to bound RHS of (32) by the following steps.

Step 1: formulating the dynamics of consensus errors and tracking errors. For any $k > 0$, let

$$\Phi^k \triangleq \begin{bmatrix} \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] \\ \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] \end{bmatrix} \quad \text{and} \quad \mathbf{e}^k \triangleq \begin{bmatrix} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] \\ \frac{\sigma^2}{b} \end{bmatrix}.$$

Assuming that $\eta \leq \frac{\lambda_p}{80L(T_o+1)}$ and $\eta_l \leq \frac{1}{8L(T_o+1)}$, we can formulate the dynamics by following Lemma 5 and Lemma 6,

$$\Phi^{k+1} \leq (1-p)\mathbf{A}\Phi^k + \frac{(1-p)\lambda^2}{\lambda_p}\mathbf{B}\mathbf{e}^k,$$

where

$$\mathbf{A} = \begin{bmatrix} \frac{1+(1+p)\lambda^2}{2} & \frac{40\lambda^2(T_o+1)^2\eta^2}{\lambda_p} \\ \frac{400\lambda^2L^2}{\lambda_p} & \frac{1+(1+p)\lambda^2}{2} \end{bmatrix}, \quad (33)$$

$$\mathbf{B} = \begin{bmatrix} 240L^2(T_o+1)^4\eta^2\eta_l^2 & 320n(T_o+1)^2\eta^2 \\ 125L^2(T_o+1)^2\eta^2 & 180n \end{bmatrix}. \quad (34)$$

By telescoping, we have

$$\Phi^k \leq (1-p)^k\mathbf{A}^k\Phi^0 + \sum_{t=0}^{k-1}((1-p)\mathbf{A})^t\frac{(1-p)\lambda^2}{\lambda_p}\mathbf{B}\mathbf{e}^{k-1-t}. \quad (35)$$

Summing (35) from $k = 0$ to K gives

$$\begin{aligned}
\sum_{k=0}^K\Phi^k &\leq \sum_{k=0}^K(1-p)^k\mathbf{A}^k\Phi^0 + \sum_{k=0}^K\sum_{t=0}^{k-1}((1-p)\mathbf{A})^t\frac{(1-p)\lambda^2}{\lambda_p}\mathbf{B}\mathbf{e}^{k-1-t} \\
&\leq \left(\sum_{k=0}^{\infty}(1-p)^k\mathbf{A}^k\right)\Phi^0 + \left(\sum_{k=0}^{\infty}((1-p)\mathbf{A})^k\right)\sum_{k=0}^{K-1}\frac{(1-p)\lambda^2}{\lambda_p}\mathbf{B}\mathbf{e}^k, \quad (36)
\end{aligned}$$

where we define $\mathbf{0}^0 = 1$.

Step 2: showing the invertibility of $\mathbf{I} - (1-p)\mathbf{A}$. To bound the sum of Φ^k , we have to control $\sum_{k=0}^{\infty} (1-p)^k \mathbf{A}^k$. To this end, the following lemma shows the spectral radius of $(1-p)\mathbf{A}$ is less than 1, implying that $\mathbf{I} - (1-p)\mathbf{A}$ is invertible.

Lemma 8 (The spectral radius of \mathbf{A}). *If $\eta \leq \frac{(1+p)\lambda_p^2}{80\sqrt{10}(T_o+1)L}$, it holds that $\rho((1-p)\mathbf{A}) < 1$, where $\rho(\mathbf{A})$ represents the spectral radius of \mathbf{A} defined in (33).*

Proof. The eigenvalue x of \mathbf{A} satisfies

$$p(x) := x^2 - (1 + (1+p)\lambda^2)x + \left(\frac{(1 + (1+p)\lambda^2)^2}{4} - \frac{16000\lambda^4(T_o+1)^2L^2\eta^2}{\lambda_p^2} \right) = 0. \quad (37)$$

