Prometheus: Taming Sample and Communication Complexities in Constrained Decentralized Stochastic Bilevel Learning

Zhuqing Liu [1] **Xin Zhang** [2] **Prashant Khanduri** [3] **Songtao Lu** [4] **Jia Liu** [1]

Abstract

In recent years, decentralized bilevel optimization has gained significant attention thanks to its versatility in modeling a wide range of multi-agent learning problems, such as multi-agent reinforcement learning and multi-agent meta-learning. However, one unexplored and fundamental problem in this area is how to solve decentralized stochastic bilevel optimization problems with *domain constraints*, while achieving low sample and communication complexities. This problem often arises from multi-agent learning problems with safety constraints. As shown in this paper, constrained decentralized bilevel optimization is far more challenging than its unconstrained counterpart due to the complex coupling structure, which necessitates new algorithm design and analysis techniques. Toward this end, we investigate a class of constrained decentralized bilevel optimization problems, where multiple agents collectively solve a nonconvex-strongly-convex bilevel problem with constraints in the upper-level variables. We propose an algorithm called Prometheus (proximal tracked stochastic recursive estimator) that achieves the first $\mathcal{O}(\epsilon^{-1})$ results in both sample and communication complexities for constrained decentralized bilevel optimization, where $\epsilon > 0$ is a desired stationarity error. Collectively, the results in this work contribute to a theoretical foundation for low sample- and communication-complexity constrained decentralized bilevel learning.

[1]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA [2]Department of Statistics, Iowa State University, Ames, IA, USA [3]Department of Computer Science, Wayne State University, Detroit, MI, USA [4]IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Jia Liu <liu@ece.osu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

In recent years, decentralized bilevel optimization has gained signifiant attention thanks to its versatility in modeling a wide range of multi-agent learning problems, such as multi-agent reinforcement learning (MARL) and multi-agent meta learning. However, one unexplored and fundamental problem in this area is how to solve decentralized stochastic bilevel optimization with *domain constraints*, while achieving low sample and communication complexities. This problem often arises from, but is not limited to, safety-constrained(Mansoor et al., 2023; 2021) MARL for autonomous driving (Bennajeh et al., 2019), sparsity-regularized multi-agent meta-learning (Poon & Peyré, 2021), and rank-constrained matrix completion for recommender systems (Pochmann & Von Zuben, 2022), etc. As shown later in this paper, constrained decentralized bilevel optimization is far more challenging than its unconstrained counterpart due to the non-smoothness and complex coupling between domain constraints and the bilevel problem structure. Also, as its name suggests, a defining feature of constrained decentralized bilevel optimization is "decentralized," which implies that all agents must rely on communications to reach a *consensus* on an optimal solution without any coordination from a server. Due to the potentially unreliable network connections and the limited computation capability at each agent, such consensus-based approaches call for low sample and communication complexities. To our knowledge, none of the existing works in the literature has considered solving domain-constrained decentralized bilevel optimization with low sample and communication complexities (e.g., (Gao et al., 2022; Yang et al., 2022; Lu et al., 2022a; Chen et al., 2022b;c; Huang et al., 2023) see Section 2 for detailed discussions). This motivates to fill in important gap in the literature in light of the growing importance of constrained decentralized bilevel optimization.

Specifically, we focus on a class of constrained decentralized multi-task bilevel optimization problems, where we aim to solve a decentralized *nonconvex-strongly-convex* bilevel optimization problem with i) multiple lower-level problems and ii) consensus and domain constrains on the upper level. Such problems naturally arise in security-constrained bilevel model for integrated natural gas and electricity system (Li

et al., 2017), multi-agent actor-critic reinforcement learning (Zhang et al., 2020) and constraint meta-learning (Liu et al., 2019). In the optimization literature, a natural approach for handling domain constraints is the *proximal operator*. However, proximal algorithm design and theoretical analysis for constrained decentralized bilevel optimization problems is far from a trivial extension of their unconstrained counterparts. In fact, in the literature, the proximal operator for constrained bilevel optimization has been under-explored even in the single-agent setting, not to mention the more complex multi-agent settings. The most related works in terms of handling domain constraints can be found in (Hong et al., 2020; Chen et al., 2022a; Ghadimi & Wang, 2018), which rely on direct projected (stochastic) gradient descent to solve the constrained *single-agent* bilevel problem. In contrast, our work considers general domain constraints that require evaluation of proximal operators in each iteration for *mutli-agent* settings. Actually, until this work, it remains unclear how to design proximal algorithms to handle domain constraints for decentralized bilevel optimization.

The main contribution of this paper is that we propose a series of new proximal-based algorithmic techniques to overcome the aforementioned challenges and achieve low sample and communication complexities for domain-constrained decentralized bilevel optimization problems. The main results of this work are summarized below:

- We propose a decentralized optimization approach called Prometheus (proximal tracked stochastic recursive estimator), which is a cleverly designed hybrid algorithm that integrates proximal operations, recursive variance reduction, lower-level gradient tracking, and upper-level consensus techniques. We show that, to achieve an ϵ-stationary point, Prometheus enjoys a convergence rate of $\mathcal{O}(1/T)$, where T is the maximum number of iterations. This implies $\mathcal{O}(\epsilon^{-1})$ communication complexity and $\mathcal{O}(\sqrt{n}K\epsilon^{-1}+n)$ sample complexity per agent.
- We reveal a new and interesting insight that the recursive variance reduction technique in Prometheus is not only sufficient but also necessary for achieving $\mathcal{O}(1/T)$ convergence rate in the sense that: a "non-variance-reduced" special version of Prometheus could only achieve a much slower $\mathcal{O}(1/\sqrt{T})$ convergence to a constant error-ball rather than an ϵ-stationary point with arbitrarily small ϵ-tolerance. This insight advances our understanding and state of the art of algorithm design for constrained decentralized bilevel optimization.
- To further lower sample complexity, we propose a new hyper-gradient estimator for the upper-level function inspired by (Agarwal et al., 2016) in the single-level optimization literature. This new estimator leads to a more accurate stochastic estimation than the conventional stochastic estimator used in (Khanduri et al., 2021; Ghadimi & Wang, 2018; Hong et al., 2020; Liu et al., 2022). We show that our new hyper-gradient stochastic estimator outperforms existing estimators both theoretically (cf. Lemma 1) and experimentally (cf. Fig. 3, Fig. 8).

Table 1. Comparisons among algorithms for bilevel optimization problems. Sample complexities (both upper and lower) as defined in the sense of achieving an ϵ-stationary point defined in (2), n is the size of dataset at each agent. Algorithms shown in shaded are decentralized learning algorithms.

Algorithms	Constriants	Sample Complex.	Commun. Complex.
SUSTAIN (Khanduri et al., 2021)	✗	$\tilde{\mathcal{O}}(\epsilon^{-1.5})$	-
RSVRB (Guo & Yang, 2021)	✗	$\mathcal{O}(\epsilon^{-1.5})$	-
VRBO (Yang et al., 2021)	✗	$\mathcal{O}(\epsilon^{-1.5})$	-
AID-BiO /ITD-BiO (Ji et al., 2021)	✗	$\mathcal{O}(n\epsilon^{-1})$	-
TTSA (Hong et al., 2020)	✓	$\mathcal{O}(\epsilon^{-5/2})$	-
STABLE (Chen et al., 2022a)	✓	$\mathcal{O}(\epsilon^{-2})$	-
DSBO (Yang et al., 2022)	✗	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
SPDB (Lu et al., 2022a)	✗	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
DSBO (Chen et al., 2022b)	✗	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$
VRDBO (Gao et al., 2022)	✗	$\mathcal{O}(\epsilon^{-1.5})$	$\mathcal{O}(\epsilon^{-1.5})$
INTERACT (Liu et al., 2022)	✗	$\mathcal{O}(n\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-1})$
INTERACT-VR (Liu et al., 2022)	✗	$\mathcal{O}(\sqrt{n}K\epsilon^{-1}+n)$	$\mathcal{O}(\epsilon^{-1})$
Prometheus [Ours.]	✓	$\mathcal{O}(\sqrt{n}K\epsilon^{-1}+n)$	$\mathcal{O}(\epsilon^{-1})$

2. Related Work

In this section, we first provide a quick overview of the state-of-the-art on single-agent constrained bilevel optimization as well as decentralized bilevel optimization.

1) Constrained Bilevel Optimization in the Single-Agent Setting: As mentioned in Section 1, various techniques have been proposed to solve single-agent bilevel optimization, such as utilizing full-gradient-based techniques (e.g., AID-based methods (Rajeswaran et al., 2019; Franceschi et al., 2018; Ji et al., 2021), ITD-based methods (Pedregosa, 2016; Maclaurin et al., 2015; Ji et al., 2021)), stochastic gradient-based techniques (Ghadimi & Wang, 2018; Khanduri et al., 2021; Guo & Yang, 2021), STORM-based techniques (Cutkosky & Orabona, 2019), and VR-based techniques (Yang et al., 2021). However, none of these existing works have considered domain constraints. To our knowledge, the only works that considered domain constraints in the single-agent setting can be found in (Hong et al., 2020; Chen et al., 2022a; Ghadimi & Wang, 2018). In (Ghadimi & Wang, 2018), the authors proposed a double-loop algorithm called BSA, where in the inner loop the lower level problem is solved to sufficient accuracy, while in the outer

loop projected (stochastic) gradient descent is utilized to update the model parameters. The double-loop structure of BSA led to slow convergence. To achieve the ϵ-stationary point, the BSA Algorithm requires $\mathcal{O}(\epsilon^{-3})$ samples of the inner function and $\mathcal{O}(\epsilon^{-2})$ samples of the outer function, respectively. In (Hong et al., 2020), a two-timescale single loop stochastic approximation (TTSA) algorithm based on projected (stochastic) gradient descent was proposed to solve the constrained bilevel optimization problems. However, TTSA has to choose step-sizes of different orders for the upper and lower level problems to ensure convergence, which leads to suboptimal complexity results. Later in (Chen et al., 2022a), an algorithm called STABLE algorithm is proposed to utilize a momentum-based gradient estimator and combines the Moreau-envelop-based analysis to achieve an $\mathcal{O}(\epsilon^{-2})$ sample-complexity. As mentioned in Section 1, however, the methods in (Ghadimi & Wang, 2018; Hong et al., 2020; Chen et al., 2022a) considered only simple constraints. Moreover, the aforementioned methods are not applicable in the decentralized setting.

2) Decentralized Bilevel Optimization: Decentralized bilevel optimization has also received increasing attention in recent years. For example, Yang et al. (2022), (Lu et al., 2022a), (Lu et al., 2022b) and Chen et al. (2022b) respectively proposed stochastic gradient (SG)-type decentralized algorithms for bilevel optimization and achieve an $\mathcal{O}(\epsilon^{-2})$ sample-communication complexity. The VRDBO method in (Gao et al., 2022) employed momentum-based techniques to achieve better $\mathcal{O}(\epsilon^{-1.5})$ complexity results. However, VRDBO updates upper- and lower-level variables in an alternating fashion. As will be shown later, our Prometheus algorithm updates upper-level and lower-level variables *simultaneously*, which renders a much lower implementation complexity than VRDBO. Besides, Prometheus achieves $\mathcal{O}(\sqrt{n}K\epsilon^{-1}+n)$ sample complexities, which is a near-optimal and outperforms existing decentralized bilevel algorithms. It is worth noting that, the in aforementioned works, consensus is required at both lower- and upper-levels. Such a formulation can be viewed as multiple agents collaboratively solving the same bilevel optimization problem. In contrast, consensus is required in the upper-level subproblem in our work, which allows multiple *different* lower-level tasks. This is a more practically-relevant formulation for many MARL and multi-agent meta-learning applications.

The most related work on decentralized bilevel optimization is (Liu et al., 2022), which also considered multiple lower-level tasks. However, our work differs from (Liu et al., 2022) in the following two key aspects: (i) The INTERACT-VR method in (Liu et al., 2022) is *unconstrained* and *cannot* handle *non-smooth* objectives considered in our work. As a result, using a straightforward proximal extension of the INTERACT-VR would not work. As shown in (Hong et al., 2022), the direct proximal extension of the algorithm may diverge in solving the decentralized minimization problem. Since conventional minimization can be viewed as a special case of bilevel optimization, following a similar line of analysis, we can conclude that the INTERACT-VR may diverge if we use the direct proximal extension method. To tackle this challenge, we propose a special proximal operator $\tilde{x}_i(x_{i,t})$. We show that this special-structured proximal operator not only makes our Prometheus algorithm numerically efficient but also renders the convergence analysis of Prometheus theoretically tractable. (ii) Our Prometheus algorithm integrates a new stochastic gradient estimator. We show that this new hyper-gradient stochastic estimator is superior to existing estimator, as demonstrated both theoretically through Lemma 1 and experimentally through Figures 3 and 8. For clearer comparisons, we summarize and compare the complexity results of all algorithms mentioned above in Table 1.

3. Problem Formulation and Applications

1) Problem Formulation: Consider an undirected connected network $\mathcal{G}=(\mathcal{N},\mathcal{L})$ that represents a peer-to-peer network, where $\mathcal{N}$ and $\mathcal{L}$ are the sets of agents (nodes) and edges, respectively, with $|\mathcal{N}|=m$. Each agent i has local computation capability and can share information with its neighboring agents denoted as $\mathcal{N}_i \triangleq \{i' \in \mathcal{N} : (i,i') \in \mathcal{L}\}$. Each agent i has access to a local dataset of size n. All agents in the network collaboratively solve the following constrained decentralized bilevel optimization problem:

$$\begin{aligned}
&\min_{\mathbf{x}_i\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^{m}[\ell(\mathbf{x}_i)+h(\mathbf{x}_i)] \\
&\qquad \triangleq \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[f\left(\mathbf{x}_i,\mathbf{y}_i^*(\mathbf{x}_i);\bar{\xi}_{ij}\right)+h(\mathbf{x}_i)] \\
&\text{s.t. } \mathbf{y}_i^*(\mathbf{x}_i)=\arg\min_{\mathbf{y}_i\in\mathbb{R}^{p_2}} g(\mathbf{x}_i,\mathbf{y}_i) \triangleq \frac{1}{n}\sum_{j=1}^{n} g(\mathbf{x}_i,\mathbf{y}_i;\zeta_{ij}), \forall i; \\
&\qquad \mathbf{x}_i=\mathbf{x}_{i'}, \quad \text{if} \quad (i,i')\in\mathcal{L}, \qquad (1)
\end{aligned}$$

where $\mathcal{X}\subseteq\mathbb{R}^{p_1}$ is a convex constraint set, and $\mathbf{x}_i\in\mathcal{X}$ and $\mathbf{y}_i\in\mathbb{R}^{p_2}$ are parameters to be trained for the upper-level and lower-level subproblems at agent i, respectively. Here, $\ell(\mathbf{x}_i) \triangleq f\left(\mathbf{x}_i,\mathbf{y}_i^*(\mathbf{x}_i)\right) = \frac{1}{n}\sum_{j=1}^{n} f\left(\mathbf{x}_i,\mathbf{y}_i^*(\mathbf{x}_i);\bar{\xi}_{ij}\right)$ is the local objective function, and $h(\mathbf{x}_i)$ is a convex proximal function (possibly non-differentiable) for regularization. The equality constraints $\mathbf{x}_i=\mathbf{x}_{i'}$ ensure that the local copies at connected agents i and i' are equal to each other, hence the name "consensus form." As shown in Eq. (1), the upper-level subproblem is to optimize the objective function $\frac{1}{m}\sum_{i=1}^{m}[\ell(\mathbf{x}_i)+h(\mathbf{x}_i)]$, where $\mathbf{x}_i$ is the decision variable. The lower-level subproblem is to obtain the optimal $\mathbf{y}_i$-solutions by minimizing the objective function $g(\mathbf{x}_i,\mathbf{y}_i)$ given a set $\mathbf{x}_i$-values. In both upper and lower levels, m

is the total number of agents and n is the size of dataset at each agent.

