Distributed Stochastic Bilevel Optimization: Improved Complexity and Heterogeneity Analysis

Youcheng Niu, Jinming Xu*, Ying Sun, Yan Huang, Li Chai

December 25, 2023

Abstract

Distributed bilevel optimization has gained increasing popularity due to its wide applications in large-scale machine learning. This paper focuses on a class of nonconvex-strongly-convex distributed stochastic bilevel optimization (DSBO) problems with personalized inner-level objectives. Existing algorithms require extra computation loops to compute inner-level solutions accurately and Hessian inverse matrices for estimating the hypergradient, which incurs high computational complexity of gradient evaluation. To address this issue, we propose a loopless personalized distributed algorithm (termed LoPA) that leverages iterative approximation for the inner-level solutions and Hessian inverse matrices. Theoretical analysis shows that LoPA has a sublinear rate of $\mathcal{O}(\frac{1}{(1-\rho)^2 K} + \frac{b^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}} K^{\frac{2}{3}}} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}}))$, where K is the total number of iterations, b quantifies the data heterogeneity across nodes, and $\sigma_{\mathrm{p}}, \sigma_{\mathrm{c}}$ represent the gradient sampling variances associated with the inner-level and out-level variables, respectively. We provide an explicit characterization of the heterogeneity, and develop a variant of LoPA based on gradient tracking to eliminate the heterogeneity, yielding a rate of $\mathcal{O}(\frac{1}{(1-\rho)^4 K} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}}))$. The computational complexity of LoPA is shown to be of the order of $\mathcal{O}(\epsilon^{-2})$ thanks to the loopless structure, outperforming existing counterparts for DSBO by an order of $\mathcal{O}(\log \epsilon^{-1})$. Numerical experiments demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Bilevel optimization is a hierarchical optimization framework that involves an outer-level and an inner-level problems, where the solution to the outer-level problem depends on the solution to 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 [1,2], neural architecture search [3,4], hyperparameter selection [5,6], and reinforcement learning [7,8]. 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 be networked together to improve learning efficiency or enable multi-task learning [9–13]. This distributed manner facilitates collaboration among nodes to achieve the overall objective. In

*Y. Niu, J. Xu, Y. Huang and L. Chai are with the College of Control Science and Engineering, Zhejiang University; Y. Sun is with School of Electrical Engineering and Computer Science, The Pennsylvania State University. Correspondence to Jinming Xu (Email: jimmyxu@zju.edu.cn).

particular, 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 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 $x \in \mathbb{R}^n$. Besides, 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 [14], adversarial learning [15], and reinforcement learning [7] to fair transceiver design [16]. 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 a 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 previously unknown tasks with a small number of task-specific gradient updates [1, 2, 17].

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

Algorithm	Setting	# of Loop	Inner Step	Batch Size	Complexity	Hetero. Analysis
MA-DSBO [10]	**G**	N-Q-Loop	SGD	$\mathcal{O}(1)$	$\mathcal{O}\left(\epsilon^{-2}\log\epsilon^{-1}\right)$	No
Gossip-DSBO [18]	**G**	Q-Loop	SGD	$\mathcal{O}(1)$	$\mathcal{O}\left(\epsilon^{-2}\log\epsilon^{-1}\right)$	No
SLAM [19]	**G**	Q-Loop	SGD	$\mathcal{O}(1)$	$\mathcal{O}\left(\epsilon^{-2}\log\epsilon^{-1}\right)$	No
VRDBO [20]	**G**	Q-Loop	STORM	$\mathcal{O}(1)$	$\mathcal{O}\left(\epsilon^{-3/2}\log\epsilon^{-1}\right)$	No
SPDB [11]	**P**	Q-Loop	SGD	$\mathcal{O}(1)$	$\mathcal{O}\left(\epsilon^{-2}\log\epsilon^{-1}\right)$	No
LoPA (this work)	**P**	No-Loop	SGD	$\mathcal{O}(1)$	$\mathcal{O}\left(\epsilon^{-2}\right)$	Yes

*$\mathbf{G}$ and $\mathbf{P}$ represent DSBO with global and inner-level personalized objectives, respectively; The complexity represents the Hessian evaluations required to attain an ϵ-stationary point.

Different from conventional single-level optimization, bilevel optimization poses additional challenges due to its nested structure and inherent non-convexity [21]. In most cases, obtaining the hypergradient $\nabla\Phi_i(x)$ is difficult since the inner-level solutions $\theta_i^*(x)$ are not explicitly known [22]. The closed-form expression of the hypergradients can be obtained by Approximate Implicit Differentiation approaches (AID), but it typically involves two nested loops [23, 24]: the N-loop to find a near-optimal solution to the inner-level function with N iterations, and the Q-loop to approximate the Hessian inverse matrices with Q iterations. This is particularly challenging in large-scale machine learning applications where running these two nested loops is prohibitively expensive [24]. As a result, it is not straightforward to integrate SGD methods with AID approaches for bilevel optimization for properly balancing the computational burden with the error due to biased estimator of hypergradients resulting from inexact computation. To address this issue, various approximation algorithms have been proposed to effectively estimate hypergradients and reduce stochastic bias [8, 22, 24–30]. 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 [11, 18–20]. However, these existing algorithms usually require extra computation loops to estimate the above inner-level solutions and Hessian inverse matrices for solving DSBO, which induces relatively high computation costs. Therefore, the following question arises naturally:

> *Can we design a new decentralized learning algorithm without involving extra computation loops while still achieving comparable or even better complexity?*

Moreover, different from standard distributed optimization, distributed bilevel optimization encounters distinct difficulties in dealing with node heterogeneity, with its exact influence on the convergence performance remaining unclear. Analyzing this impact and identifying the crucial factors associated with node heterogeneity is even more challenging due to the nested structure of DSBO problems, which leads to another important theoretical question:

> *How to characterize heterogeneity for DSBO, and how does it affect the convergence?*

Summary of Contributions. To address the above issues, 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, 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 [10, 11, 18–20], LoPA leverages iterative approximation approaches that rely on a single stochastic gradient iteration to track the value of Hessian inverse matrices and inner-level solutions.

- **Improved complexity**. We show that LoPA-LG attain an ϵ-stationary point at a sublinear rate of $\mathcal{O}(\frac{1}{(1-\rho)^2K} + \frac{b^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}}))$, where K is the total number of iterations, b denotes the heterogeneity across nodes, and $\sigma_{\mathrm{p}}, \sigma_{\mathrm{c}}$ are the gradient sampling variances associated with the inner-level and the out-level variables, respectively. Thanks to the loopless structure, LoPA-LG is shown to have a computational complexity of the order of $\mathcal{O}(\epsilon^{-2})$, improving existing works for DSBO [10, 11, 18–20] by order of $\mathcal{O}(\log \epsilon^{-1})$ (See Table 1).

- **Heterogeneity analysis**. We explicitly characterize the heterogeneity across nodes and its impact on convergence for DSBO. Leveraging the heterogeneity analysis, we are able to prove the convergence of LoPA-LG under weaker assumptions compared to [10, 11, 18–20] (refer to Remark 1 for more details). Moreover, we prove that LoPA-GT is capable of eliminating the heterogeneity, yielding a sublinear rate of $\mathcal{O}(\frac{1}{(1-\rho)^4K} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}}))$. Numerical experiments on practical machine learning tasks demonstrate the effectiveness of the proposed algorithm in dealing with heterogeneity.

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 extra computation loops to reduce the bias of approximating Hessian inverse matrices (Q-loop) and inner-level solutions (N-loop) [24], increasing the batch size [25], adopting two-timescale step-sizes to eliminate steady-state stochastic variance [8], incorporating additional correction terms [26–29], and exploring the smoothness of objectives [22, 30]. Among these stochastic approximation algorithms, works [22, 24, 26] 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. In particular, the nature of the two-timescale update incurs the sub-optimal computational complexity of $\mathcal{O}\left(\epsilon^{-5/2}\log\epsilon^{-1}\right)$ [8]. Based on warm-start strategies, a computational complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ is provided in [25,29,30]. By employing momentum accelerations in both outer- and inner-level optimization procedures such as STORM [31] or SPIDER [32], works [27, 28] improve the complexity to $\mathcal{O}\left(\epsilon^{-3/2}\log\epsilon^{-1}\right)$. While the aforementioned works provide some insights into designing algorithms for solving stochastic bilevel optimization, they cannot be directly applied to distributed problem (1) as considered in this work.

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 [12]. 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 [33], gradient tracking [9], and alternating direction multiplier methods [34], accompanied by theoretical advancements. Specifically, for stochastic nonconvex problems, these algorithms can achieve a computational complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ with SGD methods [33, 35], which provide unbiased estimations for the full gradients. However, these methods are not readily available to be adapted to tackle the interaction between two levels of functions in solving bilevel optimization problems.

There have been some efforts aimed 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) [18], MA-DSBO [10], VRDBO [20], and SLAM [19] have been proposed which aim to solve a common inner-level task in a distributed manner. On the other hand, for personalized DSBO, methods like SPDB [11] have been developed where each node has its own local inner-level task. Various algorithmic frameworks have been developed in the above-mentioned works leveraging distributed optimization methods [9, 33, 34] 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 [10], Hessian-inverse-vector product [2] or Neumann series approaches [24] to avoid explicit computation of the inverse matrices. Since the outer-level gradients obtained using stochastic gradient descent (SGD) methods are no longer unbiased estimators [22], these distributed bilevel optimization algorithms involve extra computation loops to reduce the bias due to the approximation of the inverse Hessian matrices and inner-level solutions. As a result, the computational complexity of these algorithms is typically $\mathcal{O}(\epsilon^{-2}\log\epsilon^{-1})$ [10, 11, 18, 19]. The complexity can be improved to $\mathcal{O}\left(\epsilon^{-3/2}\log\epsilon^{-1}\right)$ by utilizing variance reduced gradient estimators [20]. 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 and its impact on convergence performance in DSBO.

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 preliminaries including relevant network models and assumptions.

3.1 Preliminaries

Network models. 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 $\mathcal{N}_i = \{j \mid (i,j) \in \mathcal{E}\}$. 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 existing literature of bilevel optimization [8, 10, 11, 18, 20, 24].

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 $\theta \in \mathbb{R}^p$ and $x \in \mathbb{R}^n$, $\|\nabla_\theta f_i(x,\theta)\| \leqslant C_{f,\theta}$.

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; $\nabla^2_{x\theta} g_i(\cdot,\cdot)$, $\nabla^2_{\theta\theta} g_i(\cdot,\cdot)$ are respectively $L_{g,x\theta}$- and $L_{g,\theta\theta}$-Lipschitz-continuous;

(iii) For any $\theta \in \mathbb{R}^p$ and $x \in \mathbb{R}^n$, $\|\nabla^2_{x\theta} g_i(x,\theta)\| \leqslant C_{g,x\theta}$.

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 [33, 35, 36].

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) $\sum_{i=1}^m \|\nabla_x f_i(x,\theta) - \nabla_x f(x,\theta)\|^2 \leqslant b_f^2$, $\sum_{i=1}^m \|\nabla_\theta f_i(x,\theta) - \nabla_\theta f(x,\theta)\|^2 \leqslant b_f^2$;

(ii) $\frac{1}{m}\sum_{i=1}^m \sum_{j=1}^m \|\nabla_\theta g_i(x,\theta_j^*(x)) - \nabla_\theta g_j(x,\theta_j^*(x))\|^2 \leqslant b_g^2$;

(iii) $\frac{1}{m}\sum_{i=1}^m \sum_{j=1}^m \|\nabla^2_{x\theta} g_i(x,\theta_j^*(x)) - \nabla^2_{x\theta} g_j(x,\theta_j^*(x))\|^2 \leqslant b_g^2$,

$\frac{1}{m}\sum_{i=1}^m \sum_{j=1}^m \|\nabla^2_{\theta\theta} g_i(x,\theta_j^*(x)) - \nabla^2_{\theta\theta} g_j(x,\theta_j^*(x))\|^2 \leqslant b_g^2$;

where $\nabla_x f$ and $\nabla_\theta f$ represent the partial gradient with respect to x and θ, respectively, while $\nabla^2_{x\theta} g$ and $\nabla^2_{\theta\theta} g$ denote Jacobian and Hessian, respectively.

Remark 1 (Weaker assumptions) *The parameters b_f^2 and b_g^2 are introduced to quantify the data heterogeneity on the outer-level and inner-level functions across nodes, respectively. If the outer- or inner-level functions have the same form and data distribution, then their corresponding heterogeneity parameters are equal to zero. It is worth noting that Assumption 4(i) for the outer-level functions is weaker than those made in previous works such as [10, 11, 18–20], which assume that each $\nabla_x f_i(x,\theta)$ is bounded by a constant. As for the heterogeneity in the inner-level functions, 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 it to be bounded at any θ.*

3.2 The LoPA Algorithm

In this section, we present our algorithm, termed LoPA, for problem (1). Following the standard procedures as did in distributed optimization [33, 37], 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 $\{x_i\}$.

Hypergradient construction. By the chain rule and implicit function theorem [24], we can compute the Hypergradient $\nabla\Phi_i(x_i)$ as follows:

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

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 (cf. (4) and (6)). As such, the Hessian-inverse-vector products $\left[\nabla_{\theta\theta}^2 g_i(x_i,\theta_i^*(x_i))\right]^{-1}\nabla_\theta f_i(x_i,\theta_i^*(x))$ (abbreviated Hv) and local hypergradient thus can be approximately computed as:

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

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

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 (cf., (8)). As for step (2), notice that v_i can be regarded as the solution of the following strongly convex problem:

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

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 together the ingredients, LoPA designs the following updates for the local inner-level variable θ_i^{k+1}, local Hv estimate variable v_i^{k+1} and 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 $\theta > 0, \alpha > 0, \lambda > 0$ are step-sizes. The directions d_i^k and h_i^k are further updated as:

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

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

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

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

Here, $\nabla_\theta \hat{g}_i$ (resp. $\nabla^2_{\theta\theta}\hat{g}_i$, $\nabla_\theta \hat{f}_i$, $\nabla_x \hat{f}_i$, $\nabla^2_{x\theta}\hat{g}_i$) is a stochastic gradient estimate of $\nabla_\theta g_i$ (resp. $\nabla^2_{\theta\theta}g_i$, $\nabla_\theta f_i$, $\nabla_x f_i$, $\nabla^2_{x\theta}g_i$) depending on the random variable $\xi_{i,1}^{k+1}$ (resp. $\xi_{i,2}^{k+1}$, $\xi_{i,3}^{k+1}$, $\varsigma_{i,1}^{k+1}$, $\varsigma_{i,2}^{k+1}$) and ζ_i^{k+1} is a tuple defined as $\zeta_i^{k+1} \triangleq \{\xi_{i,1}^{k+1}, \xi_{i,2}^{k+1}, \xi_{i,3}^{k+1}, \varsigma_{i,1}^{k+1}, \varsigma_{i,2}^{k+1}\}$.

However, using the above stochastic estimators will lead to steady-state variance errors under Assumption 1-4 and single-timescale step-sizes [8, 22], unless an increasing number of batch sizes is used [25] or extra smoothness conditions are imposed on Hv variables [30] (refer to Remark 8 in Appendix for more details). To address this issue, we further introduce a gradient momentum step to reduce the impact of variances as follows:

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

where $0 < \gamma < 1$.

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

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

where $0 < \tau < 1$ and two alternative choices of the direction y_i^k are considered:

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

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

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 proposed LoPA algorithm is summarized in Algorithm 1, where we call Algorithm 1 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 methods (NS) and conjugate gradient methods (CG) to directly approximate the Hessian inverse matrices to a high-precision at each iteration k. To be more specific, the key idea of the NS [8, 24] and CG methods [2, 38] 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:*

$$\textbf{\textit{(NS):}}\; Q\lambda\prod_{t=0}^{Q}\left(I - \lambda\nabla^2_{\theta\theta}\hat{g}_i\left(x_i, \theta_i; \xi_{i,2}^t\right)\right) \approx \left[\nabla^2_{\theta\theta}g_i\left(x_i, \theta_i\right)\right]^{-1}, \tag{13}$$

$$\textbf{\textit{(CG):}}\; \sum_{t=0}^{Q}\prod_{j=t+1}^{Q}\left(I - \lambda\nabla^2_{\theta\theta}\hat{g}_i(x_i, \theta_i; \xi_{i,2}^t)\right)\nabla_\theta \hat{f}_i(x_i, \theta_i; \varsigma_{i,1}^t) \approx \left[\nabla^2_{\theta\theta}g_i\left(x_i, \theta_i\right)\right]^{-1}\nabla_\theta f_i(x_i, \theta_i).$$

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 examples, the state-of-the-art works [11, 19, 20] 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.*

Algorithm 1 LoPA

1: **Require**: Initialize θ_i^0, v_i^0, x_i^0, $s_i^0(\zeta_i^0)$, 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}, $s_i^{k+1}(\zeta_i^{k+1})$, 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**

4 Convergence Results

In this section, we respectively analyze the performance of LoPA-LG and LoPA-GT for nonconvex-strongly-convex cases. We make the following assumption on the stochastic gradients used for estimating the gradients of the outer- and inner-level functions.

Let

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

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:

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

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

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

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 1 (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*

$$\sum_{i=1}^{m} \|\nabla\Phi_i(x) - \nabla\Phi(x)\|^2 \leqslant b^2, \tag{15}$$

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.

The proof of Lemma 1 is deferred to Appendix D.1. ■

Note that the heterogeneity of overall hypergradients is constituted by two main parts: the inner-level heterogeneity b_f^2 and the outer-level heterogeneity b_g^2. Let $\kappa_g = \frac{L_{g,\theta}}{\mu_g}$ denote the condition number. It is not difficult to show that $C_1(\mu_g, C_{g,x\theta}) = \mathcal{O}(\kappa_g^2)$ and $C_2(\mu_g, L_{f,x}, L_{f,\theta}, L_{g,x\theta}, L_{g,\theta\theta}, C_{f,\theta}, C_{g,x\theta}) = \mathcal{O}(\kappa_g^6)$. According to the definition of b^2, we can observe that b^2 is of $\mathcal{O}(\kappa_g^2 b_f^2 + \kappa_g^6 b_g^2)$. This observation suggests that the heterogeneity of inner-level objective functions plays a crucial role in determining the heterogeneity of overall hypergradients.

Theorem 1 *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 $\bar{x}^k = (1/m)\sum_{i=1}^m x_i^k$ and $L_{fg,x} = 2L_{f,x}^2 + 4M^2 L_{g,x\theta}^2$ 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*

$$\frac{1}{K+1}\sum_{k=0}^{K} \mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \leqslant \frac{4(V^0 - V^K)}{d_0\alpha(K+1)} + \frac{4}{d_0}\alpha\sigma_r^2 + \frac{12\vartheta}{d_0}\alpha^2 b^2, \tag{16}$$

where $\sigma_r^2 = 2(d_1 + 2C_{g,x\theta}^2 D_r)(\sigma_{f,\theta}^2 + M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} + (d_2 + L_{fg,x}D_r)\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2} + D_r(\sigma_{f,x}^2 + M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}$ and $\vartheta = D_r\frac{3\tau^2 L_{fg,x}}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha}$ with $D_r = \frac{1}{m}d_3 + d_4$. The coefficients d_0, d_1, d_2, d_3, d_4, d_5 of the unified Lyapunov function V^k can be found in Appendix B.3.

The proof of Theorem 1 is deferred to Appendix B.3. ■

Corollary 1 *Consider the same setting as Theorem 1. 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.4, and $\gamma = \mathcal{O}(\alpha)$, $\lambda = \mathcal{O}(\alpha)$, $\beta = \mathcal{O}(\alpha)$, $\tau = \mathcal{O}(\alpha)$, then we have*[1]

$$\frac{1}{K+1}\sum_{k=0}^{K} \mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] = \mathcal{O}\left(\frac{1}{(1-\rho)^2 K} + \frac{b^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}} K^{\frac{2}{3}}} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right), \tag{17}$$

where $\sigma_{\mathrm{p}} = \mathcal{O}(\sigma_{f,\theta} + \sigma_{g,\theta\theta} + \sigma_{g,\theta}), \sigma_{\mathrm{c}} = \mathcal{O}(\sigma_{f,x} + \sigma_{g,x\theta})$.

The proof of Corollary 1 is deferred to Appendix B.4. ■

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

Remark 3 (Heterogeneity analysis) *The above corollary shows that LoPA-LG has a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{K}})$, where K represents the number of iterations required to achieve an ϵ-stationary point. It also highlights the impact of heterogeneity on the convergence rate of LoPA-LG under the local gradient scheme. Unlike previous studies that rely on the boundness of each local hypergradient to approximate heterogeneity, we define the heterogeneity factor b^2 to obtain a more precise upper bound in (17). To the best of our knowledge, this is the first study to characterize and elucidate the influence of heterogeneity on convergence in distributed bilevel optimization.*

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 2.

