Distributed Stochastic Bilevel Optimization: Improved Complexity and Heterogeneity Analysis

Youcheng Niu[†]   YCNIU@ZJU.EDU.CN

Jinming Xu[†,*]   JIMMYXU@ZJU.EDU.CN

Ying Sun[‡]   YBS5190@PSU.EDU

Yan Huang[†]   HUANGYAN5616@ZJU.EDU.CN

Li Chai[†]   CHAILI@ZJU.EDU.CN

[†] *College of Control Science and Engineering, Zhejiang University, China*

[‡] *School of Electrical Engineering and Computer Science, The Pennsylvania State University, USA*

[*] *Corresponding Author*

Editor: Peter Richtarik

Abstract

This paper considers solving a class of nonconvex-strongly-convex distributed stochastic bilevel optimization (DSBO) problems with personalized inner-level objectives. Most existing algorithms require computational loops for hypergradient estimation, leading to computational inefficiency. Moreover, the impact of data heterogeneity on convergence in bilevel problems is not explicitly characterized yet. To address these issues, we propose LoPA, a loopless personalized distributed algorithm that leverages a tracking mechanism for iterative approximation of inner-level solutions and Hessian-inverse matrices without relying on extra computation loops. Our theoretical analysis explicitly characterizes the heterogeneity across nodes (denoted by b), and establishes a sublinear rate of $\mathcal{O}\left(\frac{1}{(1-\rho)K} + \frac{(\frac{b}{\sqrt{m}})^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right)$ without the boundedness of local hypergradients, where σ_{p} and σ_{c} represent the gradient sampling variances associated with the inner- and outer-level variables, respectively. We also integrate LoPA with a gradient tracking scheme to eliminate the impact of data heterogeneity, yielding an improved rate of $\mathcal{O}\left(\frac{1}{(1-\rho)^2 K} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right)$. The computational complexity of LoPA is of $\mathcal{O}(\epsilon^{-2})$ to an ϵ-stationary point, matching the communication complexity due to the loopless structure, which outperforms existing counterparts for DSBO. Numerical experiments validate the effectiveness of the proposed algorithm.

Keywords: distributed learning, bilevel optimization, non-convex optimization, heterogeneity analysis, gradient tracking

1. Introduction

Bilevel optimization is a hierarchical optimization framework that involves an outer- and inner-level problem, where the solution of the outer-level problem depends on that of the inner-level problem. This framework has gained significant attention recently in the field of machine learning due to its wide applications in areas such as meta-learning (Finn et al., 2017; Rajeswaran et al., 2019), neural architecture search (Zhang et al., 2022; Xue

et al., 2021), hyperparameter selection (Okuno et al., 2018; Bertrand et al., 2020), and reinforcement learning (Wang et al., 2016; Hong et al., 2023). With the increasing importance of large-scale machine learning, bilevel optimization has emerged as a promising approach in distributed settings, where multiple nodes with computation and communication capabilities can collaborate to improve the learning efficiency (Xu et al., 2015; Chen et al., 2022b; Lu et al., 2022a; Nedic, 2020; Jiao et al., 2022). Achieving this goal requires properly coordinating the nodes. In this paper, we aim to address a class of distributed stochastic bilevel optimization (DSBO) problems consisting of m nodes, each with a personalized inner-level objective as follows:

$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{f_i\left(x, \theta_i^*(x)\right)}_{\Phi_i(x)}, \text{ s.t. } \theta_i^*(x) = \arg\min_{\theta_i \in \mathbb{R}^p} g_i(x, \theta_i), \tag{1}$$

where $x \in \mathbb{R}^n$ and $\theta_i \in \mathbb{R}^p$ are the global and local model parameters, respectively; $f_i : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ denotes the outer-level objective of node i which is possibly nonconvex while $g_i : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is the inner-level objective that is strongly convex in θ uniformly for all $x \in \mathbb{R}^n$. We consider the stochastic setting where $f_i(x, \theta) = \mathbb{E}_{\varsigma_i \sim \mathcal{D}_{f_i}}[\hat{f}_i(x, \theta, \varsigma_i)]$ and $g_i(x, \theta) = \mathbb{E}_{\xi_i \sim \mathcal{D}_{g_i}}[\hat{g}_i(x, \theta, \xi_i)]$, with $\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$ denoting the data distribution related to the i-th outer- and inner-level objective, respectively.

Motivating Examples. Problem (1) finds a broad range of applications in practical distributed machine learning and min-max/compositional optimization problems, ranging from few-shot learning (Goldblum et al., 2020), adversarial learning (Madry et al., 2017), and reinforcement learning (Wang et al., 2016) to fair transceiver design (Razaviyayn et al., 2020). For instance, consider the following distributed meta-learning problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \sum_{t \in \mathcal{T}_i} f_i^t(\theta_i^*(x)), \text{ s.t. } \theta_i^*(x) = \arg\min_{\theta_i \in \mathbb{R}^p} \left\{ \sum_{t \in \mathcal{T}_i} \langle \theta_i, \nabla f_i^t(x) \rangle + \frac{\nu}{2} \|x - \theta_i\|^2 \right\},$$

where x is the global model parameter to be learned, f_i^t denotes the loss function for the t-th subtask corresponding to the task set $\mathcal{T}_i$ in node i, and $\nu > 0$ is an adjustable parameter. The objective of nodes is to cooperatively learn a good initial global model x that makes use of the knowledge obtained from past experiences among nodes to better adapt to unseen tasks with a small number of task-specific gradient updates (Finn et al., 2017; Fallah et al., 2020; Rajeswaran et al., 2019).

Different from the conventional single-level problem, bilevel optimization faces additional challenges due to its hierarchical structure, often leading to non-convex objectives (Ji et al., 2021; Ghadimi and Wang, 2018). In most cases, obtaining a closed-form expression, or computing exactly the hypergradient $\nabla \Phi_i(x)$ is difficult, due to its dependency on the inner-level solutions $\theta_i^*(x)$ (Chen et al., 2021a). For those with strongly-convex inner-level problems, the expression of the hypergradients can be obtained by implicit function theorem and further approximated by Approximate Implicit Differentiation (AID) approaches, but it typically involves two nested loops (Ji et al., 2022; Ghadimi and Wang, 2018): an N-loop to find a near-optimal solution to the inner-level function, and a Q-loop to approximate the Hessian-inverse matrices. This is particularly challenging in large-scale machine learning applications where running these two nested loops is prohibitively computationally expensive, and also costly in terms of the training time (Ghadimi and Wang, 2018). Furthermore,

in the stochastic setting, integrating the SGD method with AID for bilevel optimization with low computation cost and sample complexity is challenging, due to the estimation bias of the hypergradient. To address this issue, various approximation algorithms have been proposed to estimate the hypergradients and reduce the bias (Ghadimi and Wang, 2018; Hong et al., 2023; Arbel and Mairal, 2022; Chen et al., 2022a, 2021a; Khanduri et al., 2021; Yang et al., 2021; Li et al., 2022; Dagréou et al., 2022a). To deal with a wide range of large-scale machine learning tasks, there have been some distributed algorithms recently proposed for DSBO leveraging the distributed gradient descent approach, such as (Yang et al., 2022; Lu et al., 2022b; Gao et al., 2022; Lu et al., 2022a). However, these existing approaches often require computation loops to estimate the inner-level solutions and Hessian-inverse matrices for solving DSBO, inducing a high computation cost. Therefore, the following question arises naturally: *Can we design loopless decentralized learning algorithms for DSBO problems that achieve better computational complexity and training efficiency?* Moreover, unlike the standard distributed optimizations, the bilevel structure and the heterogeneity of inner-level solutions introduce new challenge in characterizing the node heterogeneity and obtaining sharp convergence, with its influence on the convergence rate unclear. This leads to another important theoretical question: *How to characterize the heterogeneity for DSBO problems, and how does it affect the algorithm's convergence performance?*

Summary of Contributions. To address the above challenges, we propose a new loopless distributed personalized algorithm (termed LoPA) for solving problem (1) and provide improved complexity as well as heterogeneity analysis. We summarize the key contributions as follows:

- **New loopless distributed algorithms**. We propose a new loopless distributed algorithm LoPA without requiring extra computation loops for personalized DSBO, which can employ either local gradient or gradient tracking scheme, termed LoPA-LG and LoPA-GT, respectively. Different from existing distributed algorithms for DSBO problems (Chen et al., 2022b; Lu et al., 2022b; Gao et al., 2022; Lu et al., 2022a; Yang et al., 2022), LoPA leverages an iterative approximation approach that requires only a single SGD step per iteration to respectively track the Hessian-inverse matrix and the inner-level solution. To control the impact of the bias on the overall convergence rate caused by removing the computation loops, LoPA further employs a gradient momentum step coupled with a relaxation step in the consensus update, ensuring that the bias decays at a sufficient fast rate.

- **Heterogeneity analysis**. We quantify the degree of heterogeneity among nodes in personalized DSBO by analyzing that of the inner- and outer-level functions (c.f., Lemma 5). Our analysis shows that LoPA-LG is able to achieve a convergence rate of $\mathcal{O}(\frac{\kappa^8}{(1-\rho)K} + \frac{\kappa^{\frac{16}{3}}(\frac{b}{\sqrt{m}})^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}} + \frac{\kappa^{\frac{5}{2}}}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}}))$ (c.f., Corollary 7), where $\sigma_{\mathrm{p}} = \mathcal{O}(\kappa^{\frac{1}{2}}\sigma_{f,\theta} + \kappa^{\frac{3}{2}}\sigma_{g,\theta\theta} + \kappa^{\frac{5}{2}}\sigma_{g,\theta})$, $\sigma_{\mathrm{c}} = \mathcal{O}(\sigma_{f,x} + \kappa\sigma_{g,x\theta})$ are the sampling gradient variances, $b^2 = \mathcal{O}(\kappa^2 b_f^2 + \kappa^6 b_g^2)$ is the heterogeneity among nodes, and κ represents the condition number (c.f., Assumptions 4, 5, Equation (15)). Our analysis relies on the weak assumption of the bounded inner-level heterogeneity at the optimum, while avoiding the boundedness of local hypergradients (c.f., Remark 1), and the resulting rate shows the clear dependence of convergence on the network connectivity ρ, condition

number κ, sampling variances $\sigma_\mathrm{p}, \sigma_\mathrm{c}$ and node heterogeneity b. The techniques used for heterogeneity analysis allow us to explicitly characterize the impact of the heterogeneity on convergence for DSBO compared to existing works (Chen et al., 2022b; Lu et al., 2022b; Gao et al., 2022; Lu et al., 2022a; Yang et al., 2022) (c.f., Remarks 1 and 8), which is of independent interest.

- **Improved complexity**. The analysis of LoPA-LG reveals that the node heterogeneity affects the convergence rate by introducing a transient term of $\mathcal{O}(K^{-\frac{2}{3}})$ in DSBO, vanishing at a slower rate than the leading term, in which the inner-level heterogeneity plays a crucial role. This motivates us further designing LoPA-GT based on gradient tracking to eliminate the heterogeneity. In particular, we prove that LoPA-GT improves the rate to $\mathcal{O}(\frac{\kappa^8}{(1-\rho)^2 K} + \frac{\kappa^{\frac{5}{2}}}{\sqrt{K}}(\sigma_\mathrm{p} + \frac{1}{\sqrt{m}}\sigma_\mathrm{c}))$. Thanks to the loopless structure, both LoPA-LG and LoPA-GT are shown to have a computational complexity of $\mathcal{O}(\epsilon^{-2})$ to attain an ϵ-stationary point, improving existing works on DSBO (Chen et al., 2022b; Lu et al., 2022b; Gao et al., 2022; Lu et al., 2022a; Yang et al., 2022) by order of $\mathcal{O}(\log \epsilon^{-1})$ (See Table 1). Our analysis further reveals that our algorithms can achieve the best-known $\mathcal{O}(\kappa^5 m^{-1} \epsilon^{-2})$-computational complexity w.r.t. out-level gradient evaluations in existing works on DSBO without mean-square smoothness (c.f., Remark 12). Numerical experiments on machine learning tasks demonstrate the effectiveness of the proposed algorithm in dealing with node heterogeneity and reducing the computational cost.

2. Related Works

Bilevel optimization with SGD methods. There have been some efforts devoted to achieving more accurate stochastic hypergradients and ensuring convergence in solving bilevel optimization with SGD methods, such as using computation loops to reduce the error of approximating Hessian-inverse matrices (Q-loop) and increase the accuracy of inner-level solutions (N-loop) (Ghadimi and Wang, 2018), increasing the batch size (Arbel and Mairal, 2022), adopting two-timescale step-sizes to eliminate steady-state stochastic variance (Hong et al., 2023), incorporating additional correction terms (Chen et al., 2022a; Khanduri et al., 2021; Li et al., 2022; Yang et al., 2021; Chen et al., 2024), and exploring the smoothness of objectives (Chen et al., 2021a; Dagréou et al., 2022a). Among these stochastic approximation algorithms, works (Ghadimi and Wang, 2018; Chen et al., 2022a, 2021a) achieve a computational complexity of $\mathcal{O}\left(\epsilon^{-2} \log \epsilon^{-1}\right)$ due to use of extra computation loops for estimating the hypergradients, where ϵ represents the desired accuracy. Hong et al. (2023) develop a two-timescale approximation algorithm, whereas, the nature of the two-timescale update results in a sub-optimal computational complexity of $\mathcal{O}\left(\epsilon^{-5/2} \log \epsilon^{-1}\right)$ for the algorithm. Based on warm-start strategies, a computational complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ is provided in (Arbel and Mairal, 2022; Li et al., 2022; Dagréou et al., 2022a; Chen et al., 2024). By employing momentum accelerations in both outer- and inner-level optimization procedures such as STORM (Cutkosky and Orabona, 2019) or SPIDER (Fang et al., 2018) and imposing more strict assumptions, works (Khanduri et al., 2021; Yang et al., 2021) further improves the rate to $\mathcal{O}\left(\epsilon^{-3/2} \log \epsilon^{-1}\right)$. While the aforementioned works provide some insights into stochastic bilevel algorithmic design, they cannot be directly applied to distributed problem (1) as considered in this work.

Table 1: Comparison of existing distributed stochastic bilevel optimization algorithms with SGD methods.

Algorithm	Setting	# of Loop	Inner Step	Complexity	HA	Assumption	Scheme
MA-DSBO (Chen et al., 2022b)	**G**	N-Q-Loop	SGD	$\mathcal{O}\left(\epsilon^{-2}\log\epsilon^{-1}\right)$	No	f-**BG**, g-**BH**	
Gossip-DSBO (Yang et al., 2022)	**G**	Q-Loop	SGD	$\mathcal{O}\left(\epsilon^{-2}\log\epsilon^{-1}\right)$	No	f-**BG**, g-**BG**	LG
LoPA-LG (this work)	**P**	No-Loop	SGD	$\mathcal{O}\left(\epsilon^{-2}\right)$	Yes	f-**BH**, g-**BHO**	
SLAM (Lu et al., 2022b)	**G**	Q-Loop	SGD	$\mathcal{O}\left(\epsilon^{-2}\log\epsilon^{-1}\right)$	No	–	
VRDBO (Gao et al., 2022)	**G**	Q-Loop	STORM	$\mathcal{O}(\epsilon^{-3/2}\log\epsilon^{-1})$	No	**MSS**	GT
SPDB (Lu et al., 2022a)	**P**	Q-Loop	SGD	$\mathcal{O}(\epsilon^{-2}\log\epsilon^{-1})$	No	–	
LoPA-GT (this work)	**P**	No-Loop	SGD	$\mathcal{O}\left(\epsilon^{-2}\right)$	Yes	–	

*The algorithms above the dashed line use local gradient (LG) schemes, while those below use gradient tracking (GT) schemes; **G** and **P** represent DSBO with global and inner-level personalized objectives, respectively; The complexity represents the Hessian evaluations required to attain an ϵ-stationary point; HA refers to heterogeneity analysis. The column of 'assumption' outlines the *additional* assumptions beyond the standard Assumptions 1–3: f-**BG** refers to the gradient of $f_i(x,\theta)$ being bounded for any x and θ, i.e., $\|\nabla f_i(x,\theta)\|^2 \leqslant C$; g-**BH** refers to bounded data heterogeneity of $f_i(x,\theta), i \in \mathcal{V}$ for any x and θ, i.e., $\sum_{i=1}^{m}\sum_{i=j}^{m}\|\nabla g_i(x,\theta)-\nabla g_j(x,\theta)\|^2 \leqslant C$; g-**BHO** refers to bounded data heterogeneity of $g_i, i \in \mathcal{V}$ at the optimum $\theta_i^*(x)$ for any x; **MSS** refers to mean-squared smoothness, i.e., $\mathbb{E}[\|\nabla \hat{f}_i(w_1,\varsigma) - \nabla \hat{f}_i(w_2,\varsigma)\|] \leqslant l\|w_1 - w_2\|$ for any $w_1 = (x_1,\theta_1)$ and $w_2 = (x_2,\theta_2)$, which is a stronger condition than the function smoothness presented in Assumptions 2 and 3.

Distributed bilevel optimization. Compared to their centralized or parameter-server counterparts, distributed optimization offers several advantages on network scalability, system robustness, and privacy protection through peer-to-peer communication (Nedic, 2020). However, it also faces unique challenges, especially in dealing with data heterogeneity among nodes. In recent decades, various variants of distributed optimization algorithms have been developed, including distributed gradient descent (Chen et al., 2021b), gradient tracking (Xu et al., 2015), and alternating direction multiplier methods (Shi et al., 2014), accompanied by theoretical advancements. Specifically, for stochastic single-level nonconvex problems, these algorithms can achieve a computational complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ with SGD methods (Koloskova et al., 2020; Chen et al., 2021b). However, these single-level methods are not readily available to be adapted to tackle the interaction between the inner and outer levels of functions in solving bilevel optimization problems due to the absence of explicit knowledge of optimal solutions to the inner-level problem.

There have been some efforts aiming at solving distributed bilevel optimization problems, which can be generally cast into two categories: global DSBO and personalized DSBO. For global DSBO, pioneering works such as Gossip-based DSBO (termed Gossip-DSBO) (Yang et al., 2022), MA-DSBO (Chen et al., 2022b), VRDBO (Gao et al., 2022), and SLAM (Lu et al., 2022b) have been proposed which aim to solve an inner-level problem in the finite sum structure in a distributed manner. On the other hand, few works, such as SPDB (Lu et al., 2022a), have been developed for solving personalized DSBO problems where each node

has its own local inner-level problem. Various algorithmic frameworks have been developed in the above-mentioned works leveraging distributed optimization methods (Chen et al., 2021b; Xu et al., 2015; Shi et al., 2014) to minimize the outer- and inner-level functions and handle consensus constraints. To estimate the Hessian-inverse matrices, these frameworks utilize techniques such as Jacobian-Hessian-inverse product (Chen et al., 2022b), Hessian-inverse-vector product (Rajeswaran et al., 2019) or Neumann series approaches (Ghadimi and Wang, 2018) to avoid explicit computation of the inverse matrices. As the estimation of the Hessian-inverse matrices and inner-level solutions obtained using SGD methods is biased (Chen et al., 2021a), these algorithms often incorporate extra computation loops to reduce the bias of the hypergradient estimation due to the approximation of the Hessian-inverse matrices and inner-level solutions, and the computational complexity of these algorithms is typically of $\mathcal{O}(\epsilon^{-2} \log \epsilon^{-1})$ (Yang et al., 2022; Chen et al., 2022b; Lu et al., 2022b,a). The complexity can be improved to $\mathcal{O}\left(\epsilon^{-3/2} \log \epsilon^{-1}\right)$ by utilizing variance reduced gradient estimators (Gao et al., 2022) and additional mean-squared smoothness assumptions. However, it should be noted that these existing distributed algorithms for DSBO still incur high computational costs due to the extra computation loops required. Moreover, it remains unclear how one can properly characterize the heterogeneity among nodes as well as its detailed impact on convergence performance in DSBO.

Parallel works. Recently, there have been several studies (Zhang et al., 2023; Dong et al., 2023; Kong et al., 2024) concurrently exploring loopless algorithms for distributed bilevel optimization problems. However, these works all focus on global DSBO problems where the out-level objective relies on a common inner-level solution. In particular, Zhang et al. (2023) employ variance reduction techniques and GT schemes to achieve an iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ under the assumption of mean-square smoothness, while Dong et al. (2023) establish an iteration complexity of $\mathcal{O}(\epsilon^{-1})$ with the help of GT schemes for deterministic cases. Kong et al. (2024) further provide a detailed complexity analysis for their proposed method employing LG schemes. In contrast, we address personalized DSBO problems where the outer-level objective relies on *personalized* inner-level solutions, leading to a more challenging scenario for heterogeneity analysis. We thus provide unique analysis for quantifying the node heterogeneity and characterize its impact on the convergence for both LoPA-LG and LoPA-GT under a unified analytical framework, yielding tighter rates that clearly show the dependence of convergence on various factors, such as κ, b, $\sigma_{\mathrm{p}}, \sigma_{\mathrm{c}}$ and ρ, and improved complexity results in terms of out-level gradient evaluations under weaker assumptions on data heterogeneity (c.f Remarks 1, 12, Corollaries 7, 10).

3. Algorithm Design

In this section, we will present the proposed LoPA algorithm. Before delving into the details of the algorithm, we first provide some necessary preliminaries including relevant network models and assumptions.

3.1 Preliminaries

Network model. We model an undirected communication network as a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$, where $\mathcal{V} = \{1, \cdots, m\}$ is the set of nodes, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges,

and $W = [w_{ij}]_{i,j=1}^m$ is the weight matrix. The set of neighbors of node i is denoted by $\mathcal{N}_i = \{j \,|\, (i,j) \in \mathcal{E}\}$. We make the following standard assumption on graph $\mathcal{G}$.

Assumption 1 (Network connectivity) *The communication network $\mathcal{G}$ is connected and the weight matrix W satisfies i) $w_{ij} = w_{ji} > 0$ if and only if $(i,j) \in \mathcal{E}$; and $w_{ij} = 0$ otherwise; ii) W is doubly stochastic. Consequently, we have $\rho \triangleq \|W - \frac{1_m 1_m^{\mathrm{T}}}{m}\|^2 \in [0,1)$.*

In what follows, we make several assumptions on the outer- and inner-level functions of problem (1), which are common in the bilevel optimization literature (Ghadimi and Wang, 2018; Hong et al., 2023; Chen et al., 2022b; Yang et al., 2022; Chen et al., 2024; Lu et al., 2022a; Chen et al., 2023).

Assumption 2 (Outer-level functions) *Let $L_{f,x}$, $L_{f,\theta}$ and $C_{f,\theta}$ be positive constants. Each outer-level function $(x,\theta) \mapsto f_i(x,\theta)$, $i \in \mathcal{V}$ satisfies the following properties:*
(i) f_i is continuously differentiable;
(ii) For any $\theta \in \mathbb{R}^p$, $\nabla_x f_i(\cdot,\theta)$ is $L_{f,x}$-Lipschitz-continuous in x; and for any $x \in \mathbb{R}^n$, $\nabla_\theta f_i(x,\cdot)$ is $L_{f,\theta}$-Lipschitz-continuous in θ;
(iii) For any $x \in \mathbb{R}^n$, $\|\nabla_\theta f_i(x,\theta_i^*(x))\| \leqslant C_{f,\theta}$.

Assumption 3 (Inner-level functions) *Let μ_g, $L_{g,\theta}$, $L_{g,x\theta}$, $L_{g,\theta\theta}$ and $C_{g,x\theta}$ be positive constants. Each inner-level function $(x,\theta) \mapsto g_i(x,\theta)$, $i \in \mathcal{V}$ satisfies the following properties:*
(i) For any $x \in \mathbb{R}^n$, $g_i(x,\cdot)$ is μ_g-strongly convex in θ; g_i is twice continuously differentiable;
(ii) For any $x \in \mathbb{R}^n$, $\nabla_\theta g_i(x,\cdot)$ is $L_{g,\theta}$-Lipschitz-continuous in θ; $\nabla_{x\theta}^2 g_i(\cdot,\cdot)$, $\nabla_{\theta\theta}^2 g_i(\cdot,\cdot)$ are respectively $L_{g,x\theta}$- and $L_{g,\theta\theta}$-Lipschitz-continuous;
(iii) For any $\theta \in \mathbb{R}^p$ and $x \in \mathbb{R}^n$, $\|\nabla_{x\theta}^2 g_i(x,\theta)\| \leqslant C_{g,x\theta}$.

Next, recalling that $f_i(x,\theta) = \mathbb{E}_{\varsigma_i \sim \mathcal{D}_{f_i}}[\hat{f}_i(x,\theta,\varsigma_i)]$ and $g_i(x,\theta) = \mathbb{E}_{\xi_i \sim \mathcal{D}_{g_i}}[\hat{g}_i(x,\theta,\xi_i)]$, we proceed to make the following assumption regarding the data heterogeneity across nodes for problem (1), which resembles that of distributed single-level optimization (Chen et al., 2021b; Koloskova et al., 2020; Lian et al., 2017).

Assumption 4 (Bounded heterogeneity) *Let $\nabla_x f(x,\theta) \triangleq \frac{1}{m} \sum_{j=1}^m \nabla_x f_j(x,\theta)$ and $\nabla_\theta f(x,\theta) \triangleq \frac{1}{m} \sum_{j=1}^m \nabla_\theta f_j(x,\theta)$. There exist positive constants b_f^2 and b_g^2 such that:*
(i) $\sum_{i=1}^m \|\nabla_x f_i(x,\theta) - \nabla_x f(x,\theta)\|^2 \leqslant b_f^2$, $\sum_{i=1}^m \|\nabla_\theta f_i(x,\theta) - \nabla_\theta f(x,\theta)\|^2 \leqslant b_f^2$, for any x and θ;
(ii) $\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m \|\nabla_\theta g_i(x,\theta_j^*(x)) - \nabla_\theta g_j(x,\theta_j^*(x))\|^2 \leqslant b_g^2$, for any x and $\theta_i^*(x)$, $i \in \mathcal{V}$;
(iii) $\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m \|\nabla_{x\theta}^2 g_i(x,\theta_j^*(x)) - \nabla_{x\theta}^2 g_j(x,\theta_j^*(x))\|^2 \leqslant b_g^2$,
  $\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m \|\nabla_{\theta\theta}^2 g_i(x,\theta_j^*(x)) - \nabla_{\theta\theta}^2 g_j(x,\theta_j^*(x))\|^2 \leqslant b_g^2$, for any x and $\theta_i^*(x)$, $i \in \mathcal{V}$;
where $\nabla_x f$ and $\nabla_\theta f$ represent the partial gradient with respect to x and θ, respectively, while $\nabla_{x\theta}^2 g$ and $\nabla_{\theta\theta}^2 g$ denote Jacobian and Hessian, respectively.