In the context of our decentralized bilevel optimization problem in Eq. (1), the sample and communication complexities can be formally defined as follows:

Definition 1 (Sample Complexity)**.** : The sample complexity is defined as the total number of incremental first-order oracle (IFO) calls required for all agents to converge to an ϵ-stationary point. Each IFO call evaluates a pair of $(\bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \nabla g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}))$ at agent i.

Definition 2 (Communication Complexity)**.** : The communication complexity is defined as the total number of communication rounds needed to converge to an ϵ-stationary point. In each round, every node can send and receive vector-valued information to and from its neighboring nodes.

Next, we define the notion of ϵ-stationarity point for Problem (1) for convergence performance characterization. We say that $\{\mathbf{x}_i, \mathbf{y}_i, \forall i \in [m]\}$ is an ϵ-stationarity point if

$$\underbrace{\mathbb{E}\|\tilde{\mathbf{x}} - \mathbf{1} \otimes \bar{\mathbf{x}}\|^2}_{\text{Saddle point error}} + \underbrace{\mathbb{E}\|\mathbf{x} - \mathbf{1} \otimes \bar{\mathbf{x}}\|^2}_{\text{Consensus error}} + \underbrace{\mathbb{E}\|\mathbf{y} - \mathbf{y}^*\|^2}_{\text{lower problem error}} \leq \epsilon, \quad (2)$$

where $\bar{\mathbf{x}} \triangleq \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i$, $\mathbf{y} \triangleq [\mathbf{y}_1^\top, ...\mathbf{y}_m^\top]^\top$, and $\mathbf{y}^* \triangleq [\mathbf{y}_1^{*\top}, ...\mathbf{y}_m^{*\top}]^\top$, and $\tilde{\mathbf{x}}$ is a proximal point that will be defined later in Section 4. The first term in (2) quantifies the convergence of the $\bar{\mathbf{x}}$ to a proximal point of stationarity of the global objective. The second term in (2) measures the consensus error among local copies of the upper variable, while the last term in (2) quantifies the (aggregated) error in the lower problem's iterates across all agents. Thus, $\epsilon \to 0$ implies that the algorithm achieves three goals simultaneously: i) consensus of upper variables, ii) stationary point of Problem (1), and iii) solution to the lower problem. As mentioned in Section 1, two of the most important performance metrics in decentralized optimization are the sample and communication complexities.

2) Motivating Applications: Problem (1) arises naturally from many real-world applications. Here, we present two motivating applications to showcase its practical relevance:

- *Sparsity-Regularized Multi-agent Meta-Learning (Tian et al., 2020):* Sparsity-regularized optimization is widely seen in the machine learning community, which is one of the promising tools for high-dimensional machine learning with guaranteed statistical efficiency and robustness to overfitting. Meta-learning can naturally be formulated as a bilevel optimization problem because it involves optimizing two levels of learning simultaneously: i) the training of a down-stream model for a specific task based on a base model, and ii) the training to improve the performance of the base model. As a result, the optimization process of meta-learning is a nested optimization problem, where the lower-level problem corresponds to optimizing the down-stream task-specific models while the upper-level problem corresponding to training the base model. In a decentralized multi-agent setting, the sparsity-regularized multi-agent meta-learning can be written as:

$$\min_{\mathbf{x}\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^m f\left(\mathbf{x}_i, \mathbf{y}_i^*(\mathbf{x})\right) + h(\mathbf{x}_i);$$
$$\text{s.t. } \mathbf{y}_i^*(\mathbf{x}_i) \in \arg\min_{\mathbf{y}_i \in \mathbb{R}^{p_2}} g\left(\mathbf{x}_i, \mathbf{y}_i\right), i = 1, \ldots, m. \quad (3)$$

  Here, agent i has a local dataset with n samples, $\mathbf{x} \in \mathcal{X}$ denotes the base model parameters shared by all agents (hence consensus is needed), and $\mathbf{y}_i$ are task-specific model parameters computed by each agent i.

- *Decentralized Rank-Constrained Matrix Completion for Recommender Systems (Panagoda, 2021)*: Rank-Constrained Matrix Completion (RCMC) is a technique commonly used in recommender systems to predict missing values in a sparse user-item matrix. The goal of RCMC is to find a low-rank matrix that best approximates the observed data. In a decentralized setting, the rank-constrained matrix completion can be rewrite as:

$$\min_{\mathbf{x}\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^m f(\mathbf{x}_i, \mathbf{y}_i) + h(\mathbf{x}_i);$$
$$\text{s.t. } \min_{\mathbf{y}_i} g(\mathbf{x}_i, \mathbf{y}_i), i = 1, ..., m, \quad (4)$$

  Here, $f(\cdot)$ is the objective function that measures the quality of the recovered matrix, and $g(\cdot)$ is specified by the user (e.g., selecting the optimal values of the hyper-parameters that govern the behavior of the upper level problem), and $h(\cdot)$ is the sparsity regularization which are often used in RCMC to prevent the low-rank matrix approximation from overfitting.

4. Solution Approach

In this section, we first present the Prometheus algorithm for solving the constrained decentralized bilevel optimization problems in Problem (1) in Sections 4.1–4.2. Then, we provide its theoretical convergence guarantees in Section 4.3. Lastly, we will reveal a key insight on the benefit of using the proposed variance reduction techniques in Section 4.4. Due to space limitation, we relegate the proofs to supplementary material.

4.1. Preliminaries

To present the Prometheus algorithm, we first introduce several basic components as preparation.

1) Network-Consensus Matrix: Our Prometheus algorithm is based on the network-consensus mixing approach:

in each iteration, every agent exchanges and aggregates neighboring information through a consensus weight matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$. We define λ as the second largest eigenvalue of the matrix $\mathbf{M}$. Let $[\mathbf{M}]_{ii'}$ represent the element in the i-th row and the i'-th column in $\mathbf{M}$. The choice of $\mathbf{M}$ should satisfy the following properties: (a) *doubly stochastic:* $\sum_{i=1}^m [\mathbf{M}]_{ii'} = \sum_{j=1}^m [\mathbf{M}]_{ii'} = 1$; (b) *symmetric:* $[\mathbf{M}]_{ii'} = [\mathbf{M}]_{i'i}, \forall i, i' \in \mathcal{N}$; and (c) *network-defined sparsity:* $[\mathbf{M}]_{ii'} > 0$ if $(i, i') \in \mathcal{L}$; otherwise $[\mathbf{M}]_{ii'} = 0, \forall i, i' \in \mathcal{N}$.

2) Stochastic Gradient Estimators: In Prometheus, we need to estimate the stochastic gradient of the bilevel problem using the implicit function theorem. We note that in the literature of bilevel optimization with stochastic gradient, a commonly adopted stochastic gradient estimator is of the form (Khanduri et al., 2021; Ghadimi & Wang, 2018; Hong et al., 2020; Liu et al., 2022):

$$\bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{ij}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0) - \frac{1}{L_g} \nabla^2_{\mathbf{xy}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0) \hat{\mathbf{H}}_{i,k} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0), \quad (5)$$

where $\hat{\mathbf{H}}_{i,k} \triangleq K \prod_{p=1}^{k(K)} (\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^p)}{L_g})$. Here, $K \in \mathbb{N}$ is a predefined parameter and $k(K) \sim \mathcal{U}\{0, \ldots, K-1\}$ is an integer-valued random variable uniformly chosen from $\{0, \ldots, K-1\}$. It can be shown that $\hat{\mathbf{H}}_{i,k}$ is a biased estimator for the Hessian inverse $\left[\nabla^2_{\mathbf{yy}} g(\mathbf{x}, \mathbf{y}; \zeta)\right]^{-1} = \sum_{i=1}^{\infty} (\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g(\mathbf{x}, \mathbf{y}; \zeta)}{L_g})^i$. However, this estimator has the limitation that it only incorporates the *first* term in the Taylor approximation, thus resulting in a large variance and could eventually increase the communication complexity of decentralized bilevel optimizaiton.

To address this issue, in this paper, we propose a new stochastic gradient estimator inspired by (Koh & Liang, 2017) from conventional single-level optimization:

$$\bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{ij}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0) - \frac{1}{L_g} \nabla^2_{\mathbf{xy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0\right) \mathbf{H}_{i,k} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0), \quad (6)$$

where $\mathbf{H}_{i,0} = \mathbf{I}$ and

$$\mathbf{H}_{i,k} = \mathbf{I} + \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^k\right)}{L_g}\right) \mathbf{H}_{i,k-1} = \mathbf{I} + \sum_{j'=1}^{k(K)} \prod_{p=1}^{j'} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^p\right)}{L_g}\right). \quad (7)$$

Compared to the conventional estimator, the key difference in our new estimator lies in the matrix $\mathbf{H}_{i,k}$. In particular, our $\mathbf{H}_{i,k}$ is in a recursive form that is able to capture the *entire* Taylor series at once without increasing the sample complexity. Thanks to this recursive form, $\mathbf{H}_{i,k}$ utilizes $O(k^2)$ samples, as opposed to only $O(k)$ samples in the conventional $\hat{\mathbf{H}}_{i,k}$-Hessian inverse estimator, thus leading to a much smaller variance and eventually much lower communication complexity. It is worth noting that although our $\mathbf{H}_{i,k}$ estimator leverages more training samples, the computation cost is the *same* as that of $\hat{\mathbf{H}}_{i,k}$ due to the recursive structure in (6).

As mentioned earlier, our Hessian inverse estimator is inspired by ideas in stochastic second-order optimization (Agarwal et al., 2016). Interestingly, similar technique to estimate the Hessian inverse also appeared in (Koh & Liang, 2017). However, (Koh & Liang, 2017) and (Agarwal et al., 2016) are only designed for solving a conventional single-level minimization problem. In comparison, our proposed stochastic estimator can be used in bilevel learning, particularly for solving non-smooth regularizers in upper-level problems, which are far more complicated and require new proof techniques and performance analysis. More importantly, the Hessian inverse estimator technique was used in (Koh & Liang, 2017) and (Agarwal et al., 2016) as a heuristic without any performance analyis. In contrast, we theoretically and numerically demonstrate that our new estimator outperforms the conventional one in Sections 4.3 and 5. Our theoretical analysis (cf. Lemma 1) shows that the Lipschitz constant of our estimator is smaller compared to the conventional one. Our experimental results (cf. Fig. 3 and the Appendix) further confirm that our estimator has a small variance.

4.2. The Prometheus Algorithm

The algorithm design and analysis for solving Problem (1) faces a number of challenges: (i) the objective function $\mathbf{x}$ is non-convex; (ii) the objective function is non-smooth due to the proximal function; (iii) the constraint set on $\mathbf{x}$-variables; (iv) the decentralized bilevel problem structure. The main challenge comes from the *coupling* between proximal operation (for addressing challenges (ii) and (iii)) and the decentralized bi-level structure), which renders the theoretical analysis of algorithm design extremely challenging in proving our proposed algorithm to be both sample- and communication-efficient. To address these challenges, our proposed Prometheus algorithm carefully integrates proximal, gradient tracking, and variance reduction techniques, which can be viewed as a triple-hybrid approach. The procedure of Prometheus can be organized into three key steps:

- *Step 1 (Local Proximal Operations):* In each iteration t, each agent i performs proximal operations to cope with the domain constraint set $\mathcal{X}$ for the upper-level variables:

$$\widetilde{\mathbf{x}}_{i,t} = \tilde{\mathbf{x}}_i(\mathbf{x}_{i,t}) = \arg\min_{\mathbf{x}\in\mathcal{X}} [\langle \mathbf{u}_{i,t}, \mathbf{x} - \mathbf{x}_{i,t}\rangle + \tfrac{\tau}{2}\|\mathbf{x} - \mathbf{x}_{i,t}\|^2 + h(\mathbf{x})], \quad (8)$$

where $\tau > 0$ is a proximal control parameter and $\mathbf{u}_{i,t}$ is an auxiliary vector. The proximal update rule is motivated by the SONATA method (Scutari & Sun, 2019) used in a decentralized minimization.

- *Step 2 (Consensus Update in Upper-Level Variables):* Next, each agent i updates the upper and lower model parameters $\mathbf{x}_i, \mathbf{y}_i$ as follows:

$$\mathbf{x}_{i,t+1} = \sum_{i' \in \mathcal{N}_i} [\mathbf{M}]_{ii'} \mathbf{x}_{i',t} + \alpha(\tilde{\mathbf{x}}_i(\mathbf{x}_{i,t}) - \mathbf{x}_{i,t}), \quad (9)$$

$$\mathbf{y}_{i,t+1} = \mathbf{y}_{i,t} - \beta \mathbf{v}_{i,t}, \quad (10)$$

where α and β are constant step-sizes for updating $\mathbf{x}$- and $\mathbf{y}$-variables, respectively. Note that updating $\mathbf{x}_{i,t+1}$ in Eq. (9) is a local weighted average at agent i and plus a local update in the spirit of Frank-Wolfe given a proximal point. Eq. (10) performs a local stochastic gradient descent update for the $\mathbf{y}$-variable at each agent i.

It is worth pointing out that the auxiliary proximal operator $\tilde{\mathbf{x}}_{i,t}$ in (8) and (9) and the resultant local update $\alpha(\tilde{\mathbf{x}}_i(\mathbf{x}_{i,t}) - \mathbf{x}_{i,t})$ in the consensus step play an important role in helping us tackle the non-smooth objective challenge. This successive convex approximation (SCA) technique is dramatically different from the conventional algorithm design in ordinary single-level stochastic optimization. Without this new SCA technique, it will be difficult, if not entirely impossible, to achieve convergence guarantees. Moreover, the use of the above new SCA technique also necessitates many proof techniques that are quite different from the proofs in ordinary single-level stochastic optimization (see the proof details of our Lemma 5 and Lemma 7 in the Appendix).