Suppose that x_1 and x_2 are the two solutions to (37). Without the loss of generality, we have $x_1 \geq x_2$ and

$$\begin{aligned} x_1 &= \frac{1 + (1+p)\lambda^2}{2} + \frac{40\sqrt{10}\lambda^2(T_o+1)L\eta}{\lambda_p}, \\ x_2 &= \frac{1 + (1+p)\lambda^2}{2} - \frac{40\sqrt{10}\lambda^2(T_o+1)L\eta}{\lambda_p}. \end{aligned}$$

If $\eta \leq \frac{(1+p)\lambda_p^2}{80\sqrt{10}(T_o+1)L}$, we have

$$\frac{40\sqrt{10}\lambda^2(T_o+1)L\eta}{\lambda_p} \leq \frac{(1+p)\lambda_p\lambda^2}{2}.$$

Correspondingly,

$$\rho(\mathbf{A}) = \max\{|x_1|, |x_2|\} = |x_1| \leq \frac{1 + (1+p)\lambda^2(1 + \lambda_p)}{2}$$

and

$$\rho((1-p)\mathbf{A}) \leq \frac{1-p + (1+p)(1-\lambda_p^2)}{2} < 1,$$

where the last line is based on Assumption 1. □

Step 3: computing the cumulative consensus errors. Because of the invertibility of $\mathbf{I} - (1-p)\mathbf{A}$, it follows [HJ12, Corollary 5.6.16] that

$$\sum_{k=0}^{\infty} (1-p)^k \mathbf{A}^k = (\mathbf{I} - (1-p)\mathbf{A})^{-1}$$

such that (36) becomes

$$\sum_{k=0}^K \Phi^k \leq (\mathbf{I} - (1-p)\mathbf{A})^{-1} \Phi^0 + \mathbf{C} \sum_{k=0}^{K-1} \mathbf{e}^k. \quad (38)$$

where

$$\mathbf{C} = (\mathbf{I} - (1-p)\mathbf{A})^{-1} \frac{(1-p)\lambda^2}{\lambda_p} \mathbf{B}.$$

Now, we are going to control the upper bound of the $(\mathbf{I} - (1-p)\mathbf{A})^{-1}$, i.e., the upper bound of

$$\frac{1}{\det(\mathbf{I} - (1-p)\mathbf{A})} \text{adj}(\mathbf{I} - (1-p)\mathbf{A}),$$

where $\det(\mathbf{A})$ means the determinant of \mathbf{A} and $\text{adj}(\mathbf{A})$ represents the adjugate of \mathbf{A} . If the step-size further satisfies $\eta \leq \frac{(1+p)\lambda_p^2}{360(T_o+1)L}$,

$$\det(\mathbf{I} - (1-p)\mathbf{A}) = \left(1 - \frac{(1-p)(1 + (1+p)\lambda^2)}{2} \right)^2 - \frac{16000\lambda^4(T_o+1)^2L^2(1-p)^2\eta^2}{\lambda_p^2}$$

$$\begin{aligned}
&= \left(\frac{(1+p)\lambda_p}{2} \right)^2 - \frac{16000\lambda^4(T_o+1)^2L^2(1-p)^2\eta^2}{\lambda_p^2} \\
&\geq \frac{(1+p)^2\lambda_p^2}{4} - \frac{(1+p)^2\lambda_p^2}{8} = \frac{(1+p)^2\lambda_p^2}{8}.
\end{aligned}$$

Thus,

$$\begin{aligned}
(\mathbf{I} - (1-p)\mathbf{A})^{-1} &\leq \frac{8}{(1+p)^2\lambda_p^2} \begin{bmatrix} \frac{(1+p)\lambda_p}{2} & \frac{40(1-p)\lambda^2(T_o+1)^2\eta^2}{\lambda_p} \\ \frac{400(1-p)\lambda^2L^2}{\lambda_p} & \frac{(1+p)\lambda_p}{2} \end{bmatrix} \\
&\leq \frac{4}{(1+p)^2\lambda_p^3} \begin{bmatrix} (1+p)\lambda_p^2 & 80(1-\lambda_p)(T_o+1)^2\eta^2 \\ 800(1-\lambda_p)L^2 & (1+p)\lambda_p^2 \end{bmatrix}
\end{aligned}$$