Remark 1 (Weaker assumptions on data heterogeneity) *The parameters b_f^2 and b_g^2 are introduced to quantify the data heterogeneity on the outer- and inner-level functions across nodes, respectively. It is worth noting that Assumption 4(i) is weaker than that of the previous works based on LG schemes for DSBO problems, such as (Chen et al., 2022b; Yang*

et al., 2022) where the norm of $\nabla_x f_i(x, \theta)$ is assumed to be bounded by a constant. As for the heterogeneity in the inner-level functions (Assumption 4(ii) and 4(iii)), we only require that it is uniformly bounded at the optimum $\theta_i^*(x)$, $i \in \mathcal{V}$, for all x. This requirement is less restrictive than assuming the inner-level heterogeneity to be bounded at any θ as in (Kong et al., 2024).

3.2 The Proposed LoPA Algorithm

In this section, we present our algorithm, termed LoPA, for problem (1). Following the standard procedures as in distributed optimization (Chen et al., 2021b; Nedic and Ozdaglar, 2009), we let each node i maintain a local estimate x_i for the global decision variable x. At each iteration k, each node i alternates between a descent step with an estimate of $\nabla \Phi_i(x_i)$ and average consensus ensuring the consistency of the x_i's.

Hypergradient construction. By the chain rule and implicit function theorem (Ghadimi and Wang, 2018), we can compute the Hypergradient $\nabla \Phi_i(x_i)$ as follows:

$$\nabla \Phi_i(x_i) = \nabla_x f_i(x_i, \theta_i^*(x_i)) - \nabla_{x\theta}^2 g_i(x_i, \theta_i^*(x_i)) \big[\nabla_{\theta\theta}^2 g_i(x_i, \theta_i^*(x_i)) \big]^{-1} \nabla_\theta f_i(x_i, \theta_i^*(x_i)).$$

Note that computing the outer-level gradients in each iteration according to the above expression is computationally demanding. To address this issue, we first introduce an auxiliary variable θ_i to approximate the inner-level solutions $\theta_i^*(x_i)$, whose update follows a simple stochastic gradient descent step (c.f., (4) and (6)). As such, the Hessian-inverse-vector products $\big[\nabla_{\theta\theta}^2 g_i(x_i, \theta_i^*(x_i)) \big]^{-1} \nabla_\theta f_i(x_i, \theta_i^*(x))$ (abbreviated Hv) and local hypergradient thus can be approximately computed as follows (Dagréou et al., 2022b; Arbel and Mairal, 2022):

$$v_i = \big[\nabla_{\theta\theta}^2 g_i(x_i, \theta_i) \big]^{-1} \nabla_\theta f_i(x_i, \theta_i), \tag{2}$$

$$s_i = \nabla_x f_i(x_i, \theta_i) - \nabla_{x\theta}^2 g_i(x_i, \theta_i) v_i. \tag{3}$$

In the presence of stochasticity, steps (2) and (3) still cannot be computed directly as the gradient and Hessian are unknown. To overcome this issue, for step (3), we replace $\nabla_x f_i$ and $\nabla_{x\theta}^2 g_i$ with their stochastic estimates, respectively (c.f., (8)). As for step (2), notice that v_i can be regarded as the solution of the following strongly convex problem:

$$v_i = \arg\min_v \big\{ \frac{1}{2} v^{\mathrm{T}} \nabla_{\theta\theta}^2 g_i(x_i, \theta_i) v - \nabla_\theta f_i(x_i, \theta_i) v \big\}.$$

Instead of directly computing the solution using stochastic approximation methods, we propose to approximate it by performing only one stochastic gradient iteration which warm-starts with the value of v_i initialized to its value from the previous iteration. By doing so, we aim to approximate the solution more efficiently while taking advantage of the progress made in the previous iteration. Further details are provided in Remark 2 while the specific updates are given by (5) and (7).

Putting all the ingredients together, LoPA designs the following updates for the local inner-level variable θ_i^{k+1}, the local Hv estimate variable v_i^{k+1} and the local hypergradient s_i^{k+1} at iteration $k + 1$:

$$\theta_i^{k+1} = \theta_i^k - \beta d_i^k, \tag{4}$$

$$v_i^{k+1} = v_i^k - \lambda h_i^k. \tag{5}$$

where $\beta > 0$ and $\lambda > 0$ are step-sizes. The directions d_i^k and h_i^k are further updated as:

$$d_i^{k+1} = \nabla_\theta \hat{g}_i(x_i^{k+1}, \theta_i^{k+1}; \xi_{i,1}^{k+1}), \tag{6}$$

$$h_i^{k+1} = \nabla_{\theta\theta}^2 \hat{g}_i(x_i^{k+1}, \theta_i^{k+1}; \xi_{i,2}^{k+1}) v_i^{k+1} - \nabla_\theta \hat{f}_i(x_i^{k+1}, \theta_i^{k+1}; \varsigma_{i,1}^{k+1}). \tag{7}$$

With θ_i^{k+1} and v_i^{k+1} at hand, the local hypergradient is approximated by

$$s_i^{k+1} = \nabla_x \hat{f}_i(x_i^{k+1}, \theta_i^{k+1}; \varsigma_{i,2}^{k+1}) - \nabla_{x\theta}^2 \hat{g}_i(x_i^{k+1}, \theta_i^{k+1}; \xi_{i,3}^{k+1}) v_i^{k+1}. \tag{8}$$

Here, $\nabla_\theta \hat{g}_i$ (resp. $\nabla_{\theta\theta}^2 \hat{g}_i$, $\nabla_\theta \hat{f}_i$, $\nabla_x \hat{f}_i$, $\nabla_{x\theta}^2 \hat{g}_i$) denotes a stochastic gradient estimate of $\nabla_\theta g_i$ (resp. $\nabla_{\theta\theta}^2 g_i$, $\nabla_\theta f_i$, $\nabla_x f_i$, $\nabla_{x\theta}^2 g_i$) depending on the random variable $\xi_{i,1}^{k+1}$ (resp. $\xi_{i,2}^{k+1}$, $\xi_{i,3}^{k+1}$, $\varsigma_{i,1}^{k+1}$, $\varsigma_{i,2}^{k+1}$).

However, using the above stochastic estimators will lead to steady-state errors under Assumptions 1-4 and single-timescale step-sizes (Chen et al., 2021a; Hong et al., 2023), unless an increasing number of batch sizes is used (Arbel and Mairal, 2022) or extra smoothness conditions are imposed on Hv variables (Dagréou et al., 2022a) (refer to Section 4.4 for more details). To address this issue, we further introduce a gradient momentum step to reduce the impact as follows:

$$z_i^{k+1} = s_i^k + (1 - \gamma)(z_i^k - s_i^k), \tag{9}$$

where $0 < \gamma < 1$. The gradient momentum step helps control the order of the sampling variance by maintaining a moving average of the past approximated hypergradients (Chen et al., 2022b; Ghadimi et al., 2020).

Distributed gradient descent/tracking. The update of local copy x_i follows standard distributed gradient method as:

$$x_i^{k+1} = (1 - \tau)x_i^k + \tau\left(\sum_{j \in \mathcal{N}_i} w_{ij} x_j^k - \alpha y_i^k\right), \tag{10}$$

where $\alpha > 0$ denotes the step size, and a relaxation step characterized by a parameter $\tau \in (0, 1)$ is employed to smooth both the consensus and the local gradient processes. Then, two alternative choices of the direction y_i^k are considered:

$$\text{(Local gradient scheme)} \quad y_i^{k+1} = z_i^{k+1}. \tag{11}$$

$$\text{(Gradient tracking scheme)} \quad y_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} y_j^k + z_i^{k+1} - z_i^k. \tag{12}$$

Eq. (11) yields a distributed gradient descent type algorithm where y_i^k estimates the local hypergradient ∇f_i; whereas (12) employs the gradient tracking technique so that y_i^k estimates the average hypergradient $\nabla\Phi$.

The overall LoPA algorithm is summarized in Algorithm 1, where we refer to the versions with local gradient and gradient tracking schemes as LoPA-LG and LoPA-GT, respectively.

Remark 2 (Iterative approximation approach for Hv) *In estimating hypergradients, LoPA takes a different approach from existing distributed algorithms that use Neumann Series (NS) methods and Summed Hessian Inverse Approximation (SHIA) methods to directly*

Algorithm 1 LoPA

1: **Require**: Initialize θ_i^0, v_i^0, x_i^0, s_i^0, z_i^0, y_i^0, $i \in \mathcal{V}$ and set step-sizes $\{\alpha, \beta, \lambda, \gamma, \tau\}$.
2: **for** $k = 0, 1, 2..., K$, each node $i \in \mathcal{V}$ in parallel **do**
3:     Sample batch $\zeta_i^{k+1} = \{\xi_{i,1}^{k+1}, \xi_{i,2}^{k+1}, \xi_{i,3}^{k+1}, \varsigma_{i,1}^{k+1}, \varsigma_{i,2}^{k+1}\}$.
4:     Communicate with neighboring node $j \in \mathcal{N}_i$.
5:     Update state variables θ_i^{k+1}, v_i^{k+1}, x_i^{k+1} according to (4), (5), (10);
6:     Update local gradient estimates d_i^{k+1}, h_i^{k+1}, s_i^{k+1}, z_i^{k+1} according to (6), (7), (8), (9);
7:     Update the descent direction of outer-level variables y_i^{k+1} as follows:
8:     **if** gradient tracking scheme is not used **then**
9:         LoPA-LG: Update y_i^{k+1} according to (11);
10:    **else**
11:        LoPA-GT: Update y_i^{k+1} according to (12).
12:    **end if**
13: **end for**

---

approximate the Hessian-inverse matrices to a high-precision at each iteration k. To be more specific, the key idea of the NS (Ghadimi and Wang, 2018; Hong et al., 2023) and SHIA methods (Dagréou et al., 2022b; Rajeswaran et al., 2019) is to approximate the Hessian-inverse matrices and Hv in multiple iterations, respectively. The approximation process of these two methods can be summarized as follows:

$$\textbf{(NS):}\ Q\lambda \prod_{t=0}^{Q} \left(I - \lambda\nabla_{\theta\theta}^2 g_i\left(x_i, \theta_i\right)\right) \approx \left[\nabla_{\theta\theta}^2 g_i\left(x_i, \theta_i\right)\right]^{-1}, \tag{13}$$

$$\textbf{(SHIA):}\ \lambda\sum_{t=0}^{Q} \left(\prod_{j=0}^{t} \left(I - \lambda\nabla_{\theta\theta}^2 g_i(x_i, \theta_i)\right)\right)\nabla_\theta f_i(x_i, \theta_i) \approx \left[\nabla_{\theta\theta}^2 g_i\left(x_i, \theta_i\right)\right]^{-1}\nabla_\theta f_i(x_i, \theta_i).$$

We can know from the above expressions that the high-precision approximation generally requires a large Q, which leads to extra computation loops at each iteration k. For example, the state-of-the-art works (Lu et al., 2022a,b; Gao et al., 2022) require Q obeying $\mathcal{O}(\log \epsilon^{-1})$. Unlike these methods, LoPA adopts an iterative approximation approach with one stochastic gradient iteration for tracking the states of Hessian-inverse matrices and inner-level solutions. Thus, LoPA enjoys a loopless structure in the algorithmic design and achieves a computational complexity of $\mathcal{O}(1)$ with respect to the number of Hessian evaluations at each iteration k while maintaining the same complexity for outer- and inner-level gradient and Jacobian evaluations at each iteration k.

4. Convergence Results

In this section, we respectively analyze the performance of LoPA-LG and LoPA-GT for nonconvex-strongly-convex cases.

4.1 Preliminaries

We make the following assumption on the stochastic gradients used for estimating the gradients of the outer- and inner-level functions.

Let

$$\mathcal{F}^k = \sigma \left\{ \bigcup_{i=1}^m (x_i^0, \theta_i^0, v_i^0, z_i^0, y_i^0, \cdots, x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k) \right\} \tag{14}$$

be the σ-algebra generated by the random variables up to the k-th iteration.

Assumption 5 (Stochastic gradient estimates) *We assume the random variables $\xi_{i,1}^k$, $\xi_{i,2}^k$, $\xi_{i,3}^k$, $\varsigma_{i,1}^k$, $\varsigma_{i,2}^k$ are mutually independent for any iteration k; and also independent across all the iterations. Furthermore, for any $x \in \mathbb{R}^n$, $\theta \in \mathbb{R}^p$ and $k \geqslant 0$, the followings hold:*
(i) Unbiased estimators:

$\mathbb{E}[\nabla_\theta \hat{g}_i(x,\theta;\xi_{i,1}^k)] = \nabla_\theta g_i(x,\theta), \ \mathbb{E}[\nabla_{\theta\theta}^2 \hat{g}_i(x,\theta;\xi_{i,2}^k)] = \nabla_{\theta\theta}^2 g_i(x,\theta), \ \mathbb{E}[\nabla_{x\theta}^2 \hat{g}_i(x,\theta;\xi_{i,3}^k)] = \nabla_{x\theta}^2 g_i(x,\theta),$
$\mathbb{E}[\nabla_\theta \hat{f}_i(x,\theta;\varsigma_{i,1}^k)] = \nabla_\theta f_i(x,\theta), \ \mathbb{E}[\nabla_x \hat{f}_i(x,\theta;\varsigma_{i,2}^k)] = \nabla_x f_i(x,\theta).$

(ii) Bounded stochastic variances:

$$\mathbb{E}[\|\nabla_\theta g_i(x,\theta;\xi_{i,1}^k) - \nabla_\theta g_i(x,\theta)\|^2] \leqslant \sigma_{g,\theta}^2, \mathbb{E}[\|\nabla_{\theta\theta}^2 \hat{g}_i(x,\theta;\xi_{i,2}^k) - \nabla_{\theta\theta}^2 g_i(x,\theta)\|^2] \leqslant \sigma_{g,\theta\theta}^2,$$

$$\mathbb{E}[\|\nabla_{x\theta}^2 \hat{g}_i(x,\theta;\xi_{i,3}^k) - \nabla_{x\theta}^2 g_i(x,\theta)\|^2] \leqslant \sigma_{g,x\theta}^2, \mathbb{E}[\|\nabla_\theta \hat{f}_i(x,\theta;\varsigma_{i,1}^k) - \nabla_\theta f_i(x,\theta)\|^2] \leqslant \sigma_{f,\theta}^2,$$

$$\mathbb{E}[\|\nabla_x \hat{f}_i(x,\theta;\varsigma_{i,2}^k) - \nabla_x f_i(x,\theta)\|^2] \leqslant \sigma_{f,x}^2.$$

The following two propositions provide the smoothness property of $\nabla \Phi(x)$, $\theta_i^*(x)$, and $v_i^*(x)$, as well as the boundedness of $v_i^*(x)$, with the first proposition being adapted from (Ghadimi and Wang, 2018) and the second being derived from the properties of certain gradients. For completeness, we provide the complete proof in Section C.

Proposition 3 (Smoothness property) *Suppose Assumptions 2 and 3 hold. Let $\bar{\nabla} f_i(x,\theta) \triangleq \nabla_x f(x,\theta) - \nabla_{x\theta}^2 g_i(x,\theta) v_i(x,\theta)$ be a surrogate of the local hypergradient $\nabla f_i(x,\theta_i^*(x))$ and denote $v_i(x,\theta) \triangleq [\nabla_{\theta\theta}^2 g_i(x,\theta)]^{-1} \nabla_\theta f_i(x,\theta), v_i^*(x) \triangleq v_i(x,\theta_i^*(x))$. Then given any $x, x' \in \mathbb{R}^n$, it holds that: $\forall i \in \mathcal{V}$,*

$\|\theta_i^*(x) - \theta_i^*(x')\| \leqslant L_{\theta^*} \|x - x'\|, \|v_i(x,\theta_i^*(x)) - v_i(x',\theta_i^*(x'))\| \leqslant L_v(\|x - x'\| + \|\theta_i^*(x) - \theta_i^*(x')\|),$
$\|v_i^*(x) - v_i^*(x')\| \leqslant L_{v^*} \|x - x'\|, \|\bar{\nabla} f_i(x,\theta_i^*(x)) - \bar{\nabla} f_i(x',\theta_i^*(x'))\| \leqslant L_f(\|x - x'\| + \|\theta_i^*(x) - \theta_i^*(x')\|),$
$\|\nabla \Phi(x) - \nabla \Phi(x')\| \leqslant L \|x - x'\|,$

where the Lipschitz constants are provided as follows:

$$L_{\theta^*} \triangleq \frac{C_{g,x\theta}}{\mu_g} = \mathcal{O}(\kappa), L_v \triangleq \frac{L_{f,\theta}}{\mu_g} + \frac{C_{f,\theta} L_{g,\theta\theta}}{\mu_g^2} = \mathcal{O}(\kappa^2), L_{v^*} \triangleq (\frac{L_{f,\theta}}{\mu_g} + \frac{C_{f,\theta} L_{g,\theta\theta}}{\mu_g^2})(1 + L_{\theta^*}) = \mathcal{O}(\kappa^3),$$

$$L_f \triangleq L_{f,x} + C_{g,x\theta} L_v + \frac{C_{f,\theta} L_{g,x\theta}}{\mu_g} = \mathcal{O}(\kappa^2), L \triangleq (L_{f,x} + C_{g,x\theta} L_v + \frac{C_{f,\theta} L_{g,x\theta}}{\mu_g})(1 + L_{\theta^*}) = \mathcal{O}(\kappa^3), \tag{15}$$

with $\kappa = \frac{\max\{L_{f,x}, L_{f,\theta}, L_{g,\theta}, L_{g,x\theta}, L_{g,\theta\theta}\}}{\mu_g}$ denoting the condition number.

Proposition 4 (Boundness property) *Suppose Assumptions 2 and 3 hold. Then, there exists a constant $M = \frac{C_{f,\theta}}{\mu_g}$ such that the following holds for any x: $\forall i \in \mathcal{V}$,*

$$\|v_i^*(x)\| \leqslant M, \tag{16}$$

where $v_i^*(x) = v_i(x,\theta_i^*(x))$ as defined in Proposition 3.

4.2 Convergence of LoPA-LG and LoPA-GT

Convergence of LoPA-LG. We first analyze LoPA-LG that uses local gradient scheme (11). To derive the convergence results of LoPA-LG, one key step is to explore the heterogeneity on overall hypergradients. The following lemma shows the boundness and composition of the heterogeneity on overall hypergradients.

Lemma 5 (Bounded heterogeneity on overall hypergradients) *Suppose Assumptions 2, 3 and 4 hold. Let $\nabla\Phi_i(x)$ be the local hypergradient of node i evaluated at x. Then, we have*

$$\sum_{i=1}^{m} \|\nabla\Phi_i(x) - \nabla\Phi(x)\|^2 \leqslant b^2, \tag{17}$$

where $b^2 \triangleq C_1(\mu_g, C_{g,x\theta})b_f^2 + C_2(\mu_g, L_{f,x}, L_{f,\theta}, L_{g,x\theta}, L_{g,\theta\theta}, C_{f,\theta}, C_{g,x\theta})b_g^2$ with $C_1(\mu_g, C_{g,x\theta})$ and $C_2(\mu_g, L_{f,x}, L_{f,\theta}, L_{g,x\theta}, L_{g,\theta\theta}, C_{f,\theta}, C_{g,x\theta})$ being the constants defined in Appendix D.1.

Proof: See Appendix D.1. ∎

Note that the heterogeneity of the overall hypergradients consists of two main components: the inner-level heterogeneity b_f^2 and the outer-level heterogeneity b_g^2. To further quantify the effect of heterogeneity in each level, we characterize its dependency on the condition number κ. It is not difficult to show that $C_1(\mu_g, C_{g,x\theta}) = \mathcal{O}(\kappa^2)$ and $C_2(\mu_g, L_{f,x}, L_{f,\theta}, L_{g,x\theta}, L_{g,\theta\theta}, C_{f,\theta}, C_{g,x\theta}) = \mathcal{O}(\kappa^6)$. According to the definition of b^2, we can observe that b^2 is of $\mathcal{O}(\kappa^2 b_f^2 + \kappa^6 b_g^2)$. This observation suggests that the heterogeneity of inner-level objective functions plays a crucial role in determining the heterogeneity of overall hypergradients. Lemma 5 provides a novel characterization on the heterogeneity in personalized DSBO, distinguishing itself from previous works (Chen et al., 2022b, 2021b) where the out-level objective depends on a common inner-level solution and thus the out-level heterogeneity is independent of the inner-level heterogeneity. Now, we are thus ready to present the convergence of LoPA-LG.

Theorem 6 *Suppose Assumptions 1, 2, 3, 4 and 5 hold. Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1 employing local gradient scheme as depicted in (11). Let $\bar{x}^k = (1/m)\sum_{i=1}^{m} x_i^k$, $L_{fg,x} = 2L_{f,x}^2 + 4M^2 L_{g,x\theta}^2$ and $L_{fg,\theta} = 2L_{f,\theta}^2 + 4M^2 L_{g,\theta\theta}^2$ with $M = \frac{C_{f,\theta}}{\mu_g}$. There exists a proper choice of step-sizes $\alpha, \beta, \lambda, \gamma, \tau$ such that, for any total number of iterations K, we have*

$$\frac{1}{K+1}\sum_{k=0}^{K} \mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant \frac{4(V^0 - V^K)}{\tau\alpha(K+1)} + 4\alpha\sigma_{LG}^2 + \frac{4\vartheta}{m}\alpha^2 b^2, \tag{18}$$

where $\sigma_{LG}^2 = (\frac{1}{m}\frac{d_3}{\tau} + \frac{d_4}{\tau})(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2} + 2\frac{d_1}{\tau}(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} + 2\frac{d_2}{\tau}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}$ and $\vartheta = \frac{24\alpha\varphi}{(1-\rho)^2\gamma}$ with $\varphi = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2})$ and $\omega_\theta = \frac{\mu_g L_{g,\theta}}{2(\mu_g + L_{g,\theta})}$. The coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 within the unified Lyapunov function V^k as depicted in (23) are defined as follows: $d_0 = 1$, $d_1 = \frac{8C_{g,x\theta}^2\tau\alpha}{\mu_g\lambda}$, $d_2 = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})\frac{\tau\alpha}{\omega_\theta\beta}$, $d_3 = \frac{\tau\alpha}{2\gamma}$, $d_4 = \frac{24\varphi\tau\alpha^4}{(1-\rho)^2\gamma}$, $d_5 = \frac{2\varphi\alpha}{1-\rho}$, $d_6 = 0$, and the network spectral gap ρ is defined in Assumption 1.

Proof: See Section 4.3 for the proof sketch and supporting lemmas of Theorem 6 and Appendix B.1 for the complete proof. ∎

Corollary 7 *Consider the same setting as Theorem 6. If the step-sizes are properly chosen such that $\alpha = \min\left\{u, \left(\frac{a_0}{a_1(K+1)}\right)^{\frac{1}{2}}, \left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}}\right\} \leqslant \frac{1}{m}$ for a large value of K, where u, a_0, a_1 and a_2 are given in Appendix B.2, and $\gamma = \mathcal{O}(\alpha)$, $\lambda = \mathcal{O}(\alpha)$, $\beta = \mathcal{O}(\alpha)$, then we have[1]*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{\kappa^8}{(1-\rho)K} + \frac{\kappa^{\frac{16}{3}}(\frac{b}{\sqrt{m}})^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}} + \frac{1}{\sqrt{K}}\kappa^{\frac{5}{2}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right). \quad (19)$$

where $\sigma_{\mathrm{p}} = \mathcal{O}(\kappa^{\frac{1}{2}}\sigma_{f,\theta} + \kappa^{\frac{3}{2}}\sigma_{g,\theta\theta} + \kappa^{\frac{5}{2}}\sigma_{g,\theta})$ and $\sigma_{\mathrm{c}} = \mathcal{O}(\sigma_{f,x} + \kappa\sigma_{g,x\theta})$ are the gradient sampling variances associated with the inner-level and the out-level variables, respectively.

Proof: See Appendix B.2. ∎

Remark 8 (Heterogeneity analysis) *Corollary 7 characterizes the detailed impact of the heterogeneity at each level on the convergence rate in DSBO by unique heterogeneity analysis under weaker Assumption 4 on data heterogeneity, which does not require the boundedness of local hypergradients or the boundedness of inner-level heterogeneity at any point (c.f., Assumption 4). The above Corollary 7 also shows that LoPA-LG has a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{K}})$, where K denotes the total number of iterations. In particular, the analysis of LoPA-LG reveals that the data heterogeneity affects the convergence rate by introducing a transient term of $\mathcal{O}(K^{-\frac{2}{3}})$ in DSBO, vanishing at a slower rate than the leading term, in which the inner-level heterogeneity b_g^2 plays a crucial role in personalized DSBO. To the best of our knowledge, these results are first presented in our work.*

Convergence of LoPA-GT. Now, we move on to present the main results for LoPA-GT that employs gradient tracking scheme (12) in the following Theorem 9.

Theorem 9 *Suppose Assumptions 1, 2, 3, and 5 hold. Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1 employing gradient tracking scheme (12). Let $\bar{x}^k = (1/m)\sum_{i=1}^{m} x_i^k$, $L_{fg,x} = 2L_{f,x}^2 + 4M^2 L_{g,x\theta}^2$ and $L_{fg,\theta} = 2L_{f,\theta}^2 + 4M^2 L_{g,\theta\theta}^2$ with $M = \frac{C_{f,\theta}}{\mu_g}$. There exists a proper choice of the step-sizes $\alpha, \beta, \lambda, \gamma, \tau$ such that, for any total number of iterations K, we have*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant \frac{2\left(V^0 - V^K\right)}{\tau\alpha(K+1)} + 2\alpha\sigma_{GT}^2, \quad (20)$$

where $\sigma_{GT}^2 = (\frac{1}{m}\frac{d_3}{\tau} + \frac{d_4}{\tau} + \frac{2d_6}{\tau(1-\rho)})(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2} + 2\frac{d_1}{\tau}(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} + 2\frac{d_2}{\tau}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}$ with $\varphi = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2})$ and $\omega_\theta = \frac{\mu_g L_{g,\theta}}{2(\mu_g+L_{g,\theta})}$. The coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 within the unified Lyapunov function V^k as depicted in (23) are defined as follows: $d_0 = 1$, $d_1 = \frac{8C_{g,x\theta}^2\tau\alpha}{\mu_g\lambda}$, $d_2 = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})\frac{\tau\alpha}{\omega_\theta\beta}$, $d_3 = \frac{\tau\alpha}{2\gamma}$, $d_4 = \frac{64\varphi\tau\gamma\alpha^2}{(1-\rho)^4}$, $d_5 = \frac{4\varphi\alpha}{(1-\rho)}$, $d_6 = \frac{16\varphi\tau\alpha^2}{(1-\rho)^3}$, and the network spectral gap ρ is defined in Assumption 1.

---

[1]The symbol $\mathcal{O}$ hides both the constants and parameters associated with the properties of functions.