- *Step 3 (Local Variance-Reduced Stochastic Gradient Estimate):* In the local gradient estimator step, each agent i estimates its local gradients using the following stochastic gradient estimators:

$$\mathbf{p}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) = \begin{cases} \bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) = \frac{1}{n} \sum_{j=1}^{n} \bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{ij}), \\ \quad \text{if } \mathrm{mod}(t, q) = 0, \\ \mathbf{p}_i(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1}) + \frac{1}{|\mathcal{S}_{i,t}|} \sum_{j \in \mathcal{S}_{i,t}} \\ \left(\bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{ij}) - \bar{\nabla} f\left(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1}, \bar{\xi}_{ij}\right)\right), \end{cases} \quad (11a)$$

$$\mathbf{d}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) = \begin{cases} \bar{\nabla} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) = \frac{1}{n} \sum_{j=1}^{n} \bar{\nabla} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_{ij}), \\ \quad \text{if } \mathrm{mod}(t, q) = 0, \\ \mathbf{d}_i(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1}) + \frac{1}{|\mathcal{S}_{i,t}|} \sum_{j \in \mathcal{S}_{i,t}} \\ \left(\bar{\nabla} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{ij}) - \bar{\nabla} g\left(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1}, \zeta_{ij}\right)\right). \end{cases} \quad (11b)$$

Here, $\mathcal{S}_{i,t}$ is the sample mini-batch in the t-th iteration, and q is a pre-determined inner loop iteration number. The local stochastic gradient estimation is a recursive estimator that shares some structural similarity with those in SARAH (Nguyen et al., 2017), SPIDER (Fang et al., 2018), and PAGE (Li et al., 2021) used for traditional minimization problems.

Algorithm 1 The Prometheus Algorithm at i^{th} agent.

Set parameter pair $(\mathbf{x}_{i,0}, \mathbf{y}_{i,0}) = (\mathbf{x}^0, \mathbf{y}^0)$.
Calculate local gradients: $\mathbf{u}_{i,0} = \bar{\nabla} f(\mathbf{x}_{i,0}, \mathbf{y}_{i,0})$; $\mathbf{v}_{i,0} = \nabla_{\mathbf{y}} g(\mathbf{x}_{i,0}, \mathbf{y}_{i,0})$;
for $t = 1, \cdots, T$ **do**
  Update local models $(\mathbf{x}_{i,t+1}, \mathbf{y}_{i,t+1})$ as in Eqs. (8)-(10);
  if Prometheus: **then**
    Compute the $(\mathbf{p}_i(\mathbf{x}_{i,t+1}, \mathbf{y}_{i,t+1}), \mathbf{d}_i(\mathbf{x}_{i,t+1}, \mathbf{y}_{i,t+1}))$ local estimator as in Eq. (11);
  end if
  if Prometheus-SG: **then**
    Compute the $(\mathbf{p}_i(\mathbf{x}_{i,t+1}, \mathbf{y}_{i,t+1}), \mathbf{d}_i(\mathbf{x}_{i,t+1}, \mathbf{y}_{i,t+1}))$ local estimators as in Eq. (13);
  end if
  Track global gradients $(\mathbf{u}_{i,t+1}, \mathbf{v}_{i,t+1})$ as in Eq. (12);
end for

- *Step 4 (Gradient Tracking in Upper-Level Parameters): Each agent i updates $\mathbf{u}_{i,t}$ and $\mathbf{v}_{i,t}$ by averaging over its neighboring tracked gradients:*

$$\mathbf{u}_{i,t} = \sum_{i' \in \mathcal{N}_i} [\mathbf{M}]_{ii'} \mathbf{u}_{i',t-1} + \mathbf{p}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{p}_i(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1});$$

$$\mathbf{v}_{i,t} = \mathbf{d}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}). \quad (12)$$

We summarize the Prometheus algorithm in Algorithm 1.

4.3. Convergence Analysis of the Prometheus Algorithm

Now, we focus on the convergence performance analysis for the proposed Prometheus algorithm. Before presenting the main convergence results, we first state several needed technical assumptions:

Assumption 1. For all $\zeta \in \mathrm{supp}\,(\pi_g)$ where $\mathrm{supp}(\pi)$ is the support of π, $\mathbf{x} \in \mathcal{X}, \mathcal{X} \subseteq \mathbb{R}^{p_1}, \mathbf{y} \in \mathbb{R}^{p_2}$, the lower-level function g has the following properties :

1) $g(\mathbf{x}, \mathbf{y}; \zeta)$ is μ_g-strongly convex with $\mu_g > 0$, $\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}; \zeta)$ is L_g-Lipschitz continuous with $L_g > 0$;

2) $\left\|\nabla^2_{\mathbf{xy}} g(\mathbf{x}, \mathbf{y}; \zeta)\right\|^2 \leq C_{g_{xy}}$ for some $C_{g_{xy}} > 0$, $\nabla^2_{\mathbf{xy}} g(\mathbf{x}, \mathbf{y}; \zeta)$ and $\nabla^2_{\mathbf{yy}} g(\mathbf{x}, \mathbf{y}; \zeta)$ are Lipschitz continuous with constants $L_{g_{xy}} > 0$ and $L_{g_{yy}} > 0$, respectively.

Assumption 2. For all $\xi \in \mathrm{supp}\,(\pi_f)$ where $\mathrm{supp}(\pi)$ is the support of π, $\mathbf{x} \in \mathcal{X}, \mathcal{X} \subseteq \mathbb{R}^{p_1}$, the upper-level function f has the following properties : $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \xi), \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \xi)$ are Lipschitz smooth continuous with constant $L_{f_x} \geq 0, L_{f_y} \geq 0$. $\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \xi)\| \leq C_{f_y}$, for some $C_{f_y} \geq 0$.

Assumption 3. *i)* the stochastic gradient estimate of the upper-level function satisfies: $\mathbb{E}_{\bar{\xi}}[\|\bar{\nabla} f(\mathbf{x},\mathbf{y};\bar{\xi}) - \mathbb{E}_{\bar{\xi}}[\bar{\nabla} f(\mathbf{x},\mathbf{y};\bar{\xi})]\|^2] \leq \sigma_f^2$; and *ii)* the stochastic gradient estimate of the lower-level function satisfies: $\mathbb{E}_{\zeta}[\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y};\zeta) - \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\|^2] \leq \sigma_g^2$.

We note that Assumptions.1, 2 and 3(b) are standard in the literatures of bilevel optimization (see, e.g., (Ghadimi & Wang, 2018; Khanduri et al., 2021). In addition, Assumption 3(a) has been verified in (Khanduri et al., 2021).

To establish the convergence result of Prometheus, we first prove the Lipschitz-smoothness of the new gradient estimator proposed in (6), which is stated as follows:

Lemma 1. (Lipschitz-smoothness of the new stochastic gradient estimator in (6)). If the stochastic functions $f(\mathbf{x},\mathbf{y};\xi)$ and $g(\mathbf{x},\mathbf{y};\zeta)$ satisfy Assumptions 1–3, then we have (i) for a fixed $\mathbf{y} \in \mathbb{R}^{p_2}$, $\left\|\bar{\nabla} f\left(\mathbf{x}_1,\mathbf{y};\bar{\xi}\right) - \bar{\nabla} f\left(\mathbf{x}_2,\mathbf{y};\bar{\xi}\right)\right\|^2 \leq L_f^2 \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2, \forall\ \mathbf{x}_1,\ \mathbf{x}_2 \in \mathbb{R}^{p_1}$; and (ii) for a fixed $\mathbf{x} \in \mathbb{R}^{p_1}$, $\left\|\bar{\nabla} f\left(\mathbf{x},\mathbf{y}_1;\bar{\xi}\right) - \bar{\nabla} f\left(\mathbf{x},\mathbf{y}_2;\bar{\xi}\right)\right\|^2 \leq L_f^2 \left\|\mathbf{y}_1 - \mathbf{y}_2\right\|^2, \forall \mathbf{y}_1,\mathbf{y}_2 \in \mathbb{R}^{p_2}$. In the above expressions, $L_f > 0$ is defined as: $L_f^2 := 2L_{f_x}^2 + 6C_{g_{xy}}^2 L_{f_y}^2 \left(\frac{K}{2\mu_g L_g - \mu_g^2}\right) + 6C_{f_y}^2 L_{g_{xy}}^2 \left(\frac{K}{2\mu_g L_g - \mu_g^2}\right) + 6C_{g_{xy}}^2 C_{f_y}^2 \frac{K}{L_g^2} \sum_{j=1}^{K} j^2 \left(1 - \frac{\mu_g}{L_g}\right)^{2(j-1)} \frac{1}{L_g^2} L_{g_{yy}}^2$.

We note that the Lipschitz-smoothness constant L_f of Lemma 1 is smaller than that of the conventional estimator in (5), which we denote as L_{conv} here, *i.e.*, $L_f \leq L_{conv}$. This also shows superiority of our new estimator. Due to space limitation, we state the definition of L_{conv} in Lemma 4 in the appendix.

Next, we need the following Lipschitz-continuity properties of the approximate gradient $\bar{\nabla} f(\mathbf{x},\mathbf{y})$, the lower level solution $\mathbf{y}*$, and the true gradient $\nabla \ell(\mathbf{x})$, which have been proved in the literature:

Lemma 2. (Ghadimi & Wang, 2018) Under Assumptions 1–2, we have $\|\bar{\nabla} f(\mathbf{x},\mathbf{y}) - \nabla \ell(\mathbf{x})\| \leq L \|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}\|$, $\|\mathbf{y}^*(\mathbf{x}_1) - \mathbf{y}^*(\mathbf{x}_2)\| \leq L_y \|\mathbf{x}_1 - \mathbf{x}_2\|$, $\|\nabla \ell(\mathbf{x}_1) - \nabla \ell(\mathbf{x}_2)\| \leq L_\ell \|\mathbf{x}_1 - \mathbf{x}_2\|$ for all $\mathbf{x},\mathbf{x}_1,\mathbf{x}_2 \in \mathbb{R}^{p_1}, \mathbf{y} \in \mathbb{R}^{p_2}$, where the Lipschitz constants are defined as: $L \triangleq L_{f_x} + \frac{L_{f_y} C_{g_{xy}}}{\mu_g} + C_{f_y}\left(\frac{L_{g_{xy}}}{\mu_g} + \frac{L_{g_{yy}} C_{g_{xy}}}{\mu_g^2}\right)$, $L_\ell \triangleq L + \frac{L C_{g_{xy}}}{\mu_g}$, and $L_y \triangleq \frac{C_{g_{xy}}}{\mu_g}$.

Lemma 2 establishes the smoothness of the implicit function in Problem (1), which only relies on the Assumptions 1 and 2 to hold. Lastly, following the same token as in (Hong et al., 2020), we show a critical fact on the exponentially fast decay of the bias of our stochastic estimator in (6), which is stated as follows.

Lemma 3 (Exponentially Decaying Bias)**.** Under Assumptions 1–3, the stochastic gradient estimate of the upper level objective in (6) satisfies $\|\nabla f(\mathbf{x},\mathbf{y}) - \mathbb{E}[\bar{\nabla} f(\mathbf{x},\mathbf{y};\bar{\xi})]\| \leq \frac{C_{g_{xy}} C_{f_y}}{\mu_g}\left(1 - \frac{\mu_g}{L_g}\right)^K$.

The assumptions above and Lemmas 1-3 lead to the main convergence result of Prometheus as follows.

Theorem 1. Under Assumptions1-3, if the step-sizes $\alpha \leq \min\{C_{1,i}(\lambda, m, L_l, L_f, L_y, L, \mu_g, \tau, \beta), i = 1,\ldots,10\}$ and $\beta \leq \min\{C_{2,i}(L_y, L_f, \lambda, \mu_g), i = 1,\ldots,4$, where $C_{1,i}(\cdot)$ and $C_{2,i}(\cdot)$, $\forall i$, signify that these terms are constants that depend on the problem-specific parameters in $(\cdot)$ and their exact expressions can be found in the Appendix, then the sequence $\{\mathbf{x}_t\}$ outputs by Prometheus satisfies:

$$\frac{1}{T}\sum_{t=0}^{T-1}[\mathbb{E}\|\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + \mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + \mathbb{E}\|\mathbf{y}_t - \mathbf{y}_t^*\|^2] = \mathcal{O}\left(\frac{1}{T}\right).$$

It is worth noting that, compared to existing works on decentralized bilevel optimization, the major challenge in proving the convergence results in Theorem 1 stems from the *proximal operator* needed to solve the upper-level subproblem, which prevents the use of conventional descent lemma for convergence analysis (see Eq. (36) in the appendix). Also, compared to single-agent constrained bilevel optimization, one cannot provide theoretical convergence guarantee by using the direct projection method $\widetilde{\mathbf{x}}_{i,t} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_{i,t} - \tau \mathbf{u}_{i,t})\|^2$ as in (Hong et al., 2020; Chen et al., 2022a) due to the gradient tracking procedure in the decentralized learning. Instead, we use a different proximal update rule as shown in (8). If we do not use this SCA technique and use the direct proximal operator in (Hong et al., 2020; Chen et al., 2022a), we will numerically show in Section 5 that the algorithm would only converge to a neighborhood of a stationary point.

As mentioned earlier, the auxiliary proximal operator $\tilde{\mathbf{x}}_{i,t}$ in (8) and (9) and the resultant local update $\alpha(\tilde{\mathbf{x}}_i(\mathbf{x}_{i,t}) - \mathbf{x}_{i,t})$ help us tackle the *non-smooth* objective challenge. This successive convex approximation (SCA) technique is critical for achieving convergence guarantees. However, the use of the above new SCA technique also necessitates many proof techniques that are dramatically different from the proofs in ordinary single-level stochastic optimization (see the proof details of our Lemmas 5 and 7 in the Appendix. Fig.5).

We note that the graph structure of the underlying network does not change the order of convergence rate of our algorithms theoretically (i.e., the T-dependence in the Big-O result in Theorem 1). The step sizes α and β depend on the network topology through λ, where λ is the second largest eigenvalue in magnitude of the network consensus matrix M (i.e., $\lambda = \max\{|\lambda_2|, |\lambda_m|\} \in (0,1)$), which is in turn determined by the network graph topology. For a sparse network, λ is close to 1, while for a dense network, λ is

close to 0. As a result, for a sparse network with λ being close to 1, Theorem 1 implies smaller step sizes α and β ($\alpha = O(1-\lambda)$, $\beta = O((1-\lambda)^4)$, which can then lead to a slower convergence. But theoretically, these smaller step sizes only affect the hidden constants in the $O(1/T)$ convergence result in Theorem 1, but not the T-dependence. Also, experimentally, we observe that the graph structure only has a small impact on the convergence as shown in our appendix. Further, Theorem 1 implies the following sample and communication complexity results:

Corollary 2 (Sample and Communication Complexities of Prometheus)**.** Under the conditions of Theorem 1, to achieve an ϵ-stationary solution, Prometheus requires that: i) the total number of communication rounds is $\mathcal{O}(\epsilon^{-1})$, and ii) the total number of samples is $\mathcal{O}(\sqrt{n}K\epsilon^{-1}+n)$.