and

$$\mathbf{C} \leq \frac{240(1-\lambda_p)}{(1+p)^2\lambda_p^4} \begin{bmatrix} 4c_{\lambda,1}L^2(T_o+1)^4\eta^2\eta_l^2 & 6nc_{\lambda,2}(T_o+1)^2\eta^2 \\ 3c_{\lambda,3}L^2(T_o+1)^2\eta^2 & 3nc_{\lambda,4} \end{bmatrix}$$

where

$$\begin{aligned}
c_{\lambda,1} &= ((1+p)\lambda_p^2 + 60(1-\lambda_p)\eta_c^2), \\
c_{\lambda,2} &= ((1+p)\lambda_p^2 + 40(1-\lambda_p)), \\
c_{\lambda,3} &= ((1+p)\lambda_p^2 + 1100(1-\lambda_p)L^2(T_o+1)^2\eta_l^2), \\
c_{\lambda,4} &= ((1+p)\lambda_p^2 + 1600(1-\lambda_p)L^2(T_o+1)^2\eta^2).
\end{aligned}$$

Thus, if $\mathbf{X}^0 = \mathbf{x}^0 \mathbf{1}_n^\top$,

$$\begin{aligned}
\sum_{k=0}^{K-1} \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] &\leq \sum_{k=0}^K \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] \leq \frac{320(1-\lambda_p)(T_o+1)^2\eta^2}{(1+p)^2\lambda_p^3} \mathbb{E}[\|\Phi_y^0\|_{\mathbb{F}}^2] \\
&\quad + \frac{960c_{\lambda,1}(1-\lambda_p)L^2(T_o+1)^4\eta^2\eta_l^2}{(1+p)^2\lambda_p^4} \sum_{k=0}^{K-1} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] \\
&\quad + \frac{1440c_{\lambda,2}(1-\lambda_p)(T_o+1)^2\eta^2}{b(1+p)^2\lambda_p^4} nK\sigma^2,
\end{aligned} \tag{39a}$$

$$\begin{aligned}
\sum_{k=0}^{K-1} \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] &\leq \sum_{k=0}^K \mathbb{E}[\|\Phi_y^k\|_{\mathbb{F}}^2] \leq \frac{4}{(1+p)\lambda_p} \mathbb{E}[\|\Phi_y^0\|_{\mathbb{F}}^2] \\
&\quad + \frac{720c_{\lambda,3}(1-\lambda_p)L^2(T_o+1)^2\eta^2}{(1+p)^2\lambda_p^4} \sum_{k=0}^{K-1} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] \\
&\quad + \frac{720c_{\lambda,4}(1-\lambda_p)}{b(1+p)^2\lambda_p^4} nK\sigma^2.
\end{aligned} \tag{39b}$$

Step 4: finishing up. Substituting (39) into the (32), we have

$$\begin{aligned}
&\frac{\eta(T_o+1)}{4} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \\
&\leq \tilde{f} + \frac{80c_{\lambda,1}L^2(T_o+1)^3\eta\eta_l^2}{n(1+p)^2\lambda_p^3} \mathbb{E}[\|\Phi_y^0\|_{\mathbb{F}}^2] + \frac{3c_2L^2(T_o+1)^3\eta\eta_l^2}{n(1+p)^2\lambda_p^4} \sum_{k=0}^{K-1} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] \\
&\quad + \left(\underbrace{\frac{7200c_3(1-\lambda_p)L^2(T_o+1)^3\eta\eta_l^2}{(1+p)^2\lambda_p^4} + 64L^2(T_o+1)^3\eta\eta_l^2 + \frac{3LT_o\eta^2}{2n}}_{D_1} \right) K\sigma^2/b,
\end{aligned} \tag{40}$$

where

$$\begin{aligned} c_2 &= 6400 \max\{c_{\lambda,1}, c_{\lambda,3}\}(1 - \lambda_p)L^2(T_o + 1)^2\eta^2 + (1 + p)^2\lambda_p^4 \leq 5(1 + p)^2\lambda_p^4 \\ c_3 &= 2c_{\lambda,2}\eta_c^2 + c_{\lambda,4} \leq 84\eta_c^2 + 2(1 + p)\lambda_p^2. \end{aligned}$$