Theorem 2 *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 $\bar{x}^k = (1/m)\sum_{i=1}^m x_i^k$ and $L_{fg,x} = 2L_{f,x}^2 + 4M^2 L_{g,x\theta}^2$ 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*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \leqslant \frac{2\left(V^0 - V^K\right)}{d_0\alpha(K+1)} + \frac{2\alpha\sigma_{r'}^2}{d_0}. \tag{18}$$

where $\sigma_{r'}^2 \triangleq 2(d_1 + 2C_{g,x\theta}^2 D_{r'})(\sigma_{f,\theta}^2 + M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} + (d_2 + L_{fg,x}D_{r'})\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2} + D_{r'}(\sigma_{f,x}^2 + M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}$ with $D_{r'} = \frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6$. The coefficients d_0, d_1, d_2, d_3, d_4, d_6 within the unified Lyapunov function V^k are given in Appendix B.5.

The proof of Theorem 2 is deferred to Appendix B.5. ■

Corollary 2 *Consider the same setting as Theorem 2. 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.6, and $\gamma = \mathcal{O}(\alpha)$, $\lambda = \mathcal{O}(\alpha)$, $\beta = \mathcal{O}(\alpha)$, $\tau = \mathcal{O}(\alpha)$, then we have*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] = \mathcal{O}\left(\frac{1}{(1-\rho)^4 K} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right). \tag{19}$$

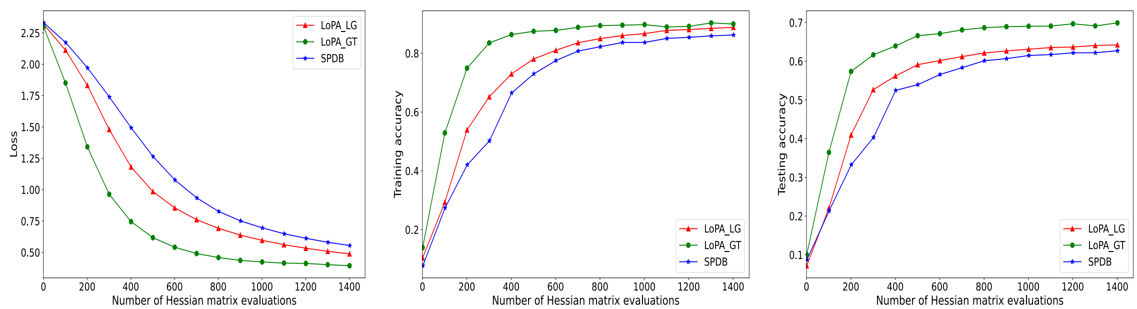
The proof of Corollary 2 is deferred to Appendix B.6. ■

Remark 4 (Improved complexity) *Corollary 2 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 heterogeneity is eliminated by gradient tracking scheme (12). 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 for 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.*

Remark 5 (Linear speed-up) *Corollaries 1 and 2 demonstrate that, in the context of DSBO problems with personalized inner-level objectives, the variance σ_{p} associated with the inner-level variable θ does not exhibit linear speed-up. In contrast, the variance σ_{c} induced by the stochastic gradients with respect to the out-level variable x is able to achieve linear speed-up. This lack of* *linear speed-up in σ_{p} is attributed to the fact that the inner-level variables are not required to achieve consensus. However, it is worthy to note that, by increasing batch size by m times when estimating the local partial gradients $\nabla_\theta f_i$, $\nabla_\theta g_i$, $\nabla_{\theta\theta} g_i$, both LoPA-LG and LoPA-GT can obtain a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{mK}})$, leading to an iteration complexity of $\mathcal{O}(m^{-1}\epsilon^{-2})$, which further implies that the communication complexity also achieves a linear speedup thanks to the loopless structure.*

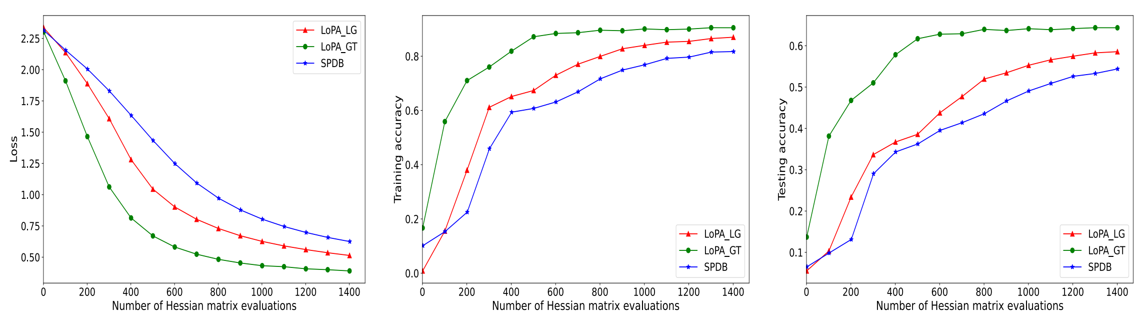
5 Numerical Experiments

In this section, we evaluate the effectiveness of our algorithms on a class of distributed classification learning problems involving heterogeneous datasets.



(a) Loss. (b) Training accuracy. (c) Testing accuracy.

Figure 1: Performance comparison of LoPA-LG, LoPA-LG and SPDB algorithms over 4 nodes.



(a) Loss. (b) Training accuracy. (c) Testing accuracy.

Figure 2: Performance comparison of LoPA-LG, LoPA-LG and SPDB algorithms over 8 nodes.

Distributed classification learning on MNIST datasets. We employ MNIST datasets to train m personalized classifiers. We construct a classifier in node i that consists of a hidden layer followed by sigmoid activation functions with parameters x shared across all nodes, and a linear layer with parameters θ_i adapted to node-specific samples. We use the cross-entropy loss 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. The experiment is conducted in two cases: $m = 4$ with each node owning 14000 samples and $m = 8$ with each node owing 6500 samples, in which communication networks are generated by random Erdős–Rényi graphs. Each node i is assigned with a random subset of overall 10 classes such that each node has different label distributions and significant data heterogeneity. The step-sizes are set as $\alpha = 0.01$, $\beta = 0.01$, $\lambda = 0.005$, $\gamma = 0.2$, $\tau = 0.2$ both for LoPA-LG and LoPA-GT. We compare our algorithms with the state-of-the-art SPDB algorithm [11]. The experiment results for loss, training accuracy, and testing accuracy are presented in Figures 1 and 2. It can be observed from 1 that the proposed LoPA-LG and LoPA-GT algorithms have lower computational complexity than SPDB [11] in terms of the number of Hessian matrices to achieve the same desired performance. 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. In this case, we conduct an 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.

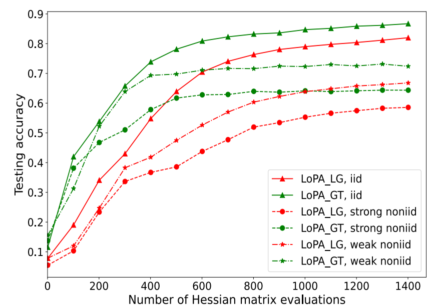
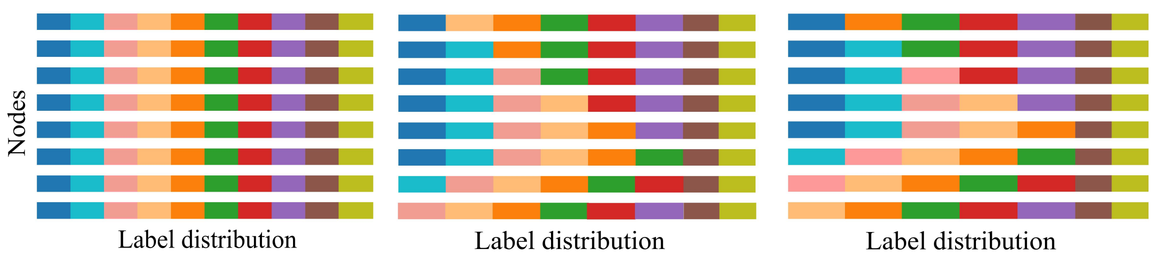


Figure 3: Testing accuracy of LoPA-LG and LoPA-GT under different data heterogeneity.

We plot the testing accuracy of LoPA-LG and LoPA-GT with the same parameter settings as 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.



(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.

6 Conclusion

In this paper, we have proposed a new loopless algorithm LoPA, for solving nonconvex-strongly-convex distributed stochastic bilevel optimization problems with personalized inner-level objectives. The proposed algorithm LoPA can significantly reduce the computational complexity of gradient evaluation. We have shown how the convergence rate of the algorithm is impacted by the heterogeneity of out and inner-level functions under the local gradient scheme, and how the heterogeneity is eliminated by gradient tracking. Moreover, we have proved that LoPA achieves a computational complexity of $\mathcal{O}(\epsilon^{-2})$, which improves existing results for DSBO by an order of $\mathcal{O}(\log \epsilon^{-1})$. Numerical experiments were conducted to verify the effectiveness of LoPA and demonstrate the impact of the heterogeneity.

References

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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, 2022.

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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*, 2022.

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

[21] 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.

[22] 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, 2021.

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

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

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

[26] 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, 2022.

[27] 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.

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

[29] 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.

[30] 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*, 2022.

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

[32] 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.

[33] 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, 2021.

[34] 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.

[35] 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.

[36] 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.

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

[38] Serge Lang. *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media, 2012.

[39] 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.

Appendix

A Technical Preliminaries and Propositions

A. 1 Technical Preliminaries

Notation. For notional convenience, we introduce necessary notations as follows:

- $v_i\left(x,\theta\right) \triangleq \left[\nabla^2_{\theta\theta} g_i\left(x,\theta\right)\right]^{-1} \nabla_\theta f_i\left(x,\theta\right)$.
- $v_i^*\left(x\right) \triangleq v_i\left(x,\theta_i^*(x)\right)$.
- $\bar{x}^k \triangleq (1/m)\sum_{i=1}^m x_i^k$.
- $x^k \triangleq \mathrm{col}\{x_i^k\}_{i=1}^m$.
- $\theta^*(\bar{x}^k) \triangleq \mathrm{col}\{\theta_i^*(\bar{x}^k)\}_{i=1}^m$.
- $v^*(\bar{x}^k) \triangleq \mathrm{col}\{v_i^*(\bar{x}^k)\}_{i=1}^m$.
- $1_m = \mathrm{col}\{1\}_{i=1}^m$.

Additionally, we define some compact notations as follows:

- $\nabla_\theta \hat{G}(x^k,\theta^k;\xi_1^{k+1}) \triangleq \mathrm{col}\{\nabla_\theta \hat{g}_i(x_i^k,\theta_i^k;\xi_{i,1}^{k+1})\}_{i=1}^m$.
- $\nabla^2_{\theta\theta} \hat{G}(x^k,\theta^k;\xi_2^{k+1}) \triangleq \mathrm{diag}\{\nabla^2_{\theta\theta} \hat{g}_i(x_i^k,\theta_i^k;\xi_{i,2}^{k+1})\}_{i=1}^m$.
- $\nabla^2_{x\theta} \hat{G}(x^k,\theta^k;\xi_3^{k+1}) \triangleq \mathrm{diag}\{\nabla^2_{x\theta} \hat{g}_i(x_i^k,\theta_i^k;\xi_{i,3}^{k+1})\}_{i=1}^m$.
- $\nabla_\theta \hat{F}(x^k,\theta^k;\varsigma_1^{k+1}) \triangleq \mathrm{col}\{\nabla_\theta \hat{f}_i(x_i^k,\theta_i^k;\varsigma_{i,1}^{k+1})\}_{i=1}^m$.
- $\nabla_x \hat{F}\left(x^k,\theta^k;\varsigma_2^{k+1}\right) \triangleq \mathrm{col}\{\nabla_x \hat{f}_i(x_i^k,\theta_i^k;\varsigma_{i,2}^{k+1})\}_{i=1}^m$.

The proposed LoPA algorithm in a compact form. For ease of subsequent analysis, by letting $\mathcal{W} = W \otimes I_n$, we can rewrite the LoPA algorithm in a compact form as follows:

$$\theta^{k+1} = \theta^k - \beta d^k \tag{20a}$$
$$v^{k+1} = v^k - \gamma h^k \tag{20b}$$
$$x^{k+1} = (1-\tau)x^k + \tau(\mathcal{W}x^k - \alpha y^k) \tag{20c}$$
$$d^{k+1} = \nabla_\theta \hat{G}(x^{k+1},\theta^{k+1};\xi_1^{k+1}) \tag{20d}$$
$$h^{k+1} = \nabla^2_{\theta\theta} \hat{G}(x^{k+1},\theta^{k+1};\xi_2^{k+1})v^{k+1} - \nabla_\theta \hat{F}(x^{k+1},\theta^{k+1};\varsigma_1^{k+1}) \tag{20e}$$
$$s^{k+1}(\zeta^{k+1}) = \nabla_x \hat{F}(x^{k+1},\theta^{k+1};\varsigma_2^{k+1}) - \nabla^2_{x\theta} \hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^{k+1} \tag{20f}$$
$$z^{k+1} = s^{k+1}(\zeta^{k+1}) + (1-\gamma)(z^k - s^k(\zeta^{k+1})) \tag{20g}$$

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

$$y^{k+1} = z^{k+1}, \tag{21}$$

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

$$y^{k+1} = \mathcal{W}y^k + z^{k+1} - z^k. \tag{22}$$

Basic inequalities. In the subsequent analysis, we will utilize a set of fundamental inequalities 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 $x_1, \cdots, x_m$:
$$\left\|\frac{1}{m}\sum_{i=1}^{m} x_i\right\|^2 \leq \frac{1}{m}\sum_{i=1}^{m}\|x_i\|^2.$$
- Standard variance decomposition for stochastic vector x:
$$\mathbb{E}\left[\|x-\mathbb{E}[x]\|^2\right] = \mathbb{E}\left[\|x\|^2\right] - \|\mathbb{E}[x]\|^2.$$

A. 2 Supporting Propositions

The following two propositions provide the smoothness property of $\nabla\Phi(x)$, $\theta_i^*(x)$, and $v_i^*(x)$, as well as the boundness of the Hv estimate v_i^k, with the first proposition being adapted from [24] and the second derived using induction arguments. For completeness, we provide a proof in Section C.

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

$$\begin{aligned}
&\left\|\theta_i^*(x) - \theta_i^*(x')\right\| \leqslant L_{\theta^*}\left\|x-x'\right\|, \quad \left\|v_i(x,\theta) - v_i(x',\theta')\right\| \leqslant L_v\left(\left\|x-x'\right\| + \left\|\theta-\theta'\right\|\right), \\
&\left\|v_i^*(x) - v_i^*(x')\right\| \leqslant L_{v^*}\left\|x-x'\right\|, \quad \left\|\bar{\nabla} f_i(x,\theta) - \bar{\nabla} f_i(x',\theta')\right\| \leqslant L_f\left(\left\|x-x'\right\| + \left\|\theta-\theta'\right\|\right), \\
&\left\|\nabla\Phi(x) - \nabla\Phi(x')\right\| \leqslant L\left\|x-x'\right\|,
\end{aligned}$$

where the Lipschitz constants are provided as follows:

$$\begin{aligned}
L_{\theta^*} &\triangleq \frac{C_{g,x\theta}}{\mu_g}, \\
L_v &\triangleq \frac{L_{f,\theta}}{\mu_g} + \frac{C_{f,\theta}L_{g,\theta\theta}}{\mu_g^2}, \\
L_{v^*} &\triangleq \left(\frac{L_{f,\theta}}{\mu_g} + \frac{C_{f,\theta}L_{g,\theta\theta}}{\mu_g^2}\right)(1+L_{\theta^*}), \\
L_f &\triangleq L_{f,x} + C_{g,x\theta}L_v + \frac{C_{f,\theta}L_{g,x\theta}}{\mu_g}, \\
L &\triangleq \left(L_{f,x} + C_{g,x\theta}L_v + \frac{C_{f,\theta}L_{g,x\theta}}{\mu_g}\right)(1+L_{\theta^*}).
\end{aligned} \tag{23}$$

Proposition 2 (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 $k+1$: $\forall i \in \mathcal{V}$,*

$$\left\|v_i^{k+1}\right\| \leqslant M. \tag{24}$$

B Proof of Theorems 1 and 2

B.1 Proof Sketch

The key idea of the proof for Theorems 1 and 2 is to characterize the dynamic of the following unified Lyapunov function with properly selected coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6:

$$\begin{aligned} V^k =& d_0\Phi(\bar{x}^k) + d_1\frac{1}{m}\underbrace{\|v^k - v^*(\bar{x}^k)\|^2}_{\text{Hv errors}} + d_2\frac{1}{m}\underbrace{\|\theta^k - \theta^*(\bar{x}^k)\|^2}_{\text{inner-level errors}} \\ &+ d_3\underbrace{\|\bar{s}^k - \bar{z}^k\|^2}_{\text{ave-variance errors}} + d_4\frac{1}{m}\underbrace{\|s^k - z^k\|^2}_{\text{variance errors}} + d_5\frac{1}{m}\underbrace{\|x^k - 1_m\otimes\bar{x}^k\|^2}_{\text{consensus errors}} + d_6\frac{1}{m}\underbrace{\|y^k - 1_m\otimes\bar{y}^k\|^2}_{\text{gradient errors}}, \end{aligned} \tag{25}$$

where $s^k \triangleq \mathbb{E}[s^k(\zeta^k)|\mathcal{F}^k]$.

To this end, we proceed to derive iterative evolutions for each term of V^k in expectation according to the following four key steps:

Step 1: 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 $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2]$ and average variance errors $\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2]$ in Lemma 2.

Step 2: Then we deal with the average variance errors $\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2]$ according to the bounded variances of different stochastic gradients and the updates of z^k. However, controlling 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 3 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 $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2]$, while the gradient increment errors $\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2]$ control the average variance errors $\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2]$. In what follows, we focus on quantifying these four error terms and establishing their recursions in Lemmas 4, 5, 6, 7.

Step 3: The next step is to control 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 8 and Lemma 9. In particular, it is shown that the gradient errors are impacted by the heterogeneity for LoPA-LG with local gradient scheme (11), whereas they decay as the iteration progresses for LoPA-GT with gradient-tracking scheme (12). Lemma 10 further provides the evolution for the variance errors $\mathbb{E}[\|s^k - z^k\|^2]$ induced by the gradient errors $\mathbb{E}[\|y^k - 1_m\otimes\bar{y}^k\|^2]$.

Step 4: Finally, by integrating the obtained results, we establish the dynamics of V^k in Section B.3 and Section B.5 for LoPA-LG and LoPA-GT, respectively, with carefully chosen coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6.

B.2 Supporting Lemmas

Lemma 2 (Descent lemma) *Suppose Assumptions 1, 2, 3 and 5 hold and recall that* $s^k = \mathbb{E}[s^k(\zeta^k)|\mathcal{F}^k]$. *Considering the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1, then we have* *the following inequality:*

$$\begin{aligned}\mathbb{E}[\Phi(\bar{x}^{k+1})] \leqslant & \mathbb{E}[\Phi(\bar{x}^k)] - \frac{\alpha}{2}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] - \frac{\alpha}{2}(1-\alpha L)\mathbb{E}[\|\bar{y}^k\|^2] \\ & + \alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2] + \alpha\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2].\end{aligned} \tag{26}$$

The proof of Lemma 2 is provided in Section D.2. ■

Lemma 3 (Hypergradient approximation errors and average variance errors) *Suppose Assumptions 1, 2, 3 and 5 hold and define* $L_{fg,x} \triangleq 2L_{f,x}^2 + 4M^2 L_{g,x\theta}^2$. *Considering the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1, if the step-size* α *satisfies* $0 < \gamma < 1$, *then we have:*

$$\begin{aligned}\mathbb{E}[\|\bar{s}^{k+1} - \bar{z}^{k+1}\|^2] \leqslant & (1-\gamma)\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2] \\ & + \frac{1}{m^2}\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2] + \frac{1}{m}\sigma_{\bar{z}}^2\alpha^2,\end{aligned} \tag{27}$$

and

$$\begin{aligned}\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2] \leqslant & \frac{L_{fg,x}}{m}\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + \frac{L_{fg,x}}{m}\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\ & + \frac{4C_{g,x\theta}^2}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2].\end{aligned} \tag{28}$$

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

The proof of Lemma 3 is provided in Section D.3. ■

Lemma 4 (Hessian-inverse-vector product errors) *Suppose Assumptions 1, 2, 3 and 5 hold and consider the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1. If the step-size* λ *satisfies*

$$\lambda < \frac{1}{\mu_g}, \tag{29}$$

then we have:

$$\begin{aligned}\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2] \leqslant & (1-\mu_g\lambda)\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2] + q_x\alpha\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] \\ & + q_x\alpha\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] + mq_s\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + m\sigma_v^2\alpha^2,\end{aligned} \tag{30}$$

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

The proof of Lemma 4 is provided in Section D.4. ■

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

$$\beta < \min\left\{\frac{2}{\mu_g + L_{g,\theta}}, \frac{\mu_g + L_{g,\theta}}{2\mu_g L_{g,\theta}}\right\}, \tag{31}$$

then we have:

$$\begin{aligned}\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2] \leqslant & (1 - \frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}}\beta)\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\ & + p_x\alpha\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + mp_s\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + m\sigma_\theta^2\alpha^2,\end{aligned} \tag{32}$$

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* $\omega_\theta \triangleq \frac{\mu_g L_{g,\theta}}{2(\mu_g + L_{g,\theta})}$.