4.4. Discussion: Variance Reduction in Prometheus

Since the variance reduction in (11) in Step 3 of Prometheus requires full gradient evaluation, it is tempting to ask what is the benefit of using the variance reduction technique. In other words, could we relinquish variance reduction (VR) in Step 3 to avoid full gradient evaluation? To answer this question, consider changing Step 3 to the following basic stochastic gradient estimator without VR:

$$\begin{aligned}\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t}) &= \bar{\nabla} f(\mathbf{x}_{i,t},\mathbf{y}_{i,t},\bar{\xi}_{i0});\\ \mathbf{d}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t}) &= \nabla g(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\zeta_{i0}).\end{aligned} \tag{13}$$

Interestingly, the following convergence result states that there always exists a *non-vanishing* constant independent of m, n, and α if Eq. (13) is used in Step 3 of Prometheus (i.e., a constant only dependent on problem instance and cannot be made arbitrarily small algorithmically).

Proposition 3. Under Assumptions1–3, with step-sizes $\alpha \le \min\{\frac{1-\lambda}{8\beta L_f}, \frac{\tau}{3L_\ell}, \frac{(1-\lambda)m}{2\sqrt{\beta}(L_\ell+\tau)}\frac{\tau}{6+3\tau}, \frac{\tau\sqrt{\beta}}{6m(1-\lambda)}, \frac{\tau(1-\lambda)}{48mL_f^2\beta}, \frac{(1-\lambda)\mu_g^2\beta^{1.5}}{23040L_y^2L^2}, \mathcal{O}(T^{-\frac{1}{2}}), \frac{(1-\lambda)m}{4\sqrt{\beta}\tau}\}, \beta \le \min\{\frac{1-\lambda}{8L_f}, \frac{(1-\lambda)^4\mu_g^2}{480^2L_y^2L_f^2}, \mathcal{O}(T^{-\frac{1}{3}})\}$, we have the following result if Eq. (13) replaces Step 3 in Prometheus,

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\right) = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+C'_\sigma,$$

where the constant C'_σ is defined as $C'_\sigma \triangleq \frac{9(6+3\tau)}{\tau^2}((\frac{C_{gxy}C_{fy}}{\mu_g}(1-\frac{\mu_g}{L_g})^K)^2+\sigma_f^2)+\frac{27(1-\lambda)}{40(8+4\alpha^2)L_y^2}\frac{\beta^{1.5}}{\alpha\tau}\sigma_g^2$.

A key insight of Proposition 3 is in order. The SG-type update in (13) is similar to the SG-type update in *unconstrained* bilevel optimization in the *single-agent* setting (Ji et al., 2021). However, unlike the SG-type method in (Ji et al., 2021) that can approach zero at an $\mathcal{O}(1/\sqrt{T})$ convergence rate, the SG-type method can *only* approach a constant error C'_σ at an $\mathcal{O}(1/\sqrt{T})$ convergence rate in the *constrained* decentralized setting. The non-vanishing constant error C'_σ is caused by the variance σ_f^2 and σ_g^2 of the stochastic gradient. Proposition 3 highlights the benefit of using the variance reduction techniques to eliminate the $\{\sigma_f,\sigma_g\}$-variance in order to approach zero asymptotically.

5. Numerical Results

In this section, we will first conduct experiments to demonstrate the small variance of our new stochastic gradient estimator. Then, we will compare Prometheus' convergence with several baselines.

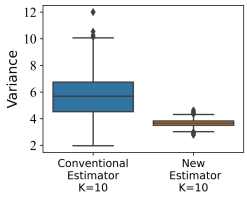

Figure 3. Hessian inverse estimator comparison.

1) New estimator vs. conventional estimator: Note that the major difference between the new and conventional estimators lies in how they estimate the Hessian inverse of the matrix $\mathbf{A}$. Thus, it suffices to compare the Hessian inverse approximations. The conventional estimator to estimate the $\mathbf{A}^{-1}$ can be denoted as $\tilde{\mathbf{A}}^{-1}_{conv} = K\prod_{p=1}^{k(K)}(\mathbf{I}-\mathbf{A}_s)$, while the new estimator can be denoted as $\tilde{\mathbf{A}}^{-1} = \sum_{j'=1}^{k(K)}\prod_{p=1}^{j'}(\mathbf{I}-\mathbf{A}_s)$. To see the benefits of our estimator and due to the high complexity of computing matrix inverse, here we consider a small example $\mathbf{A} = [[0.25, 0.0], [0.0, 0.25]]$, so that $\mathbf{A}^{-1}_{true} = [[4, 0], [0, 4]]$. Let $\mathbf{A}_s$ be a random matrix obtained from $\mathbf{A}$ plus Gaussian noise. We use $\tilde{\mathbf{A}}^{-1}_{conv}$ and $\tilde{\mathbf{A}}^{-1}$ to estimate $\mathbf{A}^{-1}$, respectively. We run 10000 independent trials with $K = 10$ and the results are shown in Fig. 3. We can see from Fig. 3 that the new Hessian inverse estimator has a much smaller variance than the conventional one. Additional experimental results on varying K and with different matrix $\mathbf{A}$ are relegated to our Appendix.

2) Convergence Performance: We verify our theoretical results of Prometheus by conducting experiments on a meta-learning problem tested on MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) datasets. Due to space limitation, we provide additional experiments on hyper-parameter optimization in the appendix. Due to the lack of existing algorithms for solving constrained decentralized bilevel optimization problem, we compare the convergence performance of Prometheus against several stripped-down version of Prometheus:

- Prometheus *with Stochastic Gradient* (Prometheus-SG): Prometheus-SG is the SG-type algo-

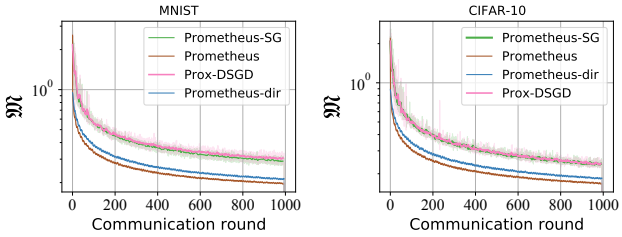

Figure 1. Five-agent network.

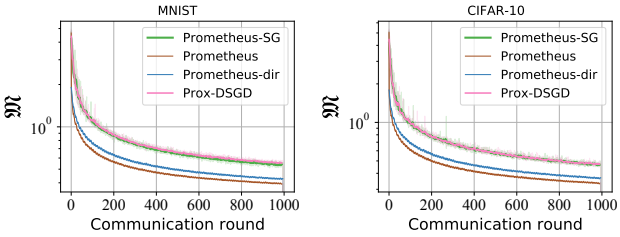

Figure 2. Ten-agent network.

rithm discussed in Section 4.4: $\mathbf{p}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) = \bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}, \bar{\xi}_{i0}); \mathbf{d}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) = \nabla g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_{i0})$.

- Prometheus *with Direct Proximal Method* (Prometheus-dir): Instead of performing $\widetilde{\mathbf{x}}_{i,t} = \arg\min_{\mathbf{x}\in\mathcal{X}}[\langle \mathbf{u}_{i,t}, \mathbf{x} - \mathbf{x}_{i,t}\rangle + \frac{\tau}{2}\|\mathbf{x} - \mathbf{x}_{i,t}\|^2 + h(\mathbf{x}_i)]$ in Prometheus, Prometheus-dir directly adds the constraints on $\mathbf{x}$: $\widetilde{\mathbf{x}}_{i,t} = \arg\min_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x} - (\mathbf{x}_{i,t} - \tau\mathbf{u}_{i,t})\|^2$.
- *Proximal Decentralized Stochastic Gradient Descent (Prox-DSGD):* This algorithm is motivated by the DSGD algorithm, which can be viewed as Prometheus without using gradient tracking. Specifically, we updates local gradient as $\mathbf{u}_{i,t} = \bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{i0}); \mathbf{v}_{i,t} = \nabla g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_{i0})$.

We also note that the Prox-DSGD algorithm can be seen as a generalization of DSBO (Yang et al., 2022), SPDB (Lu et al., 2022a), DSBO (Chen et al., 2022b) with the proximal operator. Prometheus-dir can also be seen as an extension of the algorithm INTERACT (Liu et al., 2020) to handle the constrained decentralized bilevel optimization problem. We compare Prometheus with these baselines using a two-hidden-layer neural network with 20 hidden units. The consensus matrix is chosen as $\mathbf{M} = \mathbf{I} - \frac{2\mathbf{L}}{3\lambda_{\max}(\mathbf{L})}$, where $\mathbf{L}$ is the Laplacian matrix of $\mathcal{G}$ and $\lambda_{\max}(\mathbf{L})$ denotes the largest eigenvalue of $\mathbf{L}$. Due to space limitation, we relegate the detailed parameter choices of all algorithms to the appendix. In Fig. 1, we compare the performance of Prometheus, Prometheus-SG, Prometheus-dir, and Prox-DSGD on the MNIST and CIFAR-10 datasets with with a five-agent network. The network topology can be seen in Fig. 4 in Appendix D. We note that Prometheus converges much faster than than all other algorithms in terms of the total number of communication rounds. In Fig. 2, we also observe similar results when the number of tasks (and agents) is increased to 10. Our experimental results thus verify our theoretical analysis that Prometheus has the lowest communication complexity.

6. Conclusion

In this paper, we studied the constrained decentralized nonconvex-strongly-convex bilevel optimization problems. First, we proposed an algorithm called Prometheus with a new stochastic estimator. We then showed that, to achieve an ϵ-stationary point, Prometheus achieves a sample complexity of $\mathcal{O}(K\sqrt{n}\epsilon^{-1}+n)$ and a communication complexity of $\mathcal{O}(\epsilon^{-1})$. Our numerical studies also showed the advantages of our proposed Prometheus and verified the theoretical results. Collectively, the results in this work contribute to the state of the art of low sample- and communication-complexity constrained decentralized bilevel learning.

Acknowledgments and Disclosure of Funding

This work has been supported in part by NSF grants CAREER CNS-2110259, CNS-2112471, ECCS-2140277, and CCF-2110252.

References

Agarwal, N., Bullins, B., and Hazan, E. Second-order stochastic optimization in linear time. *stat*, 1050:15, 2016.

Bennajeh, A., Bechikh, S., Said, L. B., and Aknine, S. Bilevel decision-making modeling for an autonomous driver agent: application in the car-following driving behavior. *Applied Artificial Intelligence*, 33(13):1157–1178, 2019.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Chen, T., Sun, Y., Xiao, Q., and Yin, W. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2466–2488. PMLR, 2022a.

Chen, X., Huang, M., and Ma, S. Decentralized bilevel optimization. *arXiv preprint arXiv:2206.05670*, 2022b.

Chen, X., Huang, M., Ma, S., and Balasubramanian, K. Decentralized stochastic bilevel optimization with improved per-iteration complexity. *arXiv preprint arXiv:2210.12839*, 2022c.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

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

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577. PMLR, 2018.

Gao, H., Gu, B., and Thai, M. T. Stochastic bilevel distributed optimization over a network. *arXiv preprint arXiv:2206.15025*, 2022.

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

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pp. 3748–3758. PMLR, 2020.

Guo, Z. and Yang, T. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv e-prints*, pp. arXiv–2105, 2021.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

Hong, M., Zeng, S., Zhang, J., and Sun, H. On the divergence of decentralized nonconvex optimization. *SIAM Journal on Optimization*, 32(4):2879–2908, 2022.

Huang, M., Zhang, D., and Ji, K. Achieving linear speedup in non-iid federated bilevel learning. *arXiv preprint arXiv:2302.05412*, 2023.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pp. 4882–4892. PMLR, 2021.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34, 2021.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *Available: http://yann. lecun. com/exdb/mnist*, 1998.

Li, G., Zhang, R., Jiang, T., Chen, H., Bai, L., and Li, X. Security-constrained bi-level economic dispatch model for integrated natural gas and electricity systems considering wind power and power-to-gas process. *Applied energy*, 194:696–704, 2017.

Li, Z., Bao, H., Zhang, X., and Richtarik, P. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pp. 6286–6295. PMLR, 2021.

Liu, M., Zhang, W., Mroueh, Y., Cui, X., Ross, J., Yang, T., and Das, P. A decentralized parallel algorithm for training generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, 2020.

Liu, Z., Mu, H., Zhang, X., Guo, Z., Yang, X., Cheng, K.-T., and Sun, J. Metapruning: Meta learning for automatic neural network channel pruning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3296–3305, 2019.

Liu, Z., Zhang, X., Khanduri, P., Lu, S., and Liu, J. Interact: Achieving low sample and communication complexities in decentralized bilevel learning over networks. *arXiv preprint arXiv:2207.13283*, 2022.

Lu, S., Cui, X., Squillante, M. S., Kingsbury, B., and Horesh, L. Decentralized bilevel optimization for personalized client learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5543–5547. IEEE, 2022a.

Lu, S., Zeng, S., Cui, X., Squillante, M., Horesh, L., Kingsbury, B., Liu, J., and Hong, M. A stochastic linearized augmented lagrangian method for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 35:30638–30650, 2022b.

Maclaurin, D., Duvenaud, D., and Adams, R. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pp. 2113–2122. PMLR, 2015.

Mansoor, A., Diao, X., and Smidts, C. Backward failure propagation for conceptual system design using isfa. 11 2021.

Mansoor, A., Diao, X., and Smidts, C. A method for backward failure propagation in conceptual system design. *Nuclear Science and Engineering*, 2023. doi: 10.1080/00295639.2023.2196937.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takac, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pp. 2613–2621. PMLR, 2017.

Panagoda, M. H. *Convergence Analysis and Bilevel Optimization Algorithms for Matrix Completion Problems*. PhD thesis, George Mason University, 2021.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

Pochmann, V. O. and Von Zuben, F. J. Multi-objective bilevel recommender system for food diets. In *2022 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8. IEEE, 2022.

Poon, C. and Peyré, G. Smooth bilevel programming for sparse regularization. *Advances in Neural Information Processing Systems*, 34:1543–1555, 2021.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Scutari, G. and Sun, Y. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176(1):497–544, 2019.

Tian, H., Liu, B., Yuan, X.-T., and Liu, Q. Meta-learning with network pruning. In *European Conference on Computer Vision*, pp. 675–700. Springer, 2020.

Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Yang, S., Zhang, X., and Wang, M. Decentralized gossip-based stochastic bilevel optimization over communication networks. *arXiv preprint arXiv:2206.10870*, 2022.

Zhang, H., Chen, W., Huang, Z., Li, M., Yang, Y., Zhang, W., and Wang, J. Bi-level actor-critic for multi-agent coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7325–7332, 2020.