Proof: See Section 4.3 for the proof sketch and supporting lemmas of Theorem 9 and Appendix B.3 for the complete proof. ∎

Corollary 10 *Consider the same setting as Theorem 9. If the step-sizes are properly chosen such that $\alpha = \min\left\{u', \left(\frac{a_0'}{a_1'(K+1)}\right)^{\frac{1}{2}}\right\} \leqslant \frac{1}{m}$ for a large value of K, where u', a_0' and a_1' are given in Appendix B.4, and $\gamma = \mathcal{O}(\alpha)$, $\lambda = \mathcal{O}(\alpha)$, $\beta = \mathcal{O}(\alpha)$, then we have*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\Big(\frac{\kappa^8}{(1-\rho)^2 K} + \frac{\kappa^{\frac{5}{2}}}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\Big), \qquad (21)$$

where $\sigma_{\mathrm{p}} = \mathcal{O}(\kappa^{\frac{1}{2}}\sigma_{f,\theta} + \kappa^{\frac{3}{2}}\sigma_{g,\theta\theta} + \kappa^{\frac{5}{2}}\sigma_{g,\theta})$ and $\sigma_{\mathrm{c}} = \mathcal{O}(\sigma_{f,x} + \kappa\sigma_{g,x\theta})$.

Proof: See Appendix B.4. ∎

Remark 11 (Linear speed-up) *Corollaries 7 and 10 demonstrate that, in the context of DSBO problems with personalized inner-level objectives, the term σ_{p} associated with the inner-level variable θ does not decay with m. In contrast, the term σ_{c} induced by the stochastic gradients with respect to the out-level variable x decays at order $O(\frac{1}{\sqrt{mK}})$. This lack of linear speed-up in σ_{p} is due to the fact that the inner-level solutions $\theta_i^*(x)$ are different for each node i such that neither the local inner-level variables or the local Hv variables do not need to reach a consensus in personalized DSBO, and thus the related variances can only be reduced via temporal averaging, yielding an order of $O(\frac{1}{\sqrt{K}})$. In contrast, the outer-level variable x is shared across the nodes and can achieve a speed up with respect to the network size m. Our result also implies that to achieve the right balance of the two terms, the batch size should be increased by m times when estimating the local partial gradients $\nabla_\theta\hat{f}_i$, $\nabla_\theta\hat{g}_i$, $\nabla_{\theta\theta}\hat{g}_i$. In this case, both LoPA-LG and LoPA-GT can obtain a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{mK}})$, leading to an iteration complexity of $\mathcal{O}(m^{-1}\epsilon^{-2})$.*

Remark 12 (Improved complexity) *Corollary 10 shows that LoPA-GT has a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{K}})$. It is also noted from the above result that, the effect of the heterogeneity is eliminated by gradient tracking scheme (12) in personalized DSBO even when the inner-level variables lack communication. Thanks to the loopless structure, both LoPA-LG and LoPA-GT can achieve a computational complexity of $\mathcal{O}(\epsilon^{-2})$ (w.r.t. the number of Hessian evaluations) to attain an ϵ-stationary point. This computational complexity improves existing state-of-the-art works on DSBO problems by the order of $\mathcal{O}(\log(\epsilon^{-1}))$. Note that the computational complexity for inner- and outer-level gradient and Jacobian evaluations is also of the order $\mathcal{O}(\epsilon^{-2})$ for LoPA-LG and LoPA-GT. Our analysis for both LoPA-LG and LoPA-GT provides tighter rates that clearly show the dependence of convergence on the condition number, heterogeneity at each level, sampling variances at each level, and network connectivity, in which the detailed effects of sampling variances (c.f., (19) and (21)) enable us to improve the rate by properly optimizing batch sizes. In particular, when the batch sizes of the gradient evaluations $\nabla_\theta\hat{f}$, $\nabla_\theta\hat{g}$, $\nabla_{\theta\theta}^2\hat{g}$, and $\nabla_{x\theta}^2\hat{g}$ are respectively chosen as $\mathcal{O}(m\kappa)$, $\mathcal{O}(m\kappa^5)$, $\mathcal{O}(m\kappa^3)$, and $\mathcal{O}(\kappa^2)$ times relative to $\nabla_x\hat{f}$, an improved iteration rate of $\mathcal{O}(\frac{\kappa^{\frac{5}{2}}}{\sqrt{mK}})$ can be obtained, which matches the result in centralized counterparts (Chen et al., 2021a). In*

this case, a computational complexity of $\mathcal{O}(\kappa^5 m^{-1}\epsilon^{-2})$ w.r.t. out-level gradient evaluations (also $\mathcal{O}(\kappa^8 \epsilon^{-2})$ w.r.t. Hessian evaluations[2]) can be achieved for both LoPA-LG and LoPA-GT due to the loopless structure, which are the best-known result among existing works on DSBO without mean-square smoothness assumption.

Dependency on network connectivity. Note that it follows from the results (18), (20) and the corresponding definitions of $d_0, \cdots, d_6$ in (51) and (89) that Theorems 6 and 9 have respective dependencies of $\mathcal{O}(\frac{1}{(1-\rho)^2})$ and $\mathcal{O}(\frac{1}{(1-\rho)^4})$ on network connectivity without proper choice of step-sizes. These results of dependency are aligned with that of existing single-level distributed methods (Lian et al., 2017; Alghunaim and Yuan, 2022) using LG schemes and GT schemes. By optimizing the step-size selection, we can obtain a tighter dependence of $\mathcal{O}(\frac{1}{1-\rho})$ and $\mathcal{O}(\frac{1}{(1-\rho)^2})$ for LoPA-LG and LoPA-GT (c.f., Corollary 7 and Corollary 10), respectively, where the dependence of $\mathcal{O}(\frac{1}{1-\rho})$ for LG schemes is superior to $\mathcal{O}(\frac{1}{(1-\rho)^2})$ provided in the work (Yang et al., 2022). Notably, our result also reveals that LoPA-GT has a worse dependence on the network spectral gap compared to LoPA-LG, due to the introduction of GT schemes to eliminate heterogeneity.

Convergence of LoPA-LG and LoPA-GT in deterministic case. Next, we present the convergence results in Corollaries 13 and 14 for special cases where the inner- and outer-level functions in problem (1) are deterministic. These results can be directly derived by following a similar proof to that of Corollaries 7 and 10, with the sampling variances set to zero.

Corollary 13 *Consider the deterministic case for Problem (1) under the same setting as Theorem 6 for LoPA-LG. Suppose Assumptions 1-4 hold. If the step-sizes are properly chosen such that $\alpha = \min\left\{u, \left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}}\right\}$ for a large value of K, where u, a_0 and a_2 are the same as that of Corollary 7, and $\gamma = \mathcal{O}(\alpha)$, $\lambda = \mathcal{O}(\alpha)$, $\beta = \mathcal{O}(\alpha)$, then we have*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{\kappa^8}{(1-\rho)K} + \frac{\kappa^{\frac{16}{3}}\left(\frac{b}{\sqrt{m}}\right)^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}}\right).$$

Corollary 14 *Consider the deterministic case for Problem (1) under the same setting as Theorem 9 for LoPA-GT. Suppose Assumptions 1-3 hold. If the step-sizes are chosen such that $\alpha \leqslant u'$, where u' is the same as that of Corollary 10, and $\gamma = \mathcal{O}(\alpha)$, $\lambda = \mathcal{O}(\alpha)$, $\beta = \mathcal{O}(\alpha)$, then we have*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{\kappa^8}{(1-\rho)^2 K}\right).$$

From Corollaries 13 and 14, we observe that the convergence rates of LoPA-LG and LoPA-GT for the deterministic case can be improved to $\mathcal{O}(\frac{1}{K^{2/3}})$ and $\mathcal{O}(\frac{1}{K})$, respectively.

---

[2]Here, the absence of linear speed-up of Hessian complexity is due to the fact that, different from global DSBO problems, the Hessian-inverse-vector product variables do not need to reach a consensus in personalized DSBO.

4.3 Proof Sketch and Supporting Lemmas for Theorems 6 and 9

Before presenting the proof sketch, we first introduce some essential notations as follows:

$$
\begin{aligned}
v_i(x,\theta) &\triangleq \left[\nabla^2_{\theta\theta} g_i(x,\theta)\right]^{-1} \nabla_\theta f_i(x,\theta), \; v_i^*(x) \triangleq v_i(x,\theta_i^*(x)), \\
\bar{x}^k &\triangleq (1/m)\sum\nolimits_{i=1}^m x_i^k, \; x^k \triangleq \mathrm{col}\{x_i^k\}_{i=1}^m, \; \theta^*(\bar{x}^k) \triangleq \mathrm{col}\{\theta_i^*(\bar{x}^k)\}_{i=1}^m, \\
v^*(\bar{x}^k) &\triangleq \mathrm{col}\{v_i^*(\bar{x}^k)\}_{i=1}^m, \nabla\tilde{\Phi}(\bar{x}^k) = \mathrm{col}\{\nabla\Phi_i(\bar{x}^k)\}_{i=1}^m.
\end{aligned}
\tag{22}
$$

The notations z^k, y^k, θ^k, s^k, $\bar{y}^k$, $\bar{z}^k$, $\bar{s}^k$ share the similar definitions. The key idea of the proof for Theorems 6 and 9 is to characterize the dynamics of the following unified Lyapunov function with properly selected coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6

$$
\begin{aligned}
V^k = {} & d_0\Phi(\bar{x}^k) + d_1\frac{1}{m}\underbrace{\left\|v^k - v^*(\bar{x}^k)\right\|^2}_{\text{Hv errors}} + d_2\frac{1}{m}\underbrace{\left\|\theta^k - \theta^*(\bar{x}^k)\right\|^2}_{\text{inner-level errors}} + d_3\underbrace{\left\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\right\|^2}_{\text{ave-variance errors}} \\
& + d_4\frac{1}{m}\underbrace{\left\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\right\|^2}_{\text{variance errors}} + d_5\frac{1}{m}\underbrace{\left\|x^k - 1_m \otimes \bar{x}^k\right\|^2}_{\text{consensus errors}} + d_6\frac{1}{m}\underbrace{\left\|y^k - 1_m \otimes \bar{y}^k\right\|^2}_{\text{gradient errors}},
\end{aligned}
\tag{23}
$$

where the detailed definitions of these coefficients can be founded in (51) and (89), corresponding to Theorems 6 and 9, respectively. To this end, we proceed to derive iterative evolution for each term of V^k in expectation according to the following four key steps:

Step 1 (Quantifying the descent of the overall objective function): We begin by quantifying the descent of the overall objective function $\Phi(\bar{x}^k)$ evaluated at the average point by using its smoothness and the tracking property of $\bar{y}^k$ for $\bar{z}^k$. This descent is controlled by the hypergradient approximation errors $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2]$ in Lemma 15.

Lemma 15 (Descent lemma) *Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1. Suppose Assumptions 1, 2, 3 and 5 hold. Then, we have*

$$
\begin{aligned}
\mathbb{E}[\Phi(\bar{x}^{k+1})] \leqslant {} & \mathbb{E}[\Phi(\bar{x}^k)] - \frac{\tau\alpha}{2}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] - \frac{\tau\alpha}{2}(1 - \tau\alpha L)\mathbb{E}[\|\bar{y}^k\|^2] \\
& + \frac{\tau\alpha}{2}\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2].
\end{aligned}
\tag{24}
$$

Proof: See Appendix D.2. ∎

Step 2 (Characterizing the average variance errors and hypergradient errors): We then deal with the average variance errors $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2]$ according to the bounded variances of different stochastic gradients and the updates of z^k. However, bounding the hypergradient errors $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2]$ is more challenging as it requires an investigation into how the iterative approximation strategies with one stochastic gradient iteration influence the evolution of Hv errors and inner-level errors. Lemma 16 shows that the Hv product errors $\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2]$, inner-level errors $\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2]$ and the consensus errors $\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2]$ jointly control the hypergradient approximation errors $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k]\|^2]$, while the term $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k]\|^2]$ controls the average variance errors $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2]$. In what follows, we focus on quantifying these four error terms and establishing their recursions in Lemmas 17, 18, 19.

Lemma 16 (Hypergradient approximation errors and average variance errors)
Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1. Suppose Assumptions 1, 2, 3 and 5 hold. If the step-size γ satisfies $0 < \gamma < 1$, then we have

$$
\mathbb{E}[\|\nabla\Phi(\bar{x}^{k+1}) - \bar{z}^{k+1}\|^2] \leqslant (1-\gamma)\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2] + r_z\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k]\|^2] \\
+ r_y\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + \frac{1}{m}\sigma_{\bar{z}}^2\alpha^2,
\tag{25}
$$

and

$$
\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k]\|^2] \leqslant \frac{L_{fg,x}}{m}\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + \frac{L_{fg,x}}{m}\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
+ \frac{4C_{g,x\theta}^2}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2].
\tag{26}
$$

where $r_z \triangleq \frac{2\gamma}{\alpha}$, $r_y \triangleq \frac{2L^2}{\gamma}$, and $\sigma_{\bar{z}}^2 \triangleq (\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}$ with $\hat{M}^2 \triangleq 2M^2 + 2\frac{1}{m}\|v^k - v^*(\bar{x}^k)\|^2$ and $L_{fg,x} = 2L_{f,x}^2 + 4M^2L_{g,x\theta}^2$.

    Proof: See Appendix D.3. ■

Lemma 17 (Hessian-inverse-vector product errors) *Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1. Suppose Assumptions 1, 2, 3 and 5 hold. If the step-size λ satisfies*

$$
\lambda < \frac{1}{\mu_g},
\tag{27}
$$

then we have

$$
\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2] \leqslant (1-\mu_g\lambda)\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2] + q_x\alpha\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] \\
+ q_x\alpha\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] + mq_s\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + m\sigma_v^2\alpha^2,
\tag{28}
$$

where $q_x \triangleq \frac{4L_{fg,\theta}\lambda}{\mu_g\alpha}$, $\sigma_v^2 \triangleq 2(\sigma_{f,\theta}^2 + \hat{M}^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2}$, $q_s \triangleq \frac{2L_{v*}^2}{\varpi\lambda}$ with $L_{fg,\theta} = 2L_{f,\theta}^2 + 4M^2L_{g,\theta\theta}^2$ and $\varpi = \frac{\mu_g}{3}$.

    Proof: See Appendix D.4. ■

Lemma 18 (Inner-level errors) *Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1. Suppose Assumptions 1, 2, 3 and 5 hold. If the step-size β satisfies*

$$
\beta < \min\left\{\frac{2}{\mu_g + L_{g,\theta}}, \frac{\mu_g + L_{g,\theta}}{2\mu_g L_{g,\theta}}\right\},
\tag{29}
$$

then we have

$$
\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2] \leqslant (1 - \frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}}\beta)\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
+ p_x\alpha\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + mp_s\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + m\sigma_\theta^2\alpha^2,
\tag{30}
$$

where $p_x \triangleq \frac{4L_{g,\theta}^2\beta}{\omega_\theta\alpha}$, $\sigma_\theta^2 \triangleq 2\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}$, $p_s \triangleq \frac{2L_{\theta*}^2}{\omega_\theta\beta}$ with $\omega_\theta = \frac{\mu_g L_{g,\theta}}{2(\mu_g + L_{g,\theta})}$.

Proof: See Appendix D.5. ∎

Lemma 19 (Consensus errors) *Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1. Suppose Assumptions 1, 2, 3 and 5 hold. Then, we have*

$$\mathbb{E}[\|x^{k+1} - 1_m \otimes \bar{x}^{k+1}\|^2] \leqslant (1 - \tau\frac{1-\rho}{2})\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + \frac{2\tau\alpha^2}{1-\rho}\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2], \tag{31}$$

where $0 < \tau < 1$, $\rho = \|\mathcal{W} - \mathcal{J}\|^2 \in [0, 1)$ with $\mathcal{W} \triangleq W \otimes I_n$ and $\mathcal{J} \triangleq \frac{1_m 1_m^{\mathrm{T}}}{m} \otimes I_n$.

Proof: See Appendix D.6. ∎

Step 3 (Characterizing the gradient errors): The next step is to upperbound the gradient error parts $\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2]$ in consensus errors caused by the data heterogeneity across nodes. Hence, we respectively demonstrate how the gradient errors change in LoPA-LG with $y^{k+1} = z^{k+1}$ and LoPA-GT with $y^{k+1} = \mathcal{W}y^k + z^{k+1} - z^k$ in Lemma 20 and Lemma 21. In particular, it is shown that the gradient errors are impacted by the heterogeneity for LoPA-LG with local gradient scheme (11), whereas these error terms will decay as the iteration progresses for LoPA-GT with gradient-tracking scheme (12). Lemma 22 further provides the evolution for the variance errors $\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2]$ induced by the gradient errors $\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2]$.

Lemma 20 (Gradient errors for LoPA-LG) *Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1 employing the local gradient scheme (11). Suppose Assumptions 1, 2, 3, 4 and 5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2] \\
&\leqslant 3b^2 + 3m\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] + 3\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2],
\end{aligned} \tag{32}$$

where $\nabla\tilde{\Phi}(\bar{x}^k) = \mathrm{col}\{\nabla\Phi_i(\bar{x}^k)\}_{i=1}^m$, and b^2 is the heterogeneity on overall hypergradients denoted in Lemma 5.

Proof: See Appendix D.7. ∎

Lemma 21 (Gradient errors for LoPA-GT) *Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1 employing the gradient tracking scheme (12). Suppose Assumptions 1, 2, 3 and 5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2] \\
&\leqslant \frac{1+\rho}{2}\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2] + \frac{4}{1-\rho}\gamma^2\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - \mathbb{E}[s^k]\|^2] \\
&\quad + \frac{4}{1-\rho}\gamma^2\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2] + \frac{2}{1-\rho}m\sigma_y^2\alpha^2,
\end{aligned} \tag{33}$$

where $\sigma_y^2 \triangleq (\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}$.

Proof: See Appendix D.8. ∎

18

Lemma 22 (Variance errors) *Consider the sequence $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ generated by Algorithm 1. Suppose Assumptions 1, 2, 3 and 5 hold. Recall that $\nabla\tilde{\Phi}(\bar{x}^k) = \text{col}\{\nabla\Phi_i(\bar{x}^k)\}_{i=1}^m$. Then, we have*

$$
\begin{aligned}
\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^{k+1}) - z^{k+1}\|^2] \leqslant &(1-\gamma)\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2] + r_z\alpha\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - \mathbb{E}[s^k]\|^2] \\
&+ mr_y\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + m\sigma_z^2\alpha^2,
\end{aligned}
\tag{34}
$$

and

$$
\begin{aligned}
\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - \mathbb{E}[s^k]\|^2] \leqslant &L_{fg,x}\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + L_{fg,x}\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
&+ 4C_{g,x\theta}^2\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2].
\end{aligned}
\tag{35}
$$

where $\sigma_z^2 \triangleq (\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}$.

Proof: The proof of Lemma 22 is similar to Lemma 16 and we omit it here. ∎

Step 4 (Integrating Steps 1, 2, 3 to obtain the overall dynamics with the unified Lyapunov function): Finally, by integrating the obtained results and employing small-gain-like techniques, we establish the dynamics of V^k in Sections B.1 and Section B.3 for LoPA-LG and LoPA-GT, respectively, using a set of carefully chosen coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6.

4.4 Further Discussions on Convergence Analysis

Unique heterogeneity analysis and weaker assumption on heterogeneity. We adopt unique heterogeneity analysis approaches to investigate the impact of heterogeneity. Unlike single-level optimization and global DSBO, personalized DSBO involves distinct inner-level solutions across nodes, making the overall heterogeneity depend intricately on both outer-level and inner-level heterogeneity. This results in a fundamentally different and more challenging scenario for heterogeneity analysis. Moreover, we adopt a weaker Assumption 4 on heterogeneity than that of existing works on DSBO, where our assumption requires only the boundedness of inner-level heterogeneity at the optimum $\theta_i^*(x)$ (instead of being at any point), while avoiding the boundedness of local hypergradients (c.f., Remark 1). This weaker assumption on heterogeneity, together with the different inner-level solutions, presents a unique challenge in explicitly characterizing the heterogeneity in personalized DSBO, as it requires new techniques to analyze the heterogeneity of the inner-level solutions. To tackle these challenges, based on the structure of the hypergradient (c.f., Assumption 4 (ii) and (iii)), we introduce the heterogeneity of the Hessian and Jacobian matrices to characterize the inner-level heterogeneity in DSBO. Furthermore, by establishing the intricate relationships between inner-level and outer-level heterogeneity, we explicitly characterize the overall heterogeneity (c.f., Lemma 5) that clearly shows the impact of inner- and outer-level heterogeneity and provides new insights into heterogeneity in personalized DSBO (cf. Remark 8).

Different hypergradient estimation and unified analytical framework. Our analysis for hypergradient approximation errors differs from the existing literature (Chen et al., 2022b; Lu et al., 2022b; Gao et al., 2022; Lu et al., 2022a; Yang et al., 2022) due to the additional errors from loopless approximation and network dynamic (c.f., Lemma 16). It poses an additional challenge for technique analysis in controlling the variances and analyzing heterogeneity. In

contrast, previous works (Ghadimi and Wang, 2018; Yang et al., 2021; Lu et al., 2022a) utilize NS and HSIA methods (13) along with extra loops with increasing number of iterations Q to obtain highly accurate hypergradient approximations $\tilde{\nabla}\Phi(\bar{x}^k)$ (corresponding to the term z^k in (24)), i.e., $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \tilde{\nabla}\Phi(\bar{x}^k)\|^2] \leqslant \frac{C_{g,x\theta}C_{f,\theta}}{\mu_g}(1 - \frac{\mu_g}{L_{g,\theta}})^Q$. Besides, we adopt a different technique to asymptotically control the bound of $\|v_i^k\|$ without requiring the boundedness of the stochastic gradients or the boundedness of $\|v_i^k\|$ (Gao et al., 2022; Yang et al., 2022) (c.f., Lemmas 16, 17). On the other hand, we also establish a unified analytical framework for both LoPA-LG and LopA-GT with tighter convergence rates. In particular, we design a unified Lyapunov function and employ small-gain-like techniques to establish a overall error evolution dynamic for both LoPA-LG and LoPA-GT. Besides, leveraging unique heterogeneity analysis technique, we are able to bound the gradient tracking errors via variance-related errors (c.f., Lemma 22) for LoPA-LG and LoPA-GT in a unified manner (c.f. Lemma 5, 20-22), while characterizing the effect of the heterogeneity. This above analysis, indeed, provides an unified analytical framework that enables us to establish intricate relationships among the errors as well as their detailed effects on the convergence performance (c.f., Theorems 6, 9), leading to tighter convergence rates (c.f., Corollaries 7, 10) and improved complexity (c.f., Remark 12). The above analysis and obtained results distinguish our work significantly from existing works on DSBO.

The necessity of gradient momentum steps. The gradient momentum step (9) is introduced to control stochastic variances in hypergradient estimation by leveraging the moving average of historical hypergradient information and selecting proper step-size γ. This step allows us to bound only the term $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k]\|^2]$ instead of $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2]$ in Lemma 16, thereby ensuring the convergence of LoPA. In particular, let us consider the case where we exclude the recursion (9), i.e., $\gamma = 1$ and $z^k = s^k$. In this scenario, the average variance error term $\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2]$ reduces to the term $\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2]$, which can only be bounded by a sampling variance of order $\mathcal{O}(\alpha)$. As a result, this case either leads to non-decaying variance errors or requires the proper choice of two-timescale step-sizes to control the decay of the term $\mathbb{E}[|\bar{y}^k|^2]$ in (28) and (30), where the later will lead to a sub-optimal rate of $\mathcal{O}(K^{-2/5})$ (Hong et al., 2023). Instead, with the recursion (9), when the step-size γ is properly taken as $\gamma = \mathcal{O}(\alpha)$, the term $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2]$ can be asymptotically controlled by the term $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k]\|^2]$ and the sampling variance of an order of $\mathcal{O}(\alpha^2)$ (c.f., the last term in Lemma 16). Similar techniques have been employed in (Chen et al., 2022b; Ghadimi et al., 2020; Chen et al., 2024). There are other kinds of gradient momentum methods based on double gradient evaluations (Cutkosky and Orabona, 2019; Fang et al., 2018) or unbiased averaging stochastic gradients (Defazio et al., 2014) that can help achieve a faster rate under some extra conditions (Gao et al., 2022; Dagréou et al., 2022a).

5. Numerical Experiments

In this section, we present two numerical experiments to test the performance of the proposed LoPA algorithm and verify the theoretical findings. Specifically, the first experiment considers classification problems for a 10-class classification task, while the second focuses on hyperparameter optimization in l_2-regularized binary logistic regression problems for a two-class classification task.

5.1 Distributed Classification

In this subsection, we evaluate the effectiveness of our algorithms on a class of distributed classification problems involving heterogeneous datasets. We employ MNIST datasets to train m personalized classifiers for a 10-class classification task. Specifically, we construct a classifier in node i that consists of a hidden layer with parameters x shared across all nodes followed by sigmoid activation functions and a linear layer with parameters θ_i adapted to node-specific samples. The cross-entropy loss is used as the outer-level objective f_i. Regarding the inner-level objective function g_i, we include a quadratic regularization term to the parameters θ_i based on the cross-entropy loss to avoid overfitting to local samples. Specifically, f_i and g_i take the following form

$$
\begin{aligned}
\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} f_i(x, \theta_i^*(x)) &= \frac{1}{m} \sum_{i=1}^{m} \sum_{(s_{ij}, b_{ij}) \in \mathcal{D}_i} b_{ij} \ln\left(y_j(x, \theta_i^*(x); s_{ij})\right), \\
\text{s.t. } \theta_i^*(x) &= arg\min_{\theta_i \in \mathbb{R}^p} g_i(x, \theta_i) = \sum_{(s_{ij}, b_{ij}) \in \mathcal{D}_i} b_{ij} \ln\left(y_j(x, \theta_i; s_{ij})\right) + \frac{\mu}{2} \|\theta_i\|^2,
\end{aligned}
\tag{36}
$$

where $(s_{ij}, b_{ij}) \in \mathcal{D}_i$ denotes the j-th sample assigned to the node i with s_{ij} being the j-th feature and b_{ij} being j-th label in one-shot form; μ denotes the penalty coefficient; $y_j(x, \theta; s_{ij})$ represents the output of the neural network parameterized by the layer weight x and θ.

(a) Loss. (b) Training accuracy. (c) Testing accuracy.
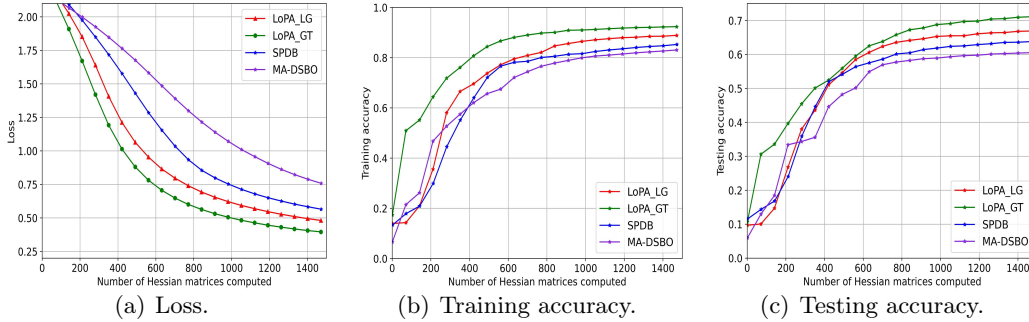
Figure 1: Performance comparison of SPDB, MA-DSBO and our LoPA-LG and LoPA-GT algorithms over 4 nodes for a 10-class classification task using MNIST dataset.