The proof of Lemma 5 is provided in Section D.5. ■

Lemma 6 (Consensus errors) *Suppose Assumptions 1, 2, 3 and 5 hold and consider the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1. Then we have:*

$$\begin{aligned} &\mathbb{E}[\|x^{k+1} - 1_m \otimes \bar{x}^{k+1}\|^2] \\ \leqslant&(1 - \tau\frac{1-\rho}{2})\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + \frac{2\tau\alpha^2}{1-\rho}\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2], \end{aligned} \tag{33}$$

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

The proof of Lemma 6 is provided in Section D.6. ■

Lemma 7 (Gradient increment errors) *Suppose Assumptions 1, 2, 3 and 5 hold and consider the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1. Then we have:*

$$\begin{aligned} \mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2] \leqslant& u_x\alpha^2\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] + u_\theta\alpha^2\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\ &+ u_v\alpha^2\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2] + u_y\tau^2\alpha^2\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2] \\ &+ mu_s\tau^2\alpha^2\mathbb{E}[\|\bar{y}^k\|^2] + m\sigma_u^2\alpha^2, \end{aligned} \tag{34}$$

where $u_x \triangleq 12L_{fg,x}\frac{\tau^2}{\alpha^2} + L_{g,\theta}^2 L_{fg,x}\frac{\beta^2}{\alpha^2} + 4C_{g,x\theta}^2 L_{fg,\theta}\frac{\lambda^2}{\alpha^2}$, $u_\theta \triangleq L_{fg,x}L_{g,\theta}^2\frac{\beta^2}{\alpha^2} + 4C_{g,x\theta}^2 L_{fg,\theta}\frac{\lambda^2}{\alpha^2}, u_v \triangleq 16C_{g,x\theta}^2 L_{g,\theta}^2\frac{\lambda^2}{\alpha^2}$, $u_s \triangleq 3L_{fg,x}, u_y \triangleq 3L_{fg,x}$, $\sigma_u^2 \triangleq 4C_{g,x\theta}^2(\sigma_{f,\theta}^2 + M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} + L_{fg,x}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}$.

The proof of Lemma 7 is provided in Section D.7. ■

Lemma 8 (Gradient errors in the case of local gradient scheme (11)**)** *Suppose Assumptions 1, 2, 3, 4 and 5 hold and consider the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1 in the case of local gradient scheme* (11)*. Then we have:*

$$\begin{aligned} &\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2] \\ \leqslant&3b^2 + 3m\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] + 6\mathbb{E}[\|s^k - z^k\|^2] \\ &+ 6\mathbb{E}[L_{fg,x}\|x^k - 1_m \otimes \bar{x}^k\|^2 + L_{fg,x}\|\theta^k - \theta^*(\bar{x}^k)\|^2 + 4C_{g,x\theta}^2\|v^k - v^*(\bar{x}^k)\|^2], \end{aligned} \tag{35}$$

where b^2 *is the heterogeneity on overall hypergradients denoted in Lemma 1.*

The proof of Lemma 8 is provided in Section D.8. ■

Lemma 9 (Gradient errors in the case of gradient tracking scheme (12)**)** *Suppose Assumptions 1, 2, 3 and 5 hold and consider the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1 in the case of gradient tracking scheme* (12)*. Then we have:*

$$\begin{aligned} &\mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2] \\ \leqslant&\frac{1+\rho}{2}\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2] + \frac{4}{1-\rho}\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2] \\ &+ \frac{4}{1-\rho}\gamma^2\mathbb{E}[\|s^k - z^k\|^2] + \frac{4}{1-\rho}m\sigma_y^2\alpha^2. \end{aligned} \tag{36}$$

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

The proof of Lemma 9 is provided in Section D.9. ■

Lemma 10 (Variance errors) *Suppose Assumptions 1, 2, 3 and 5 hold and consider the sequence* $\{x_i^k, \theta_i^k, v_i^k, z_i^k, y_i^k\}$ *generated by Algorithm 1. Then we have*

$$\begin{aligned}\mathbb{E}[\|s^{k+1} - z^{k+1}\|^2] \leqslant &(1-\gamma)\mathbb{E}[\|s^k - z^k\|^2] \\ &+ \mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2] + m\sigma_z^2\alpha^2,\end{aligned} \tag{37}$$

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

The proof of Lemma 10 is similar to Lemma 3 and we omit it here. ■

Remark 6 (Different hypergradient analysis) *Our analysis for hypergradients differs from existing distributed literature [10, 11, 18–20]. In our approach, we employ Hv errors, inner-level errors, and consensus errors to control the hypergradient approximation errors in Lemma 3. Conversely, these works utilize NS and CG methods* (13) *along with extra computation loops to obtain highly accurate approximations* $\tilde{\nabla}\Phi(\bar{x}^k)$ *(corresponding to the term* z^k *in* (26)*) of hypergradients. They bound the approximation errors of hypergradients by* $\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \tilde{\nabla}\Phi(\bar{x}^k)\|^2] \leqslant \frac{C_{g,x\theta}C_{f,\theta}}{\mu_g}(1 - \frac{\mu g}{Lg,\theta})^Q$ *[11, 24, 28], where* Q *denotes an increasing number of iterations in the computation loops.*

Remark 7 (Different heterogeneity analysis) *We adopt a distinct approach to handle the gradient errors* $\mathbb{E}[\|y^k - 1_m \otimes \bar{y}^k\|^2]$ *discussed in Lemma 6 regarding the consensus errors. Previous works on DSBO generally require each local hypergradient to be bounded by imposing a stricter assumption regarding the boundedness of* $\nabla_x f_i(x, \theta)$ *in order to control the gradient errors. Unlike these works, we derive a more accurate bound for the gradient errors by analyzing the heterogeneity on overall hypergradients in Lemma 8. In contrast to traditional distributed optimization approaches [33, 35, 36], our method takes into account not only the consensus errors but also the influences of Hv errors, inner-level errors, and variance errors on the gradient errors. These multiple sources of errors further complicate the study of the impact of the heterogeneity on convergence behavior under local gradient scheme* (21) *in DSBO problems. Furthermore, when employing the gradient tracking scheme* (22)*, Lemma 9 demonstrates that the heterogeneity is absorbed by the gradient increment errors and gradient errors in a recursive manner.*

Remark 8 (Effect of gradient momentum steps) *It follows from Lemma 3 that the gradient momentum step* (20g) *can help reduce the influence of stochastic variances by leveraging the gradient increment term* $\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2]$ *to absorb the stochastic noise of the order* $\mathcal{O}(\alpha\sigma^2)$*. To illustrate this point, let's consider the case where we exclude the recursion* (20g)*, i.e.,* $z^k = s^k(\zeta^k)$*. In this case, the average variance error term* $\alpha\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2]$ *in* (26) *reduces to* $\alpha\mathbb{E}[\|\bar{s}^k - \bar{s}^k(\zeta^k)\|^2]$ *with an order of* $\mathcal{O}(\alpha\sigma^2)$*. Consequently, this case either leads to steady-state variance errors or requires the selection of two-timescale step-sizes to ensure that the ratios* $\frac{\alpha^2}{\gamma}$ *in* (30) *and* $\frac{\alpha^2}{\beta}$ *in* (32) *approach 0 before the term* $\mathbb{E}[|\bar{y}^k|^2]$*, where the later will lead to a rate of* $\mathcal{O}(K^{-2/5})$ *[8]. It is noteworthy that similar techniques have been employed in the realm of bilevel optimization in prior works such as [10, 20, 26, 29, 39].*

In what follows, we determine the coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 and establish the descent of the Lyapunov function (25) for Theorem 1 and Theorem 2 with the help of Lemmas 1-10.

B.3 Proof of Theorem 1

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 (25) to establish the dynamic of the function. To this end, we first set the coefficients d_0, d_1 and d_2 as follows:

$$d_0 = \mu_g \omega_\theta \frac{\lambda\beta}{\alpha^2}, d_1 = 8C_{g,x\theta}^2 \omega_\theta \frac{\beta}{\alpha}, d_2 = 8C_{g,x\theta}^2 q_x + \mu_g L_{fg,x} \frac{\lambda}{\alpha}. \tag{38}$$

where the parameters $L_{fg,x} = 2L_{f,x}^2 + 4M^2 L_{g,x\theta}^2$, $q_x = \frac{4L_{fg,\theta}\lambda}{\mu_g \alpha}$ and $\omega_\theta = \frac{\mu_g L_{g,\theta}}{2(\mu_g + L_{g,\theta})}$ are defined in Lemmas 3, 4 and 5, respectively. Then considering above coefficients and combining Lemmas 2-5, we can reach the following inequality:

$$\begin{aligned}
&d_0\mathbb{E}[\Phi(\bar{x}^{k+1})] + d_1\frac{1}{m}\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2] + d_2\frac{1}{m}\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2] \\
&\leqslant d_0\mathbb{E}[\Phi(\bar{x}^{k})] + d_1\frac{1}{m}\mathbb{E}[\|v^{k} - v^*(\bar{x}^{k})\|^2] + d_2\frac{1}{m}\mathbb{E}[\|\theta^{k} - \theta^*(\bar{x}^{k})\|^2] \\
&\quad - \frac{d_0}{2}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] - (\frac{d_0}{2}\alpha(1-\alpha L) - d_1 q_s\alpha^2 - d_2 p_s \alpha^2)\mathbb{E}[\|\bar{y}^k\|^2] \\
&\quad - d_1\frac{\mu_g\lambda}{2}\frac{1}{m}\mathbb{E}[\|v^{k} - v^*(\bar{x}^{k})\|^2] - d_2\omega_\theta\beta\frac{1}{m}\mathbb{E}[\|\theta^{k} - \theta^*(\bar{x}^{k})\|^2] \\
&\quad + (d_0 L_{fg,x} + d_1 q_x + d_2 p_x)\alpha\frac{1}{m}\mathbb{E}[\|x^k - 1_m\otimes\bar{x}^k\|^2] \\
&\quad + d_0\alpha\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2] + \left(d_1\sigma_v^2 + d_2\sigma_\theta^2\right)\alpha^2.
\end{aligned} \tag{39}$$

We next deal with the average variance error term $\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2]$ in (39). Specifically, we let $d_3 = \frac{2d_0\alpha}{\gamma}$ and $d_4 = \frac{2d_0\alpha^2}{\gamma}$ and add the term $d_3\mathbb{E}[\|\bar{s}^{k+1} - \bar{z}^{k+1}\|^2]$ and $d_4\frac{1}{m}\mathbb{E}[\|s^{k+1} - z^{k+1}\|^2]$ in both sides of (39). Then, employing Lemmas 3 and 10 gives us:

$$\begin{aligned}
&d_0\mathbb{E}[\Phi(\bar{x}^{k+1})] + d_1\frac{1}{m}\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2] \\
&\quad + d_2\frac{1}{m}\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2] + d_3\mathbb{E}[\|\bar{s}^{k+1} - \bar{z}^{k+1}\|^2] + d_4\frac{1}{m}\mathbb{E}[\|s^{k+1} - z^{k+1}\|^2] \\
&\leqslant d_0\mathbb{E}[\Phi(\bar{x}^{k})] + d_1\frac{1}{m}\mathbb{E}[\|v^{k} - v^*(\bar{x}^{k})\|^2] \\
&\quad + d_2\frac{1}{m}\mathbb{E}[\|\theta^{k} - \theta^*(\bar{x}^{k})\|^2] + d_3\mathbb{E}[\|\bar{s}^{k} - \bar{z}^{k}\|^2] + d_4\frac{1}{m}\mathbb{E}[\|s^{k} - z^{k}\|^2] \\
&\quad - \frac{d_0}{2}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] - (\frac{d_0}{2}\alpha\,(1-\alpha L) - d_1 q_s\alpha^2 - d_2 p_s \alpha^2)\mathbb{E}[\|\bar{y}^k\|^2] \\
&\quad - d_1\frac{\mu_g\lambda}{2}\frac{1}{m}\mathbb{E}[\|v^{k} - v^*(\bar{x}^{k})\|^2] - d_2\omega_\theta\beta\frac{1}{m}\mathbb{E}[\|\theta^{k} - \theta^*(\bar{x}^{k})\|^2] \\
&\quad + (d_0 L_{fg,x} + d_1 q_x + d_2 p_x)\alpha\frac{1}{m}\mathbb{E}[\|x^k - 1_m\otimes\bar{x}^k\|^2] \\
&\quad - d_3\frac{\gamma}{2}\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2] - d_4\gamma\frac{1}{m}\mathbb{E}[\|s^k - z^k\|^2] + (d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_z^2)\alpha^2 \\
&\quad + (d_3\frac{1}{m} + d_4)\frac{1}{m}\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2].
\end{aligned} \tag{40}$$

We next use Lemma 7 to bound the gradient increment errors $\mathbb{E}[\|s^{k+1}(\zeta^{k+1})-s^k(\zeta^{k+1})\|^2]$ as follows:

$$
\begin{aligned}
&d_0\mathbb{E}[\Phi(\bar{x}^{k+1})]+d_1\frac{1}{m}\mathbb{E}[\|v^{k+1}-v^*(\bar{x}^{k+1})\|^2]\\
&+d_2\frac{1}{m}\mathbb{E}[\|\theta^{k+1}-\theta^*(\bar{x}^{k+1})\|^2]+d_3\mathbb{E}[\|\bar{s}^{k+1}-\bar{z}^{k+1}\|^2]+d_4\frac{1}{m}\mathbb{E}[\|s^{k+1}-z^{k+1}\|^2]\\
\leqslant & d_0\mathbb{E}[\Phi(\bar{x}^k)]+d_1\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]\\
&+d_2\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2]+d_3\mathbb{E}[\|\bar{s}^k-\bar{z}^k\|^2]+d_4\frac{1}{m}\mathbb{E}[\|s^k-z^k\|^2]\\
&-\frac{d_0}{2}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]-(\frac{1}{2}d_0\alpha(1-\alpha L)-d_1q_s\alpha^2-d_2p_s\alpha^2-(d_3\frac{1}{m}+d_4)u_s\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2]\\
&-(d_1\frac{\mu_g\lambda}{2}-(d_3+d_4)u_v\alpha^2)\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]-(d_2\omega_\theta\beta-(d_3\frac{1}{m}+d_4)u_\theta\alpha^2)\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\
&-d_3\frac{\gamma}{2}\mathbb{E}[\|\bar{s}^k-\bar{z}^k\|^2]-d_4\gamma\frac{1}{m}\mathbb{E}[\|s^k-z^k\|^2]+(d_3\frac{1}{m}+d_4)u_y\tau^2\alpha^2\frac{1}{m}\mathbb{E}[\|y^k-1_m\otimes\bar{y}^k\|^2]\\
&+((d_0L_{fg,x}+d_1q_x+d_2p_x)\alpha+(d_3\frac{1}{m}+d_4)u_x\alpha^2)\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]\\
&+(d_1\sigma_v^2+d_2\sigma_\theta^2+\frac{1}{m}d_3\sigma_{\bar{z}}^2+d_4\sigma_z^2+(d_3\frac{1}{m}+d_4)\sigma_u^2)\alpha^2.
\end{aligned}
\tag{41}
$$

We proceed in eliminating the term $\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]$. Letting

$$
d_5=\frac{\alpha}{\tau}\frac{4}{1-\rho}((d_0L_{fg,x}+d_1q_x+d_2p_x)+(d_3\frac{1}{m}+d_4)u_x\alpha),d_6=0,
\tag{42}
$$

and employing Lemma 6 and the inequality (41) yields the following inequality:

$$
\begin{aligned}
\mathbb{E}[V^{k+1}]\leqslant&\mathbb{E}[V^k]-\frac{d_0}{2}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\\
&-(\frac{1}{2}d_0\alpha(1-\alpha L)-d_1q_s\alpha^2-d_2p_s\alpha^2-(d_3+d_4)u_s\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2]\\
&-(d_1\frac{\mu_g\lambda}{2}-(d_3\frac{1}{m}+d_4)u_v\alpha^2)\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]\\
&-(d_2\omega_\theta\beta-(d_3\frac{1}{m}+d_4)u_\theta\alpha^2)\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\
&-d_3\frac{\gamma}{2}\mathbb{E}[\|\bar{s}^k-\bar{z}^k\|^2]-d_4\gamma\frac{1}{m}\mathbb{E}[\|s^k-z^k\|^2]-d_5\tau\frac{1-\rho}{4}\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]\\
&+((d_3\frac{1}{m}+d_4)\frac{\tau^2u_y}{\alpha}+d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3\frac{1}{m}\mathbb{E}[\|y^k-1_m\otimes\bar{y}^k\|^2]\\
&+(d_1\sigma_v^2+d_2\sigma_\theta^2+\frac{1}{m}d_3\sigma_{\bar{z}}^2+d_4\sigma_z^2+(d_3\frac{1}{m}+d_4)\sigma_u^2)\alpha^2.
\end{aligned}
\tag{43}
$$

We note that for LoPA-GT with local gradient scheme (11), the gradient errors can be upper bounded by (36) in Lemma 8. Substituting this boundness into the above inequality, we get:

$$
\begin{aligned}
&\mathbb{E}[V^{k+1}] \\
&\leqslant \mathbb{E}[V^k] - \frac{d_0}{4}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \\
&- (\frac{d_0}{4}\alpha - 3((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3)\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \\
&- (\frac{1}{2}d_0\alpha(1-\alpha L) - d_1 q_s\alpha^2 - d_2 p_s\alpha^2 - (d_3\frac{1}{m} + d_4)u_s\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2] \\
&- (d_1\frac{\mu_g\lambda}{2} - (d_3\frac{1}{m} + d_4)u_v\alpha^2 - 24C_{g,x\theta}^2((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3)\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2] \\
&- (d_2\omega_\theta\beta - (d_3\frac{1}{m} + d_4)u_\theta\alpha^2 - 6L_{fg,x}((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3)\frac{1}{m}\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
&- (d_4\gamma - 6((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3)\frac{1}{m}\mathbb{E}[\|s^k - z^k\|^2] - d_3\frac{\gamma}{2}\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2] \\
&- (d_5\tau\frac{1-\rho}{4} - 6L_{fg,x}((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3)\frac{1}{m}\mathbb{E}[\|x^k - 1_m\otimes\bar{x}^k\|^2] \\
&+ (d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_z^2 + (d_3\frac{1}{m} + d_4)\sigma_u^2)\alpha^2 + 3((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3 b^2.
\end{aligned} \tag{44}
$$

If the step-size α satisfies the following conditions:

$$\frac{1}{2}d_0\alpha(1-\alpha L) - d_1 q_s\alpha^2 - d_2 p_s\alpha^2 - (d_3\frac{1}{m} + d_4)u_s\tau^2\alpha^2 \geqslant 0, \tag{45}$$

$$\frac{d_0}{4}\alpha - 3((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3 \geqslant 0, \tag{46}$$

$$d_1\frac{\mu_g\lambda}{2} - (d_3\frac{1}{m} + d_4)u_v\alpha^2 - 24C_{g,x\theta}^2((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3 \geqslant 0, \tag{47}$$

$$d_2\omega_\theta\beta - (d_3\frac{1}{m} + d_4)u_\theta\alpha^2 - 6L_{fg,x}((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3 \geqslant 0, \tag{48}$$

$$d_4\gamma - 6((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3 \geqslant 0, \tag{49}$$

$$d_5\tau\frac{1-\rho}{4} - 6L_{fg,x}((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3 \geqslant 0, \tag{50}$$

then we further have:

$$\mathbb{E}[V^{k+1}] \leqslant \mathbb{E}[V^k] - \frac{d_0}{4}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \tag{51}$$

$$+ (d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_z^2 + (d_3\frac{1}{m} + d_4)\sigma_u^2)\alpha^2 \tag{52}$$

$$+ 3((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha})\alpha^3 b^2, \tag{53}$$

where the coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 of the Lyapunov function (25) are as follows:

$$\begin{aligned} & d_0 = \mu_g \omega_\theta \frac{\lambda\beta}{\alpha^2}, d_1 = 8C_{g,x\theta}^2 \omega_\theta \frac{\beta}{\alpha}, d_2 = 8C_{g,x\theta}^2 q_x + \mu_g L_{fg,x}\frac{\lambda}{\alpha}, \\ & d_3 = \frac{2d_0\alpha}{\gamma},\ d_4 = \frac{2d_0\alpha^2}{\gamma},\ d_5 = \frac{\alpha}{\tau}\frac{4}{1-\rho}((d_0L_{fg,x} + d_1q_x + d_2p_x) + (d_3\frac{1}{m} + d_4)u_x\alpha), d_6 = 0. \end{aligned} \tag{54}$$