Variable	Definition
$\mathcal{N}$	Set of nodes.
$\mathcal{L}$	Set of edges.
m	Number of agents.
i	i-th agent.
j	j-th local sample at each agent.
T	Total iteration numbers.
t	t-th iteration.
K	$K \in \mathbb{N}$ is a predefined parameter.
k	k is an integer-valued random variable uniformly chosen from $\{0, \ldots, K-1\}$.
α	Upper-level step-size.
β	Lower-level step-size.
τ	Proximal control parameter.
μ_g	Constant from the strongly-convex assumption, see details in Assumption. 1.
L_g	Constant from the Lipschitz continuous assumption, see details in Assumption. 1.
$C_{g_{xy}}$	Constant from the bounded gradient assumption, see details in Assumption. 1.
$L_{g_{xy}}$	Constant from the gradient Lipschitz continuous assumption, see details in Assumption. 1.
L_{f_x}, L_{f_y}	Constant from the Lipschitz smooth continuous assumption, see details in Assumption. 2.
C_{f_y}	Constant from the bounded gradient assumption, see details in Assumption. 2.
σ_f, σ_g	Constant from the bounded variance assumption, see details in Assumption. 3.
$\mathbf{M}$	Consensus weight matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$.
λ	Second largest eigenvalue of matrix $\mathbf{M}$.

Table 2. Notation Table.

A. Additional Theoretical Results

Lemma 4. (Lipschitz-smoothness of conventional stochastic gradient estimator). With the conventional stochastic gradient estimator $\bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{ij}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0) - \frac{K}{L_g} \nabla^2_{\mathbf{xy}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0) \cdot \prod_{p=1}^{k(K)} (I - \frac{\nabla^2_{\mathbf{yy}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^p)}{L_g}) \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)$. If the stochastic functions $f(\mathbf{x}, \mathbf{y}; \xi)$ and $g(\mathbf{x}, \mathbf{y}; \zeta)$ satisfy Assumptions 1–3, then we have

(i) For a fixed $\mathbf{y} \in \mathbb{R}^{p_2}$, $\mathbb{E}_{\bar{\xi}} \left\| \nabla f\left(\mathbf{x}_1, \mathbf{y}; \bar{\xi}\right) - \nabla f\left(\mathbf{x}_2, \mathbf{y}; \bar{\xi}\right) \right\|^2 \leq L_{conv}^2 \left\| \mathbf{x}_1 - \mathbf{x}_2 \right\|^2, \forall\, \mathbf{x}_1,\ \mathbf{x}_2 \in \mathbb{R}^{d_{p_1}}$.

(ii) For a fixed $\mathbf{x} \in \mathbb{R}^{p_1}$, $\mathbb{E}_{\bar{\xi}} \left\| \nabla f\left(\mathbf{x}, \mathbf{y}_1; \bar{\xi}\right) - \nabla f\left(\mathbf{x}, \mathbf{y}_2; \bar{\xi}\right) \right\|^2 \leq L_{conv}^2 \left\| \mathbf{y}_1 - \mathbf{y}_2 \right\|^2, \forall \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^{p_2}$.

We have

$$
\begin{aligned}
L_{conv}^2 :=& 2L_{f_x}^2 + +6C_{g_{xy}}^2 L_{f_y}^2 \left(\frac{K}{2\mu_g L_g - \mu_g^2}\right) + 6C_{f_y}^2 L_{g_{xy}}^2 \left(\frac{K}{2\mu_g L_g - \mu_g^2}\right) \\
&+ 6C_{g_{xy}}^2 C_{f_y}^2 \frac{K^2}{L_g^2} \max_{k(K)} \{k(K)^2 \left(1 - \frac{\mu_g}{L_g}\right)^{2(k(K)-1)}\} \frac{1}{L_g^2} L_{g_{yy}}^2. \qquad (14)
\end{aligned}
$$

In the above expressions, $L_{conv} \geq L_f$, L_f is the Lipschitz constant for our proposed stochastic gradient estimator and can be found in Lemma. 1.

B. Proof of Main results

Before diving in our theoretical analysis, we first introduce the following notations:

$$
\begin{aligned}
\overline{\mathbf{x}}_t &= \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_{i,t}, \mathbf{x}_t = \left[\mathbf{x}_{1,t}^\top, \cdots, \mathbf{x}_{m,t}^\top\right]^\top, \\
\mathbf{p}_t &= \left[\mathbf{p}_1(\mathbf{x}_{1,t}, \mathbf{y}_{1,t})^\top, \cdots, \mathbf{p}_m(\mathbf{x}_{m,t}, \mathbf{y}_{m,t})^\top\right]^\top,
\end{aligned}
$$

$$\mathbf{d}_t = \left[\mathbf{d}_1(\mathbf{x}_{1,t}, \mathbf{y}_{1,t})^\top, \cdots, \mathbf{d}_m(\mathbf{x}_{m,t}, \mathbf{y}_{m,t})^\top\right]^\top,$$
$$\bar{\mathbf{p}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{p}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \bar{\mathbf{d}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{d}_i(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \tag{15}$$

To prove Theorem 1, we structure our proof into the following key steps:

Step 1:

Lemma 5 (Descending Inequality for upper function)**.** Under the stated assumptions, the following descending inequality holds for Prometheus:

$$\begin{aligned}\ell(\bar{\mathbf{x}}_{t+1}) - \ell(\bar{\mathbf{x}}_t) \leq & (\frac{\alpha L_\ell}{2mr} + \frac{\alpha\tau}{2rm})\|\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + (\frac{\alpha r}{m} + \frac{\alpha\tau r}{2m} + \frac{\alpha^2 L_\ell}{2m} - \frac{\alpha\tau}{m})\|\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 \\ & + L^2\frac{3\alpha}{2rm}\sum_{i=1}^m \|\mathbf{y}_{i,t}^* - \mathbf{y}_{i,t}\|^2 + \frac{3\alpha}{2rm}\|\mathbf{u}_t - \mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2 - h\left(\bar{\mathbf{x}}_{t+1}\right) + h\left(\bar{\mathbf{x}}_t\right) \\ & + \frac{3\alpha}{2rm}\|\frac{1}{m}\sum_{i=1}^m \bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{u}}_t\|^2,\end{aligned} \tag{16}$$

where $\mathbf{y}_{i,t}^* = \arg\min_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y})$.

Proof.

$$\widetilde{\mathbf{x}}_{i,t} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}} \left\langle \mathbf{u}_{i,t}, \mathbf{x} - \mathbf{x}_{i,t}\right\rangle + \frac{\tau}{2}\left\|\mathbf{x} - \mathbf{x}_{i,t}\right\|^2 + h\left(\mathbf{x}\right), \tag{17}$$

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t + \alpha\left(\frac{1}{m}\sum_{i=1}^m \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right), \tag{18}$$

It follows form the optimal conditions of $h(\mathbf{x}_i)$ that

$$\begin{aligned}0 &\geq \left\langle \mathbf{u}_{i,t} + \tau\left(\widetilde{\mathbf{x}}_{i,t} - \mathbf{x}_{i,t}\right) + \partial h\left(\widetilde{\mathbf{x}}_{i,t}\right), \widetilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle \\ &\geq \left\langle \mathbf{u}_{i,t} + \tau\left(\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right), \widetilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle + \tau\left\|\widetilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 + \left\langle \partial h\left(\widetilde{\mathbf{x}}_{i,t}\right), \widetilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle \\ &\geq \left\langle \mathbf{u}_{i,t} + \tau\left(\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right), \widetilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle + \tau\left\|\widetilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 + h\left(\tilde{\mathbf{x}}_{i,t}\right) - h\left(\bar{\mathbf{x}}_t\right).\end{aligned} \tag{19}$$

From convexity, we have:

$$h\left(\bar{\mathbf{x}}_{t+1}\right) \leq (1-\alpha)h\left(\bar{\mathbf{x}}_t\right) + \alpha h\left(\frac{1}{m}\sum_{i=1}^m \tilde{\mathbf{x}}_{i,t}\right) \leq h\left(\bar{\mathbf{x}}_t\right) + \alpha\frac{1}{m}\sum_{i=1}^m \left(h\left(\tilde{\mathbf{x}}_{i,t}\right) - h\left(\bar{\mathbf{x}}_t\right)\right). \tag{20}$$

Therefore, it follows that

$$\alpha\frac{1}{m}\sum\left\langle \mathbf{u}_{i,t} + \tau\left(\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right), \widetilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle + \frac{\alpha\tau}{m}\left\|\tilde{\mathbf{x}}_{i,t} - 1\bar{\mathbf{x}}_t\right\|^2 + h\left(\bar{\mathbf{x}}_{t+1}\right) - h\left(\bar{\mathbf{x}}_t\right) \leq 0. \tag{21}$$

Then, we have

$$\begin{aligned}\ell(\bar{\mathbf{x}}_{t+1}) - \ell(\bar{\mathbf{x}}_t) &\overset{(a)}{\leq} \langle\nabla\ell(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\rangle + \frac{L_\ell}{2}\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2 \\ &\overset{(b)}{\leq} \left\langle \nabla\ell\left(\bar{\mathbf{x}}_t\right), \alpha\left(\frac{1}{m}\sum_{i\in m}\tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right)\right\rangle + \frac{\alpha^2 L_\ell}{2}\left\|\frac{1}{m}\sum_{i=1}^m \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 \\ &\leq \alpha\frac{1}{m}\sum_{i=1}^m \left\langle \nabla\ell\left(\bar{\mathbf{x}}_t\right), \tilde{\mathbf{x}}_{i,t} - \bar{\mathbf{x}}_t\right\rangle + \frac{\alpha^2 L_\ell}{2}\frac{1}{m}\left\|\tilde{\mathbf{x}}_t - 1\bar{\mathbf{x}}_t\right\|^2\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\leq}\alpha\frac{1}{m}\sum_{i=1}^{m}\langle\nabla\ell\left(\bar{\mathbf{x}}_t\right)-\mathbf{u}_{i,t}-\tau\left(\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right),\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\rangle+\frac{\alpha^2L_\ell}{2}\frac{1}{m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2\\
&\quad-\frac{\alpha\tau}{m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-h\left(\bar{\mathbf{x}}_{t+1}\right)+h\left(\bar{\mathbf{x}}_t\right)\\
&=\frac{\alpha}{m}\sum_{i=1}^{m}\langle\nabla\ell\left(\bar{\mathbf{x}}_t\right)-\mathbf{u}_{i,t},\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\rangle+\frac{\alpha\tau}{m}\sum_{i=1}^{m}\langle\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t,\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\rangle\\
&\quad+\frac{\alpha^2L_\ell}{2m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-\frac{\alpha\tau}{m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-h\left(\bar{\mathbf{x}}_{t+1}\right)+h\left(\bar{\mathbf{x}}_t\right)\\
&=\frac{\alpha}{m}\sum_{i=1}^{m}\left\langle\nabla\ell\left(\bar{\mathbf{x}}_t\right)-\frac{1}{m}\sum_{i_1=1}^{m}\nabla\ell(\mathbf{x}_{i_1,t}),\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\right\rangle\\
&\quad+\frac{\alpha}{m}\sum_{i=1}^{m}\left\langle\frac{1}{m}\sum_{i_1=1}^{m}\nabla\ell(\mathbf{x}_{i_1,t})-\mathbf{u}_{i,t},\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\right\rangle\\
&\quad+\frac{\alpha\tau}{m}\sum_{i=1}^{m}\langle\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t,\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\rangle+\frac{\alpha^2L_\ell}{2m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-\frac{\alpha\tau}{m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-h\left(\bar{\mathbf{x}}_{t+1}\right)+h\left(\bar{\mathbf{x}}_t\right)\\
&\overset{(d)}{\leq}\frac{\alpha}{m}\sum_{i=1}^{m}\frac{1}{2r}\|\nabla\ell\left(\bar{\mathbf{x}}_t\right)-\nabla\ell(\mathbf{x}_{i,t})\|^2+\frac{\alpha}{m}\sum_{i=1}^{m}\frac{r}{2}\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\|^2\\
&\quad+\frac{\alpha}{m}\sum_{i=1}^{m}\frac{1}{2r}\|\frac{1}{m}\sum_{i_1=1}^{m}\nabla\ell(\mathbf{x}_{i_1,t})-\mathbf{u}_{i,t}\|^2+\frac{\alpha}{m}\sum_{i=1}^{m}\frac{r}{2}\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\|^2+\frac{\alpha\tau}{m}\sum_{i=1}^{m}\frac{1}{2r}\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\|^2\\
&\quad+\frac{\alpha\tau}{m}\sum_{i=1}^{m}\frac{r}{2}\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\|^2+\frac{\alpha^2L_\ell}{2m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-\frac{\alpha\tau}{m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-h\left(\bar{\mathbf{x}}_{t+1}\right)+h\left(\bar{\mathbf{x}}_t\right)\\
&\overset{(e)}{\leq}\frac{\alpha}{m}\sum_{i=1}^{m}\frac{1}{2r}L_\ell\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2+\frac{\alpha}{m}\sum_{i=1}^{m}\frac{r}{2}\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\|^2+\frac{\alpha\tau}{m}\sum_{i=1}^{m}\frac{1}{2r}\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\|^2\\
&\quad+\frac{\alpha}{m}\sum_{i=1}^{m}\frac{3}{2r}\|\frac{1}{m}\sum_{i_1=1}^{m}\nabla\ell(\mathbf{x}_{i_1,t})-\frac{1}{m}\sum_{i_1=1}^{m}\nabla g(\mathbf{x}_{i_1,t},\mathbf{y}_{i_1,t})\|^2\\
&\quad+\frac{\alpha}{m}\sum_{i=1}^{m}\frac{3}{2r}(\|\frac{1}{m}\sum_{i_1=1}^{m}\bar{\nabla}f(\mathbf{x}_{i_1,t},\mathbf{y}_{i_1,t})-\bar{\mathbf{u}}_t\|^2+\|\bar{\mathbf{u}}_t-\mathbf{u}_{i,t}\|^2)+\frac{\alpha}{m}\sum_{i=1}^{m}\frac{r}{2}\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\|^2\\
&\quad+\frac{\alpha\tau}{m}\sum_{i=1}^{m}\frac{r}{2}\|\tilde{\mathbf{x}}_{i,t}-\bar{\mathbf{x}}_t\|^2+\frac{\alpha^2L_\ell}{2m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-\frac{\alpha\tau}{m}\left\|\tilde{\mathbf{x}}_t-1\bar{\mathbf{x}}_t\right\|^2-h\left(\bar{\mathbf{x}}_{t+1}\right)+h\left(\bar{\mathbf{x}}_t\right)\\
&\overset{(f)}{\leq}(\frac{\alpha L_\ell}{2mr}+\frac{\alpha\tau}{2rm})\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+(\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2L_\ell}{2m}-\frac{\alpha\tau}{m})\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\\
&\quad+L^2\frac{3\alpha}{2rm}\sum_{i=1}^{m}\|\mathbf{y}_{i,t}^*-\mathbf{y}_{i,t}\|^2+\frac{3\alpha}{2rm}\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2-h\left(\bar{\mathbf{x}}_{t+1}\right)+h\left(\bar{\mathbf{x}}_t\right)\\
&\quad+\frac{3\alpha}{2rm}\|\frac{1}{m}\sum_{i=1}^{m}\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\bar{\mathbf{u}}_t\|^2,
\end{aligned}
\tag{22}
$$

where (a) is because of Lipschitz continuous gradients of l, (b) is because of the updating rules. (c) is from 19. (d) and (e) are because of the triangle inequality. (f) is from the definition of $\mathbf{u}_{i,t},\ell_i(\mathbf{x}_{i,t})$.