The experiments on the 10-class classification task are conducted under two distinct scenarios: i) $m = 4$ with each node having 14000 samples; and ii) $m = 8$ with each node 6500 samples, where communication networks are generated by random Erdős–Rényi graphs, and $\mu = 0.52$. Each node i is assigned with a random subset of overall 10 classes such that each node has different label distributions and high data heterogeneity. For the classifier in each node, the input is a 784-dimensional vector. The first layer of the classifier contains 28 neurons, while the second layer contains 10 neurons. The step-sizes are set as $\alpha = 0.01$, $\beta = 0.01$, $\lambda = 0.008$, $\gamma = 0.4$, $\tau = 0.4$ both for LoPA-LG and LoPA-GT. We compare our algorithms with the state-of-the-art SPDB (Lu et al., 2022a) and MA-DSBO (Chen et al., 2022b) algorithms, where the mini-batch sizes are set to 50 for all algorithms. These two algorithms respectively utilize the NS and SHIA methods for estimating Hessian inverse matrices. To adapt the MA-DSBO algorithm for personalized settings, we modify the
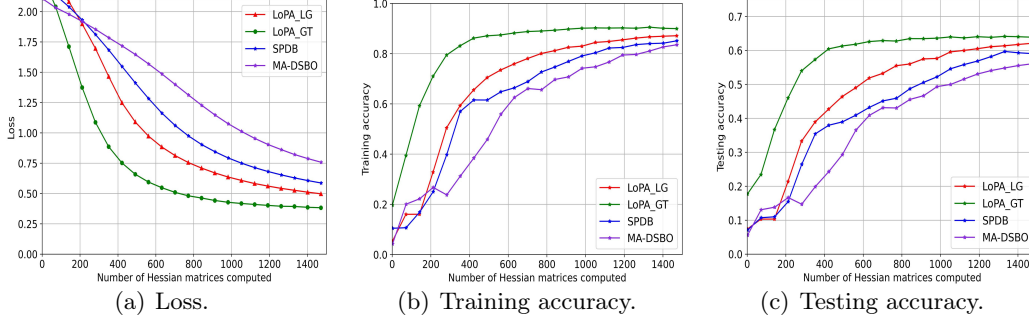
(a) Loss.   (b) Training accuracy.   (c) Testing accuracy.

Figure 2: Performance comparison of SPDB, MA-DSBO and our LoPA-LG and LoPA-GT algorithms over 8 nodes for a 10-class classification task using MNIST dataset.

updates of inner-level variables and Hessian-inverse-vector variables in MA-DSBO so that they exclusively perform local gradient updates.

The experiment results for loss, training accuracy, and testing accuracy are presented in Figures 1 and 2. It can be observed from Figure 1 that the proposed LoPA-LG and LoPA-GT algorithms have lower computational complexity than SPDB and MA-DSBO in terms of the number of Hessian matrices to achieve the same desired accuracy. When the number of nodes increases to 8, we can observe similar results as shown in Figure 1. This further demonstrates the scalability of the proposed algorithms. Moreover, the difference between LoPA-LG and LoPA-GT highlights the advantages of gradient tracking.

Impact of heterogeneity. We conduct an addtional experiment to evaluate the performance of our proposed algorithms under different settings of heterogeneous label distributions. Specifically, we consider a network of 8 nodes with each holding 7500 samples, and generate the following three different label distributions among nodes: i) independent and identically distributed (IID) datasets; ii) non-IID datasets with strong heterogeneity; iii) non-IID dataset with weak heterogeneity. The considered three label distributions are depicted in Figure 4.

We plot the testing accuracy of LoPA-LG and LoPA-GT with the same parameter settings as the previous experiments in Figure 3. The experiment results, as shown in Figure 3, demonstrate that LoPA-GT can maintain a relatively higher accuracy compared to LoPA-LG as the level of data heterogeneity increases. This suggests that LoPA-GT is more robust against data heterogeneity, verifying the theoretical results and the effectiveness of LoPA-GT in scenarios with heterogeneous datasets.
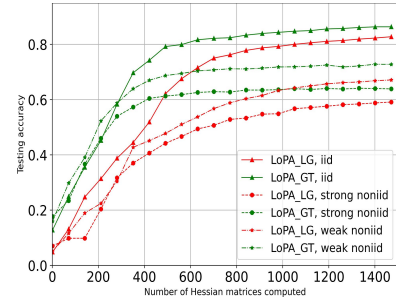
Figure 3: Testing accuracy of LoPA-LG and LoPA-GT under different data heterogeneity for a 10-class classification task using MNIST dataset.

5.2 Hyperparameter Optimization

In this subsection, we delve into hyperparameter optimization in l_2-regularized binary logistic regression problems for a two-class classification scenario. The hyperparameter optimization

(a) IID dataset.

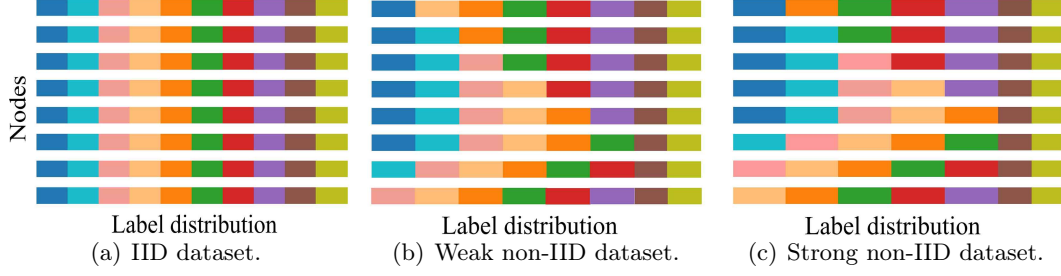(b) Weak non-IID dataset.

(c) Strong non-IID dataset.

Figure 4: Synthetic label distributions with different levels of data heterogeneity across nodes. The label classes are represented with different colors.
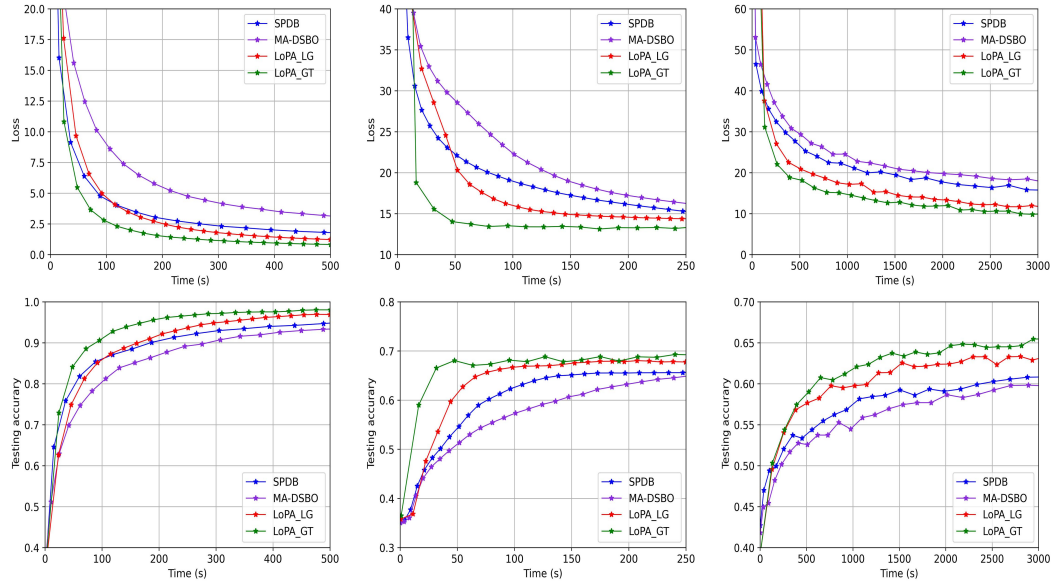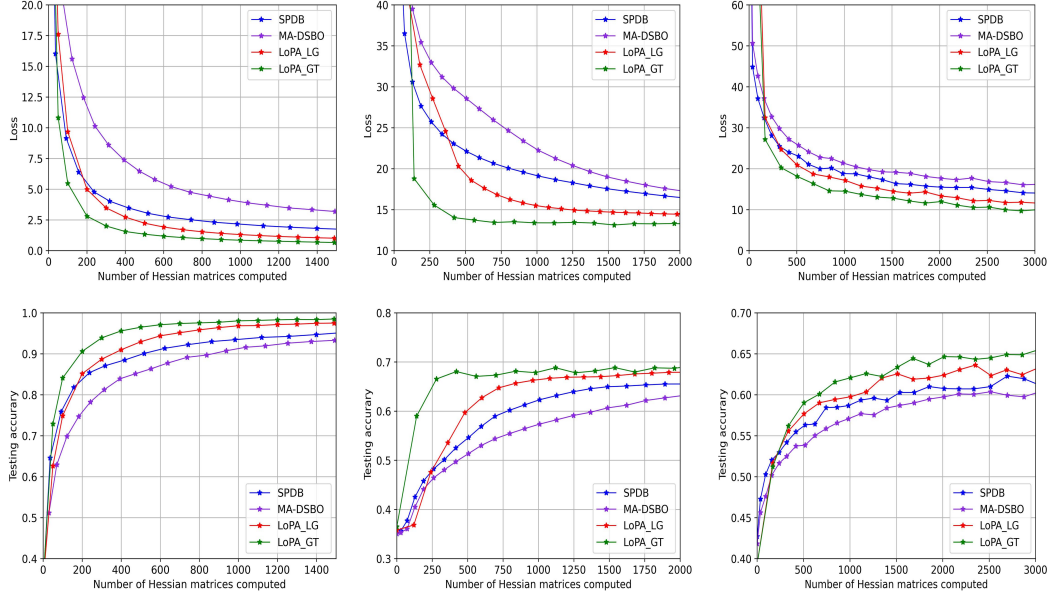

Figure 5: Performance comparison of SPDB, MA-DSBO and our LoPA-LG and LoPA-GT algorithms w.r.t. the computational time for hyperparameter optimization on binary logistic regression problems under different datasets: i) MNIST (first column); ii) covtype (second column); iii) cifar10 (third column).

problem can be framed as a class of bilevel optimization. At the inner level, the goal is to minimize the logistic regression loss on the training data, taking into account the l_2-regularization term for a given hyperparameter. Simultaneously, at the outer level, the objective is to maximize the performance of the l_2-regularized logistic regression model on a validation set by optimizing the hyperparameter. In this experiment, we consider the case that each node i has local validation and training sets, i.e., $\mathcal{D}_i^{\text{val}}$ and $\mathcal{D}_i^{\text{train}}$. The goal of nodes is to cooperatively determine an optimal regularization strength λ that can enhance

Figure 6: Performance comparison of SPDB, MA-DSBO and our LoPA-LG and LoPA-GT algorithms w.r.t. the number of Hessian matrices computed for hyperparameter optimization on binary logistic regression problems under different datasets: i) MNIST (first column); ii) covtype (second column); iii) cifar10 (third column).

overall performance, while optimizing their personalized model parameter, i.e.,

$$
\begin{aligned}
&\min_{\lambda \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} f_i(\lambda, \theta_i^*(\lambda)) = \frac{1}{m} \sum_{i=1}^{m} \sum_{(s_{ij}, b_{ij}) \in \mathcal{D}_i^{\mathrm{val}}} \log(1 + e^{-(b_{ij} s_{ij}^{\mathrm{T}} \theta_i^*(\lambda))}) \\
&\text{s.\,t.}\ \ \theta_i^*(\lambda) = \arg \min_{\theta_i \in \mathbb{R}^p} g_i(\lambda, \theta_i) = \sum_{(s_{ij}, b_{ij}) \in \mathcal{D}_i^{\mathrm{train}}} \log(1 + e^{-(b_{ij} s_{ij}^{\mathrm{T}} \theta_i)}) + \theta_i^{\mathrm{T}} \mathrm{diag}\{e^\lambda\} \theta_i.
\end{aligned}
\tag{37}
$$

where $e^\lambda = \mathrm{col}\{e^{\lambda_t}\}_{t=1}^p$ with $p = n$, $\theta_i^*(\lambda)$ is the optimal model in node i given the hyperparameter λ, and (s_{ij}, b_{ij}) represents the j-th sample in node i with $s_{ij} \in \mathbb{R}^p$ being the feature and $b_{ij} \in \mathbb{R}$ being the corresponding label.

Firstly, we conduct the experiment across various datasets including MNIST (784 features, 12000 samples for digits '0' and '1'), covtype (54 features, 90000 samples for the 'Lodgepole' and 'Ponderosa' pine classes), and cifar10 (3072 features, 6000 samples for the 'dog' and 'horse' classes). The experiment is implemented in a connected network with $m = 10$, where the validation and training sets for each node are randomly assigned with a uniform number of samples. To further validate the efficacy of our proposed algorithms, we comprehensively compare them with state-of-the-art SPDB (Lu et al., 2022a) and MA-DSBO (Chen et al., 2022b) algorithms in terms of computational time and the number of Hessian estimates. The mini-batch sizes are set to 40 for MNIST, 200 for covtype, and 15 for cifar10 across all algorithms. The step-sizes in both the proposed algorithms and the baseline algorithms are manually adjusted for optimal performance. The results are shown in Figures 5 and 6. It can be seen from Figure 5 that the proposed algorithms achieve a certain training and loss accuracy with fewer number of Hessian matrices compared to the SPDB (Lu et al.,

2022a) and MA-DSBO (Chen et al., 2022b) algorithms. Besides, the proposed algorithms demonstrate a reduced time requirement, as depicted in Figure 6. These results further verify the superiority of the proposed algorithms in terms of computational complexity and running efficiency.
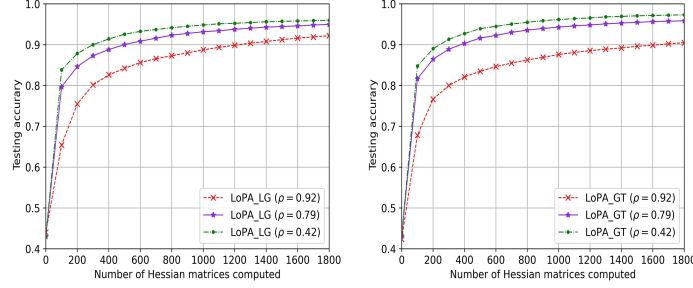

Figure 7: Performance of LoPA-LG and LoPA-GT across different topologies w.r.t. the number of Hessian matrices computed for hyperparameter optimization on binary logistic regression problems using MNIST dataset. Left: testing accuracy of LG schemes w.r.t. the network spectral gap ρ; Right: testing accuracy of GT schemes w.r.t. the network spectral gap ρ. The figure shows GT is more sensitive to the variation of network connectivity.

Additionally, we evaluate the performance of our LoPA-LG and LoPA-GT algorithms across three types of networks with the network spectral gap $\rho = 0.92$ (ring network), $\rho = 0.79$ (star network), and $\rho = 0.42$ (random network) for $m = 10$, as shown in Figure 7(a)-7(c). The doubly stochastic matrix is generated using the Metropolis rule. This experiment also considers hyperparameter optimization on the l_2-regularized binary logistic regression problem in (37) utilizing 6000 samples from the digits '0' and '1' of the MNIST dataset, with validation and training sets randomly assigned to each node in uniform quantities. The mini-batch sizes are set to 25 for both LoPA-LA and LoPA-GT. The results in Figure 7 indicate that as the network spectral gap ρ increases (i.e., a decrease in network connectivity), both the LoPA-LG and LoPA-GT algorithms exhibit a degradation in performance. Notably, when ρ shifts from 0.79 to 0.92, the LoPA-GT algorithm experiences a more significant performance drop compared to LoPA-LG, demonstrating the GT scheme is more sensitive to the variation of network connectivity. This observation corroborates our theoretical findings in Corollaries 7 and 10, which suggest that the GT scheme exhibits greater dependence on network connectivity.

Furthermore, we compare the proposed algorithms with the single-level distributed DGD algorithm (Lian et al., 2017) under a heterogeneous label distribution scenario with $m = 10$. Specifically, we conduct the experiment using three types of datasets: MNIST (with 4000 samples for the digits '1' and '3'), covtype (with 30000 samples for the 'cottonwood' and 'aspen' classes), and cifar10 (with 2000 samples for the 'plane' and 'car' classes). In this case, the proportion of positive and negative labels in half of the nodes is set to be approximately $0.8 \pm a$ with a random $a \in [-0.15, 0.15]$, and in the remaining nodes, it is around $0.35 \pm a$ with a random $a \in [-0.15, 0.15]$. The DGD algorithm is implemented to train a common regression model by solving the l_2-regularized binary logistic regression problem with a fixed l_2-regularization term $\frac{\mu}{2}\|\theta\|^2$, where the regularization coefficient μ is manually adjusted to

$0.47, 0.25, 0.15$ for the MINIST, covetype and cifar10 datasets, respectively. Our algorithms aim to train personalized regression models through the hyperparameter optimization on l_2-regularized binary logistic regression problems as outlined in (37). The mini-batch sizes are set to 20 for MNIST, 150 for covtype, and 10 for cifar10 across all algorithms. The results regarding the average testing accuracy among nodes are provided in Figure 8. It can be observed that the proposed LoPA-LG and LoPA-GT algorithms achieve superior performance under the heterogeneous scenario without manually adjusting the regularization coefficient.
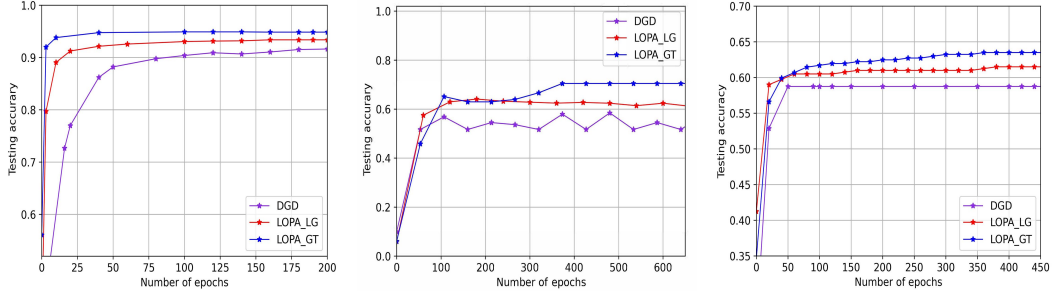
Figure 8: Performance comparison of DGD and our LoPA-LG and LoPA-GT algorithms for distributed binary logistic regression problems. i) MNIST (first column); ii) covtype (second column); iii) cifar10 (third column).

6. Conclusion

In this paper, we have proposed a new loopless algorithm LoPA for solving nonconvex-strongly-convex DSBO problems with personalized inner-level objectives. The proposed LoPA algorithm is shown to converge sublinearly, while significantly reducing the computational complexity of gradient evaluation. We have also explicitly characterized the detailed impact of each-level data heterogeneity on the convergence under local gradient schemes, and have shown how the heterogeneity is eliminated by employing gradient tracking schemes. Moreover, we have introduced a unified analytical framework that enables us to obtain a tighter convergence rate and the best known computational complexity for out-level gradient evaluations in DSBO problems, while explicitly revealing the crucial role of the inner-level heterogeneity. Numerical experiments were conducted to verify the effectiveness of LoPA and demonstrate the impact of the heterogeneity.

Acknowledgments

References

Sulaiman A Alghunaim and Kun Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing*, 70:3264–3279, 2022.

Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. *The Tenth International Conference on Learning Representations*, 2022.

Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pages 810–821. PMLR, 2020.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021a.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022a.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishnakumar Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. *arXiv preprint arXiv:2210.12839*, 2022b.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *arXiv preprint arXiv:2306.12067*, 2023.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024.

Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Accelerating gossip SGD with periodic global averaging. In *International Conference on Machine Learning*, pages 1791–1802. PMLR, 2021b.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in neural information processing systems*, 32, 2019.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *arXiv preprint arXiv:2201.13409*, 2022a.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *arXiv preprint arXiv:2201.13409*, 2022b.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Youran Dong, Shiqian Ma, Junfeng Yang, and Chao Yin. A single-loop algorithm for decentralized bilevel optimization. *arXiv preprint arXiv:2311.08945*, 2023.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Hongchang Gao, Bin Gu, and My T Thai. Stochastic bilevel distributed optimization over a network. *arXiv preprint arXiv:2206.15025*, 2022.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33: 17886–17895, 2020.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021.

Kaiyi Ji, Mingrui Liu, Yingbin Liang, and Lei Ying. Will bilevel optimizers benefit from loops. *arXiv preprint arXiv:2205.14224*, 2022.

Yang Jiao, Kai Yang, Tiancheng Wu, Dongjin Song, and Chengtao Jian. Asynchronous distributed bilevel optimization. *arXiv preprint arXiv:2212.10048*, 2022.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

Boao Kong, Shuchen Zhu, Songtao Lu, Xinmeng Huang, and Kun Yuan. Decentralized bilevel optimization over graphs: Loopless algorithmic update and transient iteration complexity. *arXiv preprint arXiv:2402.03167*, 2024.

Junyi Li, Bin Gu, and Heng Huang. A fully single loop algorithm for bilevel optimization without Hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7426–7434, 2022.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

Songtao Lu, Xiaodong Cui, Mark S Squillante, Brian Kingsbury, and Lior Horesh. Decentralized bilevel optimization for personalized client learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5543–5547. IEEE, 2022a.

Songtao Lu, Siliang Zeng, Xiaodong Cui, Mark S Squillante, Lior Horesh, Brian Kingsbury, Jia Liu, and Mingyi Hong. A stochastic linearized augmented lagrangian method for decentralized bilevel optimization. In *Advances in Neural Information Processing Systems*, 2022b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Angelia Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Takayuki Okuno, Akiko Takeda, Akihiro Kawana, and Motokazu Watanabe. On l_p-hyperparameter learning via bilevel nonsmooth optimization. *arXiv preprint arXiv:1806.01520*, 2018.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

Mengdi Wang, Ji Liu, and Ethan Fang. Accelerating stochastic composition optimization. *Advances in Neural Information Processing Systems*, 29, 2016.

Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 2055–2060. IEEE, 2015.

Chao Xue, Xiaoxing Wang, Junchi Yan, Yonggang Hu, Xiaokang Yang, and Kewei Sun. Rethinking bi-level optimization in neural architecture search: a Gibbs sampling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10551–10559, 2021.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

Shuoguang Yang, Xuezhou Zhang, and Mengdi Wang. Decentralized gossip-based stochastic bilevel optimization over communication networks. *arXiv preprint arXiv:2206.10870*, 2022.

Miao Zhang, Wei Huang, and Bin Yang. Interpreting operation selection in differentiable architecture search: A perspective from influence-directed explanations. *Advances in Neural Information Processing Systems*, 35:31902–31914, 2022.

Yihan Zhang, My T Thai, Jie Wu, and Hongchang Gao. On the communication complexity of decentralized bilevel optimization. *arXiv preprint arXiv:2311.11342*, 2023.

Appendix

Contents

Appendix A. Technical Preliminaries

Notation. For notional convenience, we define some compact notations as follows:
$\nabla_\theta \hat{G}(x^k, \theta^k; \xi_1^k) \triangleq \text{col}\{\nabla_\theta \hat{g}_i(x_i^k, \theta_i^k; \xi_{i,1}^k)\}_{i=1}^m$, $\nabla_{\theta\theta}^2 \hat{G}(x^k, \theta^k; \xi_2^k) \triangleq \text{diag}\{\nabla_{\theta\theta}^2 \hat{g}_i(x_i^k, \theta_i^k; \xi_{i,2}^k)\}_{i=1}^m$,
$\nabla_{x\theta}^2 \hat{G}(x^k, \theta^k; \xi_3^k) \triangleq \text{diag}\{\nabla_{x\theta}^2 \hat{g}_i(x_i^k, \theta_i^k; \xi_{i,3}^k)\}_{i=1}^m$, $\nabla_\theta \hat{F}(x^k, \theta^k; \varsigma_1^k) \triangleq \text{col}\{\nabla_\theta \hat{f}_i(x_i^k, \theta_i^k; \varsigma_{i,1}^k)\}_{i=1}^m$,
$\nabla_x \hat{F}(x^k, \theta^k; \varsigma_2^k) \triangleq \text{col}\{\nabla_x \hat{f}_i(x_i^k, \theta_i^k; \varsigma_{i,2}^k)\}_{i=1}^m$.

The proposed LoPA algorithm in a compact form. For the sake of subsequent analysis, let $\mathcal{W} = W \otimes I_n$, and we can then rewrite the LoPA algorithm in a more compact form as follows:

$$\theta^{k+1} = \theta^k - \beta d^k, \tag{38a}$$

$$v^{k+1} = v^k - \gamma h^k, \tag{38b}$$

$$x^{k+1} = (1 - \tau)x^k + \tau(\mathcal{W}x^k - \alpha y^k), \tag{38c}$$

$$d^{k+1} = \nabla_\theta \hat{G}(x^{k+1}, \theta^{k+1}; \xi_1^{k+1}), \tag{38d}$$

$$h^{k+1} = \nabla_{\theta\theta}^2 \hat{G}(x^{k+1}, \theta^{k+1}; \xi_2^{k+1})v^{k+1} - \nabla_\theta \hat{F}(x^{k+1}, \theta^{k+1}; \varsigma_1^{k+1}), \tag{38e}$$

$$s^{k+1} = \nabla_x \hat{F}(x^{k+1}, \theta^{k+1}; \varsigma_2^{k+1}) - \nabla_{x\theta}^2 \hat{G}(x^{k+1}, \theta^{k+1}; \xi_3^{k+1})v^{k+1}, \tag{38f}$$

$$z^{k+1} = s^k + (1 - \gamma)(z^k - s^k), \tag{38g}$$

with LoPA-LG updating y^{k+1} as

$$y^{k+1} = z^{k+1}, \tag{39}$$

while LoPA-GT updating y^{k+1} as

$$y^{k+1} = \mathcal{W}y^k + z^{k+1} - z^k. \tag{40}$$

Basic inequalities. In the subsequent analysis, we will utilize a set of fundamental inequalities and equalities to simplify the analysis as follows:

- Young's inequality with parameter $\eta > 0$: $\|a + b\|^2 \leq (1 + \frac{1}{\eta})\|a\|^2 + (1 + \eta)\|b\|^2, \forall a, b$.
- Jensen's inequality with l_2-norm for any vectors $x_1, \cdots, x_m$: $\|\frac{1}{m}\sum_{i=1}^m x_i\|^2 \leq \frac{1}{m}\sum_{i=1}^m \|x_i\|^2$.
- Variance decomposition for a stochastic vector x: $\mathbb{E}[\|x - \mathbb{E}[x]\|^2] = \mathbb{E}[\|x\|^2] - \|\mathbb{E}[x]\|^2$.