Next, we proceed to find the sufficient conditions for the step-sizes to satisfy the conditions (45) to (50). To ensure that condition (45) holds, a sufficient condition is:

$$\alpha \leqslant u_1 = \min\left\{\frac{1}{2L}, \frac{1}{16}\frac{\sqrt{\mu_g \varpi \lambda^2}}{C_{g,x\theta}L_{v^*}}, \frac{\omega_\theta}{8L_{\theta^*}}\sqrt{\frac{\beta^2}{\left(32\frac{C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g} + L_{fg,\theta}\right)}}, \sqrt{\frac{\gamma}{96L_{fg,x}\tau^2}}\right\}. \tag{55}$$

To address the conditions (46)-(50), we start by simplifying the term $d_0L_{fg,x} + d_1q_x + d_2p_x$ in d_5 as:

$$\begin{aligned} d_0L_{fg,x} + d_1q_x + d_2p_x =& (8C_{g,x\theta}^2 q_x + \mu_g L_{fg,x}\frac{\lambda}{\alpha})(\omega_\theta\frac{\beta}{\alpha} + p_x) \\ =& \underbrace{(L_{fg,x} + \frac{32C_{g,x\theta}^2 L_{fg,x}}{\mu_g^2})(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2})}_{\triangleq \varphi}\mu_g\omega_\theta\frac{\lambda\beta}{\alpha^2}, \end{aligned} \tag{56}$$

by which we can further bound the term $(d_3 + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2c_\tau}{1-\rho}$ as follows:

$$\begin{aligned} \vartheta \triangleq& ((d_3\frac{1}{m} + d_4)\frac{\tau^2 u_y}{\alpha} + d_5\frac{2}{1-\rho}\frac{\tau}{\alpha}) \\ \leqslant& ((d_3 + d_4)\frac{\tau^2 u_y}{\alpha} + \frac{8}{(1-\rho)^2}((d_0L_{fg,x} + d_1q_x + d_2p_x) + (d_3 + d_4)u_x\alpha)) \\ =& (d_3 + d_4)\frac{\tau^2 u_y}{\alpha} + \frac{8}{(1-\rho)^2}(d_3 + d_4)u_x\alpha + \frac{8}{(1-\rho)^2}(d_0L_{fg,x} + d_1q_x + d_2p_x) \\ \overset{(a)}{\leqslant}& \frac{8}{(1-\rho)^2}(d_3 + d_4)\frac{u_y\tau^2}{\alpha} + \frac{8}{(1-\rho)^2}(d_3 + d_4)u_x\alpha + \frac{8}{(1-\rho)^2}\varphi\mu_g\omega_\theta\frac{\lambda\beta}{\alpha^2}, \end{aligned} \tag{57}$$

where the step (a) uses the fact that $\frac{1}{(1-\rho)^2} \geqslant 1$. Considering the aforementioned bounds, we can derive the following sufficient selection condition for the step-sizes α, λ, β to satisfy the conditions (46)-(50):

$$\begin{aligned} \alpha \leqslant& u_2 = \min\left\{\frac{(1-\rho)}{16}\sqrt{\frac{\mu_g\omega_\theta}{\varphi}}, \frac{(1-\rho)}{24}\sqrt{\frac{1}{L_{fg,x}}(1 + \frac{4L_{g,\theta}^2}{\omega_\theta^2})},\right. \\ & \left.\frac{1}{96}\frac{(1-\rho)^2\mu_g\omega_\theta}{\varphi}, \frac{1}{12}\frac{1-\rho}{\sqrt{L_{fg,x}}}, \frac{1}{6L_{fg,x}}\sqrt{\frac{\varphi\gamma}{\mu_g\omega_\theta\tau^2}}, \frac{\sqrt{L_{g,\theta}^2 L_{fg,x}\beta^2 + 4C_{x\theta}^2 L_{fg,\theta}\lambda^2}}{3\tau L_{fg,x}}\right\}, \end{aligned} \tag{58}$$

and

$$\begin{aligned}
&\lambda \leqslant \min\left\{\frac{\gamma}{16L_{g,\theta}^2}, \sqrt{\frac{1}{64}\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x}\right)\frac{\gamma}{C_{g,x\theta}^2 L_{fg,\theta}}}\right\}, \\
&\beta \leqslant \sqrt{\frac{1}{16}(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x})\frac{\lambda}{L_{fg,x}L_{g,\theta}^2}},
\end{aligned} \tag{59}$$

where u_x, φ, ϑ are given by (7), (56), (57), respectively. Furthermore, by combining Lemmas 2-5, we have an additional condition for the step-sizes λ, β, τ, γ on the basis of (59) as follows:

$$\begin{aligned}
&\lambda \leqslant \min\left\{\frac{1}{\mu_g}, \frac{\gamma}{16L_{g,\theta}^2}, \sqrt{\frac{1}{64}\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x}\right)\frac{\gamma}{C_{g,x\theta}^2 L_{fg,\theta}}}\right\}, \\
&\beta \leqslant \min\left\{\frac{2}{\mu_g+L_{g,\theta}}, \frac{\mu_g+L_{g,\theta}}{2\mu_g L_{g,\theta}}, \sqrt{\frac{1}{16}(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x})\frac{\lambda}{L_{fg,x}L_{g,\theta}^2}}\right\}, \\
&\tau < 1, \gamma < 1.
\end{aligned} \tag{60}$$

By incorporating the aforementioned selection conditions for the step-size α, we can conclude that:

$$\alpha \leqslant u \triangleq \min\{u_1, u_2\}. \tag{61}$$

with λ, β, τ, γ satisfying the condition (60). Then, by combining the definition of d_0 and the the condition (60) and (61), we can deduce that

$$\mathbb{E}[V^{k+1}] \leqslant \mathbb{E}[V^k] - \frac{d_0}{4}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] + \alpha^2\sigma_r^2 + 3\vartheta\alpha^3 b^2, \tag{62}$$

where ϑ is denoted in (57) and σ_r^2 is given by:

$$\begin{aligned}
\sigma_r^2 \triangleq & d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_z^2 + (\frac{1}{m}d_3 + d_4)\sigma_u^2 \\
= & d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_{\bar{z}}^2 + (\frac{1}{m}d_3 + d_4)\left(2C_{g,x\theta}^2\sigma_v^2 + L_{fg,x}\sigma_\theta^2\right) \\
= & (d_1 + 2C_{g,x\theta}^2(\frac{1}{m}d_3 + d_4))\sigma_v^2 + (d_2 + L_{fg,x}(\frac{1}{m}d_3 + d_4))\sigma_\theta^2 + (\frac{1}{m}d_3 + d_4)\sigma_{\bar{z}}^2 \\
= & 2(d_1 + 2C_{g,x\theta}^2(\frac{1}{m}d_3 + d_4))(\sigma_{f,\theta}^2 + M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} + (d_2 + L_{fg,x}(\frac{1}{m}d_3 + d_4))\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2} \\
& + (\frac{1}{m}d_3 + d_4)(\sigma_{f,x}^2 + M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2},
\end{aligned} \tag{63}$$

where the first equality uses the definition of σ_u^2 in Lemma 7; the second equality uses the fact that $\sigma_{\bar{z}}^2 = \sigma_z^2$ in Lemmas 3 and 10; the last equality is derived by substituting the definitions of $\sigma_{\bar{z}}^2, \sigma_v^2, \sigma_\theta^2$ in Lemmas 3, 4, 5. Now, summing up and telescoping the above inequality from $k = 0$ to K yields:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant \frac{4(V^0 - V^K)}{d_0\alpha(K+1)} + \frac{4}{d_0}\alpha\sigma_r^2 + \frac{12\vartheta}{d_0}\alpha^2 b^2. \tag{64}$$

This completes the proof. ■

B.4 Proof of Corollary 1

In the subsequent analysis, based on Theorem 1 we will select the step-sizes α, γ, β, τ, λ in term of the number of iterations K. When the step-sizes $\gamma,\lambda,\beta,\tau$ are taken as $\gamma=c_\gamma\alpha,\lambda=c_\lambda\alpha,\beta=c_\beta\alpha,\tau=c_\tau\alpha$ with the positive parameters $c_\gamma,c_\lambda,c_\beta,c_\tau$ being independent of K, from (54) we have that $d_0=\mu_g\omega_\theta c_\lambda c_\beta$ and $d_3=\frac{2d_0}{c_\gamma}$ are independent of the step-size α. Then it follows from the inequality (57) and the condition (61) that

$$\begin{aligned}\vartheta \leqslant&\frac{8}{(1-\rho)^2}(d_3+d_4)\frac{u_y\tau^2}{\alpha}+\frac{8}{(1-\rho)^2}(d_3+d_4)u_x\alpha+\frac{8}{(1-\rho)^2}\varphi\mu_g\omega_\theta\frac{\lambda\beta}{\alpha^2}\\ \leqslant&\frac{8}{(1-\rho)^2}(d_3+ud_3)\frac{u_y\tau}{\alpha}+\frac{8}{(1-\rho)^2}(d_3+ud_3)u_xu+\frac{8}{(1-\rho)^2}\varphi\mu_g\omega_\theta\frac{\lambda\beta}{\alpha^2}\\ =&\frac{8}{(1-\rho)^2}(d_3+ud_3)u_yc_\tau+\frac{8}{(1-\rho)^2}(d_3+ud_3)u_xu+\frac{8}{(1-\rho)^2}\varphi\mu_g\omega_\theta c_\lambda c_\beta\\ \triangleq&\hat\vartheta=\mathcal{O}(\frac{1}{(1-\rho)^2}),\end{aligned}\tag{65}$$

where the second step uses the fact that $\tau<1$. Furthermore, with the above-mentioned selection condition for the step-sizes λ, β, τ, γ we also have that σ_v, σ_θ, $\sigma_{\bar z}$ and σ_u 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 $d_4=d_3\alpha$, the variance related term σ_r in (63) can be further derived as:

$$\begin{aligned}\sigma_r=&\sqrt{d_1\sigma_v^2+d_2\sigma_\theta^2+\frac{1}{m}d_3\sigma_{\bar z}^2+d_4\sigma_z^2+(d_3\frac{1}{m}+d_4)\sigma_u^2}\\ \leqslant&\sqrt{d_1\sigma_v^2+d_2\sigma_\theta^2+(d_3\frac{1}{m}+d_3\alpha)\sigma_u^2}+\sqrt{\frac{1}{m}d_3\sigma_{\bar z}^2+d_3\alpha\sigma_{\bar z}^2}\\ \leqslant&\sqrt{d_1\sigma_v^2+d_2\sigma_\theta^2+(d_3\frac{1}{m}+d_3\frac{1}{m})\sigma_u^2}+\sqrt{\frac{1}{m}d_3\sigma_{\bar z}^2+d_3\frac{1}{m}\sigma_{\bar z}^2}\\ \leqslant&\sqrt{d_1\sigma_v^2+d_2\sigma_\theta^2+2d_3\sigma_u^2}+\sqrt{\frac{2}{m}d_3\sigma_{\bar z}^2}\\ \leqslant&\underbrace{\sqrt{d_1}\sigma_v+\sqrt{d_2}\sigma_\theta+2\sqrt{d_3}\sigma_u}_{\triangleq\sigma_{\mathrm{p}}}+\frac{\sqrt{2}}{\sqrt{m}}\underbrace{\sqrt{d_3}\sigma_{\bar z}}_{\triangleq\sigma_c}\triangleq\hat\sigma_r,\end{aligned}\tag{66}$$

where the first inequality holds due to the fact that $\sigma_{\bar z}=\sigma_z$ and $d_4=d_3\alpha$ as well as the triangle inequality; and the second inequality is derived by the condition $\alpha\leqslant\frac{1}{m}$ induced by a large number of iterations K; the third inequality uses the fact that $d_3\frac{1}{m}\leqslant d_3$. By combining the definitions of $\sigma_{\bar z}$, σ_v, σ_θ, σ_u in Lemmas 3, 4, 5, 7, it follows that

$$\begin{aligned}&\sigma_{\mathrm{p}}=\sqrt{d_1}\sigma_v+\sqrt{d_2}\sigma_\theta+\sqrt{2d_3}\sigma_u=\mathcal{O}(\sigma_{f,\theta}+\sigma_{g,\theta\theta}+\sigma_{g,\theta}),\\ &\sigma_{\mathrm{c}}=\sqrt{2d_3}\sigma_{\bar z}=\mathcal{O}(\sigma_{f,x}+\sigma_{g,x\theta}).\end{aligned}\tag{67}$$

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

$$a_0\triangleq\frac{4(V^0-V^K)}{d_0},a_1\triangleq\frac{4}{d_0}\hat\sigma_r^2,a_2\triangleq\frac{12\hat\vartheta}{d_0}b^2.\tag{68}$$

Then combining the inequalities (65) and (66), the inequality (64) becomes:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\leqslant a_0\frac{1}{\alpha\,(K+1)}+a_1\alpha+a_2\alpha^2. \tag{69}$$

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 $\gamma,\lambda,\beta,\tau$ are taken as $\gamma=c_\gamma\alpha,\lambda=c_\lambda\alpha,\beta=c_\beta\alpha,\tau=c_\tau\alpha$ with $c_\gamma,c_\lambda,c_\beta,c_\tau$ being independent of K, we can proceed with the following discussion:

- 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:

$$\begin{aligned}\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\leqslant&\left(\frac{a_1a_0}{(K+1)}\right)^{\frac{1}{2}}+\left(\frac{a_1a_0}{(K+1)}\right)^{\frac{1}{2}}+a_2\left(\frac{a_0}{a_1\,(K+1)}\right)\\ \leqslant&2\left(\frac{a_1a_0}{(K+1)}\right)^{\frac{1}{2}}+a_2^{\frac{1}{3}}\left(\frac{a_0}{(K+1)}\right)^{\frac{2}{3}}.\end{aligned} \tag{70}$$

- 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:

$$\begin{aligned}\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\leqslant&2a_2^{\frac{1}{3}}\left(\frac{a_0}{(K+1)}\right)^{\frac{2}{3}}+a_1\left(\frac{a_0}{a_2\,(K+1)}\right)^{\frac{1}{3}}\\ \leqslant&2a_2^{\frac{1}{3}}\left(\frac{a_0}{(K+1)}\right)^{\frac{2}{3}}+\left(\frac{a_1a_0}{(K+1)}\right)^{\frac{1}{2}}.\end{aligned} \tag{71}$$

- 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:

$$\begin{aligned}\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\leqslant&\frac{a_0}{u\,(K+1)}+a_1\left(\frac{a_0}{a_1\,(K+1)}\right)^{\frac{1}{2}}+a_2\left(\frac{a_0}{a_2\,(K+1)}\right)^{\frac{2}{3}}\\ \leqslant&\frac{a_0}{u\,(K+1)}+\left(\frac{a_1a_0}{(K+1)}\right)^{\frac{1}{2}}+a_2^{\frac{1}{3}}\left(\frac{a_0}{(K+1)}\right)^{\frac{2}{3}}.\end{aligned} \tag{72}$$

With the above discussion regarding (70), (71), and (72), we can conclude that:

$$\begin{aligned}&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\\ \leqslant&\frac{a_0}{u\,(K+1)}+2\left(\frac{a_1a_0}{(K+1)}\right)^{\frac{1}{2}}+2a_2^{\frac{1}{3}}\left(\frac{a_0}{(K+1)}\right)^{\frac{2}{3}}\\ \leqslant&\frac{1}{d_0}\Big(\frac{4\left(V^0-V^K\right)}{u\,(K+1)}+\frac{8\hat{\sigma}_r\sqrt{(V^0-V^K)}}{\sqrt{K+1}}+\frac{(12\hat{\vartheta}b^2)^{\frac{1}{3}}\left(V^0-V^K\right)^{\frac{2}{3}}}{(K+1)^{\frac{2}{3}}}\Big).\end{aligned} \tag{73}$$

Note that $u = \mathcal{O}\left((1-\rho)^2\right)$ in (61), $\hat{\vartheta} = \mathcal{O}(\frac{1}{(1-\rho)^2})$ in (65), $\hat{\sigma}_r = \mathcal{O}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})$ in (66), and d_0 is independent of the term $\frac{1}{1-\rho}$. Then we have that:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{V^0}{(1-\rho)^2 K} + \frac{b^{\frac{2}{3}}(V^0)^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}} + \frac{\sqrt{V^0}}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right). \tag{74}$$

When we initialize the outer-level variables as $x_i^0 = x_j^0, \forall i,j \in \mathcal{V}$, we can derive that $\|x^0 - 1_m \otimes \bar{x}^0\|^2 = 0$ holds and V^0 is independent of the term $\frac{1}{1-\rho}$, which further gives that:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{1}{(1-\rho)^2 K} + \frac{b^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}K^{\frac{2}{3}}} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right). \tag{75}$$

This completes the proof. ■

B.5 Proof of Theorem 2

The proof of Theorem 2 follows similar steps to that of Theorem 1. 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 (25) for LoPA-GT. To be specific, by letting $d_0 = \mu_g\omega_\theta\frac{\lambda\beta}{\alpha^2}, d_1 = 8C_{g,x\theta}^2\omega_\theta\frac{\beta}{\alpha}, d_2 = 8C_{g,x\theta}^2 q_x + \mu_g L_{fg,x}\frac{\lambda}{\alpha}$ and combining Lemmas 2-5, the inequality (39) can also be derived for LoPA-GT. In the inequality (39), we recall that related parameters are defined in Lemmas 2-5. In what follows, we focus on dealing with the term $\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2]$. Note that the evolution of the gradient errors $[\|y^k - 1_m \otimes y^k\|^2]$ are controlled by the term $\mathbb{E}[\|s^k - z^k\|^2]$ under gradient tracking scheme (12). Combining this fact, we let $d_3 = \frac{2d_0\alpha}{\gamma}, d_4 = \frac{2d_0\alpha^2}{\gamma}, d_6 = \frac{2(1-\rho)d_0\alpha^2}{\gamma}$, and leverage Lemmas 3 and 9 to further transform the inequality (39) into:

$$\begin{aligned}
& d_0\mathbb{E}[\Phi(\bar{x}^{k+1})] + d_1\frac{1}{m}\mathbb{E}[\|v^{k+1} - v^*(\bar{x}^{k+1})\|^2] + d_2\frac{1}{m}\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2] \\
& + d_3\mathbb{E}[\|\bar{s}^{k+1} - \bar{z}^{k+1}\|^2] + d_4\frac{1}{m}\mathbb{E}[\|s^{k+1} - z^{k+1}\|^2] + d_6\frac{1}{m}\mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2] \\
\leqslant & d_0\mathbb{E}[\Phi(\bar{x}^k)] + d_1\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2] + d_2\frac{1}{m}\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
& + d_3\mathbb{E}[\|\bar{s}^{k+1} - \bar{z}^{k+1}\|^2] + d_4\frac{1}{m}\mathbb{E}[\|s^{k+1} - z^{k+1}\|^2] + d_6\frac{1}{m}\mathbb{E}[\|y^k - 1_m \otimes y^k\|^2] \\
& - \frac{1}{2}d_0\alpha\mathbb{E}[\left\|\nabla\Phi(\bar{x}^k)\right\|^2] - (\frac{1}{2}d_0\alpha(1-\alpha L) - d_1 q_s\alpha^2 - d_2 p_s\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2] \\
& - d_1\frac{\mu_g\lambda}{2}\frac{1}{m}\mathbb{E}[\|v^k - v^*(\bar{x}^k)\|^2] - d_2\omega_\theta\beta\frac{1}{m}\mathbb{E}[\|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
& - \frac{1-\rho}{2}d_6\frac{1}{m}\mathbb{E}[\|y^k - 1_m \otimes y^k\|^2] + (d_0 L_{fg,x} + d_1 q_x + d_2 p_x)\alpha\frac{1}{m}\mathbb{E}[\|x^k - 1_m \otimes \bar{x}^k\|^2] \\
& - d_3\frac{\gamma}{2}\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2] - (d_4\gamma - d_6\frac{4}{1-\rho}\gamma^2)\frac{1}{m}\mathbb{E}[\|s^k - z^k\|^2] \\
& + (d_3\frac{1}{m} + d_4 + \frac{4}{1-\rho}d_6)\frac{1}{m}\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2] \\
& + (d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_z^2 + \frac{4}{1-\rho}d_6\sigma_y^2)\alpha^2,
\end{aligned} \tag{76}$$