□

Step 2:

Lemma 6 (Error Bound on $\mathbf{y}^*(\mathbf{x})$)**.** Under the stated Assumptions 1-3, letting $\alpha \leq \frac{1}{4L_f}$, we have

$$\begin{aligned}\|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t+1}\|^2 \leq & -\frac{\mu_g\beta}{4}\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 + \frac{9\beta}{2\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 \\ & - (1 + \frac{\mu_g\beta}{4})\frac{\beta^2}{2}\|\mathbf{v}_{i,t}\|^2 + \frac{5L_y^2}{\mu_g\beta}\|\mathbf{x}_{i,t} - \mathbf{x}_{i,t+1}\|^2, \end{aligned} \tag{23}$$

Proof.

$$\|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t}\|^2 = \|\mathbf{y}_{i,t} - \beta\mathbf{v}_{i,t} - \mathbf{y}^*_{i,t}\|^2 = \|\mathbf{y}_{i,t} - \mathbf{y}^*_{i,t}\|^2 + \beta^2\|\mathbf{v}_{i,t}\|^2 - 2\beta\langle \mathbf{y}_{i,t} - \mathbf{y}^*_{i,t}, \mathbf{v}_{i,t}\rangle. \tag{24}$$

Under the Assumption 1.(a), we have:

$$\begin{aligned} & g(\mathbf{x}_{i,t}, \mathbf{y}) - g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \frac{\mu_g}{2}\|\mathbf{y} - \mathbf{y}_{i,t}\|^2 \geq \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \mathbf{y} - \mathbf{y}_{i,t}\rangle \\ =& \langle \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\rangle \\ =& \langle \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\rangle \\ & - \frac{1}{4\beta}\|\mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\|^2 + \frac{1}{4\beta}\|\mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\|^2. \end{aligned} \tag{25}$$

With $\beta \leq 1/2L_g$, it follows that

$$\begin{aligned} & \frac{1}{4\beta}\|\mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\|^2 \geq \frac{L_g}{2}\|\mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\|^2 \\ & \geq g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t+1}) - g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}), \mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\rangle. \end{aligned} \tag{26}$$

Combining (25) and (26), with the update $\mathbf{y}_{i,t+1} - \mathbf{y}_{i,t} = -\beta\mathbf{v}_{i,t}$, we have:

$$\begin{aligned} & g(\mathbf{x}_{i,t}, \mathbf{y}) - g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t+1}) - \frac{\mu_g}{2}\|\mathbf{y} - \mathbf{y}_{i,t}\|^2 \\ \geq & \langle \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle - \frac{1}{4\beta}\|\mathbf{y}_{i,t+1} - \mathbf{y}_{i,t}\|^2 \\ = & \langle \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t}\rangle + \langle \mathbf{v}_{i,t}, \mathbf{y}_{i,t} - \mathbf{y}_{i,t+1}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle - \frac{\beta}{4}\|\mathbf{v}_{i,t}\|^2 \\ = & \langle \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t}\rangle + \beta\|\mathbf{v}_{i,t}\|^2 + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle - \frac{\beta}{4}\|\mathbf{v}_{i,t}\|^2 \\ = & \langle \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}, \mathbf{y} - \mathbf{y}_{i,t+1}\rangle + \frac{3\beta}{4}\|\mathbf{v}_{i,t}\|^2. \end{aligned} \tag{27}$$

Let $\mathbf{y} = \mathbf{y}^*_{i,t}$, we have

$$\begin{aligned} & g(\mathbf{x}_{i,t}, \mathbf{y}^*_{i,t}) - g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t+1}) - \frac{\mu_g}{2}\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 \\ \geq & \langle \mathbf{v}_{i,t}, \mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\rangle + \langle \nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}, \mathbf{y}^*_{i,t} - \mathbf{y}_{i,t+1}\rangle + \frac{3\beta}{4}\|\mathbf{v}_{i,t}\|^2 \\ \overset{(a)}{\geq} & \langle \mathbf{v}_{i,t}, \mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\rangle - \frac{2}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 - \frac{\mu_g}{8}\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t+1}\| + \frac{3\beta}{4}\|\mathbf{v}_{i,t}\|^2 \\ \overset{(b)}{\geq} & \langle \mathbf{v}_{i,t}, \mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\rangle - \frac{2}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 \\ & - \frac{\mu_g}{4}\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 - \frac{\mu_g}{4}\|\mathbf{y}_{i,t} - \mathbf{y}_{i,t+1}\|^2 + \frac{3\beta}{4}\|\mathbf{v}_{i,t}\|^2 \\ \overset{(c)}{=} & \langle \mathbf{v}_{i,t}, \mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\rangle - \frac{2}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 - \frac{\mu_g}{4}\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 + (\frac{3\beta}{4} - \frac{\mu_g\beta^2}{4})\|\mathbf{v}_{i,t}\|^2, \end{aligned} \tag{28}$$

where (a) follows from $-\langle \mathbf{x}, \mathbf{y}\rangle \leq \frac{1}{2c}\|\mathbf{x}\|^2 + \frac{c}{2}\|\mathbf{y}\|^2$ and $c = \frac{\mu}{4}$, (b) is due to $\|\mathbf{x}+\mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$, and (c) is from $\mathbf{y}_{i,t+1} - \mathbf{y}_{i,t} = -\beta \mathbf{v}_{i,t}$.

Since $g(\mathbf{x}_{i,t}, \mathbf{y}^*_{i,t}) \leq g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t+1})$ and mutiplying 2β on both sides of Eqs. 28, we have

$$-\frac{\mu_g\beta}{2}\|\mathbf{y}^*_{i,t}-\mathbf{y}_{i,t}\|^2 \geq 2\beta\langle \mathbf{v}_{i,t}, \mathbf{y}^*_{i,t}-\mathbf{y}_{i,t}\rangle - \frac{4\beta}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 + (\frac{3\beta^2}{2} - \frac{\mu_g\beta^3}{2})\|\mathbf{v}_{i,t}\|^2. \tag{29}$$

Then, we have

$$-2\beta\langle \mathbf{v}_{i,t}, \mathbf{y}_{i,t} - \mathbf{y}^*_{i,t}\rangle \leq -\frac{\mu_g\beta}{2}\|\mathbf{y}^*_{i,t}-\mathbf{y}_{i,t}\|^2 + \frac{4\beta}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 - (\frac{3\beta^2}{2} - \frac{\mu_g\beta^3}{2})\|\mathbf{v}_{i,t}\|^2. \tag{30}$$

Next, combining (24) and (30) and setting β, we have

$$\begin{aligned}
\|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t}\|^2 \leq& (1 - \frac{\mu_g\beta}{2})\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 + \frac{4\beta}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 \\
&+ (\frac{\beta^2}{2} - \frac{\mu_g\beta^3}{2})\|\mathbf{v}_{i,t}\|^2 \\
\leq& (1 - \frac{\mu_g\beta}{2})\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 + \frac{4\beta}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 + \frac{\beta^2}{2}\|\mathbf{v}_{i,t}\|^2.
\end{aligned} \tag{31}$$

Then, it holds that

$$\begin{aligned}
&\|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t+1}\|^2 = \|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t} + \mathbf{y}^*_{i,t} - \mathbf{y}^*_{i,t+1}\|^2 \\
\overset{(a)}{\leq}& (1 + \frac{\mu_g\beta}{4})\|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t}\|^2 + (1 + \frac{4}{\mu_g\beta})\|\mathbf{y}^*_{i,t} - \mathbf{y}^*_{i,t+1}\|^2 \\
\overset{(b)}{\leq}& (1 + \frac{\mu_g\beta}{4})\|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t}\|^2 + (1 + \frac{4}{\mu_g\beta})L_y^2\|\mathbf{x}_{i,t} - \mathbf{x}_{i,t+1}\|^2 \\
\overset{(c)}{\leq}& (1 + \frac{\mu_g\beta}{4})(1 - \frac{\mu_g\beta}{2})\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 + (1 + \frac{\mu_g\beta}{4})\frac{4\beta}{\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 \\
&- (1 + \frac{\mu_g\beta}{4})\frac{\beta^2}{2}\|\mathbf{v}_{i,t}\|^2 + (1 + \frac{4}{\mu_g\beta})L_y^2\|\mathbf{x}_{i,t} - \mathbf{x}_{i,t+1}\|^2 \\
\overset{(d)}{\leq}& (1 - \frac{\mu_g\beta}{4})\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 + \frac{9\beta}{2\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2 - (1 + \frac{\mu_g\beta}{4})\frac{\beta^2}{2}\|\mathbf{v}_{i,t}\|^2 \\
&+ \frac{5L_y^2}{\mu_g\beta}\|\mathbf{x}_{i,t} - \mathbf{x}_{i,t+1}\|^2,
\end{aligned} \tag{32}$$

where (a) follows from $\|\mathbf{x}+\mathbf{y}\|^2 \leq (1 + 1/c)\|\mathbf{x}\|^2 + (1 + c)\|\mathbf{y}\|^2$ and $c = \mu_g\beta/4$, (b) follows from Assumption 3, (c) follows from plugging (31), and (d) due to the facts that:

$$\begin{aligned}
&(1 + \frac{\mu_g\beta}{4})(1 - \frac{\mu_g\beta}{2}) = 1 + \frac{\mu_g\beta}{4} - \frac{\mu_g\beta}{2} - \frac{\mu_g^2\beta^2}{8} \leq 1 - \frac{\mu_g\beta}{4}, \\
&(1 + \frac{\mu_g\beta}{4})\frac{4\beta}{\mu_g} \leq (1 + \frac{\mu_g}{4}\cdot\frac{1}{2\mu_g})\frac{4\beta}{\mu_g} = \frac{9\beta}{2\mu_g}, \\
&1 + \frac{4}{\mu_g\beta} \leq \frac{1}{\mu_g\beta} + \frac{4}{\mu_g\beta} = \frac{5}{\mu_g\beta}\ .
\end{aligned} \tag{33}$$

Plugging (33) into (32) yields:

$$\begin{aligned}
&\|\mathbf{y}_{i,t+1} - \mathbf{y}^*_{i,t+1}\|^2 - \|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 \\
\leq& -\frac{\mu_g\beta}{4}\|\mathbf{y}^*_{i,t} - \mathbf{y}_{i,t}\|^2 + \frac{9\beta}{2\mu_g}\|\nabla_{\mathbf{y}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \mathbf{v}_{i,t}\|^2
\end{aligned}$$

$$- (1 + \frac{\mu_g \beta}{4}) \frac{\beta^2}{2} \|\mathbf{v}_{i,t}\|^2 + \frac{5L_y^2}{\mu_g \beta} \|\mathbf{x}_{i,t} - \mathbf{x}_{i,t+1}\|^2. \tag{34}$$

This completes the proof of the lemma. □

Step 3:

Lemma 7 (Iterates Contraction)**.** The following contraction properties of the iterates hold:

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{1} \otimes \bar{\mathbf{x}}_t\|^2 \leq & (1 + c_1)\lambda^2 \|\mathbf{x}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{x}}_{t-1}\|^2 + (1 + \frac{1}{c_1})\alpha^2 \|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2, \\
\|\mathbf{u}_t - \mathbf{1} \otimes \bar{\mathbf{u}}_t\|^2 \leq & (1 + c_2)\lambda^2 \|\mathbf{u}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{u}}_{t-1}\|^2 \\
& + (1 + \frac{1}{c_2}) \|\mathbf{p}_t - \mathbf{p}_{t-1}\|^2.
\end{aligned} \tag{35}$$

where c_1 and c_2 are arbitrary positive constants. Additionally, we have

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \leq & 8\|(\mathbf{x}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2 \|(\mathbf{x}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{x}}_{t-1})\|^2 \\
& + 4\alpha^2 \|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{x}}_{t-1})\|^2. \\
\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 \leq & \beta^2 \|\mathbf{v}_{t-1}\|^2,
\end{aligned} \tag{36}$$

Proof. Define $\widetilde{\mathbf{M}} = \mathbf{M} \otimes \mathbf{I}_m$. First for the iterates $\mathbf{x}_t$, we have the following contraction:

$$\|\widetilde{\mathbf{M}}\mathbf{x}_t - \mathbf{1} \otimes \bar{\mathbf{x}}_t\|^2 = \|\widetilde{\mathbf{M}}(\mathbf{x}_t - \mathbf{1} \otimes \bar{\mathbf{x}}_t)\|^2 \leq \lambda^2 \|\mathbf{x}_t - \mathbf{1} \otimes \bar{\mathbf{x}}_t\|^2, \tag{37}$$

This is because $\mathbf{x}_t - \mathbf{1} \otimes \mathbf{x}_t$ is orthogonal $\mathbf{1}$, which is the eigenvector corresponding to the largest eigenvalue of $\widetilde{\mathbf{M}}$, and $\lambda = \max\{|\lambda_2|, |\lambda_m|\}$. Hence,

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{1} \otimes \bar{\mathbf{x}}_t\|^2 &= \|\widetilde{\mathbf{M}}\mathbf{x}_{t-1} + \alpha(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}) - 1[\bar{\mathbf{x}}_{t-1} + \alpha(\frac{1}{m}\sum_{i=1}^m \tilde{\mathbf{x}}_i - \mathbf{x}_{t-1})]\|^2 \\
&\leq (1 + c_1)\lambda^2 \|\mathbf{x}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{x}}_{t-1}\|^2 + (1 + \frac{1}{c_1})\alpha^2 \|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2.
\end{aligned} \tag{38}$$

For $\mathbf{u}_t$, we have

$$\begin{aligned}
& \|\mathbf{u}_t - \mathbf{1} \otimes \bar{\mathbf{u}}_t\|^2 \\
=& \|\widetilde{\mathbf{M}}\mathbf{u}_{t-1} + \mathbf{p}_t - \mathbf{p}_{t-1} - \mathbf{1} \otimes \big(\bar{\mathbf{u}}_{t-1} + \bar{\mathbf{p}}(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{p}}(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1})\big)\|^2 \\
\leq & (1 + c_2)\lambda^2 \|\mathbf{u}_{t-1} - \mathbf{1} \otimes \mathbf{u}_{t-1}\|^2 + (1 + \frac{1}{c_2})\|\mathbf{p}_t - \mathbf{p}_{t-1} \\
& - \mathbf{1} \otimes \big(\bar{\mathbf{p}}(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}) - \bar{\mathbf{p}}(\mathbf{x}_{i,t-1}, \mathbf{y}_{i,t-1})\big)\|^2 \\
\leq & (1 + c_2)\lambda^2 \|\mathbf{u}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{u}}_{t-1}\|^2 \\
& + (1 + \frac{1}{c_2})\|\big(\mathbf{I} - \frac{1}{n}(\mathbf{1}\mathbf{1}^\top) \otimes \mathbf{I}\big)\big(\mathbf{p}_t - \mathbf{p}_{t-1}\big)\|^2 \\
\overset{(a)}{\leq} & (1 + c_2)\lambda^2 \|\mathbf{u}_{t-1} - \mathbf{1} \otimes \bar{\mathbf{u}}_{t-1}\|^2 + (1 + \frac{1}{c_2})\|\mathbf{p}_t - \mathbf{p}_{t-1}\|^2.
\end{aligned} \tag{39}$$

where (a) is due to $\|\mathbf{I} - \frac{1}{m}(\mathbf{1}\mathbf{1}^\top) \otimes \mathbf{I}\| \leq 1$.