Appendix B. Proof of Theorems and Corollaries

B.1 Proof of Theorem 6

To analyze the convergence of LoPA-LG, we need to properly select the coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 of the Lyapunov function (23) and establish the dynamic of the function based on the results in Section 4.3. To this end, we first set the coefficients d_0, d_1 and d_2 as follows:

$$d_0 = 1, d_1 = \frac{8C_{g,x\theta}^2 \tau\alpha}{\mu_g\lambda}, d_2 = \left(\frac{8C_{g,x\theta}^2 \tau\alpha}{\mu_g\lambda}q_x + L_{fg,x}\tau\right)\frac{\alpha}{\omega_\theta\beta}, d_3 = \frac{\tau\alpha}{2\gamma}, \tag{41}$$

where the parameters $L_{fg,x} = 2L_{f,x}^2 + 4M^2L_{g,x\theta}^2$, $q_x = \frac{4L_{fg,\theta}\lambda}{\mu_g\alpha}$ and $\omega_\theta = \frac{\mu_g L_{g,\theta}}{2(\mu_g+L_{g,\theta})}$ are defined in Lemmas 16, 17 and 18, respectively. Then, considering above coefficients and combining Lemmas 15-18, we can reach the following inequality

$$
\begin{aligned}
&d_0\mathbb{E}[\Phi(\bar{x}^{k+1})]+d_1\frac{1}{m}\mathbb{E}[\|v^{k+1}-v^*(\bar{x}^{k+1})\|^2] \\
&+d_2\frac{1}{m}\mathbb{E}[\|\theta^{k+1}-\theta^*(\bar{x}^{k+1})\|^2]+d_3\frac{1}{m}\mathbb{E}[\|\nabla\Phi(\bar{x}^{k+1})-\bar{z}^{k+1}\|^2] \\
\leqslant&d_0\mathbb{E}[\Phi(\bar{x}^k)]+d_1\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]+d_2\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2]+d_3\frac{1}{m}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)-\bar{z}^k\|^2] \\
&-\frac{d_0}{2}\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]-(\frac{d_0}{2}\tau\alpha(1-\tau\alpha L)-d_1q_s\tau^2\alpha^2-d_2p_s\tau^2\alpha^2-d_3r_y\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2] \\
&-d_1\frac{\mu_g\lambda}{2}\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]-d_2\omega_\theta\beta\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2] \\
&+(L_{fg,x}d_3r_z+d_1q_x+d_2p_x)\frac{1}{m}\alpha\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2] \\
&+\frac{1}{m}d_3\sigma_{\bar{z}}^2\alpha^2+d_1\sigma_v^2\alpha^2+d_2\sigma_\theta^2\alpha^2.
\end{aligned}
\tag{42}
$$

We next deal with the gradient error term $\mathbb{E}[\|y^k-1_m\otimes\bar{y}^k\|^2]$ induced by the consensus errors $\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]$ under local gradient scheme (11). Specifically, we let

$$
\begin{aligned}
d_4 &= (L_{fg,x}d_3r_z+d_1q_x+d_2p_x)\frac{24}{(1-\rho)^2}\frac{\alpha}{\gamma}\alpha^3, \\
d_5 &= 2(L_{fg,x}d_3r_z+d_1q_x+d_2p_x)\frac{2\alpha}{\tau(1-\rho)}.
\end{aligned}
$$

Then, employing Lemmas 20 and 22 gives us

$$
\begin{aligned}
&d_4\frac{1}{m}\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^{k+1})-z^{k+1}\|^2]+d_5\frac{1}{m}\mathbb{E}[\|x^{k+1}-1_m\otimes\bar{x}^{k+1}\|^2] \\
\leqslant&d_4\frac{1}{m}\mathbb{E}\left[\|\nabla\tilde{\Phi}(\bar{x}^k)-z^k\|^2\right]+d_5\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]-d_5\tau\frac{1-\rho}{2}\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2] \\
&+d_5\frac{6\tau\alpha^2}{1-\rho}\frac{b^2}{m}+d_5\frac{6\tau\alpha^2}{1-\rho}[\|\nabla\Phi(\bar{x}^k)\|^2]+d_4r_z\alpha\frac{1}{m}\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k)-s^k\|^2]+d_4r_y\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2]+d_4\sigma_{\bar{z}}^2\alpha^2 \\
\leqslant&d_4\frac{1}{m}\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k)-z^k\|^2]+d_5\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]-d_5\tau\frac{1-\rho}{2}\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2] \\
&+d_5\frac{6\tau\alpha^2}{1-\rho}[\|\nabla\Phi(\bar{x}^k)\|^2] \\
&+d_4r_y\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2]+d_4\sigma_{\bar{z}}^2\alpha^2+(L_{fg,x}d_3r_z+d_1q_x+d_2p_x)\frac{24\alpha}{(1-\rho)^2\gamma}\frac{b^2\alpha^3}{m} \\
&+4C_{g,x\theta}^2d_4r_z\alpha\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]+L_{fg,x}d_4r_z\alpha\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2] \\
&+L_{fg,x}d_4r_z\alpha\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2],
\end{aligned}
\tag{43}
$$

where the last step uses the boundedness of the term $\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - s^k\|^2]$ in Lemma 22. We proceed in eliminating the term $\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2]$ in (42). It is noted that for the term $\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2]$ in (43) we have $-d_5\tau\frac{1-\rho}{2} = -2\left(L_{fg,x}d_3r_z + d_1q_x + d_2p_x\right)\alpha$. Then, incorporating the inequality (43) into the inequality (42) and letting $d_6 = 0$, we obtain

$$
\begin{aligned}
&\mathbb{E}[V^{k+1}]\\
&\leqslant \mathbb{E}[V^k] - \frac{d_0}{4}\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] - (\frac{d_0}{4}\tau\alpha - d_5\frac{6\tau\alpha^2}{1-\rho})\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\\
&\quad - ((\frac{d_0}{2}\tau\alpha\,(1-\tau\alpha L) - d_1q_s\tau^2\alpha^2 - d_2p_s\tau^2\alpha^2 - d_3r_y\tau^2\alpha^2 - d_4r_y\tau^2\alpha^2))\mathbb{E}[\|\bar{y}^k\|^2]\\
&\quad - (d_1\frac{\mu_g\lambda}{2} - 4C_{g,x\theta}^2d_4r_z\alpha)\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2] - (d_2\omega_\theta\beta - L_{fg,x}d_4r_z\alpha)\frac{1}{m}\alpha\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2]\\
&\quad - (d_5\frac{\tau(1-\rho)}{4} - L_{fg,x}d_4r_z\alpha)\frac{1}{m}\mathbb{E}[\|x^k - 1_m\otimes\bar{x}^k\|^2]\\
&\quad + (\frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_1\sigma_v^2 + d_2\sigma_\theta^2 + d_4\sigma_z^2)\alpha^2 + (L_{fg,x}d_3r_z + d_1q_x + d_2p_x)\frac{24\alpha}{(1-\rho)^2\gamma}\frac{b^2\alpha^3}{m}\\
&= \mathbb{E}[V^k] - \frac{d_0}{4}\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] - (\frac{d_0}{4}\tau\alpha - d_5\frac{6\tau\alpha^2}{1-\rho})\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\\
&\quad - ((\frac{d_0}{2}\tau\alpha\,(1-\tau\alpha L) - d_1q_s\tau^2\alpha^2 - d_2p_s\tau^2\alpha^2 - d_3r_y\tau^2\alpha^2 - d_4r_y\tau^2\alpha^2))\mathbb{E}[\|\bar{y}^k\|^2]\\
&\quad - (d_1\frac{\mu_g\lambda}{2} - 4C_{g,x\theta}^2d_4r_z\alpha - 2(\frac{1}{m}d_3\frac{\gamma^2}{\alpha^2} + d_1\frac{\lambda^2}{\alpha^2} + d_4\frac{\gamma^2}{\alpha^2})\sigma_{g,x\theta}^2\alpha^2))\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2]\\
&\quad - (d_2\omega_\theta\beta - L_{fg,x}d_4r_z\alpha)\frac{1}{m}\alpha\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2]\\
&\quad - (d_5\frac{\tau(1-\rho)}{4} - L_{fg,x}d_4r_z\alpha)\frac{1}{m}\mathbb{E}[\|x^k - 1_m\otimes\bar{x}^k\|^2]\\
&\quad + ((\frac{1}{m}d_3 + d_4)(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2} + 2d_1\left(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2\right)\frac{\lambda^2}{\alpha^2} + 2d_2\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2})\alpha^2\\
&\quad + (L_{fg,x}d_3r_z + d_1q_x + d_2p_x)\frac{24\alpha}{(1-\rho)^2\gamma}\frac{b^2\alpha^3}{m},
\end{aligned}
\tag{44}
$$

where the last equality uses the fact that $\sigma_{\bar{z}}^2 = \sigma_z^2$ in Lemmas 16 and 22 and the definitions of $\sigma_{\bar{z}}^2, \sigma_v^2, \sigma_\theta^2$ in Lemmas 16, 17, 18. If the step-size α satisfies the following conditions

$$
\frac{d_0}{2}\tau\alpha\,(1-\tau\alpha L) - d_1q_s\tau^2\alpha^2 - d_2p_s\tau^2\alpha^2 - d_3r_y\tau^2\alpha^2 - d_4r_y\tau^2\alpha^2 \geqslant 0,
\tag{45}
$$

$$
\frac{d_0}{4}\alpha - d_5\frac{6\tau\alpha^2}{1-\rho} \geqslant 0,
\tag{46}
$$

$$
d_1\frac{\mu_g\lambda}{2} - 4C_{g,x\theta}^2d_4r_z\alpha - 2(\frac{1}{m}d_3\frac{\gamma^2}{\alpha^2} + d_1\frac{\lambda^2}{\alpha^2} + d_4\frac{\gamma^2}{\alpha^2})\sigma_{g,x\theta}^2\alpha^2 \geqslant 0,
\tag{47}
$$

$$
d_2\omega_\theta\beta - L_{fg,x}d_4r_z\alpha \geqslant 0,
\tag{48}
$$

$$
d_5\frac{\tau(1-\rho)}{4} - L_{fg,x}d_4r_z\alpha \geqslant 0,
\tag{49}
$$

then we further have

$$\mathbb{E}[V^{k+1}] \leqslant \mathbb{E}[V^k] - \frac{d_0}{4}\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] + (L_{fg,x}d_3r_z + d_1q_x + d_2p_x)\frac{24\alpha}{(1-\rho)^2\gamma}\frac{b^2\alpha^3}{m} \tag{50}$$

$$+ ((\frac{1}{m}d_3 + d_4)(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2} + 2d_1\left(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2\right)\frac{\lambda^2}{\alpha^2} + 2d_2\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2})\alpha^2,$$

where the coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 of the Lyapunov function (23) are as follows:

$$d_0 = 1, d_1 = \frac{8C_{g,x\theta}^2\tau\alpha}{\mu_g\lambda}, d_2 = (\frac{8C_{g,x\theta}^2\tau\alpha}{\mu_g\lambda}q_x + L_{fg,x}\tau)\frac{\alpha}{\omega_\theta\beta}, d_3 = \frac{\tau\alpha}{2\gamma}, \tag{51}$$

$$d_4 = (L_{fg,x}d_3r_z + d_1q_x + d_2p_x)\frac{24}{(1-\rho)^2}\frac{\alpha}{\gamma}\alpha^3, d_5 = 2\left(L_{fg,x}d_3r_z + d_1q_x + d_2p_x\right)\frac{2\alpha}{\tau(1-\rho)}, d_6 = 0.$$

Next, we proceed to find the sufficient conditions for the step-sizes to satisfy the conditions (45) to (49). To address the conditions (45)-(49), we start by simplifying the term $L_{fg,x}d_3r_z + d_1q_x + d_2p_x$ in d_4 and d_5 as

$$L_{fg,x}d_3r_z + d_1q_x + d_2p_x = (L_{fg,x} + \frac{32C_{g,x\theta}^2L_{fg,\theta}}{\mu_g\mu_g} + \left(\frac{32C_{g,x\theta}^2L_{fg,\theta}}{\mu_g^2} + L_{fg,x}\right)\frac{4L_{g,\theta}^2}{\omega_\theta^2})\tau$$

$$= \underbrace{(L_{fg,x} + \frac{32C_{g,x\theta}^2L_{fg,\theta}}{\mu_g^2})(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2}))}_{\triangleq\varphi}\tau. \tag{52}$$

Then, d_2, d_4 and d_5 can be simplified as follows:

$$d_2 = (L_{fg,x} + \frac{32C_{g,x\theta}^2L_{fg,\theta}}{\mu_g^2})\frac{\tau\alpha}{\omega_\theta\beta}, d_4 = \frac{24\varphi\tau\alpha^4}{(1-\rho)^2\gamma}, d_5 = \frac{2\varphi\alpha}{(1-\rho)}.$$

To ensure that condition (45) holds, a sufficient selection condition for the step-sizes α, β, λ, γ, τ is given as

$$\alpha \leqslant u_1 \triangleq \frac{1}{2\tau L}, \tag{53}$$

$$\lambda \geqslant \frac{48L_{v^*}C_{g,x\theta}}{\mu_g}\tau\alpha, \tag{54}$$

$$\beta \geqslant \frac{6L_{\theta^*}}{\omega_\theta}(\frac{32C_{g,x\theta}^2L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/2}\tau\alpha, \tag{55}$$

$$\gamma \geqslant \max\{4L\tau\alpha, \frac{32L\varphi^{1/2}}{(1-\rho)}\tau\alpha^2\}, \tag{56}$$

with $0 < \tau < 1$. Furthermore, we can derive the following sufficient selection condition for the step-sizes α, γ and λ to satisfy the conditions (46)-(49)

$$\alpha \leqslant u_2 \triangleq \min \left\{ \frac{1-\rho}{12\varphi^{1/2}}, \frac{(1-\rho)^{2/3}}{4\varphi^{1/3}}, \frac{\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x}\right)^{1/3}(1-\rho)^{2/3}}{4\varphi^{1/3}}, \right. \tag{57}$$

$$\left. \frac{(1-\rho)^{1/2}}{3(L_{fg,x})^{1/4}}, \frac{C_{g,x\theta}^{2/3}(1-\rho)^{2/3}}{4\varphi^{1/3}\sigma_{g,x\theta}^{2/3}} \right\},$$

$$\gamma \leqslant \frac{C_{g,x\theta}^2}{\sigma_{g,x\theta}^2}, \quad \lambda \leqslant \frac{\mu_g}{16\sigma_{g,x\theta}^2}, \tag{58}$$

where $L_{fg,x}$, $L_{fg,\theta}$ and φ are given by (16), (17) and (52), respectively. It is noted that $\frac{\varphi^{1/2}\alpha}{1-\rho} \leqslant \frac{1}{12}$ by the condition (58), which implies that the term $\frac{32L^{1/2}\varphi^{1/2}}{(1-\rho)}\tau\alpha^2$ in (56) can be bounded by

$$\frac{32L^{1/2}\varphi^{1/2}}{(1-\rho)}\tau\alpha^2 \leqslant \frac{32}{12}\tau L\alpha < 4L\tau\alpha.$$

By combining Lemmas 15-18 and the above inequality, we have the following condition for the step-sizes λ, β and γ on the basis of (54)-(58)

$$\min\{\frac{1}{\mu_g}, \frac{\mu_g}{8\sigma_{g,x\theta}^2}\} > \lambda \geqslant \frac{48L_{v^*}C_{g,x\theta}}{\mu_g}\tau\alpha, \tag{59}$$

$$\min\left\{\frac{2}{\mu_g + L_{g,\theta}}, \frac{\mu_g + L_{g,\theta}}{2\mu_g L_{g,\theta}}\right\} > \beta \geqslant \frac{6L_{\theta^*}}{\omega_\theta}\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x}\right)^{1/2}\tau\alpha, \tag{60}$$

$$\min\{1, \frac{2C_{g,x\theta}^2}{\sigma_{g,x\theta}^2}\} > \gamma \geqslant 4L\tau\alpha, \tag{61}$$

where the selection for λ, β and γ can be guaranteed to be non-empty when the step-size α satisfies the following condition:

$$\alpha \leqslant u_3 \triangleq \min\left\{\frac{\min\{1, \frac{\mu_g^2}{8\sigma_{g,x\theta}^2}\}}{96\tau L_{v^*}C_{g,x\theta}}, \frac{\min\left\{\frac{2}{\mu_g + L_{g,\theta}}, \frac{\mu_g + L_{g,\theta}}{2\mu_g L_{g,\theta}}\right\}w_\theta}{12\tau(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/2}L_{\theta^*}}, \frac{\min\{1, \frac{2C_{g,x\theta}^2}{\sigma_{g,x\theta}^2}\}}{8L\tau}\right\}. \tag{62}$$

As a result, to ensure that the conditions (45)-(49) hold, with $0 < \tau < 1$, we can properly select the step-size α as follows:

$$\alpha \leqslant u \triangleq \min\{u_1, u_2, u_3\}, \tag{63}$$

where u_1, u_2, and u_3 are defined by (53), (58), and (62), respectively, and are clearly independent of λ, β, and γ. Subsequently, we can determine the step sizes λ, β and γ according to (59)-(61). Then, by combining the definitions of d_0 in (51) and φ in (52) and considering condition (59)-(63), the inequality (50) can be derived as

$$\frac{\mathbb{E}[V^{k+1}]}{\tau} \leqslant \frac{\mathbb{E}[V^k]}{\tau} - \frac{d_0}{4}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] + \alpha^2\sigma_{LG}^2 + \frac{\vartheta}{m}\alpha^3 b^2, \tag{64}$$

where ϑ and σ_{LG}^2 are given by

$$\vartheta \triangleq \frac{24\alpha\varphi}{(1-\rho)^2\gamma}, \tag{65}$$

$$\sigma_{LG}^2 \triangleq (\frac{1}{m}\frac{d_3}{\tau} + \frac{d_4}{\tau})(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2} + 2\frac{d_1}{\tau}\left(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2\right)\frac{\lambda^2}{\alpha^2} + 2\frac{d_2}{\tau}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}, \tag{66}$$

where $d_0 = 1$, $d_2 = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})\frac{\tau\alpha}{\omega_\theta\beta}$, $d_3 = \frac{\tau\alpha}{2\gamma}$, $d_4 = \frac{24\varphi\tau\alpha^4}{(1-\rho)^2\gamma}$, $d_5 = \frac{2\varphi\alpha}{(1-\rho)}$, $d_6 = 0$ with $\varphi = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2})$. Now, summing up and telescoping the above inequality from $k = 0$ to K yields

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant \frac{4(V^0 - V^K)}{\tau\alpha(K+1)} + 4\alpha\sigma_{LG}^2 + \frac{4\vartheta}{m}\alpha^2 b^2, \tag{67}$$

where we use the fact that $d_0 = 1$. This completes the proof. ∎

B.2 Proof of Corollary 7

In the subsequent analysis, based on Theorem 6 we will select the step-sizes α, γ, β, λ in term of the number of iterations K. When the conditions (59)-(61) and (63) hold and the step-sizes γ, λ, β are taken as $\gamma = c_\gamma\alpha, \lambda = c_\lambda\alpha, \beta = c_\beta\alpha$ with the positive parameters $c_\gamma, c_\lambda, c_\beta$ being independent of the terms K, $1 - \rho$ and satisfying

$$c_\lambda \geqslant \frac{48L_{v^*}C_{g,x\theta}}{\mu_g}\tau, c_\beta \geqslant \frac{6L_{\theta^*}}{\omega_\theta}(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/2}\tau, c_\gamma \geqslant 4L\tau, \tag{68}$$

then from (65) we have that $d_0 = 1, d_1 = \frac{8C_{g,x\theta}^2\tau}{\mu_g c_\lambda}, d_2 = (\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})\frac{\tau}{\omega_\theta c_\beta}, d_3 = \frac{\tau}{2c_\gamma}$ are independent of the step-size α. When the parameters $c_\gamma, c_\lambda, c_\beta$ satisfy (68) and their values are properly selected, the conditions (59)-(61) always hold. For examples, a sufficient condition for the upper bounds of the parameters $c_\gamma, c_\lambda, c_\beta$ could be $\frac{96L_{v^*}C_{g,x\theta}}{\mu_g}\tau > c_\lambda$, $\frac{12L_{\theta^*}}{\omega_\theta}(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/2}\tau > c_\beta$, and $8L\tau > c_\gamma$, if we consider a sufficient case $\alpha \leqslant u_3$ under the condition (63) and substitute it into the conditions (59)-(61). Then it follows from the definition of φ in (52) that ϑ in (65) can be rewritten as follows:

$$\vartheta = \frac{24\varphi}{(1-\rho)^2 c_\gamma} = \mathcal{O}(\frac{1}{(1-\rho)^2}). \tag{69}$$

Furthermore, with the above-mentioned selection condition for the step-sizes λ, β, γ, we also have that the terms σ_v, σ_θ, $\sigma_{\bar{z}}$ and σ_z are independent of the step-size α. Therefore, by combining the results that d_1, d_2, d_3 are independent of the step-size α as well as

$d_4 = \mathcal{O}(\alpha^3 d_3)$, the variance related term σ_{LG} in (66) can be further derived as

$$
\begin{aligned}
\sigma_{LG} = & \sqrt{(\frac{1}{m}\frac{d_3}{\tau} + \frac{d_4}{\tau})(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2} + 2\frac{d_1}{\tau}\left(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2\right)\frac{\lambda^2}{\alpha^2} + 2\frac{d_2}{\tau}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}} \\
\leqslant & \sqrt{\frac{2}{m}\frac{d_3}{\tau}(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\lambda^2}{\alpha^2} + 2\frac{d_1}{\tau}\left(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2\right)\frac{\gamma^2}{\alpha^2} + 2\frac{d_2}{\tau}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}} \qquad (70) \\
\leqslant & \sqrt{2}(\underbrace{\sqrt{\frac{d_1}{\tau}}(\sigma_{f,\theta} + 2M\sigma_{g,\theta\theta})c_\lambda + \sqrt{\frac{d_2}{\tau}}\sigma_{g,\theta}c_\beta}_{\triangleq \sigma_{\mathrm{p}}}) + \frac{\sqrt{2}}{\sqrt{m}}\underbrace{\sqrt{\frac{d_3}{\tau}}(\sigma_{f,x} + 2M\sigma_{g,x\theta})c_\gamma}_{\triangleq \sigma_c} \triangleq \hat{\sigma}_{LG},
\end{aligned}
$$

where the first inequality holds as d_4 is high-order term w.r.t α such that $d_4 \ll \frac{1}{m}d_3$ under a large number of iterations K, and the second inequality is derived by the triangle inequality. Meanwhile, from (70), it follows that

$$
\begin{aligned}
\sigma_{\mathrm{p}} &= \mathcal{O}(\sigma_{f,\theta} + \sigma_{g,\theta\theta} + \sigma_{g,\theta}), \\
\sigma_{\mathrm{c}} &= \mathcal{O}(\sigma_{f,x} + \sigma_{g,x\theta}).
\end{aligned} \qquad (71)
$$

By the condition (63), we have that $\alpha \leqslant u = \mathcal{O}(1-\rho)$. Then, combining the definition of φ in (52) and the condition that $c_\gamma, c_\lambda, c_\beta$ are independent of the term $1-\rho$, it follows from (65) that the coefficients d_4 and d_5 can be bounded as follows:

$$
d_4 = \frac{24\varphi\tau\alpha^4}{(1-\rho)^2\gamma} \leqslant \frac{24\varphi\tau}{(1-\rho)^2 c_\gamma}u^3 \triangleq \hat{d}_4 = \mathcal{O}(1-\rho), \qquad (72)
$$

$$
d_5 = \frac{2\varphi}{(1-\rho)}\alpha \leqslant \frac{2\varphi}{(1-\rho)}u \triangleq \hat{d}_5 = \mathcal{O}(1), \qquad (73)
$$

which implies that

$$
\begin{aligned}
\frac{V^0 - V^K}{\tau} \leqslant & \frac{V^0}{\tau} \\
\leqslant & \frac{d_0}{\tau}\Phi(\bar{x}^0) + \frac{d_1}{\tau}\frac{1}{m}\left\|v^0 - v^*(\bar{x}^0)\right\|^2 + \frac{d_2}{\tau}\frac{1}{m}\left\|\theta^0 - \theta^*(\bar{x}^0)\right\|^2 + \frac{d_3}{\tau}\left\|\nabla\Phi(\bar{x}^0) - \bar{z}^0\right\|^2 \\
& + \frac{\hat{d}_4}{\tau}\frac{1}{m}\left\|\nabla\tilde{\Phi}(\bar{x}^0) - z^0\right\|^2 + \frac{\hat{d}_5}{\tau}\frac{1}{m}\left\|x^0 - 1_m \otimes \bar{x}^0\right\|^2 \\
\triangleq & \hat{V}^0 = \mathcal{O}(1-\rho),
\end{aligned} \qquad (74)
$$

where the last equality holds due to the fact that the coefficients d_0, d_1, d_2, and d_3 are also independent of the term $1-\rho$.