Together with (14) and (39a), we have

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[\|\bar{\mathbf{Y}}^k\|_{\mathbb{F}}^2] &\leq 3K\sigma^2/b + 3L^2 \sum_{k=0}^{K-1} \mathbb{E}[\|\Phi_x^k\|_{\mathbb{F}}^2] + 3n \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \\ &\leq 6n \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] + \frac{1920(1 - \lambda_p)L^2(T_o + 1)^2\eta^2}{(1 + p)^2\lambda_p^3} \mathbb{E}[\|\Phi_y^0\|_{\mathbb{F}}^2] \\ &\quad + \underbrace{\left(\frac{1440c_{\lambda,2}(1 - \lambda_p)L^2(T_o + 1)^2\eta^2}{(1 + p)^2\lambda_p^4} + \frac{1}{n} \right)}_{D_2} 6nK\sigma^2/b \end{aligned}$$

where the last inequality is due to the step-size conditions, i.e., $\eta \leq \frac{(1+p)\lambda_p^2}{360L(T_o+1)}$ and $\eta_l \leq \frac{1}{8L(T_o+1)}$ s.t.

$$\frac{3 \cdot 960c_{\lambda,1}L^4(T_o + 1)^4\eta^2\eta_l^2}{(1 + p)^2\lambda_p^4} \leq \frac{1}{2}.$$

Because $c_2 \leq 5(1 + p)^2\lambda_p^4$, by further assuming that $\eta_l \leq \frac{1}{27L(T_o+1)}$ we have

$$\begin{aligned} \frac{18c_2L^2(T_o + 1)^3\eta\eta_l^2}{(1 + p)^2\lambda_p^4} &\leq \frac{\eta(T_o + 1)}{8}, \\ \frac{5760c_2(1 - p)\lambda^2L^4(T_o + 1)^5\eta^3\eta_l^2}{n(1 + p)^4\lambda_p^7} &\leq \alpha_0 \frac{(1 - p)L^2(T_o + 1)^3\eta^3}{n(1 + p)^2\lambda_p^3} = \frac{\alpha_0(1 - p)(T_o + 1)\eta}{n}, \\ \frac{18c_2L^2(T_o + 1)^3\eta\eta_l^2}{(1 + p)^2\lambda_p^4} D_2 &\leq \alpha_1 L^2(T_o + 1)^3\eta\eta_l^2, \end{aligned}$$

for some positive absolute constant α_0, α_1 .

Assume that $\eta_c = \alpha\sqrt{1 + p}\lambda_p$ and $\eta_l \leq \frac{\sqrt{1+p}\lambda_p}{360\alpha L(T_o+1)}$ for any positive $\alpha > 1$. Then,

$$\begin{aligned} \frac{80c_{\lambda,1}L^2(T_o + 1)^3\eta\eta_l^2}{n(1 + p)^2\lambda_p^3} &\leq \alpha_2 \left(\frac{(T_o + 1)\eta}{n} + \frac{(1 - p)L^2(T_o + 1)^3\eta^3}{n(1 + p)^2\lambda_p^3} \right) = \alpha_2 \left(\frac{(T_o + 1)\eta}{n} \right) \\ D_1 &\leq \alpha_3 \left(\frac{(1 - p)L^2(T_o + 1)^3\eta^3}{(1 + p)^2\lambda_p^4} + L^2(T_o + 1)^3\eta\eta_l^2 \right) \end{aligned}$$

for some positive absolute constant α_2, α_3 .