According to the update, we have

$$\begin{aligned}
& \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 = \|\widetilde{\mathbf{M}}\mathbf{x}_{t-1} + \alpha(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}) - \mathbf{x}_{t-1}\|^2 \\
=& \|(\widetilde{\mathbf{M}} - \mathbf{I})\mathbf{x}_{t-1} + \alpha(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1})\|^2 \leq 2\|(\widetilde{\mathbf{M}} - \mathbf{I})\mathbf{x}_{t-1}\|^2 + 2\alpha^2 \|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2
\end{aligned}$$

$$
\begin{aligned}
=&2\|(\widetilde{\mathbf{M}}-\mathbf{I})(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+2\alpha^2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}_{t-1}\|^2\\
\leq&8\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+2\alpha^2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}_{t-1}\|^2\\
\leq&8\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+4\alpha^2\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+4\alpha^2\|(\tilde{\mathbf{x}}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2,
\end{aligned}
\tag{40}
$$

$$
\|\mathbf{y}_{i,t}-\mathbf{y}_{i,t-1}\|^2\leq\beta^2\|\mathbf{v}_{i,t-1}\|^2. \tag{41}
$$

□

Step 4: With the results from Step 1, we have

$$
\begin{aligned}
&\ell(\bar{\mathbf{x}}_{t+1})-\ell(\bar{\mathbf{x}}_t)\\
\leq&(\frac{\alpha L_\ell}{2mr}+\frac{\alpha\tau}{2rm})\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+(\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2L_\ell}{2m}-\frac{\alpha\tau}{m})\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\\
&+L^2\frac{3\alpha}{2rm}\sum_{i=1}^m\|\mathbf{y}_{i,t}^*-\mathbf{y}_{i,t}\|^2+\frac{3\alpha}{2rm}\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2-h\left(\bar{\mathbf{x}}_{t+1}\right)+h\left(\bar{\mathbf{x}}_t\right)\\
&+\frac{3\alpha}{2rm}\|\frac{1}{m}\sum_{i=1}^m\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\bar{\mathbf{u}}_t\|^2.
\end{aligned}
\tag{42}
$$

With the results from Step 2, we have

$$
\begin{aligned}
&\|\mathbf{y}_{i,t+1}-\mathbf{y}_{i,t+1}^*\|^2-\|\mathbf{y}_{i,t}-\mathbf{y}_{i,t}^*\|^2\\
&\leq-\frac{\mu_g\beta}{4}\|\mathbf{y}_{i,t}^*-\mathbf{y}_{i,t}\|^2+\frac{9\beta}{2\mu_g}\|\nabla_{\mathbf{y}}g(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbf{v}_{i,t}\|^2\\
&-(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2}\|\mathbf{v}_{i,t}\|^2+\frac{5L_y^2}{\mu_g\beta}\|\mathbf{x}_{i,t}-\mathbf{x}_{i,t+1}\|^2.
\end{aligned}
\tag{43}
$$

Combing (42) and (43) and telescoping the inequality, we have

$$
\begin{aligned}
&\ell(\bar{\mathbf{x}}_{T+1})-\ell(\bar{\mathbf{x}}_0)+h\left(\bar{\mathbf{x}}_{T+1}\right)-h\left(\bar{\mathbf{x}}_0\right)+\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\big[\|\mathbf{y}_{T+1}-\mathbf{y}_{T+1}^*\|^2-\|\mathbf{y}_0^*-\mathbf{y}_0\|^2\big]\\
\leq&(\frac{\alpha L_\ell}{2mr}+\frac{\alpha\tau}{2rm})\sum_{t=0}^T\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+(\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2L_\ell}{2m}-\frac{\alpha\tau}{m})\sum_{t=0}^T\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\\
&+L^2\frac{3\alpha}{2rm}\sum_{i=1}^m\sum_{t=0}^T\|\mathbf{y}_t^*-\mathbf{y}_t\|^2+\frac{3\alpha}{2rm}\sum_{t=0}^T\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2\\
&+\frac{3\alpha}{2rm}\sum_{t=0}^T\|\frac{1}{m}\sum_{i=1}^m\bar{\nabla}f(\mathbf{x}_t,\mathbf{y}_t)-\bar{\mathbf{u}}_t\|^2-\frac{\mu_g\beta^{3/2}}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sum_{t=0}^T\|\mathbf{y}_t^*-\mathbf{y}_t\|^2\\
&+\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}[\frac{9\beta}{2\mu_g}\sum_{t=0}^T\|\nabla_{\mathbf{y}}g(\mathbf{x}_t,\mathbf{y}_t)-\mathbf{v}_t\|^2-(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2}\sum_{t=0}^T\|\mathbf{v}_t\|^2\\
&+\frac{5L_y^2}{\mu_g\beta}\sum_{t=0}^T\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2]\\
\leq&(\frac{\alpha L_\ell}{2mr}+\frac{\alpha\tau}{2rm})\sum_{t=0}^T\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+(\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2L_\ell}{2m}-\frac{\alpha\tau}{m})\sum_{t=0}^T\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\\
&+L^2\frac{3\alpha}{2rm}\sum_{i=1}^m\sum_{t=0}^T\|\mathbf{y}_t^*-\mathbf{y}_t\|^2+\frac{3\alpha}{2rm}\sum_{t=0}^T\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2
\end{aligned}
$$

$$+\frac{3\alpha}{2rm}\sum_{t=0}^{T}\|\frac{1}{m}\sum_{i=1}^{m}\bar{\nabla}f(\mathbf{x}_t,\mathbf{y}_t)-\bar{\mathbf{u}}_t\|^2-\frac{\mu_g\beta^{3/2}}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sum_{t=0}^{T}\|\mathbf{y}_t^*-\mathbf{y}_t\|^2$$
$$+\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}[\frac{9\beta}{2\mu_g}\sum_{t=0}^{T}\|\nabla_{\mathbf{y}}g(\mathbf{x}_t,\mathbf{y}_t)-\mathbf{v}_t\|^2-(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2}\sum_{t=0}^{T}\|\mathbf{v}_t\|^2$$
$$+\frac{5L_y^2}{\mu_g\beta}\sum_{t=0}^{T}(8\|(\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2+4\alpha^2\|(\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2+4\alpha^2\|(\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2)], \tag{44}$$

where the last inequality follows from Eqs. (36).

Proof of Theorem 1

From the variance technique in Prometheus, we have

$$\begin{aligned}
&\mathbb{E}_t\|\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})\|^2\\
=&\mathbb{E}_t\|\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})+\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]\|^2\\
=&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]\|^2+\mathbb{E}_t\|\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]\|^2\\
\leq&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]\|^2+(\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2,
\end{aligned}\tag{45}$$

Moreover, with $t\in((n_t-1)\,q,n_tq-1]\cap\mathbb{Z}$, we have

$$\begin{aligned}
&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]\|^2\\
=&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})+\frac{1}{|\mathcal{S}|}\sum_{i=1}^{\mathcal{S}}\left[\bar{\nabla}g(\mathbf{x}_{i,k},\mathbf{y}_{i,k};\bar{\xi}_{i,t})-\bar{\nabla}g(\mathbf{x}_{i,k-1},\mathbf{y}_{i,k-1};\bar{\xi}_{i,t})\right]\\
&-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]+\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})]-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})]\|^2\\
=&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})]\|^2+\|\frac{1}{|\mathcal{S}|}\sum_{i=1}^{\mathcal{S}}\left[\bar{\nabla}g(\mathbf{x}_{i,k},\mathbf{y}_{i,k};\bar{\xi}_{i,t})\right.\\
&\left.-\bar{\nabla}g(\mathbf{x}_{i,k-1},\mathbf{y}_{i,k-1};\bar{\xi}_{i,t})\right]-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]+\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})]\|^2\\
\overset{(b}{\leq}&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})]\|^2\\
&+\frac{1}{|\mathcal{S}|}L_f^2\mathbb{E}_t(\|\mathbf{x}_{i,t}-\mathbf{x}_{i,t-1}\|^2+\|\mathbf{y}_{i,t}-\mathbf{y}_{i,t-1}\|^2),
\end{aligned}\tag{46}$$

where the last inequality use the mean variance theorem.

Telescoping over t from $((n_t-1)q+1$ to t, where $t\leq n_tq-1$, we obtain that

$$\begin{aligned}
&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})]\|^2\\
\leq&\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q})]\|^2\\
&+\frac{1}{|\mathcal{S}|}L_f^2\sum_{t=(n_t-1)q}^{t-1}\mathbb{E}_t(\|\mathbf{x}_{i,t}-\mathbf{x}_{i,t-1}\|^2+\|\mathbf{y}_{i,t}-\mathbf{y}_{i,t-1}\|^2
\end{aligned}\tag{47}$$

Next, for $\|\mathbf{p}_t-\mathbf{p}_{t-1}\|^2$, we have the following cases:

Case 1: $t\in((n_t-1)\,q,n_tq-1]\cap\mathbb{Z}$:

$$\mathbb{E}\|\mathbf{p}_t-\mathbf{p}_{t-1}\|^2=\sum_{i=1}^{m}\mathbb{E}\left\|\frac{1}{|\mathcal{S}_{i,t}|}\sum_{j\in\mathcal{S}_{i,t}}\nabla_{\mathbf{x}}f\left(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{j,t}\right)-\nabla_{\mathbf{x}}f\left(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1};\bar{\xi}_{j,t}\right)\right\|^2\tag{48}$$

$$\leq \frac{1}{|\mathcal{S}_{i,t}|^2}\sum_{i=1}^{m}\sum_{j\in\mathcal{S}_{i,t}}\mathbb{E}\left\|\nabla_{\mathbf{x}} f\left(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{j,t}\right)-\nabla_{\mathbf{x}} f\left(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1};\bar{\xi}_{j,t}\right)\right\|^2 \tag{49}$$

$$\leq L_f^2\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{x}_{i,t-1}-\mathbf{x}_{i,t}\right\|^2+L_f^2\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{y}_{i,t-1}-\mathbf{y}_{i,t}\right\|^2$$

$$\leq L_f^2(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\beta^2\|\mathbf{v}_{t-1}\|^2). \tag{50}$$

Case 2: $t=n_tq$:

$$\begin{aligned}
&\mathbb{E}\|\mathbf{p}_t-\mathbf{p}_{t-1}\|^2\\
=&\mathbb{E}\|\mathbf{p}_t-\mathbf{p}_{t-1}-\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})+\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})\\
&-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})+\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})\|^2\\
\leq&3\mathbb{E}\|\mathbf{p}_t-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})\|^2+3\mathbb{E}\|\mathbf{p}_{t-1}-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,t-1},\mathbf{y}_{i,t-1})\|^2\\
&+3L_f^2\mathbb{E}(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\beta^2\|\mathbf{v}_{t-1}\|^2)\\
\overset{(a)}{\leq}&3\mathbb{E}\left\|\mathbf{p}_{n_tq}-\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,n_tq},\mathbf{y}_{i,n_tq})\right\|^2+3L_f^2\beta^2\mathbb{E}\left\|\mathbf{v}_{n_tq-1}\right\|^2\\
&+3\mathbb{E}\left\|\mathbf{p}(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q}-\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q})\right\|^2\\
&+\sum_{r'=(n_t-1)q+1}^{n_tq-1}\frac{3L_f^2}{|\mathcal{S}|}\mathbb{E}\left(\|\mathbf{x}_{r'}-\mathbf{x}_{r'-1}\|^2+\|\mathbf{y}_{r'}-\mathbf{y}_{r'-1}\|^2\right)+3L_f^2\mathbb{E}\left\|\mathbf{x}_{n_tq-1}-\mathbf{x}_{n_tq}\right\|^2,
\end{aligned} \tag{51}$$

where (a) is from (47) and set $t=n_tq$.

Telescoping from $r=(n_t-1)\,q+1$ to n_tq and set $|\mathcal{S}|=q$, we have

$$\begin{aligned}
&\sum_{r=(n_t-1)q+1}^{n_tq}\mathbb{E}\|\mathbf{p}_r-\mathbf{p}_{r-1}\|^2\\
\leq&3(q+1)\mathbb{E}\left\|\mathbf{p}_{n_tq}-\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,n_tq},\mathbf{y}_{i,n_tq})\right\|^2\\
&+3(q+1)\mathbb{E}\left\|\mathbf{p}_{(n_t-1)q}-\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q})\right\|^2\\
&+\sum_{r=(n_t-1)q+1}^{n_tq}\frac{4L_f^2}{q}\mathbb{E}\left(\|\mathbf{x}_r-\mathbf{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\right)\\
=&\sum_{r=(n_t-1)q+1}^{n_tq}\frac{4L_f^2}{q}\mathbb{E}\left(\|\mathbf{x}_r-\mathbf{x}_{r-1}\|^2+\|\mathbf{y}_r-\mathbf{y}_{r-1}\|^2\right).
\end{aligned} \tag{52}$$

Since $\mathbb{E}\left\|\mathbf{p}_{n_tq}-\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,n_tq},\mathbf{y}_{i,n_tq})\right\|^2=\mathbb{E}\left\|\mathbf{p}_{(n_t-1)q}-\mathbb{E}_{\bar{\xi}_{i,t}}\mathbf{p}_i(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q})\right\|^2=0$, and with eqs.(48),we have

$$\sum_{t=1}^{T}\|\mathbf{p}_t-\mathbf{p}_{t-1}\|^2\leq\sum_{t=1}^{T}\left[4L_f^2\mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}_{t-1}\right\|^2+4L_f^2\left\|\mathbf{y}_t-\mathbf{y}_{t-1}\right\|^2\right] \tag{53}$$

Since $\mathbb{E}_t\|\mathbf{p}_i(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q})-\mathbb{E}_{\bar{\xi}_{i,t}}[\mathbf{p}_i(\mathbf{x}_{i,(n_t-1)q},\mathbf{y}_{i,(n_t-1)q})]\|^2=0$, $|\mathcal{S}|=q$, we can conclude that

$$\sum_{t=0}^{T}\|\frac{1}{m}\sum_{i=1}^{m}\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\bar{\mathbf{u}}_t\|^2$$

$$
\begin{aligned}
=&\sum_{t=0}^{T}\|\frac{1}{m}\sum_{i=1}^{m}\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\bar{\mathbf{p}}_t\|^2=\sum_{t=0}^{T}\|\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})\|^2\\
\leq& L_f^2\sum_{t=0}^{T}(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2)+(\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2\cdot T\\
\leq& L_f^2\sum_{t=0}^{T}(8\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+4\alpha^2\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2\\
&+4\alpha^2\|(\tilde{\mathbf{x}}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+\beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2)+(\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2\cdot T,
\end{aligned}\tag{54}
$$

where the last inequality follows from (36).