By the upperboundness of the gradient increment errors $\mathbb{E}[\|s^{k+1}\left(\zeta^{k+1}\right)-s^k\left(\zeta^{k+1}\right)\|^2]$ in Lemma 7, the above inequality can be derived as follows:

$$
\begin{aligned}
&d_0\mathbb{E}[\Phi(\bar{x}^{k+1})]+d_1\frac{1}{m}\mathbb{E}[\left\|v^{k+1}-v^*(\bar{x}^{k+1})\right\|^2]+d_2\frac{1}{m}\mathbb{E}[\|\theta^{k+1}-\theta^*(\bar{x}^{k+1})\|^2]\\
&+d_3\mathbb{E}[\|\bar{s}^{k+1}-\bar{z}^{k+1}\|^2]+d_4\frac{1}{m}\mathbb{E}[\|s^{k+1}-z^{k+1}\|^2]+d_6\frac{1}{m}\mathbb{E}[\|y^{k+1}-1_m\otimes\bar{y}^{k+1}\|^2]\\
\leqslant& d_0\mathbb{E}[\Phi(\bar{x}^k)]+d_1\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]+d_2\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\
&+d_3\mathbb{E}[\|\bar{s}^k-\bar{z}^k\|^2]+d_4\frac{1}{m}\mathbb{E}[\|s^k-z^k\|^2]+d_6\frac{1}{m}\mathbb{E}[\|y^k-1_m\otimes y^k\|^2]-\frac{1}{2}d_0\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\\
&-(\frac{1}{2}d_0\alpha(1-\alpha L)-d_1q_s\alpha^2-d_2p_s\alpha^2-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_s\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2]\\
&-(d_1\frac{\mu_g\lambda}{2}-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_v\alpha^2)\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]\\
&-(d_2\omega_\theta\beta-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_\theta\alpha^2)\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\
&-(\frac{1-\rho}{2}d_6-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_y\tau^2\alpha^2)\frac{1}{m}\mathbb{E}[\|y^k-1_m\otimes y^k\|^2]\\
&-d_3\frac{\gamma}{2}\mathbb{E}[\|\bar{s}^k-\bar{z}^k\|^2]-(d_4\gamma-d_6\frac{4}{1-\rho}\gamma^2)\frac{1}{m}\mathbb{E}[\|s^k-z^k\|^2]\\
&+((d_0L_{fg,x}+d_1q_x+d_2p_x)+(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_x\alpha)\alpha\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]\\
&+(d_1\sigma_v^2+d_2\sigma_\theta^2+\frac{1}{m}d_3\sigma_{\bar{z}}^2+d_4\sigma_z^2+\frac{4}{1-\rho}d_6\sigma_y^2+(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)\sigma_u^2)\alpha^2.
\end{aligned}
\tag{77}
$$

Now we let $d_5=\frac{\alpha}{\tau}\frac{2}{1-\rho}(d_0L_{fg,x}+d_1q_x+d_2p_x+(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_x\alpha)$, and add the consensus error term $d_5\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2]$ on both sides of the inequality (77). Then combining in Lemma 7, we establish the following dynamic:

$$
\begin{aligned}
\mathbb{E}[V^{k+1}]\leqslant&\mathbb{E}[V^k]-\frac{1}{2}d_0\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\\
&-(\frac{1}{2}d_0\alpha(1-\alpha L)-d_1q_s\alpha^2-d_2p_s\alpha^2-(d_3+d_4+\frac{4}{1-\rho}d_6)u_s\tau^2\alpha^2)\mathbb{E}[\|\bar{y}^k\|^2]\\
&-(d_1\frac{\mu_g\lambda}{2}-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_v\alpha^2)\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2]\\
&-(d_2\omega_\theta\beta-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_\theta\alpha^2)\frac{1}{m}\mathbb{E}[\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\
&-(\frac{1-\rho}{2}d_6-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_y\tau^2\alpha^2-d_5\frac{2\tau\alpha^2}{1-\rho})\frac{1}{m}\mathbb{E}[\|y^k-1_m\otimes y^k\|^2]\\
&-(d_4\gamma-d_6\frac{4}{1-\rho}\gamma^2)\frac{1}{m}\mathbb{E}[\|s^k-z^k\|^2]\\
&+(d_1\sigma_v^2+d_2\sigma_\theta^2+\frac{1}{m}d_3\sigma_{\bar{z}}^2+d_4\sigma_z^2+\frac{4}{1-\rho}d_6\sigma_y^2+(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)\sigma_u^2)\alpha^2.
\end{aligned}
\tag{78}
$$

Furthermore, when the following conditions hold:

$$\frac{1}{2}d_0\alpha\left(1-\alpha L\right)-d_1q_s\alpha^2-d_2p_s\alpha^2-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_s\tau^2\alpha^2\geqslant 0, \tag{79}$$

$$d_1\frac{\mu_g\lambda}{2}-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_v\alpha^2\geqslant 0, \tag{80}$$

$$d_2\omega_\theta\beta-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_\theta\alpha^2\geqslant 0, \tag{81}$$

$$\frac{1-\rho}{2}d_6-(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_y\tau^2\alpha^2-d_5\frac{2\tau\alpha^2}{1-\rho}\geqslant 0, \tag{82}$$

$$d_4\gamma-d_6\frac{4}{1-\rho}\gamma^2\geqslant 0, \tag{83}$$

we derive that:

$$\begin{aligned}\mathbb{E}[V^{k+1}]\leqslant&\mathbb{E}[V^k]-\frac{d_0}{2}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]\\&+(d_1\sigma_v^2+d_2\sigma_\theta^2+\frac{1}{m}d_3\sigma_{\bar{z}}^2+d_4\sigma_z^2+\frac{4}{1-\rho}d_6\sigma_y^2+(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)\sigma_u^2)\alpha^2,\end{aligned} \tag{84}$$

where we recall that the coefficients d_0, d_1, d_2, d_3, d_4, d_5, d_6 of the Lyapunov function (25) are given by:

$$\begin{aligned}&d_0=\mu_g\omega_\theta\frac{\lambda\beta}{\alpha^2},d_1=8C_{g,x\theta}^2\omega_\theta\frac{\beta}{\alpha},d_2=8C_{g,x\theta}^2q_x+\mu_gL_{fg,x}\frac{\lambda}{\alpha},d_3=\frac{2d_0\alpha}{\gamma},d_4=\frac{2d_0\alpha^2}{\gamma},\\&d_5=\frac{\alpha}{\tau}\frac{2}{1-\rho}(d_0L_{fg,x}+d_1q_x+d_2p_x+(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)u_x\alpha),d_6=\frac{2(1-\rho)d_0\alpha^2}{\gamma},\end{aligned} \tag{85}$$

and $\sigma_{r_0'}$ is denoted as:

$$\sigma_{r'}^2\triangleq d_1\sigma_v^2+d_2\sigma_\theta^2+\frac{1}{m}d_3\sigma_{\bar{z}}^2+d_4\sigma_z^2+\frac{4}{1-\rho}d_6\sigma_y^2+(d_3\frac{1}{m}+d_4+\frac{4}{1-\rho}d_6)\sigma_u^2.$$

Next, we proceed to find the sufficient conditions for the step-sizes to make the conditions (79)-(82) hold. To this end, we first recall that like (56) the term $d_0L_{fg,x}+d_1q_x+d_2p_x$ in d_4 can be simplified as follows:

$$d_0L_{fg,x}+d_1q_x+d_2p_x=\underbrace{(L_{fg,x}+\frac{32C_{g,x\theta}^2L_{fg,x}}{\mu_g^2})(1+\frac{4L_{g,\theta}^2}{\omega_\theta^2})}_{\triangleq\varphi}\mu_g\omega_\theta\frac{\lambda\beta}{\alpha^2}, \tag{86}$$

Then the condition (79) holds if

$$\alpha<u_1'\triangleq\min\Big\{\frac{1}{2L},\frac{1}{16}\frac{\sqrt{\mu_g\varpi\lambda^2}}{C_{g,x\theta}L_{v^*}},\frac{\omega_\theta}{8L_{\theta^*}}\sqrt{\frac{\beta^2}{\left(32\frac{C_{g,x\theta}^2L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,\theta}\right)}},\sqrt{\frac{\gamma}{16L_{fg,x}\tau^2}}\Big\}. \tag{87}$$

Besides, a sufficient condition to make the inequalities (80)-(82) hold is:

$$\begin{aligned}\alpha<u_2'\triangleq\min\Big\{&\frac{\gamma}{80L_{g,\theta}^2\lambda^2},\sqrt{(\frac{32C_{g,x\theta}^2L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x})\frac{\gamma}{20\left(L_{fg,x}L_{g,\theta}^2\beta^2+4C_{g,x\theta}^2L_{fg,\theta}\lambda^2\right)}},\\&\frac{(1-\rho)^2}{6L_{fg,x}\tau^2},\frac{(1-\rho)^4\mu_g\omega_\theta}{32\varphi},\frac{(1-\rho)^4}{16(12\tau^2L_{fg,x}+L_{g,\theta}^2L_{fg,x}\beta^2+4C_{x\theta}^2L_{fg,\theta}\lambda^2)}\Big\},\end{aligned} \tag{88}$$

and

$$\begin{aligned}
&\lambda \leqslant \min\left\{\sqrt{\frac{\gamma}{16L_{g,\theta}^2}}, \sqrt{\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x}\right)\frac{\gamma}{32C_{g,x\theta}^2 L_{fg,\theta}}}\right\}, \\
&\beta \leqslant \sqrt{\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x}\right)\frac{\gamma}{8L_{fg,x}L_{g,\theta}^2}}, \gamma < \frac{1}{4},
\end{aligned} \tag{89}$$

where φ is given by (86). Combining Lemmas 2-5, based on (89) we further have the following condition for the step-sizes $\lambda, \beta, \tau, \gamma$:

$$\begin{aligned}
&\lambda \leqslant \min\left\{\frac{1}{\mu_g}, \sqrt{\frac{\gamma}{16L_{g,\theta}^2}}, \sqrt{\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x}\right)\frac{\gamma}{32C_{g,x\theta}^2 L_{fg,\theta}}}\right\} \\
&\beta \leqslant \min\left\{\frac{2}{\mu_g+L_{g,\theta}}, \frac{\mu_g+L_{g,\theta}}{2\mu_g L_{g,\theta}}, \sqrt{\left(\frac{32C_{g,x\theta}^2 L_{fg,\theta}}{\mu_g\mu_g}+L_{fg,x}\right)\frac{\gamma}{8L_{fg,x}L_{g,\theta}^2}}\right\}, \\
&\tau < 1, \gamma < \frac{1}{4}.
\end{aligned} \tag{90}$$

Hence, the selection condition of the step-size α is given by:

$$\alpha \leqslant u' \triangleq \min\left\{u_1', u_2'\right\}. \tag{91}$$

Then, under the condition (90) and (91), it holds that

$$\mathbb{E}[V^{k+1}] \leqslant \mathbb{E}[V^k] - \frac{d_0}{2}\alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] + \alpha^2\sigma_{r'}^2, \tag{92}$$

where $\sigma_{r'}^2$ is denoted as:

$$\begin{aligned}
\sigma_{r'}^2 \triangleq & d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_z^2 + \frac{4}{1-\rho}d_6\sigma_y^2 + (d_3\frac{1}{m} + d_4 + \frac{4}{1-\rho}d_6)\sigma_u^2 \\
= & d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_{\bar{z}}^2 + \frac{4}{1-\rho}d_6\sigma_{\bar{z}}^2 + (\frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6)(2C_{g,x\theta}^2\sigma_v^2 + L_{fg,x}\sigma_\theta^2) \\
= & (d_1 + 2C_{g,x\theta}^2(\frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6))\sigma_v^2 + (d_2 + L_{fg,x}(\frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6))\sigma_\theta^2 \\
& + (\frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6)\sigma_{\bar{z}}^2 \\
= & 2(d_1 + 2C_{g,x\theta}^2(\frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6))(\sigma_{f,\theta}^2 + M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} \\
& + (d_2 + L_{fg,x}(\frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6))\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2} + (\frac{1}{m}d_3 + d_4 + \frac{4}{1-\rho}d_6)(\sigma_{f,x}^2 + M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}.
\end{aligned} \tag{93}$$

where we use the fact that $\sigma_{\bar{z}}^2 = \sigma_{\bar{y}}^2 = \sigma_z^2$ in the second equality. In what follows, by summing up and telescoping the inequality (93) from $k=0$ to K, it follows that:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant \frac{2\left(V^0 - V^K\right)}{d_0\alpha(K+1)} + \frac{2\alpha\sigma_{r'}^2}{d_0}. \tag{94}$$

This completes the proof. ■

B.6 Proof of Corollary 2

Similar to the proof of Corollary 1, when the step-sizes $\gamma, \lambda, \beta, \tau$ are taken as $\gamma = c_\gamma\alpha, \lambda = c_\lambda\alpha, \beta = c_\beta\alpha, \tau = c_\tau\alpha$ with positive parameters $c_\gamma, c_\lambda, c_\beta, c_\tau$ being independent of K, it follows from (85) that $d_0 = \mu_g\omega_\theta c_\lambda c_\beta$ and $d_3 = \frac{2d_0}{c_\gamma}$ are independent of the step-size α and $d_4 = \frac{1}{1-\rho}d_6 = d_3\alpha$. In addition, we also have that σ_v, σ_θ, $\sigma_{\bar{z}}$ and σ_u are independent of the step-size α. Then the variance related term $\sigma_{r'}$ in (98) can be further derived as:

$$\begin{aligned}
\sigma_{r'} =& \sqrt{d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + d_4\sigma_z^2 + \frac{4}{1-\rho}d_6\sigma_y^2 + (d_3\frac{1}{m} + d_4 + \frac{4}{1-\rho}d_6)\sigma_u^2} \\
=& \sqrt{d_1\sigma_v^2 + d_2\sigma_\theta^2 + (d_3\frac{1}{m} + 2d_3\alpha)\sigma_u^2 + \frac{1}{m}d_3\sigma_{\bar{z}}^2 + 2d_3\alpha\sigma_{\bar{z}}^2} \\
\leqslant& \sqrt{d_1\sigma_v^2 + d_2\sigma_\theta^2 + \frac{3}{m}d_3\sigma_u^2} + \sqrt{\frac{3}{m}d_3\sigma_{\bar{z}}^2} \\
\leqslant& \sqrt{d_1\sigma_v^2 + d_2\sigma_\theta^2 + 3d_3\sigma_u^2} + \sqrt{\frac{3}{m}d_3\sigma_{\bar{z}}^2} \\
\leqslant& \underbrace{\sqrt{d_1}\sigma_v + \sqrt{d_2}\sigma_\theta + 2\sqrt{d_3}\sigma_u}_{\triangleq \sigma_{\mathrm{p}}} + \frac{\sqrt{3}}{\sqrt{m}}\underbrace{\sqrt{d_3}\sigma_{\bar{z}}}_{\triangleq \sigma_{\mathrm{c}}} \triangleq \hat{\sigma}_{r'}
\end{aligned} \tag{95}$$

where the first equality uses the fact that $d_4 = \frac{1}{1-\rho}d_6 = d_3\alpha$ and $\sigma_{\bar{z}} = \sigma_z = \sigma_y$; the first inequality holds due to the triangle inequality and the condition $\alpha \leqslant \frac{1}{m}$ induced by a large number of iterations K. By (67), it is known that $\sigma_{\mathrm{p}} = \mathcal{O}(\sigma_{f,\theta} + \sigma_{g,\theta\theta} + \sigma_{g,\theta})$ and $\sigma_{\mathrm{c}} = \mathcal{O}(\sigma_{f,x} + \sigma_{g,x\theta})$.

In what follows, letting

$$a'_0 \triangleq \frac{2}{d_0}, a'_1 \triangleq \frac{2}{d_0}\hat{\sigma}_{r'}^2. \tag{96}$$

and combining the inequality (95), the inequality (94) can be further derived as:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \leqslant a'_0\frac{1}{\alpha(K+1)} + a'_1\alpha. \tag{97}$$

When we take the step-size α as $\alpha = \min\left\{u', \left(\frac{a'_0}{a'_1(K+1)}\right)^{\frac{1}{2}}\right\}$ and the step-sizes $\gamma, \lambda, \beta, \tau$ as $\gamma = c_\gamma\alpha, \lambda = c_\lambda\alpha, \beta = c_\beta\alpha, \tau = c_\tau\alpha$ with $c_\gamma, c_\lambda, c_\beta, c_\tau$ being independent of K, we have:

$$\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\Phi(\bar{x}^k)\|^2] \leqslant& \frac{a'_0}{u'(K+1)} + 2\left(\frac{a'_1a'_0}{K+1}\right)^{\frac{1}{2}} \\
\leqslant& \frac{1}{d_0}(\frac{2(V^0 - V^K)}{u'(K+1)} + \frac{4\hat{\sigma}_{r'}\sqrt{(V^0 - V^K)}}{\sqrt{K+1}}).
\end{aligned} \tag{98}$$

Then, by combining the fact that $u' = \mathcal{O}((1-\rho)^4)$ in (91) and $\hat{\sigma}_r = \mathcal{O}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})$ in (95), it follows from (98) that:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{V^0}{(1-\rho)^4K} + \frac{\sqrt{V^0}}{K}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right). \tag{99}$$

When we initialize the outer-level variables as $x_i^0 = x_j^0, \forall i, j \in \mathcal{V}$, we can derive that $\|x^0 - 1_m \otimes \bar{x}^0\|^2 =$

0 and V^0 is independent of the term $\frac{1}{1-\rho}$, which further gives that:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] = \mathcal{O}\left(\frac{1}{(1-\rho)^4 K} + \frac{1}{\sqrt{K}}(\sigma_{\mathrm{p}} + \frac{1}{\sqrt{m}}\sigma_{\mathrm{c}})\right). \tag{100}$$

This completes the proof. ■

C Proof of Supporting Propositions

C.1 Proof of Proposition 1

Lipschitz continuity of $\theta_i^*(x)$. Recalling the definition of $\theta_i^*(x)$, the expression of $\theta_i^*(x)$ is given by [24]:

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

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

$$\|\nabla\theta_i^*(x)\| = \left\|\nabla_{x\theta}g_i\left(x,\theta_i^*(x)\right)\left[\nabla_{\theta\theta}g_i\left(x,\theta_i^*(x)\right)\right]^{-1}\right\| \leqslant \frac{C_{g,x\theta}}{\mu_g}. \tag{101}$$

which implies that for any x, x':

$$\left\|\theta_i^*(x) - \theta_i^*(x')\right\| \leqslant \underbrace{\frac{C_{g,x\theta}}{\mu_g}}_{\triangleq L_{\theta^*}}\left\|x - x'\right\|. \tag{102}$$

Lipschitz continuity of $v_i\left(x,\theta\right)$. Recalling the definition of $v_i\left(x,\theta\right)$, we known that $v_i\left(x,\theta\right)$ admits the following expression:

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

Letting (x,θ) and (x',θ') be any two points in $\mathbb{R}^n\times\mathbb{R}^p$, it follows that

$$\begin{aligned}
&\|v_i(x,\theta) - v_i(x',\theta')\|\\
\overset{\text{(a)}}{\leqslant}&\|[\nabla^2_{\theta\theta}g_i(x,\theta)]^{-1}\nabla_\theta f_i(x,\theta) - [\nabla^2_{\theta\theta}g_i(x,\theta)]^{-1}\nabla_\theta f_i(x',\theta')\|\\
&+\|[\nabla^2_{\theta\theta}g_i\left(x,\theta\right)]^{-1}\nabla_\theta f_i(x',\theta') - [\nabla^2_{\theta\theta}g_i(x',\theta')]^{-1}\nabla_\theta f_i(x',\theta')\|\\
\overset{\text{(b)}}{\leqslant}&\frac{1}{\mu_g}\|\nabla_\theta f_i(x,\theta) - \nabla_\theta f_i(x',\theta')\|\\
&+C_{f,\theta}\|[\nabla^2_{\theta\theta}g_i(x',\theta')]^{-1}(\nabla^2_{\theta\theta}g_i(x,\theta) - \nabla^2_{\theta\theta}g_i(x',\theta'))[\nabla^2_{\theta\theta}g_i(x,\theta)]^{-1}\|\\
\overset{\text{(c)}}{\leqslant}&\frac{1}{\mu_g}L_{f,\theta}(\|x-x'\| + \|\theta-\theta'\|) + \frac{C_{f,\theta}L_{g,\theta\theta}}{\mu_g^2}(\|x-x'\| + \|\theta-\theta'\|)\\
=&\underbrace{(\frac{L_{f,\theta}}{\mu_g} + \frac{C_{f,\theta}L_{g,\theta\theta}}{\mu_g^2})}_{\triangleq L_v}(\|x-x'\| + \|\theta-\theta'\|),
\end{aligned} \tag{103}$$

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; step (c) follows from the Lipschitz continuity of $\nabla_\theta f_i$ in Assumption 2 and $\nabla^2_{\theta\theta}g_i$ in Assumption 5.

Lipschitz continuity of $v_i^*\left(x\right)$. Recalling the definition of $v_i^*\left(x\right)$, we know that $v_i^*\left(x\right) =$

$v_i\left(x, \theta_i^*(x)\right)$. Given any two x, x', by taking $\theta=\theta_i^*(x)$ and $\theta'=\theta_i^*(x')$ in (103), it follows that

$$\begin{aligned} \left\|v_i^*(x)-v_i^*(x')\right\| &= \left\|v_i(x, \theta_i^*(x))-v_i(x', \theta_i^*(x'))\right\| \\ &\leqslant L_v\left(\left\|x-x'\right\|+\left\|\theta_i^*(x)-\theta_i^*(x')\right\|\right) \\ &\leqslant L_v\left(1+L_{\theta^*}\right)\left\|x-x'\right\|, \end{aligned} \tag{104}$$

where the last inequality is obtained employing the Lipschitz continuity of $\theta_i^*(x)$ as mentioned earlier.