Therefore,

$$\begin{aligned} &\frac{\eta(T_o + 1)}{8} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \\ &\leq \tilde{f} + O\left(\frac{(T_o + 1)\eta}{n}\right) \mathbb{E}[\|\Phi_y^0\|_{\mathbb{F}}^2] + O\left(\frac{(1 - p)L^2(T_o + 1)^2\eta^2}{(1 + p)^2\lambda_p^4} + L^2(T_o + 1)^2\eta_l^2 + \frac{3L\eta}{2n}\right) \frac{K(T_o + 1)\eta\sigma^2}{b}, \end{aligned}$$

which is equivalent to

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \leq O\left(\frac{\tilde{f}}{\eta T_o K}\right) + O\left(\frac{1}{nK}\right) \mathbb{E}[\|\Phi_y^0\|_{\mathbb{F}}^2] + O\left(\frac{(1 - p)L^2 T_o^2 \eta^2}{(1 + p)^2 \lambda_p^4} + L^2 T_o^2 \eta_l^2 + \frac{L\eta}{n}\right) \frac{\sigma^2}{b}.$$

B Proof of Corollary 1

Lemma 9 (Modified Lemma 17 in [KLB⁺20]). *For any parameters $r_0 \geq 0, a_1 \geq 0, a_2 \geq 0, a_3 \geq 0$, if K is sufficiently large s.t. $\eta = \min \left\{ \left(\frac{r_0}{a_1 K} \right)^{\frac{1}{2}}, \left(\frac{r_0}{a_2 K} \right)^{\frac{1}{3}}, \left(\frac{r_0}{a_3 K} \right)^{\frac{1}{3}} \right\} \leq \bar{\eta}$, i.e.,*

$$K \geq \max \left\{ \frac{r_0}{a_1 \bar{\eta}^2}, \frac{r_0}{a_2 \bar{\eta}^3}, \frac{r_0}{a_3 \bar{\eta}^3} \right\}$$

we have

$$\Psi^K = \frac{r_0}{\eta K} + a_1 \eta + a_2 \eta^2 + a_3 \eta^2 \leq 2 \left(\frac{a_1 r_0}{K} \right)^{\frac{1}{2}} + 2 \left(\frac{\sqrt{a_2 r_0}}{K} \right)^{\frac{2}{3}} + 2 \left(\frac{\sqrt{a_3 r_0}}{K} \right)^{\frac{2}{3}}.$$

Proof. We mainly follow the proof of [KLB⁺20, Lemma 17], except the first discussed case in which $\bar{\eta}$ scales on the order of \sqrt{K} or $\sqrt[3]{K}$.

- If $\eta = \left(\frac{r_0}{a_1 K} \right)^{\frac{1}{2}} \leq \min \left\{ \left(\frac{r_0}{a_2 K} \right)^{\frac{1}{3}}, \left(\frac{r_0}{a_3 K} \right)^{\frac{1}{3}} \right\}$,

$$\Psi^K \leq 2 \left(\frac{r_0 a_1}{K} \right)^{\frac{1}{2}} + a_2 \left(\frac{r_0}{a_1 K} \right) + a_3 \left(\frac{r_0}{a_1 K} \right) \leq 2 \left(\frac{r_0 a_1}{K} \right)^{\frac{1}{2}} + \left(\frac{\sqrt{a_2 r_0}}{K} \right)^{\frac{2}{3}} + \left(\frac{\sqrt{a_3 r_0}}{K} \right)^{\frac{2}{3}},$$

- If $\eta = \left(\frac{r_0}{a_2 K} \right)^{\frac{1}{3}} \leq \min \left\{ \left(\frac{r_0}{a_1 K} \right)^{\frac{1}{2}}, \left(\frac{r_0}{a_3 K} \right)^{\frac{1}{3}} \right\}$,

$$\Psi^K \leq 2 \left(\frac{\sqrt{a_2 r_0}}{K} \right)^{\frac{2}{3}} + a_1 \left(\frac{r_0}{a_2 K} \right)^{\frac{1}{3}} + a_3 \left(\frac{r_0}{a_2 K} \right)^{\frac{2}{3}} \leq 2 \left(\frac{\sqrt{a_2 r_0}}{K} \right)^{\frac{2}{3}} + \left(\frac{r_0 a_1}{K} \right)^{\frac{1}{2}} + \left(\frac{\sqrt{a_3 r_0}}{K} \right)^{\frac{2}{3}},$$