For the sake of simplicity, let us consider the following notations

$$
a_0 \triangleq 4\hat{V}^0, a_1 \triangleq 4\hat{\sigma}_{LG}^2, a_2 \triangleq \frac{4\vartheta}{m}b^2. \qquad (75)
$$

Then, combining (69), (70) and (74), the inequality (67) becomes

$$
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant a_0\frac{1}{\alpha(K+1)} + a_1\alpha + a_2\alpha^2. \qquad (76)
$$

38

When the step-size α is taken as $\alpha = \min\left\{u, \left(\frac{a_0}{a_1(K+1)}\right)^{\frac{1}{2}}, \left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}}\right\}$ and the step-sizes γ, λ, β are taken as $\gamma = c_\gamma \alpha, \lambda = c_\lambda \alpha, \beta = c_\beta \alpha$ with $c_\gamma, c_\lambda, c_\beta$ being independent of the terms K and $1 - \rho$, we can proceed with the following discussion:

i) When $\left(\frac{a_0}{a_1(K+1)}\right)^{\frac{1}{2}}$ is smallest, we set $\alpha = \left(\frac{a_0}{a_1(K+1)}\right)^{\frac{1}{2}}$. According to the fact that $\left(\frac{a_0}{a_1(K+1)}\right)^{\frac{1}{2}} \leqslant \left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}}$, we have

$$
\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] &\leqslant \left(\frac{a_1 a_0}{K+1}\right)^{\frac{1}{2}} + \left(\frac{a_1 a_0}{K+1}\right)^{\frac{1}{2}} + a_2\left(\frac{a_0}{a_1(K+1)}\right) \\
&\leqslant 2\left(\frac{a_1 a_0}{K+1}\right)^{\frac{1}{2}} + a_2^{\frac{1}{3}}\left(\frac{a_0}{K+1}\right)^{\frac{2}{3}}.
\end{aligned}
\tag{77}
$$

ii) When $\left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}}$ is smallest, we set $\alpha = \left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}}$. According to the fact that $\left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}} \leqslant \left(\frac{a_0}{a_1(K+1)}\right)^{\frac{1}{2}}$, we have

$$
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant 2a_2^{\frac{1}{3}}\left(\frac{a_0}{K+1}\right)^{\frac{2}{3}} + \left(\frac{a_1 a_0}{K+1}\right)^{\frac{1}{2}}.
\tag{78}
$$

iii) When u is smallest, we set $\alpha = u$. According to $u \leqslant \left(\frac{a_0}{a_1(K+1)}\right)^{\frac{1}{2}}, u \leqslant \left(\frac{a_0}{a_2(K+1)}\right)^{\frac{1}{3}}$, we have

$$
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant \frac{a_0}{u(K+1)} + \left(\frac{a_1 a_0}{K+1}\right)^{\frac{1}{2}} + a_2^{\frac{1}{3}}\left(\frac{a_0}{K+1}\right)^{\frac{2}{3}}.
\tag{79}
$$

According to the above discussion regarding (77), (78), and (79), we can conclude that

$$
\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] &\leqslant \frac{a_0}{u(K+1)} + 2\left(\frac{a_1 a_0}{K+1}\right)^{\frac{1}{2}} + 2a_2^{\frac{1}{3}}\left(\frac{a_0}{K+1}\right)^{\frac{2}{3}} \\
&\leqslant \frac{4(\hat{V}^0)}{u(K+1)} + \frac{8\hat{\sigma}_{LG}(\hat{V}^0)^{\frac{1}{2}}}{(K+1)^{\frac{1}{2}}} + \frac{8\left(\frac{\vartheta b^2}{m}\right)^{\frac{1}{3}}(\hat{V}^0)^{\frac{2}{3}}}{(K+1)^{\frac{2}{3}}}.
\end{aligned}
\tag{80}
$$

In what follows, we will examine how the parameters in (80) depend on the terms κ and $1 - \rho$. To this end, we first note that with the Lipschitz constants defined in (15) and the auxiliary coefficients $L_{fg,x}, L_{fg,\theta}, \omega_\theta$ defined in Lemma 16, 17, 18 respectively. Then it follows from (52) that $\varphi = \mathcal{O}(\kappa^6)$ and $\vartheta = \mathcal{O}\left(\frac{\kappa^6}{(1-\rho)^2}\right)$ in (69). To obtain the best κ-dependence, we select $c_\gamma = \mathcal{O}(1), c_\beta = \mathcal{O}(1), c_\lambda = \mathcal{O}(1), \tau = \mathcal{O}(\kappa^{-4})$ and $u = \mathcal{O}\left(\kappa^{-3}(1-\rho)\right)$ in terms of κ and $1 - \rho$. When we initialize the outer-level variables as $x_i^0 = x_j^0, \forall i, j \in \mathcal{V}$, it follows that $\|x^0 - 1_m \otimes \bar{x}^0\|^2 = 0$ holds. Furthermore, from (74) we know that $\hat{V}^0 = \mathcal{O}(\kappa^5)$ and $\hat{V}^0$ is independent of the term $\frac{1}{1-\rho}$. Therefore, recalling the definition of $\hat{\sigma}_{LG} = \mathcal{O}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})$ in (70), we have

$$
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{\kappa^8}{(1-\rho)K} + \frac{\kappa^{\frac{16}{3}}\left(\frac{b}{\sqrt{m}}\right)^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}} + \frac{1}{\sqrt{K}}\kappa^{\frac{5}{2}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right), \tag{81}
$$

where we use the fact that $\sigma_{\mathrm{p}} = \mathcal{O}(\kappa^{\frac{1}{2}}\sigma_{f,\theta} + \kappa^{\frac{3}{2}}\sigma_{g,\theta\theta} + \kappa^{\frac{5}{2}}\sigma_{g,\theta})$ and $\sigma_{\mathrm{c}} = \mathcal{O}(\sigma_{f,x} + \kappa\sigma_{g,x\theta})$ by (70). This completes the proof. ∎

B.3 Proof of Theorem 9

The proof of Theorem 9 follows the similar steps to that of Theorem 6. Particularly, in addition to d_0, d_1, d_2, d_3, d_4, d_5, we need to properly select the coefficient d_6 in order to establish the dynamic of the Lyapunov function (23) for LoPA-GT based on the results of Section 4.3. To be specific, by letting

$$d_0 = 1, d_1 = \frac{8C_{g,x\theta}^2 \tau\alpha}{\mu_g\lambda}, d_2 = (\frac{8C_{g,x\theta}^2 \tau\alpha}{\mu_g\lambda}q_x + L_{fg,x}\tau)\frac{\alpha}{\omega_\theta\beta}, d_3 = \frac{\tau\alpha}{2\gamma}$$

and combining Lemmas 15-18, the inequality (42) can also be derived for LoPA-GT. In the inequality (42), we recall that the related parameters are defined in Lemmas 15-18. In what follows, we focus on dealing with the consensus errors $\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2]$. Note that the term $\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2]$ is controlled by the evolution of the gradient errors $\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2]$ under gradient tracking scheme (12), while the errors $\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2]$ are further influenced by the variance errors $\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - s^k\|^2]$. Motivated by this fact, we first let

$$d_4 = (L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x)\frac{64\gamma\alpha^2}{(1-\rho)^4},$$

$$d_5 = 2(L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x)\frac{2\alpha}{\tau(1-\rho)},$$

$$d_6 = (L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x)\frac{16\alpha^2}{(1-\rho)^3}.$$

Then, by integrating Lemmas 16, 21, 22, the inequality (42) becomes:

$$d_4\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^{k+1}) - z^{k+1}\|^2] + d_5\mathbb{E}[\|x^{k+1} - 1_m \otimes \bar{x}^{k+1}\|^2] + d_6\mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2]$$

$$\leqslant d_4\mathbb{E}\left[\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2\right] + d_5(1 - \tau\frac{1-\rho}{2})\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + d_6\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2]$$

$$+ (d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2)\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{x}^k) - s^k\|^2]$$

$$+ md_4 r_y\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + md_4\sigma_z^2\alpha^2 + m\frac{2d_6}{1-\rho}\sigma_y^2\alpha^2. \tag{82}$$

Now, noting that

$$d_5\tau\frac{1-\rho}{2} = 2(L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x)\alpha.$$

and combining the inequalities (42) and (82), we can establish the following dynamic

$$\mathbb{E}[V^{k+1}] \leqslant \mathbb{E}[V^k] - \frac{1}{2}d_0\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]$$

$$- (\frac{d_0}{2}\tau\alpha(1 - \tau\alpha L) - d_1 q_s\tau^2\alpha^2 - d_2 p_s\tau^2\alpha^2 - d_3 r_y\tau^2\alpha^2 - d_4 r_y\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2]$$

$$- (d_1\frac{\mu_g\lambda}{2} - 4C_{g,x\theta}^2(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2))\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2]$$

$$\tag{83}$$

$$- (d_2\omega_\theta\beta - L_{fg,x}(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2))\frac{1}{m}\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2]$$

$$- (d_5\tau\frac{1-\rho}{4} - L_{fg,x}(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2))\frac{1}{m}\mathbb{E}[\|x^k - 1_m\otimes\bar{x}^k\|^2]$$

$$- (d_4\gamma - d_6\frac{4}{1-\rho}\gamma^2)\frac{1}{m}\mathbb{E}[\|s^k - z^k\|^2]$$

$$+ \frac{1}{m}d_3\sigma_{\bar{z}}^2\alpha^2 + d_1\sigma_v^2\alpha^2 + d_2\sigma_\theta^2\alpha^2 + d_4\sigma_z^2\alpha^2 + \frac{2d_6}{1-\rho}\sigma_y^2\alpha^2$$

$$=\mathbb{E}[V^k] - \frac{1}{2}d_0\tau\alpha\mathbb{E}[\|\Phi(\bar{x}^k)\|^2]$$

$$- (\frac{d_0}{2}\tau\alpha(1-\tau\alpha L) - d_1 q_s\tau^2\alpha^2 - d_2 p_s\tau^2\alpha^2 - d_3 r_y\tau^2\alpha^2 - d_4 r_y\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2]$$

$$- (d_1\frac{\mu_g\lambda}{2} - 4C_{g,x\theta}^2(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2))\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2]$$

$$+ 2(\frac{1}{m}d_3\frac{\gamma^2}{\alpha^2} + d_1\frac{\lambda^2}{\alpha^2} + d_4\frac{\gamma^2}{\alpha^2} + \frac{2d_6}{1-\rho}\frac{\gamma^2}{\alpha^2})\sigma_{g,\theta\theta}^2\alpha^2\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2]$$

$$- (d_2\omega_\theta\beta - L_{fg,x}(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2))\frac{1}{m}\alpha\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2]$$

$$- (d_5\tau\frac{1-\rho}{4} - L_{fg,x}(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2))\frac{1}{m}\mathbb{E}[\|x^k - 1_m\otimes\bar{x}^k\|^2]$$

$$+ (\frac{1}{m}d_3 + d_4 + \frac{2d_6}{1-\rho})(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}\alpha^2$$

$$+ 2d_1(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2}\alpha^2 + 2d_2\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}\alpha^2,$$

where the last equality uses the fact that $\sigma_{\bar{z}}^2 = \sigma_z^2 = (\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}$ as shown in Lemmas 16, 21, 22, and the definitions of $\sigma_{\bar{z}}^2, \sigma_v^2, \sigma_\theta^2$ in Lemmas 16, 17, 18. Furthermore, when the following conditions hold

$$\frac{d_0}{2}\tau\alpha(1-\tau\alpha L) - d_1 q_s\tau^2\alpha^2 - d_2 p_s\tau^2\alpha^2 - d_3 r_y\tau^2\alpha^2 - d_4 r_y\tau^2\alpha^2 \geqslant 0, \quad (84)$$

$$d_1\frac{\mu_g\lambda}{2} - 4C_{g,x\theta}^2(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2) - 2(\frac{1}{m}d_3\frac{\gamma^2}{\alpha^2} + d_1\frac{\lambda^2}{\alpha^2} + d_4\frac{\gamma^2}{\alpha^2} + \frac{2d_6}{1-\rho}\frac{\gamma^2}{\alpha^2})\sigma_{g,\theta\theta}^2\alpha^2 \geqslant 0, \quad (85)$$

$$d_2\omega_\theta\beta - L_{fg,x}(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2) \geqslant 0, \quad (86)$$

$$d_5\tau\frac{1-\rho}{4} - L_{fg,x}(d_4 r_z\alpha + d_6\frac{4}{1-\rho}\gamma^2) \geqslant 0, \quad (87)$$

we derive that

$$\mathbb{E}[V^{k+1}] \leqslant \mathbb{E}[V^k] - \frac{d_0}{2}\tau\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]$$

$$+ (\frac{1}{m}d_3 + d_4 + \frac{2d_6}{1-\rho})(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}\alpha^2 \qquad (88)$$

$$+ 2d_1(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2}\alpha^2 + 2d_2\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}\alpha^2,$$

41

where we recall that the coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 of the Lyapunov function (23) are given by

$$d_0 = 1, d_1 = \frac{8C_{g,x\theta}^2 \tau \alpha}{\mu_g \lambda}, d_2 = (\frac{8C_{g,x\theta}^2 \tau \alpha}{\mu_g \lambda} q_x + L_{fg,x}\tau)\frac{\alpha}{\omega_\theta \beta}, d_3 = \frac{\tau \alpha}{2\gamma},$$

$$d_4 = (L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x)\frac{64\gamma \alpha^2}{(1-\rho)^4}, d_5 = 2(L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x)\frac{2\alpha}{\tau(1-\rho)}, \quad (89)$$

$$d_6 = (L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x)\frac{16\alpha^2}{(1-\rho)^3}.$$

Next, we proceed to find the sufficient conditions for the step-sizes to make the conditions (84)-(87) hold. To this end, we first recall that like (52) the term $L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x$ in d_4 can be simplified as follows:

$$L_{fg,x}d_3 r_z + d_1 q_x + d_2 p_x = \underbrace{(L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2})}_{\triangleq \varphi} \tau. \quad (90)$$

Then, d_2, d_4, d_5 and d_6 can be simplified as follows:

$$d_2 = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})\frac{\tau \alpha}{\omega_\theta \beta}, d_4 = \frac{64\varphi\tau\gamma\alpha^2}{(1-\rho)^4}, d_5 = \frac{4\varphi\alpha}{(1-\rho)}, d_6 = \frac{16\varphi\tau\alpha^2}{(1-\rho)^3}.$$

Now, we can derive that the condition (84) holds if

$$\alpha \leqslant u_1' \triangleq \min\left\{\frac{1}{2\tau L}, \frac{(1-\rho)^2}{32\varphi^{1/2}(\tau L)^{1/2}}\right\}, \quad (91)$$

$$\lambda \geqslant \frac{48L_{v^*}C_{g,x\theta}}{\mu_g}\tau\alpha, \quad (92)$$

$$\beta \geqslant \frac{6L_{\theta^*}}{\omega_\theta}(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/2}\tau\alpha, \quad (93)$$

$$\gamma \geqslant 4L\tau\alpha, \quad (94)$$

where φ is given by (90) and $0 < \tau < 1$. Besides, based on the condition (94), a sufficient condition to make the inequalities (85)-(87) hold is

$$\gamma \leqslant u_2' = \min\left\{\frac{4}{5}\frac{C_{g,x\theta}^2}{\sigma_{g,\theta\theta}^2}, \frac{(1-\rho)^{4/3}(\tau L)^{1/3}}{6\varphi^{1/3}}, \frac{(1-\rho)^{4/3}(\tau L)^{1/3}C_{g,x\theta}^{2/3}}{3\varphi^{1/3}\sigma_{g,\theta\theta}^{2/3}}, \right.$$

$$\left. \frac{(1-\rho)^{4/3}(L\tau)^{1/3}}{4\varphi^{1/3}L_{fg,x}^{1/3}}(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/3}, \frac{(1-\rho)^{4/3}(L\tau)^{1/3}}{4L_{fg,x}^{1/3}}\right\}, \quad (95)$$

$$\lambda \leqslant \frac{\mu_g}{20\sigma_{g,\theta\theta}^2}.$$

By combining Lemmas 15-18 and the conditions (92)-(95), we have the following condition for the step-sizes λ, β, γ

$$\min\left\{\frac{1}{\mu_g}, \frac{\mu_g}{20\sigma_{g,\theta\theta}^2}\right\} > \lambda \geqslant \frac{48L_{v^*}C_{g,x\theta}}{\mu_g}\tau\alpha, \tag{96}$$

$$\min\left\{\frac{2}{\mu_g + L_{g,\theta}}, \frac{\mu_g + L_{g,\theta}}{2\mu_g L_{g,\theta}}\right\} > \beta \geqslant \frac{6L_{\theta^*}}{\omega_\theta}\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x}\right)^{1/2}\tau\alpha, \tag{97}$$

$$\min\{1, u_2'\} > \gamma \geqslant 4L\tau\alpha, \tag{98}$$

where the selection for λ, β and γ can be guaranteed to be nonempty, if the step-size α satisfy the following condition

$$\alpha \leqslant \frac{\min\{1, u_2'\}}{8L\tau}, \alpha \leqslant u_3' \triangleq \min\left\{\frac{\min\{1, \frac{\mu_g^2}{8\sigma_{g,x\theta}^2}\}}{96\tau L_{v^*}C_{g,x\theta}}, \frac{\min\left\{\frac{2}{\mu_g + L_{g,\theta}}, \frac{\mu_g + L_{g,\theta}}{2\mu_g L_{g,\theta}}\right\}w_\theta}{12\tau\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x}\right)^{1/2}L_{\theta^*}}\right\}. \tag{99}$$

As a result, to ensure that the conditions (84)-(87) hold, with $0 < \tau < 1$, we can properly select the step-size α as follows:

$$\alpha \leqslant u' \triangleq \min\left\{u_1', \frac{\min\{1, u_2'\}}{8\tau L}, u_3'\right\}, \tag{100}$$

where u_1', u_2', and u_3' are defined by (91), (95), and (93), respectively, and are clearly independent of λ, β, and γ. Subsequently, we can determine the step-sizes λ, β and γ according to (96)-(98). Therefore, under the condition (96)-(100), it holds that

$$\frac{\mathbb{E}[V^{k+1}]}{\tau} \leqslant \frac{\mathbb{E}[V^k]}{\tau} - \frac{d_0}{2}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] + \alpha^2\sigma_{GT}^2, \tag{101}$$

where σ_{GT}^2 is denoted as

$$\sigma_{GT}^2 \triangleq \left(\frac{1}{m}\frac{d_3}{\tau} + \frac{d_4}{\tau} + \frac{2d_6}{\tau(1-\rho)}\right)(\sigma_{f,x}^2 + 2M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2} + 2\frac{d_1}{\tau}\left(\sigma_{f,\theta}^2 + 2M^2\sigma_{g,\theta\theta}^2\right)\frac{\lambda^2}{\alpha^2} + 2\frac{d_2}{\tau}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}, \tag{102}$$

where $d_0 = 1$, $d_1 = \frac{8C_{g,x\theta}^2\tau\alpha}{\mu_g\lambda}$, $d_2 = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})\frac{\tau\alpha}{\omega_\theta\beta}$, $d_3 = \frac{\tau\alpha}{2\gamma}$, $d_4 = \frac{64\varphi\tau\gamma\alpha^2}{(1-\rho)^4}$, $d_5 = \frac{4\varphi\alpha}{(1-\rho)}$, $d_6 = \frac{16\varphi\tau\alpha^2}{(1-\rho)^3}$ with $\varphi = (L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2})(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2})$. In what follows, by telescoping the inequality (101) from $k = 0$ to K and combining the fact $d_0 = 1$, it follows that

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant \frac{2\left(V^0 - V^K\right)}{\tau\alpha(K+1)} + 2\alpha\sigma_{GT}^2. \tag{103}$$

This completes the proof. ∎

B.4 Proof of Corollary 10

Similar to the proof of Corollary 7, when the conditions (96)-(98) and (100) hold and the step-sizes γ, λ, β are taken as $\gamma = c_\gamma \alpha, \lambda = c_\lambda \alpha, \beta = c_\beta \alpha$ with positive parameters $c_\gamma, c_\lambda, c_\beta$ being independent of the terms K and $1 - \rho$ and satisfying

$$c_\lambda \geqslant \frac{48 L_{v^*} C_{g,x\theta}}{\mu_g} \tau, c_\beta \geqslant \frac{6 L_{\theta^*}}{\omega_\theta} (\frac{32 C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/2} \tau, c_\gamma \geqslant 4 L \tau, \tag{104}$$

it follows from (102) that $d_0 = 1, d_1 = \frac{8 C_{g,x\theta}^2 \tau}{\mu_g c_\lambda}, d_2 = (\frac{32 C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x}) \frac{\tau}{\omega_\theta c_\beta}, d_3 = \frac{\tau}{2 c_\gamma}$ are independent of the step-size α and $1 - \rho$. When the parameters $c_\gamma, c_\lambda, c_\beta$ satisfy (104) and their values are properly selected, the conditions (96)-(98) always hold. For example, a sufficient condition for the upper bounds of the parameters $c_\gamma, c_\lambda, c_\beta$ could be $\frac{96 L_{v^*} C_{g,x\theta}}{\mu_g} \tau > c_\lambda$, $\frac{12 L_{\theta^*}}{\omega_\theta} (\frac{32 C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g^2} + L_{fg,x})^{1/2} \tau > c_\beta$, and $8 L \tau \alpha > c_\gamma$. In addition, from (89) we note that $d_4 = O(\gamma^2 \alpha d_3), d_6 = O(\gamma \alpha d_3)$, which ensures that $d_4 \ll \frac{1}{m} d_3, d_6 \ll \frac{1}{m} d_3$ under a large number of iterations K. Thus, the term σ_{GT} in (102) can be further derived as

$$\begin{aligned}
\sigma_{GT} &\leqslant \sqrt{\frac{4}{m} \frac{d_3}{\tau} (\sigma_{f,x}^2 + 2 M^2 \sigma_{g,x\theta}^2) \frac{\gamma^2}{\alpha^2} + 2 \frac{d_1}{\tau} (\sigma_{f,\theta}^2 + 2 M^2 \sigma_{g,\theta\theta}^2) \frac{\lambda^2}{\alpha^2} + 2 \frac{d_2}{\tau} \sigma_{g,\theta}^2 \frac{\beta^2}{\alpha^2}} \\
&\leqslant \sqrt{2} (\underbrace{\sqrt{\frac{d_1}{\tau}} (\sigma_{f,\theta} + 2 M \sigma_{g,\theta\theta}) c_\lambda + \sqrt{\frac{d_2}{\tau}} \sigma_{g,\theta} c_\beta}_{\triangleq \sigma_{\mathrm{p}}}) + \frac{2}{\sqrt{m}} \underbrace{\sqrt{\frac{d_3}{\tau}} (\sigma_{f,x} + 2 M \sigma_{g,x\theta}) c_\gamma}_{\triangleq \sigma_c} \triangleq \hat{\sigma}_{GT},
\end{aligned} \tag{105}$$

where the last inequality follows from the triangle inequality. By (71), we have that $\sigma_{\mathrm{p}} = \mathcal{O}(\sigma_{f,\theta} + \sigma_{g,\theta\theta} + \sigma_{g,\theta})$ and $\sigma_c = \mathcal{O}(\sigma_{f,x} + \sigma_{g,x\theta})$. By (91), (95), (93), and (100), we have that $u' = \mathcal{O}((1 - \rho)^2)$. Then, by this fact and the condition that $\gamma < 1$ in (98), the coefficients d_4, d_5, d_6 in (102) can be bounded as follows

$$\begin{aligned}
d_4 &= \frac{64 \varphi \tau \gamma \alpha^2}{(1 - \rho)^4} \leqslant \frac{64 \varphi \tau (u')^2}{(1 - \rho)^4} \triangleq \hat{d}_4 = O(1), \\
d_5 &= \frac{4 \varphi \alpha}{(1 - \rho)} \leqslant \frac{4 \varphi u'}{(1 - \rho)} \triangleq \hat{d}_5 = O(1 - \rho), \\
d_6 &= \frac{16 \varphi \tau \alpha^2}{(1 - \rho)^3} \leqslant \frac{16 \varphi \tau (u')^2}{(1 - \rho)^3} \triangleq \hat{d}_6 = O(1 - \rho).
\end{aligned} \tag{106}$$

Then, for the term $\frac{V^0 - V^K}{\tau}$ in (103), it follows that

$$\begin{aligned}
\frac{V^0 - V^K}{\tau} &\leqslant \frac{V^0}{\tau} \\
&\leqslant \frac{1}{\tau} d_0 \Phi(\bar{x}^0) + \frac{1}{\tau} d_1 \frac{1}{m} \|v^0 - v^*(\bar{x}^0)\|^2 + \frac{1}{\tau} d_2 \frac{1}{m} \|\theta^0 - \theta^*(\bar{x}^0)\|^2 + \frac{1}{\tau} d_3 \|\nabla \Phi(\bar{x}^0) - \bar{z}^0\|^2 \\
&\quad + \frac{1}{\tau} \hat{d}_4 \frac{1}{m} \|\nabla \tilde{\Phi}(\bar{x}^0) - z^0\|^2 + \frac{1}{\tau} \hat{d}_5 \frac{1}{m} \|x^0 - 1_m \otimes \bar{x}^0\|^2 + \frac{1}{\tau} \hat{d}_6 \frac{1}{m} \|y^0 - 1_m \otimes \bar{y}^0\|^2 \\
&\triangleq \hat{V}^0 = O(1 - \rho).
\end{aligned} \tag{107}$$

In what follows, letting

$$a_0' \triangleq 2\hat{V}^0, a_1' \triangleq 2\hat{\sigma}_{GT}^2, \tag{108}$$

and combining the inequality (105), the inequality (103) can be further derived as:

$$\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}[\|\nabla \Phi(\bar{x}^k)\|^2] \leqslant a_0' \frac{1}{\alpha(K+1)} + a_1' \alpha. \tag{109}$$

If we take the step-size $\alpha = \min\left\{u', \left(\frac{a_0'}{a_1'(K+1)}\right)^{\frac{1}{2}}\right\}$ and the step-sizes $\gamma = c_\gamma \alpha, \lambda = c_\lambda \alpha, \beta = c_\beta \alpha$ with $c_\gamma, c_\lambda, c_\beta$ being independent of the terms K and $1-\rho$, then we have

$$\begin{aligned}
\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}[\|\Phi(\bar{x}^k)\|^2] &\leqslant \frac{a_0'}{u'(K+1)} + 2\left(\frac{a_1' a_0'}{K+1}\right)^{\frac{1}{2}} \\
&\leqslant \frac{2\hat{V}^0}{u'(K+1)} + \frac{4\hat{\sigma}_{GT}\sqrt{\hat{V}^0}}{\sqrt{K+1}}.
\end{aligned} \tag{110}$$

Now, we proceed to analyze how the parameters in (80) depend on the terms κ and $1-\rho$. Firstly, combining the fact that $\varphi = O(\kappa^6)$ by (90), we set $\tau = \mathcal{O}(\kappa^{-4}), u' = \mathcal{O}(\kappa^{-3}(1-\rho)^2)$ and $c_\lambda = \mathcal{O}(1), c_\beta = \mathcal{O}(1), c_\gamma = \mathcal{O}(1)$ in terms of κ and $1-\rho$. Furthermore, when we initialize the outer-level variables as $x_i^0 = x_j^0, \forall i, j \in \mathcal{V}$, i.e., $\|x^0 - 1_m \otimes \bar{x}^0\|^2 = 0$, we can derive that $\hat{V}^0 = O(\kappa^5)$, and $\hat{V}^0$ is independent of the term $\frac{1}{1-\rho}$ by (107). Thus, with the result $\hat{\sigma}_{GT} = \mathcal{O}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})$ in (105), it follows from (110) that

$$\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}[\|\nabla \Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{\kappa^8}{(1-\rho)^2 K} + \frac{\kappa^{\frac{5}{2}}}{\sqrt{K}}\left(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}}\right)\right), \tag{111}$$

where, similar with Corollary 7, the terms σ_{p} and σ_{c} can be derived as $\sigma_{\mathrm{p}} = \mathcal{O}(\kappa^{\frac{1}{2}}\sigma_{f,\theta} + \kappa^{\frac{3}{2}}\sigma_{g,\theta\theta} + \kappa^{\frac{5}{2}}\sigma_{g,\theta})$ and $\sigma_{\mathrm{c}} = \mathcal{O}(\sigma_{f,x} + \kappa\sigma_{g,x\theta})$ by (105). This completes the proof. $\blacksquare$

Appendix C. Proof of Supporting Propositions

C.1 Proof of Proposition 3

Lipschitz continuity of $\theta_i^*(x)$. Recalling the definition of $\theta_i^*(x)$, the expression of $\theta_i^*(x)$ is given by (Ghadimi and Wang, 2018)

$$\nabla \theta_i^*(x) = -\nabla_{x\theta}^2 g_i\left(x, \theta_i^*(x)\right) \left[\nabla_{\theta\theta}^2 g_i\left(x, \theta_i^*(x)\right)\right]^{-1}.$$

Following the strong convexity of g_i and bounded Jacibian of $\nabla_{x\theta}^2 g_i$ in Assumption 5, we have

$$\|\nabla \theta_i^*(x)\| = \left\|\nabla_{x\theta}^2 g_i\left(x, \theta_i^*(x)\right)\left[\nabla_{\theta\theta}^2 g_i\left(x, \theta_i^*(x)\right)\right]^{-1}\right\| \leqslant \frac{C_{g,x\theta}}{\mu_g}, \tag{112}$$

which implies that for any x, x',

$$\left\|\theta_i^*(x) - \theta_i^*(x')\right\| \leqslant \underbrace{\frac{C_{g,x\theta}}{\mu_g}}_{\triangleq L_{\theta^*}} \left\|x - x'\right\|. \tag{113}$$

Lipschitz continuity of $v_i^*(x)$. Recalling the definition of $v_i(x, \theta)$, we have that $v_i(x, \theta)$ admits the following expression

$$v_i(x, \theta) = \left[\nabla_{\theta\theta}^2 g_i(x, \theta)\right]^{-1} \nabla_x f_i(x, \theta).$$

Letting x and x' be any two points in $\mathbb{R}^n$ and considering the case $\theta = \theta_i^*(x)$ and $\theta' = \theta_i^*(x')$, it follows that

$$
\begin{aligned}
&\|v_i(x, \theta) - v_i(x', \theta')\| \\
&\overset{(a)}{\leqslant} \|[\nabla_{\theta\theta}^2 g_i(x, \theta)]^{-1} \nabla_\theta f_i(x, \theta) - [\nabla_{\theta\theta}^2 g_i(x, \theta)]^{-1} \nabla_\theta f_i(x', \theta')\| \\
&\quad + \|[\nabla_{\theta\theta}^2 g_i(x, \theta)]^{-1} \nabla_\theta f_i(x', \theta') - [\nabla_{\theta\theta}^2 g_i(x', \theta')]^{-1} \nabla_\theta f_i(x', \theta')\| \\
&\overset{(b)}{\leqslant} \frac{1}{\mu_g} \|\nabla_\theta f_i(x, \theta) - \nabla_\theta f_i(x', \theta')\| \\
&\quad + C_{f,\theta} \|[\nabla_{\theta\theta}^2 g_i(x', \theta')]^{-1} (\nabla_{\theta\theta}^2 g_i(x, \theta) - \nabla_{\theta\theta}^2 g_i(x', \theta'))[\nabla_{\theta\theta}^2 g_i(x, \theta)]^{-1}\| \\
&\overset{(c)}{\leqslant} \frac{1}{\mu_g} L_{f,\theta}(\|x - x'\| + \|\theta - \theta'\|) + \frac{C_{f,\theta} L_{g,\theta\theta}}{\mu_g^2}(\|x - x'\| + \|\theta - \theta'\|) \\
&= \underbrace{\left(\frac{L_{f,\theta}}{\mu_g} + \frac{C_{f,\theta} L_{g,\theta\theta}}{\mu_g^2}\right)}_{\triangleq L_v}(\|x - x'\| + \|\theta - \theta'\|),
\end{aligned}
\tag{114}
$$

where step (a) follows from the triangle inequality; step (b) uses the upper bound of the Hessian-inverse matrix with the parameter $\frac{1}{\mu_g}$ related to the strong convexity constant as well as the boundness of $\nabla_\theta f_i(x, \theta_i^*(x))$ under the case $\theta = \theta_i^*(x)$ and $\theta' = \theta_i^*(x')$ with Assumption 2; step (c) follows from $\nabla_{\theta\theta}^2 g_i$ in Assumption 5. Then, recalling the definition of $v_i^*(x)$, we know that $v_i^*(x) = v_i(x, \theta_i^*(x))$. Thus, given any two x, x', by taking $\theta = \theta_i^*(x)$ and $\theta' = \theta_i^*(x')$ in (114), it follows that

$$
\begin{aligned}
\|v_i^*(x) - v_i^*(x')\| &= \|v_i(x, \theta_i^*(x)) - v_i(x', \theta_i^*(x'))\| \\
&\leqslant L_v\left(\|x - x'\| + \|\theta_i^*(x) - \theta_i^*(x')\|\right) \\
&\leqslant L_v(1 + L_{\theta^*})\|x - x'\|,
\end{aligned}
\tag{115}
$$

where the last inequality is due to the Lipschitz continuity of $\theta_i^*(x)$ (c.f., (113)).