Lipschitz continuity of $\bar{\nabla} f_i(x, \theta)$. Given any any two points (x, θ) and (x', θ') in $\mathbb{R}^n \times \mathbb{R}^p$, following the definition of $\bar{\nabla} f_i(x, \theta)$ and the triangle inequality yield that

$$\begin{aligned} &\|\bar{\nabla} f_i(x, \theta)-\bar{\nabla} f_i(x', \theta')\| \\ \leqslant &\|\nabla_x f_i(x, \theta)-\nabla_x f_i(x', \theta')\| \\ &+\|\nabla_{x\theta}^2 g_i(x, \theta)(v_i(x, \theta)-v_i(x', \theta'))\| \\ &+\|(\nabla_{x\theta}^2 g_i(x, \theta)-\nabla_{x\theta}^2 g_i(x', \theta'))v_i(x', \theta')\|. \end{aligned} \tag{105}$$

Utilizing Lipschitz continuity of $\nabla_x f_i$, the boundness of $\nabla_{x\theta}^2 g_i$ and Lipschitz continuity of $v_i(x, \theta)$, the first two terms on right hand of (105) can be bounded by:

$$\begin{aligned} &\left\|\nabla_x f_i(x, \theta)-\nabla_x f_i\left(x', \theta'\right)\right\|+\left\|\nabla_{x\theta}^2 g_i(x, \theta)\left(v_i(x, \theta)-v_i\left(x', \theta'\right)\right)\right\| \\ \leqslant &(L_{f,x}+C_{g,x\theta}L_v)\left(\left\|x-x'\right\|+\left\|\theta-\theta'\right\|\right). \end{aligned} \tag{106}$$

Noting that $\|v_i(x', \theta')\| \leqslant \frac{C_{f,\theta}}{\mu_g}$ by the boundness of the Hessian inverse matrix and $\nabla_\theta f_i$ and combining the Lipschitz continuity of $\nabla_{x\theta}^2 g_i$, one can get

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

Then the inequality $\left\|\bar{\nabla} f_i(x, \theta)-\bar{\nabla} f_i(x', \theta')\right\| \leqslant L_f(\|x-x'\|+\|\theta-\theta'\|)$ can be derived by integrating above inequalities, with L_f defined in (23).

Lipschitz continuity of $\nabla\Phi(x)$. It is noted from the expression of $\nabla\Phi_i(x)$ that $\nabla\Phi_i(x)=\bar{\nabla} f_i(x, \theta_i^*(x))$. Given any two x, x', by taking $\theta=\theta_i^*(x)$ and $\theta'=\theta_i^*(x')$ in (105), we can employ the result (105) to derive the following inequality:

$$\begin{aligned} &\left\|\nabla\Phi_i(x)-\nabla\Phi_i(x')\right\| \\ &=\left\|\bar{\nabla} f_i(x, \theta_i^*(x))-\bar{\nabla} f_i\left(x', \theta_i^*(x')\right)\right\| \\ &\leqslant (L_{f,x}+C_{g,x\theta}L_v+\frac{C_{f,\theta}L_{g,x\theta}}{\mu_g})\left(\left\|x-x'\right\|+\left\|\theta_i^*(x)-\theta_i^*(x')\right\|\right) \\ &\leqslant \underbrace{(L_{f,x}+C_{g,x\theta}L_v+\frac{C_{f,\theta}L_{g,x\theta}}{\mu_g})(1+L_{\theta^*})}_{\triangleq L}\left\|x-x'\right\|, \end{aligned}$$

where the last step uses the Lipschitz continuity of $v_i^*(x)$ as shown earlier. 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 2

We will prove Propositions 2 by induction arguments. To this end, we first assume that there exists a constant $M=\frac{C_{f,\theta}}{\mu_g}$ such that $\|v_i^0\| \leqslant M$. Then, we discuss by induction that, at iteration $k+1$ the recursion v_i^{k+1} will be bounded by M when $\|v_i^k\| \leqslant M$. Specifically, at iteration $k+1$, recalling the update rule of v_i^{k+1} it follows that

$$
\begin{aligned}
\|v_i^{k+1}\| =&\|(I-\lambda\nabla^2_{\theta\theta}\hat{g}_i(x_i^k,\theta_i^k;\xi_{i,2}^{k+1}))v_i^k+\lambda\nabla_\theta\hat{f}_i(x_i^k,\theta_i^k;\varsigma_{i,1}^{k+1})\| \\
\overset{(a)}{\leqslant}&\|I-\lambda\nabla^2_{\theta\theta}\hat{g}_i(x_i^k,\theta_i^k;\xi_{i,i}^{k+1})\|\|v_i^k\|+\lambda\|\nabla_\theta\hat{f}_i(x_i^k,\theta_i^k;\varsigma_{i,1}^{k+1})\| \\
\overset{(b)}{\leqslant}&(1-\lambda\mu_g)\|v_i^k\|+\lambda C_{f,\theta} \\
=&\|v_i^k\|-\lambda(\mu_g\|v_i^k\|-C_{f,\theta}) \\
\overset{(c)}{\leqslant}&M,
\end{aligned}
\tag{107}
$$

where the step (a) follows from the triangle inequality and Cauchy-Schwartz inequality; step (b) uses the lower bound of the Hessian matrix with μ_g and the Lipschitz continuity of $\nabla_\theta f_i$; the step (c) holds due to $\mu_g\left\|v_i^k\right\|-C_{f,\theta}\geqslant 0$ imposed by $\|v_i^k\|\leqslant M$. Above inequality demonstrates that, we have $\|v_i^{k+1}\|\leqslant M$ for any $k+1$. This completes the proof. ■

D. Proof of Supporting Lemmas

D.1 Proof of Lemma 1

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

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

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

$$
\begin{aligned}
&\sum_{i=1}^m\left\|\nabla\Phi_i\left(x\right)-\nabla\Phi\left(x\right)\right\|^2 \\
\leqslant&2\sum_{i=1}^m\left\|\nabla\theta_i^*(x)\nabla_\theta f_i(x,\theta_i^*(x))-\frac{1}{m}\sum_{j=1}^m\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))\right\|^2 \\
&+2\sum_{i=1}^m\left\|\nabla_x f_i(x,\theta_i^*(x))-\frac{1}{m}\sum_{j=1}^m\nabla_x f_j(x,\theta_j^*(x))\right\|^2.
\end{aligned}
\tag{108}
$$

For the first term on the right hand of (108), it follows that

$$
\begin{aligned}
&\sum_{i=1}^{m}\big\|\nabla\theta_i^*(x)\nabla_\theta f_i(x,\theta_i^*(x))-\frac{1}{m}\sum_{j=1}^{m}\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))\big\|^2\\
&\overset{(a)}{\leqslant}2\sum_{i=1}^{m}\big\|\nabla\theta_i^*(x)\nabla_\theta f_i(x,\theta_i^*(x))-\frac{1}{m}\sum_{j=1}^{m}\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\big\|^2\\
&\quad+2\sum_{i=1}^{m}\big\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))-\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\big\|^2\\
&\overset{(b)}{\leqslant}2\max_i\left\{\big\|\nabla\theta_i^*(x)\big\|^2\right\}\sum_{i=1}^{m}\big\|\nabla_\theta f_i(x,\theta_i^*(x))-\frac{1}{m}\sum_{j=1}^{m}\nabla_\theta f_j(x,\theta_i^*(x))\big\|^2\\
&\quad+2\sum_{i=1}^{m}\big\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))-\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\big\|^2\\
&\overset{(c)}{\leqslant}2\max_i\left\{\|\nabla\theta_i^*(x)\|^2\right\}b_f^2+2\sum_{i=1}^{m}\big\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))-\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\big\|^2,
\end{aligned}
\tag{109}
$$

where step (a) introduces the term $\frac{1}{m}\sum_{i=1}^{m}\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))$ and uses Young's inequality; step (b) comes from Cauchy-Schwartz inequality; step (c) use the bounded data heterogeneity in Assumption 4.

Now we proceed in providing the upper bound for the last term on the right hand of (109) as follows:

$$
\begin{aligned}
&\sum_{i=1}^{m}\big\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))-\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\big\|^2\\
&\leqslant 2\sum_{i=1}^{m}\big\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_j^*(x))-\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\big\|^2\\
&\quad+2\sum_{i=1}^{m}\big\|\frac{1}{m}\sum_{j=1}^{m}\left[\nabla\theta_j^*(x)\nabla_\theta f_j(x,\theta_i^*(x))-\nabla\theta_i^*(x)\nabla_\theta f_j(x,\theta_i^*(x))\right]\big\|^2\\
&\overset{(a)}{\leqslant}2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\big\|\nabla\theta_j^*(x)\big\|^2\big\|\nabla_\theta f_j(x,\theta_j^*(x))-\nabla_\theta f_j(x,\theta_i^*(x))\big\|^2\\
&\quad+2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\nabla_\theta f_j(x,\theta_i^*(x))\|^2\big\|\nabla\theta_j^*(x)-\nabla\theta_i^*(x)\big\|^2\\
&\overset{(b)}{\leqslant}2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}L_{f,\theta}^2\big\|\nabla\theta_j^*(x)\big\|^2\big\|\theta_j^*(x)-\theta_i^*(x)\big\|^2+2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}C_{f,\theta}^2\big\|\nabla\theta_j^*(x)-\nabla\theta_i^*(x)\big\|^2\\
&\leqslant 2\max_i\left\{L_{f,\theta}^2\|\nabla\theta_i^*(x)\|^2\right\}\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\big\|\theta_j^*(x)-\theta_i^*(x)\big\|^2\\
&\quad+2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}C_{f,\theta}^2\big\|\nabla\theta_j^*(x)-\nabla\theta_i^*(x)\big\|^2,
\end{aligned}
\tag{110}
$$

where step (a) follows from Jensen inequality and Cauchy-Schwartz inequality; step (b) employs Lipschitz continuity and boundness of $\nabla_\theta f_i$.

Now, we deal with the last term of (110). Noting that

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

we arrive at

$$\begin{aligned}
&\left\|\nabla\theta_j^*(x) - \nabla\theta_i^*(x)\right\|^2 \\
=&\left\|\nabla^2_{x\theta} g_j\left(x, \theta_j^*(x)\right)\left[\nabla_{\theta\theta} g_j\left(x, \theta_j^*(x)\right)\right]^{-1} - \nabla^2_{x\theta} g_i\left(x, \theta_i^*(x)\right)\left[\nabla_{\theta\theta} g_i\left(x, \theta_i^*(x)\right)\right]^{-1}\right\|^2 \\
\leqslant& 2\left\|\nabla^2_{x\theta} g_j\left(x, \theta_j^*(x)\right)\left[\nabla_{\theta\theta} g_j\left(x, \theta_j^*(x)\right)\right]^{-1}\left[\nabla^2_{\theta\theta} g_i\left(x, \theta_i^*(x)\right) - \nabla^2_{\theta\theta} g_j\left(x, \theta_j^*(x)\right)\right]\left[\nabla^2_{\theta\theta} g_i(x, \theta_i^*(x))\right]^{-1}\right\|^2 \\
&+ 2\left\|\left[\nabla^2_{x\theta} g_j\left(x, \theta_j^*(x)\right) - \nabla^2_{x\theta} g_i\left(x, \theta_i^*(x)\right)\right]\left[\nabla^2_{\theta\theta} g_i\left(x, \theta_i^*(x)\right)\right]^{-1}\right\|^2 \\
\leqslant& 2\frac{C^2_{g,x\theta}}{\mu_g^4}\left\|\nabla^2_{\theta\theta} g_i\left(x, \theta_i^*(x)\right) - \nabla^2_{\theta\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2 + 2\frac{1}{\mu_g^2}\left\|\nabla^2_{x\theta} g_i\left(x, \theta_i^*(x)\right) - \nabla^2_{x\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2 \\
\leqslant& 2\frac{C^2_{g,x\theta}}{\mu_g^4}\left(2L^2_{g,\theta\theta}\left\|\theta_i^*(x) - \theta_j^*(x)\right\|^2 + 2\left\|\nabla^2_{\theta\theta} g_i\left(x, \theta_j^*(x)\right) - \nabla^2_{\theta\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2\right) \\
&+ 2\frac{1}{\mu_g^2}\left(2L^2_{g,x\theta}\left\|\theta_i^*(x) - \theta_j^*(x)\right\|^2 + 2\left\|\nabla^2_{x\theta} g_i\left(x, \theta_j^*(x)\right) - \nabla^2_{x\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2\right) \\
=&(4\frac{C^2_{g,x\theta}L^2_{g,\theta\theta}}{\mu_g^4} + 4\frac{L^2_{g,x\theta}}{\mu_g^2})\left\|\theta_i^*(x) - \theta_j^*(x)\right\|^2 + 4\frac{C^2_{g,x\theta}}{\mu_g^4}\left\|\nabla^2_{\theta\theta} g_i\left(x, \theta_j^*(x)\right) - \nabla^2_{\theta\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2 \\
&+ 4\frac{1}{\mu_g^2}\left\|\nabla^2_{x\theta} g_i\left(x, \theta_j^*(x)\right) - \nabla^2_{x\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2,
\end{aligned} \tag{111}$$

which implies that the last term of (110) can be bounded by:

$$\begin{aligned}
&2C^2_{f,\theta}\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m\left\|\nabla\theta_j^*(x) - \nabla\theta_i^*(x)\right\|^2 \\
\leqslant&(8\frac{C^2_{f,\theta}C^2_{g,x\theta}L^2_{g,\theta\theta}}{\mu_g^4} + 8\frac{C^2_{f,\theta}L^2_{g,x\theta}}{\mu_g^2})\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m\left\|\theta_i^*(x) - \theta_j^*(x)\right\|^2 \\
&+ 8\frac{C^2_{f,\theta}C^2_{g,x\theta}}{\mu_g^4}\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m\left\|\nabla^2_{\theta\theta} g_i\left(x, \theta_j^*(x)\right) - \nabla^2_{\theta\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2 \\
&+ 8\frac{C^2_{f,\theta}}{\mu_g^2}\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m\left\|\nabla^2_{x\theta} g_i\left(x, \theta_j^*(x)\right) - \nabla^2_{x\theta} g_j\left(x, \theta_j^*(x)\right)\right\|^2 \\
\leqslant&(8\frac{C^2_{f,\theta}C^2_{g,x\theta}L^2_{g,\theta\theta}}{\mu_g^4} + 8\frac{C^2_{f,\theta}L^2_{g,x\theta}}{\mu_g^2})\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m\left\|\theta_i^*(x) - \theta_j^*(x)\right\|^2 + 8\frac{C^2_{f,\theta}C^2_{g,x\theta}}{\mu_g^4}b_g^2 + 8\frac{C^2_{f,\theta}}{\mu_g^2}b_g^2.
\end{aligned} \tag{112}$$

As for the second term on the right hand of (108), we bound it by:

$$
\begin{aligned}
&\sum_{i=1}^{m}\|\nabla_x f_i(x,\theta_i^*(x))-\frac{1}{m}\sum_{j=1}^{m}\nabla_x f_j(x,\theta_j^*(x))\|^2\\
\leqslant&2\sum_{i=1}^{m}\|\nabla_x f_i(x,\theta_i^*(x))-\frac{1}{m}\sum_{j=1}^{m}\nabla_x f_j(x,\theta_i^*(x))\|^2\\
&+2\sum_{i=1}^{m}\|\frac{1}{m}\sum_{j=1}^{m}\left(\nabla_x f_j(x,\theta_j^*(x))-\nabla_x f_j(x,\theta_i^*(x))\right)\|^2\\
\overset{(a)}{\leqslant}&2b_f^2+2\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|\nabla_x f_j(x,\theta_j^*(x))-\nabla_x f_j(x,\theta_i^*(x))\right\|^2\\
\overset{(b)}{\leqslant}&2b_f^2+2L_{f,x}^2\frac{1}{m}\sum_{i=1}^{m}\sum_{i=j}^{m}\left\|\theta_j^*(x)-\theta_i^*(x)\right\|^2,
\end{aligned}
\tag{113}
$$

where step (a) uses the bounded data heterogeneity in Assumption 4, while step (b) comes from Lipschitz continuity of $\nabla_x f_i$. In addition, utilizing the strong convexity of g_i in θ, we further derive that:

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|\theta_j^*(x)-\theta_i^*(x)\right\|^2\\
\leqslant&\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{\mu_g^2}\left\|\nabla_\theta g_i(x,\theta_j^*(x))-\nabla_\theta g_i(x,\theta_i^*(x))\right\|^2\\
\overset{(a)}{=}&\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{\mu_g^2}\left\|\nabla_\theta g_i(x,\theta_j^*(x))-\nabla_\theta g_j(x,\theta_j^*(x))\right\|^2\\
\leqslant&\frac{1}{\mu_g^2}b_g^2,
\end{aligned}
\tag{114}
$$

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

Substituting the results (109), (110), (111), (112), (113), (114) into (108) and rearranging the terms, we reach the following inequality:

$$
\begin{aligned}
&\sum_{i=1}^{m}\|\nabla\Phi_i(x)-\nabla\Phi(x)\|^2\\
\leqslant&4b_f^2+4\max_i\left\{\|\nabla\theta_i^*(x)\|^2\right\}b_f^2+16\frac{C_{f,\theta}^2C_{g,x\theta}^2}{\mu_g^4}b_g^2+16\frac{C_{f,\theta}^2}{\mu_g^2}b_g^2\\
&+4\left(L_{f,x}^2+\max_i\left\{L_{f,\theta}^2\|\nabla\theta_i^*(x)\|^2\right\}+4\frac{C_{f,\theta}^2C_{g,x\theta}^2L_{g,\theta\theta}^2}{\mu_g^4}+4\frac{C_{f,\theta}^2L_{g,x\theta}^2}{\mu_g^2}\right)\frac{1}{\mu_g^2}b_g^2\\
\leqslant&\underbrace{(4+4\frac{C_{g,x\theta}^2}{\mu_g^2})}_{\triangleq C_1(\mu_g,C_{g,x\theta})}b_f^2+\underbrace{4\left(4\frac{C_{f,\theta}^2C_{g,x\theta}^2}{\mu_g^4}+4\frac{C_{f,\theta}^2}{\mu_g^2}+\frac{L_{f,x}^2}{\mu_g^2}+\frac{L_{f,\theta}^2C_{g,x\theta}^2}{\mu_g^4}+4\frac{C_{f,\theta}^2C_{g,x\theta}^2L_{g,\theta\theta}^2}{\mu_g^6}+4\frac{C_{f,\theta}^2L_{g,x\theta}^2}{\mu_g^4}\right)}_{\triangleq C_2(\mu_g,L_{f,x},L_{f,\theta},L_{g,x\theta},L_{g,\theta\theta},C_{f,\theta},C_{g,x\theta})}b_g^2,
\end{aligned}
\tag{115}
$$

where the last step uses the following bound of $\nabla\theta_i^*(x)$:

$$\|\nabla\theta_i^*(x)\|^2 = \left\|\nabla_{x\theta} g_i\left(x,\theta_i^*(x)\right)\left[\nabla_{\theta\theta} g_i\left(x,\theta_i^*(x)\right)\right]^{-1}\right\|^2 \leqslant \frac{C_{g,x\theta}^2}{\mu_g^2}. \tag{116}$$

This completes the proof. ■

D.2 Proof of Lemma 2

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:

$$\bar{y}^{k+1} = \bar{y}^k + \bar{z}^{k+1} - \bar{z}^k. \tag{117}$$

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 (21). Thus, in both the local gradient and tracking gradient schemes, it can be deduced that $\bar{y}^k = \bar{z}^k$. Furthermore, by multiplying the matrix J on both sides of (20a), we can determine the average state of the update (20a) across all nodes as follows:

$$\bar{x}^{k+1} = \bar{x}^k - \alpha\bar{y}^k. \tag{118}$$

Since the overall objective function Φ is smooth by Proposition 1, it holds that:

$$\begin{aligned}
\mathbb{E}[\Phi(\bar{x}^{k+1})|\mathcal{F}^k] \leqslant& \Phi(\bar{x}^k) + \mathbb{E}[\langle\nabla\Phi(\bar{x}^k), \bar{x}^{k+1} - \bar{x}^k\rangle|\mathcal{F}^k] + \frac{\alpha^2 L}{2}\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k] \\
\overset{\text{(a)}}{=}& \Phi(\bar{x}^k) - \alpha\mathbb{E}[\langle\nabla\Phi(\bar{x}^k), \bar{y}^k\rangle|\mathcal{F}^k] + \frac{\alpha^2 L}{2}\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k] \\
\overset{\text{(b)}}{=}& \Phi(\bar{x}^k) + \frac{\alpha}{2}\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{y}^k\|^2|\mathcal{F}^k] - \frac{\alpha}{2}\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k] \\
&- \frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 + \frac{\alpha^2 L}{2}\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k] \\
\overset{\text{(c)}}{\leqslant}& \Phi(\bar{x}^k) + \alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2|\mathcal{F}^k] + \alpha\mathbb{E}[\|\bar{s}^k - \bar{y}^k\|^2|\mathcal{F}^k] \\
&- \frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 - \frac{\alpha}{2}\left(1 - \alpha L\right)\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k] \\
\overset{\text{(d)}}{\leqslant}& \Phi(\bar{x}^k) + \alpha\mathbb{E}[\|\nabla\Phi(\bar{x}^k) - \bar{s}^k\|^2|\mathcal{F}^k] + \alpha\mathbb{E}[\|\bar{s}^k - \bar{z}^k\|^2|\mathcal{F}^k] \\
&- \frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 - \frac{\alpha}{2}(1 - \alpha L)\mathbb{E}[\|\bar{y}^k\|^2|\mathcal{F}^k],
\end{aligned} \tag{119}$$

where step (a) uses the recursion (118); step (b) holds due to the equation $2a^{\mathrm{T}}b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$; step (c) uses Young's inequality; step (d) is obtained by utilizing 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 3