- If $\eta = \left(\frac{r_0}{a_3 K} \right)^{\frac{1}{3}} \leq \min \left\{ \left(\frac{r_0}{a_1 K} \right)^{\frac{1}{2}}, \left(\frac{r_0}{a_2 K} \right)^{\frac{1}{3}} \right\}$,

$$\Psi^K \leq 2 \left(\frac{\sqrt{a_3 r_0}}{K} \right)^{\frac{2}{3}} + a_1 \left(\frac{r_0}{a_3 K} \right)^{\frac{1}{3}} + a_2 \left(\frac{r_0}{a_3 K} \right)^{\frac{2}{3}} \leq 2 \left(\frac{\sqrt{a_3 r_0}}{K} \right)^{\frac{2}{3}} + \left(\frac{r_0 a_1}{K} \right)^{\frac{1}{2}} + \left(\frac{\sqrt{a_2 r_0}}{K} \right)^{\frac{2}{3}}.$$

□

By rearranging (5), we have

$$\begin{aligned} & \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^k)\|_2^2] \\ & \leq O \left(\frac{\tilde{f}}{\eta \eta_c T_o K} + \frac{\eta_l \eta_c L \sigma^2}{nb} + \frac{(1-p)}{(1+p)^2 \lambda_p^4} \frac{L^2 T_o^2 \eta_l^2 \eta_c^2 \sigma^2}{b} + \frac{L^2 T_o^2 \eta_l^2 \sigma^2}{b} + \frac{1}{nK} \mathbb{E}[\|\Phi_y^0\|_F^2] \right). \end{aligned} \quad (41)$$

Letting

$$r_o = \frac{\tilde{f}}{\sqrt{1+p} \lambda_p}, \quad a_1 = \frac{\sqrt{1+p} \lambda_p L \sigma^2}{n T_o b}, \quad a_2 = \frac{(1-p) L^2 \sigma^2}{(1+p) \lambda_p^2 b}, \quad a_3 = \frac{L^2 \sigma^2}{\alpha^2 b},$$

and applying Lemma 9, then the RHS of (41) can be bounded by

$$O \left(\left(\frac{L \tilde{f} \sigma^2}{n T_o b K} \right)^{\frac{1}{2}} + \left(\frac{\sqrt{1-p} L \tilde{f} \sigma}{(1+p) \lambda_p^2 \sqrt{b} K} \right)^{\frac{2}{3}} + \left(\frac{L \tilde{f} \sigma}{\sqrt{(1+p) \lambda_p} \alpha \sqrt{b} K} \right)^{\frac{2}{3}} + \frac{1}{nK} \mathbb{E}[\|\Phi_y^0\|_F^2] \right),$$

if K is sufficiently large such that $\min \left\{ \left(\frac{r_o}{a_1 K} \right)^{\frac{1}{2}}, \left(\frac{r_o}{a_2 K} \right)^{\frac{1}{3}}, \left(\frac{r_o}{a_3 K} \right)^{\frac{1}{3}} \right\} \leq \frac{\sqrt{1+p} \lambda_p}{360L}$, i.e.,

$$K \geq \max \left\{ \frac{360^2 n T_o b L \tilde{f}}{(1+p)^2 \lambda_p^4 \sigma^2}, \frac{360^3 b L \tilde{f}}{(1-p^2) \lambda_p^2 \sigma^2} \mathbf{1}_{p < 1}, \frac{360^3 \alpha^2 b L \tilde{f}}{(1+p)^2 \lambda_p^4 \sigma^2} \right\}, \quad (42)$$

where $\mathbf{1}_{p < 1} = 1$ if $p < 1$ and is set to be 0 otherwise.