Lipschitz continuity of $\nabla\Phi(x)$. Letting x and x' be any two points in $\mathbb{R}^n$ as well as $\theta = \theta_i^*(x)$ and $\theta' = \theta_i^*(x')$ and following the definition of $\bar{\nabla} f_i(x, \theta)$ and the triangle inequality, we have

$$
\begin{aligned}
\|\bar{\nabla} f_i(x, \theta) - \bar{\nabla} f_i(x', \theta')\| &\leqslant \|\nabla_x f_i(x, \theta) - \nabla_x f_i(x', \theta')\| \\
&\quad + \|\nabla_{x\theta}^2 g_i(x, \theta)(v_i(x, \theta) - v_i(x', \theta'))\| \\
&\quad + \|(\nabla_{x\theta}^2 g_i(x, \theta) - \nabla_{x\theta}^2 g_i(x', \theta'))v_i(x', \theta')\|.
\end{aligned}
\tag{116}
$$

Utilizing Lipschitz continuity of $\nabla_x f_i$ in Assimption 2, the boundness of $\nabla_{x\theta}^2 g_i$ in Assimption 3 and Lipschitz continuity of $v_i(x, \theta)$ (c.f., (115)), the first two terms on right hand of (116)

can be bounded by

$$\left\|\nabla_x f_i\left(x,\theta\right) - \nabla_x f_i\left(x',\theta'\right)\right\| + \left\|\nabla_{x\theta}^2 g_i\left(x,\theta\right)\left(v_i\left(x,\theta\right) - v_i\left(x',\theta'\right)\right)\right\|$$
$$\leqslant \left(L_{f,x} + C_{g,x\theta}L_v\right)\left(\left\|x - x'\right\| + \left\|\theta - \theta'\right\|\right). \tag{117}$$

Note that for the case $\theta = \theta_i^*(x)$ and $\theta' = \theta_i^*(x')$, we have $\|v_i\left(x',\theta'\right)\| \leqslant \frac{C_{f,\theta}}{\mu_g}$ by employing the boundness of the Hessian-inverse matrix $\nabla_{\theta\theta}^2 g_i$ and the gradient $\nabla_\theta f_i(x,\theta_i^*(x))$ in Assumptions 2 and 3. Then, combining the Lipschitz continuity of $\nabla_{x\theta}^2 g_i$ in Assumption 3, one can obtain

$$\left\|\left(\nabla_{x\theta}^2 g_i\left(x,\theta\right) - \nabla_{x\theta}^2 g_i\left(x',\theta'\right)\right)v_i\left(x',\theta'\right)\right\| \leqslant \frac{C_{f,\theta}L_{g,x\theta}}{\mu_g}\left(\left\|x - x'\right\| + \left\|\theta - \theta'\right\|\right).$$

Then the inequality $\left\|\bar{\nabla} f_i\left(x,\theta\right) - \bar{\nabla} f_i\left(x',\theta'\right)\right\| \leqslant L_f\left(\left\|x - x'\right\| + \left\|\theta - \theta'\right\|\right)$ can be derived by integrating above inequalities, with L_f defined in (15). It is noted from the expression of $\nabla\Phi_i(x)$ that $\nabla\Phi_i(x) = \bar{\nabla} f_i\left(x,\theta_i^*(x)\right)$. Then, given any two x, x', by taking $\theta = \theta_i^*(x)$ and $\theta' = \theta_i^*(x')$ in (116), we can employ the result (116) to derive the following inequality

$$\left\|\nabla\Phi_i(x) - \nabla\Phi_i(x')\right\| \leqslant \left(L_{f,x} + C_{g,x\theta}L_v + \frac{C_{f,\theta}L_{g,x\theta}}{\mu_g}\right)\left(\left\|x - x'\right\| + \left\|\theta_i^*(x) - \theta_i^*(x')\right\|\right)$$
$$\leqslant \underbrace{\left(L_{f,x} + C_{g,x\theta}L_v + \frac{C_{f,\theta}L_{g,x\theta}}{\mu_g}\right)\left(1 + L_{\theta^*}\right)}_{\triangleq L}\left\|x - x'\right\|, \tag{118}$$

where the last step uses the Lipschitz continuity of $v_i^*(x)$ (c.f., (115)). Combining the fact that $\Phi(x) = \frac{1}{m}\sum_{i=1}^m \Phi_i(x)$, we can derive that $\nabla\Phi(x)$ is also L-Lipschitz continuous. This completes the proof. ∎

C.2 Proof of Proposition 4

By the strong convexity of g_i in θ in Assumption 3 and the boundedness of $\nabla f_\theta(x,\theta_i^*(x))$ in Assumption 2, it follows that

$$\left\|v_i\left(x,\theta_i^*(x)\right)\right\| = \left\|\left[\nabla_{\theta\theta}^2 g_i\left(x,\theta_i^*(x)\right)\right]^{-1}\nabla_\theta f_i\left(x,\theta_i^*(x)\right)\right\| \leqslant \frac{C_{f,\theta}}{\mu_g}.$$

This completes the proof. ∎

Appendix D. Proof of Supporting Lemmas

D.1 Proof of Lemma 5

By the definition of $\Phi_i(x)$, we can compute its gradient as

$$\nabla\Phi_i(x) = \nabla_x f_i(x,\theta_i^*(x)) + \nabla\theta_i^*(x)\nabla_\theta f_i(x,\theta_i^*(x)).$$

47

Then the term $\sum_{i=1}^{m} \|\nabla\Phi_i(x) - \nabla\Phi(x)\|^2$ can be bounded as

$$
\sum_{i=1}^{m} \|\nabla\Phi_i(x) - \nabla\Phi(x)\|^2 \leqslant 2\sum_{i=1}^{m} \left\|\nabla\theta_i^*(x)\nabla_\theta f_i(x,\theta_i^*(x)) - \frac{1}{m}\sum_{j=1}^{m}\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))\right\|^2
$$
$$
+ 2\sum_{i=1}^{m}\left\|\nabla_x f_i(x,\theta_i^*(x)) - \frac{1}{m}\sum_{j=1}^{m}\nabla_x f_j(x,\theta_j^*(x))\right\|^2.
$$
$$(119)$$

For the first term on the right hand of (119), it follows that

$$
\sum_{i=1}^{m}\left\|\nabla\theta_i^*(x)\nabla_\theta f_i(x,\theta_i^*(x)) - \frac{1}{m}\sum_{j=1}^{m}\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))\right\|^2
$$
$$
\overset{(a)}{\leqslant} 2\sum_{i=1}^{m}\left\|\nabla\theta_i^*(x)\nabla_\theta f_i(x,\theta_i^*(x)) - \frac{1}{m}\sum_{j=1}^{m}\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right\|^2
$$
$$
+ 2\sum_{i=1}^{m}\left\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x)) - \nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\right\|^2 \qquad (120)
$$
$$
\overset{(b)}{\leqslant} 2\max_i\left\{\|\nabla\theta_i^*(x)\|^2\right\}b_f^2 + 2\sum_{i=1}^{m}\left\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x)) - \nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\right\|^2,
$$

where in step (a), we add and subtract the term $\frac{1}{m}\sum_{i=1}^{m}\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))$ and apply Young's inequality; step (b) leverages the bounded data heterogeneity stated in Assumption 4. In what follows, we derive an upper bound for the last term on the right-hand side of (120) as follows:

$$
\sum_{i=1}^{m}\left\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x)) - \nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\right\|^2
$$
$$
\overset{(a)}{\leqslant} 2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\nabla\theta_i^*(x)\|^2\left\|\nabla_\theta f_j(x,\theta_j^*(x)) - \nabla_\theta f_j(x,\theta_i^*(x))\right\|^2
$$
$$
+ 2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|\nabla_\theta f_j(x,\theta_j^*(x))\right\|^2\left\|\nabla\theta_j^*(x) - \nabla\theta_i^*(x)\right\|^2 \qquad (121)
$$
$$
\overset{(b)}{\leqslant} 2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}L_{f,\theta}^2\|\nabla\theta_i^*(x)\|^2\left\|\theta_j^*(x) - \theta_i^*(x)\right\|^2 + 2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}C_{f,\theta}^2\left\|\nabla\theta_j^*(x) - \nabla\theta_i^*(x)\right\|^2
$$
$$
\leqslant 2\max_i\left\{L_{f,\theta}^2\|\nabla\theta_i^*(x)\|^2\right\}\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|\theta_j^*(x) - \theta_i^*(x)\right\|^2 + 2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}C_{f,\theta}^2\left\|\nabla\theta_j^*(x) - \nabla\theta_i^*(x)\right\|^2,
$$

where step (a) adds and subtracts the term $\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_j^*(x))$, applies Young's inequality, and then proceeds by using Jensen's inequality and the Cauchy–Schwarz inequality. step (b) employs Lipschitz continuity of $\nabla_\theta f_i$ and the boundness of $\nabla_\theta f_i(x,\theta_i^*(x))$ in Assumption 2.

For the last term of (121), Noting that $\nabla \theta_i^*(x) = -\nabla_{x\theta}^2 g_i\left(x, \theta_i^*(x)\right)\left[\nabla_{\theta\theta} g_i\left(x, \theta_i^*(x)\right)\right]^{-1}$, we arrive at

$$
\|\nabla \theta_j^*(x) - \nabla \theta_i^*(x)\|^2
$$

$$
\leqslant 2\|\nabla_{x\theta}^2 g_j\left(x, \theta_j^*(x)\right)\left[\nabla_{\theta\theta} g_j(x, \theta_j^*(x))\right]^{-1}\left[\nabla_{\theta\theta}^2 g_i\left(x, \theta_i^*(x)\right) - \nabla_{\theta\theta}^2 g_j\left(x, \theta_j^*(x)\right)\right]\left[\nabla_{\theta\theta}^2 g_i(x, \theta_i^*(x))\right]^{-1}\|^2
$$

$$
+ 2\|\left[\nabla_{x\theta}^2 g_j\left(x, \theta_j^*(x)\right) - \nabla_{x\theta}^2 g_i\left(x, \theta_i^*(x)\right)\right]\left[\nabla_{\theta\theta}^2 g_i\left(x, \theta_i^*(x)\right)\right]^{-1}\|^2
$$

$$
\leqslant 2\frac{C_{g,x\theta}^2}{\mu_g^4}\|\nabla_{\theta\theta}^2 g_i\left(x, \theta_i^*(x)\right) - \nabla_{\theta\theta}^2 g_j\left(x, \theta_j^*(x)\right)\|^2 + 2\frac{1}{\mu_g^2}\|\nabla_{x\theta}^2 g_i\left(x, \theta_i^*(x)\right) - \nabla_{x\theta}^2 g_j\left(x, \theta_j^*(x)\right)\|^2
$$

$$
\leqslant (4\frac{C_{g,x\theta}^2 L_{g,\theta\theta}^2}{\mu_g^4} + 4\frac{L_{g,x\theta}^2}{\mu_g^2})\|\theta_i^*(x) - \theta_j^*(x)\|^2 + 4\frac{C_{g,x\theta}^2}{\mu_g^4}\|\nabla_{\theta\theta}^2 g_i(x, \theta_j^*(x)) - \nabla_{\theta\theta}^2 g_j(x, \theta_j^*(x))\|^2
$$

$$
+ 4\frac{1}{\mu_g^2}\|\nabla_{x\theta}^2 g_i(x, \theta_j^*(x)) - \nabla_{x\theta}^2 g_j(x, \theta_j^*(x))\|^2, \tag{122}
$$

which implies that the last term of (121) can be bounded by

$$
2C_{f,\theta}^2 \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\|\nabla \theta_j^*(x) - \nabla \theta_i^*(x)\right\|^2 \tag{123}
$$

$$
\leqslant (8\frac{C_{f,\theta}^2 C_{g,x\theta}^2 L_{g,\theta\theta}^2}{\mu_g^4} + 8\frac{C_{f,\theta}^2 L_{g,x\theta}^2}{\mu_g^2})\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\|\theta_i^*(x) - \theta_j^*(x)\right\|^2 + 8\frac{C_{f,\theta}^2 C_{g,x\theta}^2}{\mu_g^4}b_g^2 + 8\frac{C_{f,\theta}^2}{\mu_g^2}b_g^2.
$$

As for the second term on the right hand of (119), we bound it by

$$
\sum_{i=1}^{m} \|\nabla_x f_i(x, \theta_i^*(x)) - \frac{1}{m} \sum_{j=1}^{m} \nabla_x f_j(x, \theta_j^*(x))\|^2 \tag{124}
$$

$$
\overset{(a)}{\leqslant} 2\sum_{i=1}^{m} \|\nabla_x f_i(x, \theta_i^*(x)) - \frac{1}{m} \sum_{j=1}^{m} \nabla_x f_j(x, \theta_i^*(x))\|^2
$$

$$
+ 2\sum_{i=1}^{m} \|\frac{1}{m} \sum_{j=1}^{m} \left(\nabla_x f_j(x, \theta_j^*(x)) - \nabla_x f_j(x, \theta_i^*(x))\right)\|^2
$$

$$
\overset{(b)}{\leqslant} 2b_f^2 + 2\frac{1}{m} \sum_{i=1}^{m}\sum_{j=1}^{m} \|\nabla_x f_j(x, \theta_j^*(x)) - \nabla_x f_j(x, \theta_i^*(x))\|^2 \overset{(c)}{\leqslant} 2b_f^2 + 2L_{f,x}^2 \frac{1}{m} \sum_{i=1}^{m}\sum_{i=j}^{m} \|\theta_j^*(x) - \theta_i^*(x)\|^2,
$$

where step (a) adds and subtracts the term $\frac{1}{m}\sum_{j=1}^m \nabla_x f_j(x, \theta_i^*(x))$ and applies Young's inequality; step (b) employs the bounded data heterogeneity in Assumption 4; step (c) comes from Lipschitz continuity of $\nabla_x f_i$ in Assumption 2. In addition, utilizing the strong convexity of g_i in θ in Assumption 3, we further derive that

$$
\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\|\theta_j^*(x) - \theta_i^*(x)\right\|^2 \leqslant \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{\mu_g^2}\left\|\nabla_\theta g_i(x, \theta_j^*(x)) - \nabla_\theta g_i(x, \theta_i^*(x))\right\|^2 \tag{125}
$$

$$
\overset{(a)}{=} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{\mu_g^2}\left\|\nabla_\theta g_i(x, \theta_j^*(x)) - \nabla_\theta g_j(x, \theta_j^*(x))\right\|^2 \leqslant \frac{1}{\mu_g^2}b_g^2,
$$

where step (a) uses the fact that $\nabla_\theta g_i(x, \theta_i^*(x)) = 0, i \in \mathcal{V}$.

Then, by substituting the results (120), (121), (122), (123), (124), (125) into (119) and rearranging the terms, we have

$$\sum_{i=1}^{m} \|\nabla\Phi_i(x) - \nabla\Phi(x)\|^2$$

$$\leqslant 4b_f^2 + 4\max_i\{\|\nabla\theta_i^*(x)\|^2\}b_f^2 + 16\frac{C_{f,\theta}^2 C_{g,x\theta}^2}{\mu_g^4}b_g^2 + 16\frac{C_{f,\theta}^2}{\mu_g^2}b_g^2$$

$$+ 4\left(L_{f,x}^2 + \max_i\{L_{f,\theta}^2\|\nabla\theta_i^*(x)\|^2\} + 4\frac{C_{f,\theta}^2 C_{g,x\theta}^2 L_{g,\theta\theta}^2}{\mu_g^4} + 4\frac{C_{f,\theta}^2 L_{g,x\theta}^2}{\mu_g^2}\right)\frac{1}{\mu_g^2}b_g^2 \qquad (126)$$

$$\leqslant \underbrace{(4 + 4\frac{C_{g,x\theta}^2}{\mu_g^2})b_f^2}_{\triangleq C_1(\mu_g, C_{g,x\theta})} + \underbrace{4\left(4\frac{C_{f,\theta}^2 C_{g,x\theta}^2}{\mu_g^4} + 4\frac{C_{f,\theta}^2}{\mu_g^2} + \frac{L_{f,x}^2}{\mu_g^2} + \frac{L_{f,\theta}^2 C_{g,x\theta}^2}{\mu_g^4} + 4\frac{C_{f,\theta}^2 C_{g,x\theta}^2 L_{g,\theta\theta}^2}{\mu_g^6} + 4\frac{C_{f,\theta}^2 L_{g,x\theta}^2}{\mu_g^4}\right)b_g^2}_{\triangleq C_2(\mu_g, L_{f,x}, L_{f,\theta}, L_{g,x\theta}, L_{g,\theta\theta}, C_{f,\theta}, C_{g,x\theta})},$$

where the last step is due to the bound of $\nabla\theta_i^*(x)$ as follows

$$\|\nabla\theta_i^*(x)\|^2 = \|\nabla_{x\theta}g_i(x, \theta_i^*(x))[\nabla_{\theta\theta}g_i(x, \theta_i^*(x))]^{-1}\|^2 \leqslant \frac{C_{g,x\theta}^2}{\mu_g^2}. \qquad (127)$$

This completes the proof. ∎

D.2 Proof of Lemma 15

For ease of presentation, we recall that the update of y^{k+1} in the case with gradient tracking scheme in Algorithm 1 is given by

$$y^{k+1} = \mathcal{W}y^k + z^{k+1} - z^k.$$

Let the matrix $J \triangleq \frac{1_m^{\mathrm{T}}}{m} \otimes I_n$ denote the average operator among nodes. Since the weighted matrix $\mathcal{W}$ is doubly stochastic, multiplying the matrix J in both sides of above equality yields

$$\bar{y}^{k+1} = \bar{y}^k + \bar{z}^{k+1} - \bar{z}^k. \qquad (128)$$

By applying induction, we can establish that $\bar{y}^k = \bar{z}^k$ when the initial condition is $y^0 = z^0$. This implies that each node is capable of tracking the full gradient $\bar{z}^k$. On the other hand, considering the case with the local gradient scheme in Algorithm 1, we can directly observe that $y^k = z^k$ based on (39). Thus, in both the local gradient and gradient tracking schemes, it can be deduced that $\bar{y}^k = \bar{z}^k$. Furthermore, by multiplying the matrix J on both sides of (38a), we can determine the average state of the update (38a) across all nodes as follows:

$$\bar{x}^{k+1} = \bar{x}^k - \tau\alpha\bar{y}^k. \qquad (129)$$

Since the overall objective function Φ is smooth by Proposition 3, it holds that

$$\mathbb{E}[\Phi(\bar{x}^{k+1})|\mathcal{F}^k] \overset{(a)}{\leqslant} \Phi(\bar{x}^k) - \tau\alpha\mathbb{E}[\langle\nabla\Phi(\bar{x}^k), \bar{y}^k\rangle|\mathcal{F}^k] + \frac{\tau^2\alpha^2 L}{2}\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k]$$

$$\overset{(b)}{\leqslant} \Phi(\bar{x}^k) + \frac{\tau\alpha}{2}\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2|\mathcal{F}^k] \qquad (130)$$

$$- \frac{\tau\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 - \frac{\tau\alpha}{2}(1 - \tau\alpha L)\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k],$$

where step (a) uses the recursion (129); step (b) holds due to the equation $2a^{\mathrm{T}}b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$; and Young's inequality as well as the fact that $\bar{y}^k = \bar{z}^k$. Then taking the total expectation yields the desired result. This completes the proof. ∎

D.3 Proof of Lemma 16

Recall that the update of z^{k+1} in (38g) is given by: $z^k = s^k + (1-\gamma)(z^k - s^k)$. Then, considering the update (38g), we can obtain a recursive expression for the term $\nabla\Phi(\bar{x}^{k+1}) - \bar{z}^{k+1}$ by introducing the terms $\mathbb{E}[\bar{s}^k|\mathcal{F}^k]$ and $\nabla\Phi(\bar{x}^k)$ as follows:

$$
\begin{aligned}
\nabla\Phi(\bar{x}^{k+1}) - \bar{z}^{k+1} &= \nabla\Phi(\bar{x}^{k+1}) - (\bar{s}^k + (1-\gamma)(\bar{z}^k - \bar{s}^k)) \\
&= (1-\gamma)(\nabla\Phi(\bar{x}^k) - \bar{z}^k) + \gamma(\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k|\mathcal{F}^k]) \\
&\quad + \gamma(\mathbb{E}[\bar{s}^k|\mathcal{F}^k] - \bar{s}^k) + \nabla\Phi(\bar{x}^{k+1}) - \nabla\Phi(\bar{x}^k).
\end{aligned}
\tag{131}
$$

Taking the square norm on both sides under the conditional expectation of $\mathcal{F}^k$, it follows from (131) that

$$
\begin{aligned}
&\mathbb{E}[\|\nabla\Phi(\bar{x}^{k+1}) - \bar{z}^{k+1}\|^2|\mathcal{F}^k] \\
&\overset{(a)}{=} [\|(1-\gamma)(\nabla\Phi(\bar{x}^k) - \bar{z}^k) + \gamma(\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k|\mathcal{F}^k]) + \nabla\Phi(\bar{x}^{k+1}) - \nabla\Phi(\bar{x}^k)\|^2|\mathcal{F}^k] \\
&\quad + \gamma^2\mathbb{E}[\|\mathbb{E}[\bar{s}^k|\mathcal{F}^k] - \bar{s}^k\|^2|\mathcal{F}^k] \\
&\overset{(b)}{\leqslant} (1-\gamma)\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2 + 2\gamma\left\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k|\mathcal{F}^k]\right\|^2 + \frac{2}{\gamma}\|\nabla\Phi(\bar{x}^{k+1}) - \nabla\Phi(\bar{x}^k)\|^2 \\
&\quad + \gamma^2\mathbb{E}[\|\mathbb{E}[\bar{s}^k|\mathcal{F}^k] - \bar{s}^k\|^2|\mathcal{F}^k] \\
&\overset{(c)}{\leqslant} (1-\gamma)\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2 + 2\gamma\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k|\mathcal{F}^k]\|^2 \\
&\quad + \frac{2L^2}{\gamma}\tau^2\alpha^2\|\bar{y}^k\|^2 + \gamma^2\mathbb{E}[\|\mathbb{E}[\bar{s}^k|\mathcal{F}^k] - \bar{s}^k\|^2|\mathcal{F}^k],
\end{aligned}
\tag{132}
$$

where step (a) holds due to the fact that $\bar{s}^k$ is independent of $\bar{z}^k$ and $\mathbb{E}[\bar{s}^k|\mathcal{F}^k]$ is an unbiased estimate of $\bar{s}^k$, with the additional assurance that the samples in each node are independent of each other; step (b) is derived by the convexity of l_2-norm and Young's inequality with the condition $0 < \gamma < 1$; step (c) follows from Lipschitz continuity of $\nabla\Phi$ derived in Proposition 3 and the recursion (38b). Next, we will bound the last term in (132). Note that from the recursion (38f) that

$$
\begin{aligned}
\mathbb{E}[\bar{s}^k|\mathcal{F}^k] - \bar{s}^k &= J\nabla_x F(x^k, \theta^k) - J\nabla_{x\theta}^2 G(x^k, \theta^k)v^k \\
&\quad - J\nabla_x \hat{F}(x^k, \theta^{k+1}; \varsigma_2^k) + J\nabla_{x\theta}^2 \hat{G}(x^k, \theta^k; \xi_3^k)v^k.
\end{aligned}
\tag{133}
$$