Recall that the update of z^{k+1} in (20g) is given by:

$$z^{k+1} = s^{k+1}(\zeta^{k+1}) + (1-\gamma)(z^k - s^k(\zeta^{k+1})).$$

Considering the above equality, we can obtain a recursive expression for the term $\bar{s}^{k+1} - \bar{z}^{k+1}$ by introducing the term $\bar{s}^k$ as follows:

$$\begin{aligned}
\bar{s}^{k+1} - \bar{z}^{k+1} =& \bar{s}^{k+1} - (\bar{s}^{k+1}(\zeta^{k+1}) + (1-\gamma)(\bar{z}^k - \bar{s}^k(\zeta^{k+1}))) \\
=&(1-\gamma)(\bar{s}^k - \bar{z}^k) + \bar{s}^{k+1} - \bar{s}^{k+1}(\zeta^{k+1}) - (1-\gamma)(\bar{s}^k - \bar{s}^k(\zeta^{k+1})) \\
=&(1-\gamma)(\bar{s}^k - \bar{z}^k) + \gamma(\bar{s}^{k+1} - \bar{s}^{k+1}(\zeta^{k+1})) \\
&+ (1-\gamma)(\bar{s}^{k+1} - \bar{s}^k - (\bar{s}^{k+1}(\zeta^{k+1}) - \bar{s}^k(\zeta^{k+1}))).
\end{aligned} \tag{120}$$

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

$$\begin{aligned}
&\mathbb{E}[\|\bar{s}^{k+1} - \bar{z}^{k+1}\|^2|\mathcal{F}^k] \\
\overset{(a)}{=}&(1-\gamma)^2\|\bar{s}^k - \bar{z}^k\|^2 + (1-\gamma)^2\mathbb{E}[\|\bar{s}^{k+1} - \bar{s}^{k+1}(\zeta^{k+1}) - \bar{s}^k + \bar{s}^k(\zeta^{k+1})\|^2|\mathcal{F}^k] \\
&+ \gamma^2\mathbb{E}[\|\bar{s}^{k+1} - \bar{s}^{k+1}(\zeta^{k+1})\|^2|\mathcal{F}^k] \\
\overset{(b)}{=}&(1-\gamma)^2\|\bar{s}^k - \bar{z}^k\|^2 + (1-\gamma)^2\frac{1}{m^2}\mathbb{E}[\|s^{k+1} - s^{k+1}(\zeta^{k+1}) - s^k + s^k(\zeta^{k+1})\|^2|\mathcal{F}^k] \\
&+ \gamma^2\mathbb{E}\left[\|\bar{s}^{k+1} - \bar{s}^{k+1}(\zeta^{k+1})\|^2|\mathcal{F}^k\right] \\
\overset{(c)}{\leqslant}&(1-\gamma)\|\bar{s}^k - \bar{z}^k\|^2 + \frac{1}{m^2}\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2|\mathcal{F}^k] \\
&+ \gamma^2\mathbb{E}[\|\bar{s}^{k+1} - \bar{s}^{k+1}(\zeta^{k+1})\|^2|\mathcal{F}^k],
\end{aligned} \tag{121}$$

where step (a) holds due to that $\bar{s}^{k+1}(\zeta^{k+1})$ is independent of $\bar{z}^k$ and $\bar{s}^{k+1}$ is the unbiased estimate of $\bar{s}^{k+1}(\zeta^{k+1})$; step (b) uses the fact that the samples ζ_i^{k+1}, $i = 1, \cdots, m$ are independent of each other; step (c) is derived by the condition $0 < \gamma < 1$. Next, we will bound the last term in (121). Note that from the recursion (20f) that:

$$\begin{aligned}
&\bar{s}^{k+1} - \bar{s}^{k+1}(\zeta^{k+1}) \\
=&J\nabla_x F(x^{k+1}, \theta^{k+1}) - J\nabla^2_{x\theta} G(x^{k+1}, \theta^{k+1})v^{k+1} \\
&- J\nabla_x \hat{F}(x^{k+1}, \theta^{k+1}; \varsigma_2^{k+1}) + J\nabla^2_{x\theta}\hat{G}(x^{k+1}, \theta^{k+1}; \xi_3^{k+1})v^{k+1}.
\end{aligned} \tag{122}$$

Taking the square norm on both sides of (122) under the total expectation, it follows that

$$\begin{aligned}
&\mathbb{E}[\|\bar{s}^{k+1} - \bar{s}^k(\zeta^{k+1})\|^2] \\
\overset{(a)}{=}&\mathbb{E}[\|J\nabla_x F(x^{k+1}, \theta^{k+1}) - J\nabla_x \hat{F}(x^{k+1}, \theta^{k+1}; \varsigma_2^{k+1})\|^2] \\
&+ \mathbb{E}[\|J\nabla^2_{x\theta} G(x^{k+1}, \theta^{k+1})v^{k+1} - J\nabla^2_{x\theta}\hat{G}(x^{k+1}, \theta^{k+1}; \xi_3^{k+1})v^{k+1}\|^2] \\
\overset{(b)}{\leqslant}&\frac{1}{m}(\sigma^2_{f,x} + M^2\sigma^2_{g,x\theta}),
\end{aligned} \tag{123}$$

where step (a) uses the independence between the samples ς_2^{k+1} and ξ_3^{k+1} and unbiased estimates of stochastic gradients in Assumption 5; step (b) follows from the bounded variances in Assumption 5 and the fact that $\|v_i^{k+1}\| \leq M$ in Proposition 2. Taking total expectation and combining the upper bound (123), we reach

$$\begin{aligned}\mathbb{E}[\|\bar{s}^{k+1}-\bar{z}^{k+1}\|^2] \leqslant & (1-\gamma)\mathbb{E}[\|\bar{s}^{k}-\bar{z}^{k}\|^2] \\ & + \frac{1}{m}\mathbb{E}[\|s^{k+1}(\zeta^{k+1})-s^{k}(\zeta^{k+1})\|^2] + \frac{1}{m}\underbrace{(\sigma_{f,x}^2+M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}}_{\triangleq\sigma_{\bar{z}}^2}\alpha^2.\end{aligned} \tag{124}$$

This derives the result (27).

By the definition of $\nabla\Phi(\bar{x}^k)$, we 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 (20d) and the fact $s^k=\mathbb{E}[s^k(\zeta^k)|\mathcal{F}^k]$ that:

$$s^k = \nabla_x F(x^k,\theta^k) - \nabla_{x\theta}^2 G(x^k,\theta^k)v^k.$$

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

$$\begin{aligned}\nabla\Phi(\bar{x}^k)-\bar{s}^k = & J\nabla_x F(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k)) - J\nabla_x F(x^k,\theta^k) \\ & + J\nabla_{x\theta}^2 G(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k))(v^k-v^*(\bar{x}^k)) \\ & + J(\nabla_{x\theta}^2 G(x^k,\theta^k)-\nabla_{x\theta}^2 G(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k)))v^k.\end{aligned} \tag{125}$$

Then taking the square norm on both sides of (125) under the total expectation and employing the inequality $\|a+b\|^2\leqslant 2\|a\|^2+2\|b\|^2$ twice and Jensen inequality, we get:

$$\begin{aligned}& \mathbb{E}[\|\nabla\Phi(\bar{x}^k)-\bar{s}^k\|^2] \\ \leqslant & 2\frac{1}{m}\mathbb{E}[\|\nabla_x F(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k))-\nabla_x F(x^k,\theta^k)\|^2] \\ & +4\frac{1}{m}\mathbb{E}[\|-\nabla_{x\theta}^2 G(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k))v^*(\bar{x}^k)+\nabla_{x\theta}^2 G(\bar{x}^k,\theta^*(\bar{x}^k))v^k\|^2] \\ & +4\frac{1}{m}\mathbb{E}[\|-\nabla_{x\theta}^2 G(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k))v^k+\nabla_{x\theta}^2 G(x^k,\theta^k)v^k\|^2] \\ \leqslant & \underbrace{(2L_{f,x}^2+4M^2L_{g,x\theta}^2)}_{\triangleq L_{fg,x}}\frac{1}{m}\mathbb{E}[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2]+4C_{g,x\theta}^2\frac{1}{m}\mathbb{E}[\|v^k-v^*(\bar{x}^k)\|^2].\end{aligned} \tag{126}$$

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 $\|v_i^k\|$ in Proposition 2. This completes the proof. ■

D.4 Proof of Lemma 4

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

$$\begin{aligned}\mathbb{E}[\|v^{k+1}-v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k] = & \underbrace{\mathbb{E}[\|v^{k+1}-v^*(\bar{x}^{k})\|^2|\mathcal{F}^k]}_{\triangleq A_1^v} + \mathbb{E}[\|v^*(\bar{x}^{k})-v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k] \\ & + \underbrace{\mathbb{E}[2\langle v^{k+1}-v^*(\bar{x}^k), v^*(\bar{x}^k)-v^*(\bar{x}^{k+1})\rangle|\mathcal{F}^k]}_{\triangleq A_2^v}.\end{aligned} \tag{127}$$

We first 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 (20b) by combining (20e) 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:

$$
\begin{aligned}
A_1^v =& \mathbb{E}[\|v^{k+1} - v^*(\bar{x}^k)\|^2 | \mathcal{F}^k] \\
=& \mathbb{E}[\|(I - \lambda \nabla^2_{\theta\theta} \hat{G}(x^k, \theta^k; \xi_2^k)) v^k + \lambda \nabla_\theta \hat{F}(x^k, \theta^k; \varsigma_1^k) - v^*(\bar{x}^k)\|^2 | \mathcal{F}^k] \\
=& \|(I - \lambda \nabla^2_{\theta\theta} G(x^k, \theta^k)) v^k + \lambda \nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k)\|^2 \\
&+ \lambda^2 \mathbb{E}\|\nabla^2_{\theta\theta} \hat{G}(x^k, \theta^k; \xi_2^k) v^k - \nabla^2_{\theta\theta} G(x^k, \theta^k) v^k + \nabla_\theta F(x^k, \theta^k) - \nabla_\theta \hat{F}(x^k, \theta^k; \varsigma_1^k)\|^2 | \mathcal{F}^k] \\
\leqslant& \|(I - \lambda \nabla^2_{\theta\theta} G(x^k, \theta^k)) v^k + \lambda \nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k)\|^2 + m(\sigma_{f,\theta}^2 + M^2 \sigma_{g,\theta\theta}^2) \lambda^2,
\end{aligned}
\tag{128}
$$

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

$$
v^*(\bar{x}^k) = [\nabla^2_{\theta\theta} G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))]^{-1} \nabla_\theta F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)),
\tag{129}
$$

then we rewrite that:

$$
\begin{aligned}
&(I - \lambda \nabla^2_{\theta\theta} G(x^k, \theta^k)) v^k + \lambda \nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k) \\
&\quad = (I - \lambda \nabla^2_{\theta\theta} G(x^k, \theta^k)) v^k + \lambda \nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k) \\
&\qquad + \lambda(\nabla^2_{\theta\theta} G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) v^*(\bar{x}^k) - \nabla_\theta F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))) \\
&\quad = \lambda(\nabla_\theta F(x^k, \theta^k) - \nabla_\theta F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))) \\
&\qquad + (I - \lambda \nabla^2_{\theta\theta} G(x^k, \theta^k))(v^k - v^*(\bar{x}^k)) \\
&\qquad + \lambda(\nabla^2_{\theta\theta} G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) - \nabla^2_{\theta\theta} G(x^k, \theta^k)) v^*(\bar{x}^k).
\end{aligned}
\tag{130}
$$

With the above expression, the first term on the right hand of (128) can be bounded by:

$$
\begin{aligned}
&\|(I - \lambda \nabla^2_{\theta\theta} G(x^k, \theta^k)) v^k + \lambda \nabla_\theta F(x^k, \theta^k) - v^*(\bar{x}^k)\|^2 \\
\overset{\text{(a)}}{\leqslant}& (1 + \frac{\mu_g \lambda}{3})^2 \|(I - \lambda \nabla^2_{\theta\theta} G(x^k, \theta^k))(v^k - v^*(\bar{x}^k))\|^2 \\
&+ (1 + \frac{3}{\mu_g \lambda}) \lambda^2 \|\nabla_\theta F(x^k, \theta^k) - \nabla_\theta F(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))\|^2 \\
&+ (1 + \frac{\mu_g \lambda}{3})(1 + \frac{3}{\mu_g \lambda}) \lambda^2 \|(\nabla^2_{\theta\theta} G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)) - \nabla^2_{\theta\theta} G(x^k, \theta^k)) v^*(\bar{x}^k)\|^2 \\
\overset{\text{(b)}}{\leqslant}& (1 + \frac{\mu_g \lambda}{3})^2 (1 - \mu_g \lambda)^2 \|v^k - v^*(\bar{x}^k)\|^2 \\
&+ (1 + \frac{3}{\mu_g \lambda}) L_{f,\theta}^2 \lambda^2 [\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2] \\
&+ (1 + \frac{3}{\mu_g \lambda})(1 + \frac{\mu_g \lambda}{3}) L_{g,\theta\theta}^2 M^2 \lambda^2 [\|x^k - 1_m \otimes \bar{x}^k\|^2 + \|\theta^k - \theta^*(\bar{x}^k)\|^2],
\end{aligned}
\tag{131}
$$

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 $\|v_i^k\|$.

By considering the condition $1 - \lambda \mu_g > 0$ and rearranging the terms, we can further derive from the inequality (131) the following results:

$$
\begin{aligned}
&\|(I-\lambda\nabla^2_{\theta\theta}G(x^k,\theta^k))v^k+\lambda\nabla_\theta F(x^k,\theta^k)-v^*(\bar{x}^k)\|^2\\
\leqslant&(1-\frac{\mu_g\lambda}{3})(1-\mu_g\lambda)\|v^k-v^*(\bar{x}^k)\|^2\\
&+\frac{2\lambda}{\mu_g}\underbrace{(2L^2_{f,\theta}+4M^2L^2_{g,\theta\theta})}_{\triangleq L_{fg,\theta}}[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2].
\end{aligned}
\tag{132}
$$

Substituting (132) into (128), we reach an upper bound for the term A_1^v in (127) as follows:

$$
\begin{aligned}
A_1^v\leqslant&(1-\frac{\mu_g\lambda}{3})\,(1-\mu_g\lambda)\,\|v^k-v^*(\bar{x}^k)\|^2+m(1+M^2)\lambda^2\sigma^2\\
&+\frac{2\lambda}{\mu_g}L_{fg,\theta}[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2].
\end{aligned}
\tag{133}
$$

Now we analyze the term $\mathbb{E}[\|v^*(\bar{x}^k)-v^*(\bar{x}^{k+1})\|^2]$ in (127). Employing Lipschitz continuity of $v_i^*(x)$ in Proposition 1, the recursion (20a) and the upper bound of the variances in (27), we have that:

$$
\mathbb{E}[\|v^*(\bar{x}^k)-v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]\leqslant mL^2_{v^*}\mathbb{E}[\|\bar{x}^k-\bar{x}^{k+1}\|^2|\mathcal{F}^k]\leqslant m\alpha^2L^2_{v^*}\|\bar{y}^k\|^2. \tag{134}
$$

It remains to analyze the term A_2^v in (127), which can be rewritten as the following expression by leveraging Cauchy-Schwartz inequality:

$$
A_2^v\leqslant\frac{\mu_g\lambda}{3}A_1^v+\frac{3}{\mu_g\lambda}\mathbb{E}[\|v^*(\bar{x}^k)-v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]. \tag{135}
$$

In combination with (127), (133), (134), (135), we reach an evolution for $\mathbb{E}[\|v^{k+1}-v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]$ as follows:

$$
\begin{aligned}
&\mathbb{E}[\|v^{k+1}-v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]\\
\leqslant&(1+\frac{\mu_g\lambda}{3})A_1^v+(1+\frac{3}{\mu_g\lambda})\mathbb{E}[\|v^*(\bar{x}^k)-v^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]\\
\overset{(a)}{\leqslant}&(1+\frac{\mu_g\lambda}{3})(1-\frac{\mu_g\lambda}{3})(1-\mu_g\lambda)\|v^k-v^*(\bar{x}^k)\|^2\\
&+(1+\frac{\mu_g\lambda}{3})\frac{2\lambda}{\mu_g}L_{fg,\theta}[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\
&+\frac{2L^2_{v^*}\alpha^2}{\varpi\lambda}m\|\bar{y}^k\|^2+(1+\frac{\mu_g\lambda}{3})m(1+M^2)\lambda^2\sigma^2\\
\overset{(b)}{\leqslant}&(1-\mu_g\lambda)\|v^k-v^*(\bar{x}^k)\|^2+\underbrace{\frac{4L_{fg,\theta}\lambda}{\mu_g\alpha}}_{\triangleq q_x}\alpha[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\
&+\underbrace{\frac{2L^2_{v^*}}{\varpi\lambda}}_{\triangleq q_s}m\alpha^2\|\bar{y}^k\|^2+m\underbrace{2(\sigma^2_{f,\theta}+M^2\sigma^2_{g,\theta\theta})\frac{\lambda^2}{\alpha^2}}_{\triangleq\sigma_v^2}\alpha^2,
\end{aligned}
\tag{136}
$$

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 5

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 Hv errors. We start by decomposing the inner-level errors as follows:

$$\begin{aligned}\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k] = &\underbrace{\mathbb{E}[\|\theta^{k+1} - \theta^*(\bar{x}^k)\|^2|\mathcal{F}^k]}_{\triangleq A_1^\theta} + \mathbb{E}[\left\|\theta^*(\bar{x}^k) - \theta^*(\bar{x}^{k+1})\right\|^2|\mathcal{F}^k] \\ &+ \underbrace{2\mathbb{E}[\langle\theta^{k+1} - \theta^*(\bar{x}^k), \theta^*(\bar{x}^k) - \theta^*(\bar{x}^{k+1})\rangle|\mathcal{F}^k]}_{\triangleq A_2^\theta}.\end{aligned} \tag{137}$$

As for the term A_1^θ, we have that:

$$\begin{aligned}\theta^{k+1} - \theta^*(\bar{x}^k) =&\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k) \\ &+ \beta(\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta \hat{G}(x^k, \theta^k; \xi_1^k)).\end{aligned} \tag{138}$$

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

$$\begin{aligned}A_1^\theta =&\mathbb{E}[\|\theta^k - \beta\nabla_\theta \hat{G}(x^k, \theta^k; \xi_1^k) - \theta^*(\bar{x}^k)\|^2|\mathcal{F}^k] \\ =&\|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2 + \beta^2\mathbb{E}[\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta \hat{G}(x^k, \theta^k; \xi_1^k)|\mathcal{F}^k] \\ &+ 2\beta\langle\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k), \mathbb{E}[\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta \hat{G}(x^k, \theta^k; \xi_1^k)|\mathcal{F}^k]\rangle \\ \overset{(a)}{\leqslant}&(1 + \beta\omega_\theta)\|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2 + (\beta + \frac{1}{\omega_\theta})\beta L_{g,\theta}^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + m\beta^2\sigma_{g,\theta}^2 \\ \overset{(b)}{\leqslant}& (1 + \beta\omega_\theta)\|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2 + \frac{2}{\omega_\theta}\beta L_{g,\theta}^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + m\beta^2\sigma_{g,\theta}^2,\end{aligned} \tag{139}$$

where step (a) uses the variance decomposition for the term $\mathbb{E}[\|\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta \hat{G}(x^k, \theta^k; \xi_1^k)\|^2|\mathcal{F}^k]$ 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 $\nabla_\theta g_i$ and the condition that $\beta < \frac{1}{\omega_\theta}$ in (31). Next, we proceed in providing an upper bound for the first term on the right hand of (139) as follows:

$$\begin{aligned}&\|\theta^k - \beta\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \theta^*(\bar{x}^k)\|^2 \\ \overset{(a)}{=}&\|\theta^k - \theta^*(\bar{x}^k)\|^2 + \beta^2\|\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))\|^2 \\ &- 2\beta\langle\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k)), \theta^k - \theta^*(\bar{x}^k)\rangle \\ \overset{(b)}{\leqslant}&(1 - 2\beta\frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}})\|\theta^k - \theta^*(\bar{x}^k)\|^2 \\ &+ (\beta - \frac{2}{\mu_g + L_{g,\theta}})\beta\|\nabla_\theta G(1_m \otimes \bar{x}^k, \theta^k) - \nabla_\theta G(1_m \otimes \bar{x}^k, \theta^*(\bar{x}^k))\|^2 \\ \overset{(c)}{\leqslant}&(1 - 2\beta\frac{\mu_g L_{g,\theta}}{\mu_g + L_{g,\theta}})\|\theta^k - \theta^*(\bar{x}^k)\|^2,\end{aligned} \tag{140}$$