Taking the square norm on both sides of (133) under the total expectation, it follows that

$$
\begin{aligned}
\mathbb{E}[\|\mathbb{E}[\bar{s}^k|\mathcal{F}^k] - \bar{s}^k\|^2] &\overset{(a)}{=} \mathbb{E}[\|J\nabla_x F(x^k, \theta^k) - J\nabla_x \hat{F}(x^k, \theta^k; \varsigma_2^k)\|^2] \\
&\quad + \mathbb{E}[\|J\nabla_{x\theta}^2 G(x^k, \theta^k)v^k - J\nabla_{x\theta}^2 \hat{G}(x^k, \theta^k; \xi_3^k)v^k\|^2] \\
&\overset{(b)}{\leqslant} \frac{1}{m^2}(m\sigma_{f,x}^2 + \|v^k\|^2\sigma_{g,x\theta}^2) \overset{(c)}{\leqslant} \frac{1}{m}(\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2),
\end{aligned}
\tag{134}
$$

where step (a) uses the independence between the samples ς_2^k and ξ_3^k and unbiased estimates of stochastic gradients in Assumption 5; step (b) follows from the bounded variances in Assumption 5; step (c) is derived by the following fact

$$\frac{1}{m}\|v^k\|^2 \leqslant 2\frac{1}{m}\|v^*(\bar{x}^k)\|^2 + 2\frac{1}{m}\|v^k - v^*(\bar{x}^k)\|^2 \leqslant 2M^2 + 2\frac{1}{m}\|v^k - v^*(\bar{x}^k)\|^2 \triangleq \hat{M}^2, \quad (135)$$

with the last step following the result that $\|v_i^*(\bar{x}^k)\| \leq M$ in Proposition 4. Then, taking total expectation and combining the upper bound (134), we reach the result (25) as follows:

$$\mathbb{E}[\|\nabla\Phi(\bar{x}^{k+1}) - \bar{z}^{k+1}\|^2] \leqslant (1-\gamma)\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{z}^k\|^2] + \underbrace{2\frac{\gamma}{\alpha}}_{\triangleq r_z}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k|\mathcal{F}^k]\|^2]$$

$$+ \underbrace{\frac{2L^2}{\gamma}}_{\triangleq r_y}\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + \frac{1}{m}\underbrace{(\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}}_{\triangleq \sigma_{\bar{z}}^2}\alpha^2.$$

(136)

In what follows, we will derive the remaining result (26). To this end, by the definition of $\nabla\Phi(\bar{x}^k)$, we first have

$$\nabla\Phi(\bar{x}^k) = J\nabla_x F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) - J\nabla_{x\theta}^2 G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))(v^*(\bar{x}^k)).$$

Additionally, it follows from the recursion (38d) that

$$\mathbb{E}[\bar{s}^k|\mathcal{F}^k] = J\nabla_x F(x^k, \theta^k) - J\nabla_{x\theta}^2 G(x^k, \theta^k)v^k.$$

Thus, the term $\nabla\Phi(\bar{x}^k) - \bar{s}^k$ can be expressed as

$$\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k|\mathcal{F}^k] = J\nabla_x F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) - J\nabla_x F(x^k, \theta^k) + J\nabla_{x\theta}^2 G(x^k, \theta^k)(v^*(\bar{x}^k) - v^k)$$

$$+ J(\nabla_{x\theta}^2 G(x^k, \theta^k) - \nabla_{x\theta}^2 G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)))v^*(\bar{x}^k). \quad (137)$$

Then taking the square norm on both sides of (137) under the total expectation and employing the inequality $\|a + b\|^2 \leqslant 2\|a\|^2 + 2\|b\|^2$ as well as Jensen inequality, we get

$$\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \mathbb{E}[\bar{s}^k|\mathcal{F}^k]\|^2]$$

$$\leqslant 2\frac{1}{m}\mathbb{E}[\|\nabla_x F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) - \nabla_x F(x^k, \theta^k)\|^2]$$

$$+ 4\frac{1}{m}\mathbb{E}[\|-\nabla_{x\theta}^2 G(x^k, \theta^k)v^*(\bar{x}^k) + \nabla_{x\theta}^2 G(x^k, \theta^k)v^k\|^2] \quad (138)$$

$$+ 4\frac{1}{m}\mathbb{E}[\|-\nabla_{x\theta}^2 G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))v^*(\bar{x}^k) + \nabla_{x\theta}^2 G(x^k, \theta^k)v^*(\bar{x}^k)\|^2]$$

$$\leqslant \underbrace{(2L_{f,x}^2 + 4M^2 L_{g,x\theta}^2)}_{\triangleq L_{fg,x}}\frac{1}{m}\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2] + 4C_{g,x\theta}^2\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2].$$

To obtain the last step, we use Lipschitz continuity of $\nabla_x f_i$ and $\nabla_{x\theta}^2 g_i$ in Assumptions 2 and 3 as well as the boundedness of $\|v_i^*(\bar{x}^k)\|$ in Proposition 4. This completes the proof. ∎

D.4 Proof of Lemma 17

We note that the term $\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]$ can be expanded as

$$
\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k] = \underbrace{\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^k)\|^2|\mathcal{F}^k]}_{\triangleq A_1^v} + \mathbb{E}[\|v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]
$$
$$
+ \underbrace{\mathbb{E}[2\langle v^{k+1} - v^*(\bar{x}^k), v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})\rangle|\mathcal{F}^k]}_{\triangleq A_2^v}. \tag{139}
$$

We then aim to bound the term A_1^v. To this end, for reader's convenience, we repeat the argument on the recursion of v^{k+1} in (38b) by combining (38e) as follows:

$$
v^{k+1} = (I - \lambda\nabla_{\theta\theta}^2\hat{G}(x^k, \theta^k; \xi_2^k))v^k + \lambda\hat{F}(x^k, \theta^k; \varsigma_1^k).
$$

Substituting the above expression into the term A_1^v, we get

$$
\begin{aligned}
A_1^v &= \mathbb{E}[\|(I - \lambda\nabla_{\theta\theta}^2\hat{G}(x^k, \theta^k; \xi_2^k))v^k + \lambda\nabla_\theta\hat{F}(x^k, \theta^k; \varsigma_1^k) - v^*(\bar{x}^k)\|^2|\mathcal{F}^k] \\
&= \|(I - \lambda\nabla_{\theta\theta}^2 G(x^k, \theta^k))v^k + \lambda\nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k)\|^2 \\
&\quad + \lambda^2\mathbb{E}\|\nabla_{\theta\theta}^2\hat{G}(x^k, \theta^k; \xi_2^k)v^k - \nabla_{\theta\theta}^2 G(x^k, \theta^k)v^k + \nabla_\theta F(x^k, \theta^k) - \nabla_\theta\hat{F}(x^k, \theta^k; \varsigma_1^k)\|^2|\mathcal{F}^k] \\
&\leqslant \|(I - \lambda\nabla_{\theta\theta}^2 G(x^k, \theta^k))v^k + \lambda\nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k)\|^2 + m(\sigma_{f,\theta}^2 + \hat{M}^2\sigma_{g,\theta\theta}^2)\lambda^2,
\end{aligned} \tag{140}
$$

where the last step follows from the bounded variances in Assumption 5 and the inequality (135). Noting that $v^*(\bar{x}^k)$ admits the following expression

$$
v^*(\bar{x}^k) = [\nabla_{\theta\theta}^2 G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))]^{-1}\nabla_\theta F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)), \tag{141}
$$

then we have that

$$
\begin{aligned}
&(I - \lambda\nabla_{\theta\theta}^2 G(x^k, \theta^k))v^k + \lambda\nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k) \\
&= \lambda(\nabla_\theta F(x^k, \theta^k) - \nabla_\theta F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))) + (I - \lambda\nabla_{\theta\theta}^2 G(x^k, \theta^k))(v^k - v^*(\bar{x}^k)) \\
&\quad + \lambda(\nabla_{\theta\theta}^2 G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) - \nabla_{\theta\theta}^2 G(x^k, \theta^k))v^*(\bar{x}^k).
\end{aligned} \tag{142}
$$

With the above expression, the first term on the right hand of (140) can be bounded by

$$
\begin{aligned}
&\|(I - \lambda\nabla_{\theta\theta}^2 G(x^k, \theta^k))v^k + \lambda\nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k)\|^2 \\
&\overset{(a)}{\leqslant} (1 + \frac{\mu_g\lambda}{3})^2\|(I - \lambda\nabla_{\theta\theta}^2 G(x^k, \theta^k))(v^k - v^*(\bar{x}^k))\|^2 \\
&\quad + (1 + \frac{3}{\mu_g\lambda})\lambda^2\|\nabla_\theta F(x^k, \theta^k) - \nabla_\theta F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))\|^2 \\
&\quad + (1 + \frac{\mu_g\lambda}{3})(1 + \frac{3}{\mu_g\lambda})\lambda^2\|(\nabla_{\theta\theta}^2 G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) - \nabla_{\theta\theta}^2 G(x^k, \theta^k))v^*(\bar{x}^k)\|^2 \\
&\overset{(b)}{\leqslant} (1 + \frac{\mu_g\lambda}{3})^2(1 - \mu_g\lambda)^2\|v^k - v^*(\bar{x}^k)\|^2 \\
&\quad + (1 + \frac{3}{\mu_g\lambda})L_{f,\theta}^2\lambda^2[\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
&\quad + (1 + \frac{3}{\mu_g\lambda})(1 + \frac{\mu_g\lambda}{3})L_{g,\theta\theta}^2 M^2\lambda^2[\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2],
\end{aligned} \tag{143}
$$

53

where step (a) uses Young's inequality twice and step (b) follows from the strong convexity of g_i, Lipschitz continuity of $\nabla_\theta f_i$ and $\nabla^2_{\theta\theta} g_i$, and the boundedness of $\|v^*_i(\bar{x}^k)\|$. By considering the condition $1 - \lambda\mu_g > 0$ and rearranging the terms, we can further derive from (143) that

$$\left\|(I - \lambda\nabla^2_{\theta\theta}G(x^k, \theta^k))v^k + \lambda\nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k)\right\|^2 \tag{144}$$

$$\leqslant (1 - \frac{\mu_g\lambda}{3})(1 - \mu_g\lambda)\|v^k - v^*(\bar{x}^k)\|^2 + \frac{2\lambda}{\mu_g}\underbrace{(2L^2_{f,\theta} + 4M^2 L^2_{g,\theta\theta})}_{\triangleq L_{fg,\theta}}[\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2].$$

Besides, substituting (144) into (140), we reach an upper bound for the term A^v_1 in (139) as follows

$$A^v_1 \leqslant (1 - \frac{\mu_g\lambda}{3})(1 - \mu_g\lambda)\|v^k - v^*(\bar{x}^k)\|^2 + m(1 + M^2)\lambda^2\sigma^2$$
$$+ \frac{2\lambda}{\mu_g}L_{fg,\theta}[\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2]. \tag{145}$$

Now, we analyze the term $\mathbb{E}[\|v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})\|^2]$ in (139). Employing Lipschitz continuity of $v^*_i(x)$ in Proposition 3, the recursion (38a) and the upper bound of the variances in (25), we have that

$$\mathbb{E}[\|v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k] \leqslant mL^2_{v^*}\mathbb{E}[\|\bar{x}^k - \bar{x}^{k+1}\|^2|\mathcal{F}^k] \leqslant m\tau^2\alpha^2 L^2_{v^*}\|\bar{y}^k\|^2. \tag{146}$$

It remains to analyze the term A^v_2 in (139), which can be rewritten as the following expression by leveraging Cauchy-Schwartz inequality

$$A^v_2 \leqslant \frac{\mu_g\lambda}{3}A^v_1 + \frac{3}{\mu_g\lambda}\mathbb{E}[\|v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]. \tag{147}$$

Furthermore, together with the results (139), (145), (146), (147), we reach an evolution for $\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]$ as follows:

$$\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]$$
$$\overset{(a)}{\leqslant} (1 + \frac{\mu_g\lambda}{3})(1 - \frac{\mu_g\lambda}{3})(1 - \mu_g\lambda)\|v^k - v^*(\bar{x}^k)\|^2$$
$$+ (1 + \frac{\mu_g\lambda}{3})\frac{2\lambda}{\mu_g}L_{fg,\theta}[\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2]$$
$$+ \frac{2L^2_{v^*}\tau^2\alpha^2}{\varpi\lambda}m\|\bar{y}^k\|^2 + (1 + \frac{\mu_g\lambda}{3})m(1 + \hat{M}^2)\lambda^2\sigma^2 \tag{148}$$
$$\overset{(b)}{\leqslant} (1 - \mu_g\lambda)\|v^k - v^*(\bar{x}^k)\|^2 + \underbrace{\frac{4L_{fg,\theta}\lambda}{\mu_g\alpha}}_{\triangleq q_x}\alpha[\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2]$$
$$+ \underbrace{\frac{2L^2_{v^*}}{\varpi\lambda}}_{\triangleq q_s}m\tau^2\alpha^2\|\bar{y}^k\|^2 + m\underbrace{2(\sigma^2_{f,\theta} + \hat{M}^2\sigma^2_{g,\theta\theta})\frac{\lambda^2}{\alpha^2}}_{\triangleq \sigma^2_v}\alpha^2,$$

where in step (a) we denote $\varpi \triangleq \frac{\mu_g}{3}$, and in step (b) we use the fact that $\mu_g\lambda < 1$. This completes the proof. ∎

D.5 Proof of Lemma 18

The main idea to prove the evolution of the inner-level errors $\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]$ is similar to the one used for the Hessian-inverse-vector product errors. We start by decomposing the inner-level errors as follows

$$\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k] = \underbrace{\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^k)\|^2|\mathcal{F}^k]}_{\triangleq A_1^\theta} + \mathbb{E}[\|\theta^*(\bar{x}^k) - \theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]$$
$$+ \underbrace{2\mathbb{E}[\langle\theta^{k+1} - \theta^*(\bar{x}^k), \theta^*(\bar{x}^k) - \theta^*(\bar{x}^{k+1})\rangle|\mathcal{F}^k]}_{\triangleq A_2^\theta}. \tag{149}$$

As for the term A_1^θ, we have that

$$\theta^{k+1} - \theta^*(\bar{x}^k) = \theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k) + \beta(\nabla_\theta G(1_m\otimes\bar{x}^k, \theta^k) - \nabla_\theta\hat{G}(x^k, \theta^k; \xi_1^k)). \tag{150}$$

Taking square norm on both sides under the conditional expectation $\mathcal{F}^k$, we have

$$A_1^\theta = \mathbb{E}[\|\theta^k - \beta\nabla_\theta\hat{G}(x^k, \theta^k; \xi_1^k) - \theta^*(\bar{x}^k)\|^2|\mathcal{F}^k]$$
$$= \|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2 + \beta^2\mathbb{E}[\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta\hat{G}(x^k, \theta^k; \xi_1^k)|\mathcal{F}^k]$$
$$+ 2\beta\langle\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k), \mathbb{E}[\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta\hat{G}(x^k, \theta^k; \xi_1^k)|\mathcal{F}^k]\rangle$$
$$\overset{(a)}{\leqslant}(1+\beta\omega_\theta)\|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2 + (\beta + \frac{1}{\omega_\theta})\beta L_{g,\theta}^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + m\beta^2\sigma_{g,\theta}^2$$
$$\overset{(b)}{\leqslant}(1 + \beta\omega_\theta)\|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2 + \frac{2}{\omega_\theta}\beta L_{g,\theta}^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + m\beta^2\sigma_{g,\theta}^2, \tag{151}$$

where step (a) uses the variance decomposition and Cauchy-Schwartz inequality with parameter $\omega_\theta = \frac{\mu_g L_{g,\theta}}{2(\mu_g + L_{g,\theta})}$, and the step (b) comes from Lipschitz continuity of $\nabla_\theta g_i$ in Assumption 3 and the condition that $\beta < \frac{1}{\omega_\theta}$ in (29). Next, we proceed in providing an upper bound for the first term on the right hand of (151) as follows:

$$\|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2$$
$$\overset{(a)}{=}\|\theta^k - \theta^*(\bar{x}^k)\|^2 + \beta^2\|\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))\|^2$$
$$- 2\beta\langle\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)), \theta^k - \theta^*(\bar{x}^k)\rangle$$
$$\overset{(b)}{\leqslant}(1 - 2\beta\frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}})\|\theta^k - \theta^*(\bar{x}^k)\|^2 + (\beta - \frac{2}{\mu_g + L_{g,\theta}})\beta\|\nabla_\theta G(1_m\otimes\bar{x}^k, \theta^k) - \nabla_\theta G(1_m\otimes\bar{x}^k, \theta^*(\bar{x}^k))\|^2$$
$$\overset{(c)}{\leqslant}(1 - 2\beta\frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}})\|\theta^k - \theta^*(\bar{x}^k)\|^2, \tag{152}$$

where step (a) uses the fact that $\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) = 0$; step (b) come from the strong convexity and smoothness of g_i in Assumption 3; step (c) holds due to the step-size condition

55

$\beta = c_\beta \alpha < \frac{2}{\mu_g + L_{g,\theta}}$. Then, plugging (152) into (151) yields

$$
\begin{aligned}
A_1^\theta &\leqslant (1 + \beta\omega_\theta)(1 - 2\beta\frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}})\|\theta^k - \theta^*(\bar{x}^k)\|^2 + (\beta + \frac{1}{\omega_\theta})\beta L_{g,\theta}^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + m\beta^2\sigma_{g,\theta}^2 \\
&\leqslant (1 - \frac{3}{2}\beta\frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}})\|\theta^k - \theta^*(\bar{x}^k)\|^2 + \frac{2}{\omega_\theta}\beta L_{g,\theta}^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + m\beta^2\sigma_{g,\theta}^2. \quad (153)
\end{aligned}
$$

In addition, note that the inner-product term in (149) follows that

$$
\mathbb{E}[\|\theta^*(\bar{x}^k) - \theta^*(\bar{x}^{k+1})\|^2 | \mathcal{F}^k] \leqslant mL_{\theta*}^2 \tau^2 \alpha^2 \|\bar{y}^k\|^2. \quad (154)
$$

Next, for the term A_2^θ in (149), by Cauchy-Schwartz inequality, we have

$$
A_2^\theta \leqslant \omega_\theta \beta A_1^\theta + \frac{1}{\omega_\theta \beta}\mathbb{E}[\|\theta^*(\bar{x}^k) - \theta^*(\bar{x}^{k+1})\|^2 | \mathcal{F}^k]. \quad (155)
$$

By leveraging the results (153), (154), (155), the inner-level errors $\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2 | \mathcal{F}^k]$ can be bounded as follows:

$$
\begin{aligned}
&\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2 | \mathcal{F}^k] \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (156)\\
&\leqslant (1 + \omega_\theta\beta)(1 - \frac{3}{2}\beta\frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}})\|\theta^k - \theta^*(\bar{x}^k)\|^2 \\
&\quad + (1 + \omega_\theta\beta)\frac{2}{\omega_\theta}\beta L_{g,\theta}^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + (1 + \omega_\theta\beta)\beta^2\sigma^2 + (1 + \frac{1}{\omega_\theta\beta})L_{\theta*}^2\tau^2\alpha^2\|\bar{y}^k\|^2 \\
&\leqslant (1 - \frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}}\beta)\|\theta^k - \theta^*(\bar{x}^k)\|^2 + m\underbrace{\frac{2L_{\theta*}^2}{\omega_\theta\beta}}_{\triangleq q_s}\tau^2\alpha^2\|\bar{y}^k\|^2 \qquad\qquad\qquad (157)\\
&\quad + \underbrace{\frac{4L_{g,\theta}^2\beta}{\omega_\theta\alpha}}_{\triangleq q_x}\alpha\|x^k - 1_m \otimes \bar{x}^k\|^2 + m\underbrace{2\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}}_{\triangleq \sigma_\theta^2}\alpha^2,
\end{aligned}
$$

where the last inequality is due to the fact $\omega_\theta\beta < 1$ and $c_\beta = \frac{\beta}{\alpha}$. This completes the proof. ∎

D.6 Proof of Lemma 19

First, let us recall the recursion of x^{k+1} as follows

$$
x^{k+1} = (1 - \tau)x^k + \tau(\mathcal{W}x^k - \alpha y^k),
$$

by which, we have

$$
x^{k+1} - 1_m \otimes \bar{x}^{k+1} = (1 - \tau)(x^k - 1_m \otimes \bar{x}^k) + \tau((\mathcal{W} - \mathcal{J})(x^k - 1_m \otimes \bar{x}^k) - \alpha(y^k - 1_m \otimes \bar{y}^k)) \quad (158)
$$

where $\mathcal{J} = \frac{1_m 1_m^{\mathrm{T}}}{m} \otimes I_n$. Then employing Young's inequality with parameter $\eta > 0$ yields

$$
\begin{aligned}
&\mathbb{E}[\|x^{k+1} - 1_m \otimes \bar{x}^{k+1}\|^2 | \mathcal{F}^k] \\
&\leqslant (1-\tau)^2 (1 + \frac{\tau}{1-\tau}) \|x^k - 1_m \otimes \bar{x}^k\|^2 + \tau^2 (1 + \frac{1-\tau}{\tau}) \|Wx^k - \alpha y^k - (1_m \otimes \bar{x}^k - \alpha \bar{y}^k)\|^2 \\
&\leqslant (1-\tau) \|x^k - 1_m \otimes \bar{x}^k\|^2 + \tau \|(\mathcal{W} - \mathcal{J})(x^k - 1_m \otimes \bar{x}^k) - \alpha(y^k - 1_m \otimes \bar{y}^k)\|^2 \\
&\leqslant (1-\tau) \|x^k - 1_m \otimes \bar{x}^k\|^2 + \tau(1+\eta) \|\mathcal{W} - \mathcal{J}\|^2 \|x^k - 1_m \otimes \bar{x}^k\|^2 + \tau(1 + \frac{1}{\eta}) \|y^k - 1_m \otimes \bar{y}^k\|^2 \\
&\leqslant (1 - \tau \frac{1-\rho}{2}) \|x^k - 1_m \otimes \bar{x}^k\|^2 + \frac{2\tau\alpha^2}{1-\rho} \|y^k - 1_m \otimes \bar{y}^k\|^2, \quad\quad\quad (159)
\end{aligned}
$$

where the last step takes $\eta = \frac{1-\rho}{2\rho}$ and uses the fact that $\|\mathcal{W} - \mathcal{J}\|^2 = \rho$. Taking the total expectation, we get the desired result. This completes the proof. ∎

D.7 Proof of Lemma 20

For the term $\|y^k - 1_m \otimes \bar{y}^k\|^2$, we can further split it into

$$
\begin{aligned}
\|y^k - 1_m \otimes \bar{y}^k\|^2 &\leqslant \|y^k\|^2 \\
&= \sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k) + \nabla\Phi_i(\bar{x}^k) - \nabla\Phi(\bar{x}^k) + \nabla\Phi(\bar{x}^k)\|^2 \\
&\leqslant 3\sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k)\|^2 + 3\sum_{i=1}^m \|\nabla\Phi_i(\bar{x}^k) - \nabla\Phi(\bar{x}^k)\|^2 + 3m\|\nabla\Phi(\bar{x}^k)\|^2 \\
&\leqslant 3\sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k)\|^2 + 3b^2 + 3m\|\nabla\Phi(\bar{x}^k)\|^2 \\
&= 3\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2 + 3b^2 + 3m\|\nabla\Phi(\bar{x}^k)\|^2,
\end{aligned}
$$
$$(160)$$

where the second inequality follows from Lemma 5 and the last step holds due to the fact that $y_i^k = z_i^k$ under local gradient scheme (39) and $\nabla\tilde{\Phi}(\bar{x}^k) = \mathrm{col}\{\nabla\Phi_i(\bar{x}^k)\}_{i=1}^m$. Then, taking the total expectation, we get the desired result. This completes the proof. ∎

D.8 Proof of Lemma 21

According to the recursion (40), we have that the update of y^{k+1} can be derived as

$$
y^{k+1} = \mathcal{W}y^k + z^{k+1} - z^k,
$$

which further gives that

$$
\begin{aligned}
y^{k+1} - 1_m \otimes \bar{y}^{k+1} &= \mathcal{W}y^k + z^{k+1} - z^k - 1_m \otimes (\bar{y}^k + \bar{z}^{k+1} - \bar{z}^k) \\
&= (\mathcal{W} - \mathcal{J})(y^k - 1_m \otimes \bar{y}^k) + (I - \mathcal{J})(z^{k+1} - z^k).
\end{aligned}
$$
$$(161)$$

Then taking the square norm on both sides of the above expression under the conditional expectation $\mathcal{F}^k$ and using Young's inequality with the parameter $\eta = \frac{1-\rho}{2\rho}$, we have

$$
\begin{aligned}
&\mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2 | \mathcal{F}^k] \\
&\leqslant (1 + \eta)\|\mathcal{W} - \mathcal{J}\|^2 \|y^k - 1_m \otimes \bar{y}^k\|^2 + (1 + \frac{1}{\eta})\|I - \mathcal{J}\|^2 \mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathcal{F}^k] \\
&\leqslant \frac{1+\rho}{2}\|y^k - 1_m \otimes \bar{y}^k\|^2 + \frac{2}{1-\rho}\mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathcal{F}^k],
\end{aligned}
\tag{162}
$$

where the last inequality uses the fact that $\|\mathcal{W} - \mathcal{J}\|^2 = \rho$ and $\|I - \mathcal{J}\|^2 \leqslant 1$. For the last term in (162), it follows from (38g) that

$$
z^{k+1} - z^k = \gamma(\nabla\tilde{\Phi}(\bar{x}^k) - z^k) + \gamma(\mathbb{E}[s^k|\mathcal{F}^k] - \nabla\tilde{\Phi}(\bar{x}^k)) + \gamma(s^k - \mathbb{E}[s^k|\mathcal{F}^k]),
\tag{163}
$$

which further implies that

$$
\begin{aligned}
\mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathcal{F}^k] \leqslant\ & 2\gamma^2 \|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2 \\
& + 2\gamma^2 \|\nabla\tilde{\Phi}(\bar{x}^k) - \mathbb{E}[s^k|\mathcal{F}^k]\|^2 + m(\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2)\gamma^2.
\end{aligned}
\tag{164}
$$

Substituting the above inequality into (162) leads to

$$
\begin{aligned}
&\mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2 | \mathcal{F}^k] \\
&\leqslant \frac{1+\rho}{2}\|y^k - 1_m \otimes \bar{y}^k\|^2 + \frac{4}{1-\rho}\gamma^2\|\nabla\tilde{\Phi}(\bar{x}^k) - \mathbb{E}[s^k|\mathcal{F}^k]\|^2 \\
&\quad + \frac{4}{1-\rho}\gamma^2\|\nabla\tilde{\Phi}(\bar{x}^k) - z^k\|^2 + \frac{2}{1-\rho}m\underbrace{(\sigma_{f,x}^2 + \hat{M}^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}}_{\triangleq \sigma_y^2}\alpha^2,
\end{aligned}
\tag{165}
$$

which completes the proof. ∎