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; step (c) holds due to the step-size condition $\beta = c_\beta\alpha < \frac{2}{\mu_g + L_{g,\theta}}$. Then, plugging (140) into (139) yields

$$\begin{aligned}A_1^\theta \leqslant&(1+\beta\omega_\theta)(1-2\beta\frac{\mu_g L_{g,\theta}}{\mu_g+L_{g,\theta}})\|\theta^k-\theta^*(\bar{x}^k)\|^2+(\beta+\frac{1}{\omega_\theta})\beta L_{g,\theta}^2\|x^k-1_m\otimes\bar{x}^k\|^2+m\beta^2\sigma_{g,\theta}^2\\ \leqslant&(1-\frac{3}{2}\beta\frac{\mu_g L_{g,\theta}}{\mu_g+L_{g,\theta}})\|\theta^k-\theta^*(\bar{x}^k)\|^2+\frac{2}{\omega_\theta}\beta L_{g,\theta}^2\|x^k-1_m\otimes\bar{x}^k\|^2+m\beta^2\sigma_{g,\theta}^2.\end{aligned} \tag{141}$$

In addition, note that the inner-product term in (137) follows that

$$\mathbb{E}[\|\theta^*(\bar{x}^k)-\theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]\leqslant mL_{\theta^*}^2\alpha^2\|\bar{y}^k\|^2. \tag{142}$$

Next, we will control the term A_2^θ in (137) by Cauchy-Schwartz inequality as follows:

$$A_2^\theta\leqslant\omega_\theta\beta A_1^\theta+\frac{1}{\omega_\theta\beta}\mathbb{E}[\|\theta^*(\bar{x}^k)-\theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]. \tag{143}$$

Then leveraging results (141), (142), (143), we can control the inner-level errors, i.e., $\mathbb{E}[\|\theta^{k+1}-\theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]$ as follows:

$$\begin{aligned}\mathbb{E}[\|\theta^{k+1}-\theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k]\leqslant&(1+\omega_\theta\beta)A_1^\theta+(1+\frac{1}{\omega_\theta\beta})\mathbb{E}\left[\|\theta^*(\bar{x}^k)-\theta^*(\bar{x}^{k+1})\|^2|\mathcal{F}^k\right]\\ \leqslant&(1+\omega_\theta\beta)(1-\frac{3}{2}\beta\frac{\mu_g L_{g,\theta}}{\mu_g+L_{g,\theta}})\|\theta^k-\theta^*(\bar{x}^k)\|^2\\ &+(1+\omega_\theta\beta)\frac{2}{\omega_\theta}\beta L_{g,\theta}^2\|x^k-1_m\otimes\bar{x}^k\|^2+(1+\omega_\theta\beta)\beta^2\sigma^2\\ &+(1+\frac{1}{\omega_\theta\beta})L_{\theta^*}^2\alpha^2\|\bar{y}^k\|^2\\ \leqslant&(1-\frac{\mu_g L_{g,\theta}}{\mu_g+L_{g,\theta}}\beta)\|\theta^k-\theta^*(\bar{x}^k)\|^2+m\underbrace{\frac{2L_{\theta^*}^2}{\omega_\theta\beta}}_{\triangleq q_s}\alpha^2\|\bar{y}^k\|^2\\ &+\underbrace{\frac{4L_{g,\theta}^2\beta}{\omega_\theta\alpha}}_{\triangleq q_x}\alpha\|x^k-1_m\otimes\bar{x}^k\|^2+m\underbrace{2\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2}}_{\triangleq\sigma_\theta^2}\alpha^2,\end{aligned} \tag{144}$$

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

D.6 Proof of Lemma 6

First, 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

$$\begin{aligned}&x^{k+1}-1_m\otimes\bar{x}^{k+1}\\ =&(1-\tau)x^k+\tau(\mathcal{W}x^k-\alpha y^k)-1_m\otimes(\bar{x}^k-\tau\alpha\bar{y}^k)\\ =&(1-\tau)(x^k-1_m\otimes\bar{x}^k)+\tau(\mathcal{W}x^k-\alpha y^k-(1_m\otimes\bar{x}^k-\alpha\bar{y}^k))\\ =&(1-\tau)\left(x^k-1_m\otimes\bar{x}^k\right)+\tau((\mathcal{W}-\mathcal{J})(x^k-1_m\otimes\bar{x}^k)-\alpha(y^k-1_m\otimes\bar{y}^k))\end{aligned} \tag{145}$$

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

$$
\begin{aligned}
&\mathbb{E}[\|x^{k+1}-1_m\otimes\bar{x}^{k+1}\|^2|\mathcal{F}^k]\\
\leqslant&(1-\tau)^2(1+\frac{\tau}{1-\tau})\|x^k-1_m\otimes\bar{x}^k\|^2+\tau^2(1+\frac{1-\tau}{\tau})\|Wx^k-\alpha y^k-(1_m\otimes\bar{x}^k-\alpha\bar{y}^k)\|^2\\
\leqslant&(1-\tau)\|x^k-1_m\otimes\bar{x}^k\|^2+\tau\|(\mathcal{W}-\mathcal{J})\left(x^k-1_m\otimes\bar{x}^k\right)-\alpha(y^k-1_m\otimes\bar{y}^k)\|^2\\
\leqslant&(1-\tau)\|x^k-1_m\otimes\bar{x}^k\|^2+\tau(1+\eta)\|\mathcal{W}-\mathcal{J}\|^2\|x^k-1_m\otimes\bar{x}^k\|^2+\tau(1+\frac{1}{\eta})\|y^k-1_m\otimes\bar{y}^k\|^2\\
\leqslant&(1-\tau\frac{1-\rho}{2})\|x^k-1_m\otimes\bar{x}^k\|^2+\frac{2\tau\alpha^2}{1-\rho}\|y^k-1_m\otimes\bar{y}^k\|^2
\end{aligned}
\tag{146}
$$

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 7

In what follows, we proceed in analyzing the term $\mathbb{E}[\|s^{k+1}(\zeta^{k+1})-s^k(\zeta^{k+1})\|^2|\mathcal{F}^k]$. Note from the recursion (20f) that the successive difference of $s^{k+1}(\zeta^{k+1})$ can be expressed as:

$$
\begin{aligned}
s^{k+1}(\zeta^{k+1})-s^k(\zeta^{k+1})=&\nabla_x\hat{F}(x^{k+1},\theta^{k+1};\varsigma_2^{k+1})-\nabla_x\hat{F}(x^k,\theta^k;\varsigma_2^{k+1})\\
&-\nabla^2_{x\theta}\hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^{k+1}+\nabla^2_{x\theta}\hat{G}(x^k,\theta^k;\xi_3^{k+1})v^k\\
=&\nabla_x\hat{F}(x^{k+1},\theta^{k+1};\varsigma_2^{k+1})-\nabla_x\hat{F}(x^k,\theta^k;\varsigma_2^{k+1})\\
&-\nabla^2_{x\theta}\hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^{k+1}+\nabla^2_{x\theta}\hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^k\\
&-\nabla^2_{x\theta}\hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^k+\nabla^2_{x\theta}\hat{G}(x^k,\theta^k;\xi_3^{k+1})v^k.
\end{aligned}
\tag{147}
$$

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

$$
\begin{aligned}
&\mathbb{E}[\|s^{k+1}(\zeta^{k+1})-s^k(\zeta^{k+1})\|^2|\mathcal{F}^k]\\
\leqslant&2\mathbb{E}[\|\nabla_x\hat{F}(x^{k+1},\theta^{k+1};\varsigma_2^{k+1})-\nabla_x\hat{F}(x^k,\theta^k;\varsigma_2^{k+1})\|^2|\mathcal{F}^k]\\
&+4\mathbb{E}[\|-\nabla^2_{x\theta}\hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^{k+1}+\nabla^2_{x\theta}\hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^k\|^2|\mathcal{F}^k]\\
&+4\mathbb{E}[\|-\nabla^2_{x\theta}\hat{G}(x^{k+1},\theta^{k+1};\xi_3^{k+1})v^k+\nabla^2_{x\theta}\hat{G}(x^k,\theta^k;\xi_3^{k+1})v^k\|^2|\mathcal{F}^k]\\
\leqslant&\underbrace{(2L_{f,x}^2+4M^2L_{g,x\theta}^2)}_{L_{fg,x}}\mathbb{E}\left[\|x^{k+1}-x^k\|^2+\|\theta^{k+1}-\theta^k\|^2|\mathcal{F}^k\right]+4C_{g,x\theta}^2\mathbb{E}[\|v^{k+1}-v^k\|^2|\mathcal{F}^k].
\end{aligned}
\tag{148}
$$

Next, we proceed in bounding the last three terms in (148). Firstly, for the term $\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}^k]$, it can be bounded by:

$$
\begin{aligned}
&\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}^k]\\
=&\|\tau(I-\mathcal{W})x^k+\tau\alpha y^k\|^2\\
=&\|\tau(I-\mathcal{W})(x^k-1_m\otimes\bar{x}^k)+\tau\alpha y^k-\tau\alpha 1_m\otimes\bar{y}^k+\tau\alpha 1_m\otimes\bar{y}^k\|^2\\
\leqslant&3\tau^2\|I-\mathcal{W}\|^2\|x^k-1_m\otimes\bar{x}^k\|^2+3\tau^2\alpha^2\|y^k-1_m\otimes\bar{y}^k\|^2+3\tau^2\alpha^2\|1_m\otimes\bar{y}^k\|^2\\
\leqslant&12\tau^2\|x^k-1_m\otimes\bar{x}^k\|^2+3\tau^2\alpha^2\|y^k-1_m\otimes\bar{y}^k\|^2+3m\tau^2\alpha^2\|\bar{y}^k\|^2
\end{aligned}
\tag{149}
$$

where the last step uses the fact that $\|\mathcal{W}-I\|^2\leqslant 4$.

For the term $\mathbb{E}[\|\theta^{k+1}-\theta^k\|^2|\mathcal{F}^k]$, combining the recursion (20a) it can be rewritten as:

$$\begin{aligned}\mathbb{E}[\|\theta^{k+1}-\theta^k\|^2|\mathcal{F}^k] =& \beta^2\mathbb{E}[\|\nabla_\theta \hat{G}(x^k,\theta^k;\xi_1^k)\|^2|\mathcal{F}^k]\\ \leqslant& \beta^2\|\nabla_\theta G(x^k,\theta^k)-\nabla_\theta G(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k))\|^2+m\beta^2\sigma_{g,\theta}^2\\ \leqslant& L_{g,\theta}^2\beta^2[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2]+m\beta^2\sigma_{g,\theta}^2.\end{aligned} \tag{150}$$

As for the term $\mathbb{E}[\|v^{k+1}-v^k\|^2|\mathcal{F}^k]$, we note that

$$\begin{aligned}v^{k+1}-v^k =& (I-\lambda\nabla_{\theta\theta}^2\hat{G}(x^k,\theta^k;\xi_2^k))v^k+\lambda\nabla_\theta\hat{F}(x^k,\theta^k;\varsigma_1^k)-v^k\\ =& -\lambda\nabla_{\theta\theta}^2\hat{G}(x^k,\theta^k;\xi_2^k)v^k+\lambda\nabla_\theta\hat{F}(x^k,\theta^k;\varsigma_1^k).\end{aligned} \tag{151}$$

Then, it follows that

$$\begin{aligned}&\mathbb{E}[\|v^{k+1}-v^k\|^2|\mathcal{F}^k]\\ \overset{(a)}{\leqslant}& \lambda^2\|\nabla_{\theta\theta}^2 G(x^k,\theta^k)v^k-\nabla_\theta F(x^k,\theta^k)\|^2+m(\sigma_{f,\theta}^2+M^2\sigma_{g,\theta\theta}^2)\lambda^2\\ \overset{(b)}{\leqslant}& 2\lambda^2\|\nabla_\theta F(x^k,\theta^k)-\nabla_\theta F(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k))\|^2+4\lambda^2\|\nabla_{\theta\theta}^2G(x^k,\theta^k)(v^*(\bar{x}^k)-v^k)\|^2\\ &+4\lambda^2\|(\nabla_{\theta\theta}^2G(1_m\otimes\bar{x}^k,\theta^*(\bar{x}^k))-\nabla_{\theta\theta}^2G(x^k,\theta^k))v^*(\bar{x}^k)\|^2+m(1+M^2)\lambda^2\sigma^2\\ \overset{(c)}{\leqslant}& 2L_{f,\theta}^2\lambda^2[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2]+4L_{g,\theta}^2\lambda^2\|v^k-v^*(\bar{x}^k)\|^2\\ &+4M^2L_{g,\theta\theta}^2\lambda^2[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2]+m(\sigma_{f,\theta}^2+M^2\sigma_{g,\theta\theta}^2)\lambda^2\\ =& \underbrace{(2L_{f,\theta}^2+4M^2L_{g,\theta\theta}^2)}_{L_{fg,\theta}}\lambda^2[\|x^k-1_m\otimes\bar{x}^k\|^2+\|\theta^k-\theta^*(\bar{x}^k)\|^2]\\ &+4L_{g,\theta}^2\lambda^2\|v^k-v^*(\bar{x}^k)\|^2+m(\sigma_{f,\theta}^2+M^2\sigma_{g,\theta\theta}^2)\lambda^2,\end{aligned} \tag{152}$$

where step (a) uses the bounded variances in Assumption 5 and the fact that $\|v_i^{k+1}\|\le M$ in Proposition 2; step (b) is derived by introducing the expression of $v^*(\bar{x}^k)$ in (129); step (c) follows from Lipschitz continuity of the involved functions. Then, combining (148), (149), (150), (152), the term $\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2|\mathcal{F}^k]$ can be bounded by:

$$
\begin{aligned}
&\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2|\mathcal{F}^k] \\
\leqslant&(12L_{fg,x}\tau^2 + L_{g,\theta}^2 L_{fg,x}\beta^2 + 4C_{g,x\theta}^2 L_{fg,\theta}\lambda^2)\|x^k - 1_m \otimes \bar{x}^k\|^2 \\
&+ (L_{fg,x}L_{g,\theta}^2\beta^2 + 4C_{x\theta}^2 L_{fg,\theta}\lambda^2)\|\theta^k - \theta^*(\bar{x}^k)\|^2 + 16C_{g,x\theta}^2 L_{g,\theta}^2\lambda^2\|v^k - v^*(\bar{x}^k)\|^2 \\
&+ (4C_{g,x\theta}^2(1+M^2)\lambda^2 + L_{fg,x}\beta^2)m\sigma^2 + 3L_{fg,x}\tau^2\alpha^2\|y^k - 1_m \otimes \bar{y}^k\|^2 + 3mL_{fg,x}\tau^2\alpha^2\|\bar{y}^k\|^2 \\
=&\underbrace{(12L_{fg,x}\frac{\tau^2}{\alpha^2} + L_{g,\theta}^2 L_{fg,x}\frac{\beta^2}{\alpha^2} + 4C_{g,x\theta}^2 L_{fg,\theta}\frac{\lambda^2}{\alpha^2})}_{\triangleq u_x}\alpha^2\|x^k - 1_m \otimes \bar{x}^k\|^2 + \underbrace{(16C_{g,x\theta}^2 L_{g,\theta}^2\frac{\lambda^2}{\alpha^2})}_{\triangleq u_v}\alpha^2\|v^k - v^*(\bar{x}^k)\|^2 \\
&+ \underbrace{(L_{fg,x}L_{g,\theta}^2\frac{\beta^2}{\alpha^2} + 4C_{g,x\theta}^2 L_{fg,\theta}\frac{\lambda^2}{\alpha^2})}_{\triangleq u_\theta}\alpha^2\|\theta^k - \theta^*(\bar{x}^k)\|^2 + \underbrace{3L_{fg,x}}_{\triangleq u_y}\tau^2\alpha^2\|y^k - 1_m \otimes \bar{y}^k\|^2 \\
&+ m\underbrace{(4C_{g,x\theta}^2(\sigma_{f,\theta}^2 + M^2\sigma_{g,\theta\theta}^2)\frac{\lambda^2}{\alpha^2} + L_{fg,x}\sigma_{g,\theta}^2\frac{\beta^2}{\alpha^2})}_{\triangleq \sigma_u^2}\alpha^2 + m\underbrace{3L_{fg,x}}_{\triangleq u_s}\tau^2\alpha^2\|\bar{y}^k\|^2,
\end{aligned} \tag{153}
$$

This completes the proof. ■

D.8 Proof of Lemma 8

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

$$
\begin{aligned}
\|y^k - 1_m \otimes \bar{y}^k\|^2 \leqslant&\|y^k\|^2 \\
=&\sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k) + \nabla\Phi_i(\bar{x}^k) - \nabla\Phi(\bar{x}^k) + \nabla\Phi(\bar{x}^k)\|^2 \\
\leqslant&3\sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k)\|^2 + 3\sum_{i=1}^m \|\nabla\Phi_i(\bar{x}^k) - \nabla\Phi(\bar{x}^k)\|^2 + 3m\|\nabla\Phi(\bar{x}^k)\|^2 \\
\leqslant&3\sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k)\|^2 + 3b^2 + 3m\|\nabla\Phi(\bar{x}^k)\|^2,
\end{aligned} \tag{154}
$$

where the last step follows from Lemma 1.

Note that following the inequality (28) and combining the update of y_i^k under local gradient scheme (21), the term $\sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k)\|^2$ can bounded by:

$$
\begin{aligned}
&\sum_{i=1}^m \|y_i^k - \nabla\Phi_i(\bar{x}^k)\|^2 \\
\leqslant&2\sum_{i=1}^m \|s_i^k - \nabla\Phi_i\left(\bar{x}^k\right)\|^2 + 2\sum_{i=1}^m \|z_i^k - s_i^k\|^2 \\
\leqslant&2L_{fg,x}\|x^k - 1_m \otimes \bar{x}^k\|^2 + 2L_{fg,x}\|\theta^k - \theta^*(\bar{x}^k)\|^2 + 8C_{g,x\theta}^2\left\|v^k - v^*(\bar{x}^k)\right\|^2 + 2\|s^k - z^k\|^2.
\end{aligned} \tag{155}
$$

Substituting the above inequality into (154), we reach the desired result. This completes the proof. ■

D.9 Proof of Lemma 9

According to the recursion (22), we known 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

$$\begin{aligned} y^{k+1} - 1_m \otimes \bar{y}^{k+1} =& \mathcal{W}y^k + z^{k+1} - z^k - 1_m \otimes (\bar{y}^k + \bar{z}^{k+1} - \bar{z}^k) \\ =& (\mathcal{W} - \mathcal{J})(y^k - 1_m \otimes \bar{y}^k) + (I - \mathcal{J})(z^{k+1} - z^k). \end{aligned} \tag{156}$$

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 get that:

$$\begin{aligned} \mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2 | \mathcal{F}^k] \leqslant& (1+\eta)\|\mathcal{W} - \mathcal{J}\|^2 \|y^k - 1_m \otimes \bar{y}^k\|^2 \\ &+ (1 + \frac{1}{\eta})\|I - \mathcal{J}\|^2 \mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathcal{F}^k] \\ \leqslant& \frac{1+\rho}{2}\|y^k - 1_m \otimes \bar{y}^k\|^2 + \frac{2}{1-\rho}\mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathcal{F}^k], \end{aligned} \tag{157}$$

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 (157), it follows from (20g) that:

$$\begin{aligned} z^{k+1} - z^k =& s^{k+1}(\zeta^{k+1}) + (1-\gamma)(z^k - s^k(\zeta^{k+1})) - z^k \\ =& \gamma(s^k - z^k) + \gamma(s^k(\zeta^{k+1}) - s^k) + s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1}), \end{aligned} \tag{158}$$

which further implies that:

$$\begin{aligned} \mathbb{E}[\|z^{k+1} - z^k\|^2 | \mathcal{F}^k] \leqslant& 2\gamma^2\|z^k - s^k\|^2 \\ &+ 2\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2 \mathcal{F}^k] + 2m(\sigma_{f,x}^2 + M^2\sigma_{g,x\theta}^2)\gamma^2\sigma^2. \end{aligned} \tag{159}$$

Substituting the above inequality into (157), we reach

$$\begin{aligned} &\mathbb{E}[\|y^{k+1} - 1_m \otimes \bar{y}^{k+1}\|^2 | \mathcal{F}^k] \\ \leqslant& \frac{1+\rho}{2}\|y^k - 1_m \otimes \bar{y}^k\|^2 + \frac{4}{1-\rho}\mathbb{E}[\|s^{k+1}(\zeta^{k+1}) - s^k(\zeta^{k+1})\|^2 \mathcal{F}^k] \\ &+ \frac{4}{1-\rho}\gamma^2\|s^k - z^k\|^2 + \frac{4}{1-\rho}m\underbrace{(\sigma_{f,x}^2 + M^2\sigma_{g,x\theta}^2)\frac{\gamma^2}{\alpha^2}}_{\triangleq \sigma_y^2}\alpha^2, \end{aligned} \tag{160}$$

This completes the proof. ■