Similarly, we have

$$
\begin{aligned}
&\sum_{t=0}^{T}\|\frac{1}{m}\sum_{i=1}^{m}\nabla g(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbf{v}_{i,t}\|^2\\
\leq& L_f^2\sum_{t=0}^{T}(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2)\\
\leq& L_f^2\sum_{t=0}^{T}(8\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+4\alpha^2\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2\\
&+4\alpha^2\|(\tilde{\mathbf{x}}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+\beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2).
\end{aligned}\tag{55}
$$

Besides, with the results from Step 3, the update rule of $\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})$ and (53) , we have

$$
\begin{aligned}
&\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\leq(1+c_1)\lambda^2\|\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1}\|^2+(1+\frac{1}{c_1})\alpha^2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}_{t-1}\|^2,\\
&\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2-\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2\leq((1+c_2)\lambda^2-1)\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2\\
&\quad+(1+\frac{1}{c_2})4L_f^2(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2)\\
&\overset{(a)}{\leq}((1+c_2)\lambda^2-1)\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2+(1+\frac{1}{c_2})4L_f^2(8\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2\\
&\quad+4\alpha^2\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+4\alpha^2\|(\tilde{\mathbf{x}}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+\beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2),
\end{aligned}\tag{56}
$$

where (a) follows from eqs.(40).

Then, we have

$$
\begin{aligned}
&\|\mathbf{x}_{T+1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{T+1}\|^2-\|\mathbf{x}_0-\mathbf{1}\otimes\bar{\mathbf{x}}_0\|^2\\
\leq&((1+c_1)\lambda^2-1)\sum_{t=1}^{T+1}\|\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1}\|^2+(1+\frac{1}{c_1})\alpha^2\sum_{t=1}^{T+1}\|\tilde{\mathbf{x}}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1}\|^2.
\end{aligned}\tag{57}
$$

$$
\begin{aligned}
&\|\mathbf{u}_{T+1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{T+1}\|^2-\|\mathbf{u}_0-\mathbf{1}\otimes\bar{\mathbf{u}}_0\|^2\\
\leq&((1+c_2)\lambda^2-1)\sum_{t=1}^{T+1}\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2+(1+\frac{1}{c_2})4L_f^2\sum_{t=1}^{T+1}(8\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2
\end{aligned}
$$

$$+ 4\alpha^2\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\|\mathbf{v}_{t-1}\|^2). \tag{58}$$

Combing (57), (58), and (44), we have

$$\begin{aligned}
&\mathbb{E}\{\ell(\bar{\mathbf{x}}_{T+1}) - \ell(\bar{\mathbf{x}}_0) + h(\bar{\mathbf{x}}_{T+1}) - h(\bar{\mathbf{x}}_0) + \frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\big[\|\mathbf{y}_{T+1} - \mathbf{y}^*_{T+1}\|^2 - \|\mathbf{y}^*_0 - \mathbf{y}_0\|^2\big] \\
&+ \frac{1}{\sqrt{\beta}}[\|\mathbf{x}_{T+1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{T+1}\|^2 - \|\mathbf{x}_0 - \mathbf{1}\otimes\bar{\mathbf{x}}_0\|^2] + \beta[\|\mathbf{u}_{T+1} - \mathbf{1}\otimes\bar{\mathbf{u}}_{T+1}\|^2 - \|\mathbf{u}_0 - \mathbf{1}\otimes\bar{\mathbf{u}}_0\|^2]\} \\
\le&(\frac{\alpha L_\ell}{2mr} + \frac{\alpha\tau}{2rm})\sum_{t=0}^{T}\|\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + (\frac{\alpha r}{m} + \frac{\alpha\tau r}{2m} + \frac{\alpha^2 L_\ell}{2m} - \frac{\alpha\tau}{m})\sum_{t=0}^{T}\|\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 \\
&+ L^2\frac{3\alpha}{2rm}\sum_{i=1}^{m}\sum_{t=0}^{T}\|\mathbf{y}^*_t - \mathbf{y}_t\|^2 + \frac{3\alpha}{2rm}\sum_{t=0}^{T}\|\frac{1}{m}\sum_{i=1}^{m}\bar{\nabla}f(\mathbf{x}_t,\mathbf{y}_t) - \bar{\mathbf{u}}_t\|^2 \\
&+ \frac{3\alpha}{2rm}\sum_{t=0}^{T}\|\frac{1}{m}\sum_{i=1}^{m}\nabla g(\mathbf{x}_t,\mathbf{y}_t) - \mathbf{v}_t\|^2 - \frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\sum_{t=0}^{T}\|\mathbf{y}^*_t - \mathbf{y}_t\|^2 \\
&+ \frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}[\frac{9\beta}{2\mu}\sum_{t=0}^{T}(L_f^2\sum_{t=0}^{T}(8\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 \\
&+ 4\alpha^2\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2)) - (1 + \frac{\mu_g\beta}{4})\frac{\beta^2}{4}\sum_{t=0}^{T}\|\mathbf{v}_t\|^2 \\
&+ \frac{5L_y^2}{\mu_g\beta}\sum_{t=0}^{T}(8\|(\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2 + 4\alpha^2\|(\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2 + 4\alpha^2\|(\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2)] \\
&+ \frac{1}{\sqrt{\beta}}((1+c_1)\lambda^2 - 1)\sum_{t=1}^{T+1}\mathbb{E}\|\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1}\|^2 + \frac{1}{\sqrt{\beta}}(1 + \frac{1}{c_1})\alpha^2\sum_{t=1}^{T+1}\mathbb{E}\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2 \\
&+ \beta((1+c_2)\lambda^2 - 1)\sum_{t=1}^{T+1}\mathbb{E}\|\mathbf{u}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2 \\
&+ (1 + \frac{1}{c_2})\beta L_f^2\sum_{t=1}^{T+1}(8\mathbb{E}\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 \\
&+ 4\alpha^2\mathbb{E}\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\mathbb{E}\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\mathbb{E}\|\mathbf{v}_{t-1}\|^2) \\
\overset{(a)}{\le}&(\frac{\alpha L_\ell}{2mr} + \frac{\alpha\tau}{2rm})\sum_{t=0}^{T}\|\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + (\frac{\alpha r}{m} + \frac{\alpha\tau r}{2m} + \frac{\alpha^2 L_\ell}{2m} - \frac{\alpha\tau}{m})\sum_{t=0}^{T}\|\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 \\
&+ L^2\frac{3\alpha}{2rm}\sum_{i=1}^{m}\sum_{t=0}^{T}\|\mathbf{y}^*_t - \mathbf{y}_t\|^2 + \frac{3\alpha}{2rm}[L_f^2\sum_{t=0}^{T}(8\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 \\
&+ 4\alpha^2\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2) + (\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1 - \frac{\mu_g}{L_g}\right)^K)^2\cdot T] \\
&+ \frac{3\alpha}{2rm}\sum_{t=0}^{T}[L_f^2\sum_{t=0}^{T}(8\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 \\
&+ 4\alpha^2\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2)] - \frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\sum_{t=0}^{T}\|\mathbf{y}^*_t - \mathbf{y}_t\|^2 \\
&+ \frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}[\frac{9\beta}{2\mu}\sum_{t=0}^{T}(L_f^2\sum_{t=0}^{T}(8\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ 4\alpha^2\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2)) - (1+\frac{\mu_g\beta}{4})\frac{\beta^2}{4}\sum_{t=0}^{T}\|\mathbf{v}_t\|^2\\
&+ \frac{5L_y^2}{\mu_g\beta}\sum_{t=0}^{T}(8\|(\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2 + 4\alpha^2\|(\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2 + 4\alpha^2\|(\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2)]\\
&+ \frac{1}{\sqrt{\beta}}((1+c_1)\lambda^2 - 1)\sum_{t=1}^{T+1}\mathbb{E}\|\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1}\|^2 + \frac{1}{\sqrt{\beta}}(1+\frac{1}{c_1})\alpha^2\sum_{t=1}^{T+1}\mathbb{E}\|\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\|^2\\
&+ \beta((1+c_2)\lambda^2 - 1)\sum_{t=1}^{T+1}\mathbb{E}\|\mathbf{u}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2\\
&+ (1+\frac{1}{c_2})\beta L_f^2\sum_{t=1}^{T+1}(8\mathbb{E}\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2\\
&+ 4\alpha^2\mathbb{E}\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\mathbb{E}\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\mathbb{E}\|\mathbf{v}_{t-1}\|^2),
\end{aligned}
\tag{59}
$$

where (a) follows from (54) and (55). Next, choosing $c_1 = c_2 = \frac{1}{\lambda} - 1$, we have

$$
\begin{aligned}
&\ell(\bar{\mathbf{x}}_{T+1}) - \ell(\bar{\mathbf{x}}_0) + h(\bar{\mathbf{x}}_{T+1}) - h(\bar{\mathbf{x}}_0) + \frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\big[\|\bar{\mathbf{y}}_{T+1} - \mathbf{y}^*_{T+1}\|^2 - \|\mathbf{y}^*_0 - \bar{\mathbf{y}}_0\|^2\big]\\
&+ \frac{1}{\sqrt{\beta}}[\|\mathbf{x}_{T+1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{T+1}\|^2 - \|\mathbf{x}_0 - \mathbf{1}\otimes\bar{\mathbf{x}}_0\|^2] + \beta[\|\mathbf{u}_{T+1} - \mathbf{1}\otimes\bar{\mathbf{u}}_{T+1}\|^2 - \|\mathbf{u}_0 - \mathbf{1}\otimes\bar{\mathbf{u}}_0\|^2]\\
\le& C_1'\sum_{t=0}^{T}\|\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + C_2'\sum_{t=0}^{T}\|\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + C_3'\sum_{t=0}^{T}\|\mathbf{y}_t - \mathbf{y}_t^*\|^2\\
&+ C_4'\sum_{t=0}^{T}\|\mathbf{u}_t - \mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2 + C_5'\sum_{t=0}^{T}\|\mathbf{v}_t\|^2\\
&+ \frac{3\alpha}{2rm}(\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2\cdot T,
\end{aligned}
\tag{60}
$$

where the constants are

$$
\begin{aligned}
C_1' =& (\frac{\alpha L_\ell}{2mr} + \frac{\alpha\tau}{2rm}) + (8+4\alpha^2)[\frac{5L_y^2}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta} + \frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2\mu}\\
&+ \frac{3\alpha}{2rm}L_f^2 + \frac{1}{1-\lambda}L_f^2\beta] + (\lambda-1)\frac{1}{\sqrt{\beta}},
\end{aligned}
\tag{61}
$$

$$
\begin{aligned}
C_2' =& (\frac{\alpha r}{m} + \frac{\alpha\tau r}{2m} + \frac{\alpha^2 L_\ell}{2m} - \frac{\alpha\tau}{m}) + 4\alpha^2[\frac{5L_y^2}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta} + \frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2\mu}\\
&+ \frac{3\alpha}{2rm}L_f^2 + \frac{1}{1-\lambda}L_f^2\beta] + \frac{1}{1-\lambda}\alpha^2\frac{1}{\sqrt{\beta}},
\end{aligned}
\tag{62}
$$

$$
C_3' = L^2\frac{3\alpha}{2rm} - \frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta},
\tag{63}
$$

$$
C_4' = \frac{3\alpha}{2rm} + (\lambda-1)\beta,
\tag{64}
$$

$$
\begin{aligned}
C_5' =& -\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2} + \beta^3\frac{1}{1-\lambda}L_f^2\\
&+ \beta^2(\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2\mu} + \frac{3\alpha}{2rm}L_f^2).
\end{aligned}
\tag{65}
$$

To ensure $C_1' \leq \frac{1-\lambda}{4}$, we have

$$\begin{aligned}
C_1' =&(\frac{\alpha L_\ell}{2mr}+\frac{\alpha\tau}{2rm})+(8+4\alpha^2)[\frac{5L_y^2}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}+\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2\mu}\\
&+\frac{3\alpha}{2rm}L_f^2+\frac{1}{1-\lambda}L_f^2\beta]+(\lambda-1)\frac{1}{\sqrt{\beta}}\\
\leq&(\frac{\alpha L_\ell}{2mr}+\frac{\alpha\tau}{2rm})+(8+4\alpha^2)[\frac{5L_y^2}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}+\frac{\mu_g(1-\lambda)}{160L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2\mu}\\
&+\frac{3\alpha}{2rm}L_f^2+\frac{1}{1-\lambda}L_f^2\beta]+(\lambda-1)\frac{1}{\sqrt{\beta}}\\
\leq&[\frac{1-\lambda}{4\sqrt{\beta}}+(\frac{1-\lambda}{4\sqrt{\beta}}+\frac{1-\lambda}{8\sqrt{\beta}}+\frac{1-\lambda}{16\sqrt{\beta}}+\frac{1-\lambda}{16\beta})+(\lambda-1)\frac{1}{\sqrt{\beta}}]\\
\leq&[\frac{1-\lambda}{4\sqrt{\beta}}+\frac{1-\lambda}{4\sqrt{\beta}}+\frac{1-\lambda}{8\sqrt{\beta}}+\frac{1-\lambda}{16\sqrt{\beta}}+\frac{1-\lambda}{16\sqrt{\beta}}+(\lambda-1)\frac{1}{\sqrt{\beta}}]=-\frac{1-\lambda}{4}\frac{1}{\sqrt{\beta}},
\end{aligned}\tag{66}$$

where $\alpha \leq \min\{\frac{(1-\lambda)m}{2\sqrt{\beta}(L_\ell+\tau)}\frac{\tau}{6+3\tau}, \frac{(1-\lambda)m}{8\sqrt{\beta}L_f^2}\frac{\tau}{6+3\tau}\}, \beta \leq \min\{\frac{\sqrt{40}L_y}{3L_f}, \frac{1-\lambda}{16L_f}\}, r=\frac{\tau}{6+3\tau}$.

To ensure $C_2' \leq 0$, we have

$$\begin{aligned}
C_2' =& (\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2 L_\ell}{2m}-\frac{\alpha\tau}{m})+4\alpha^2[\frac{5L_y^2}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}+\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2\mu}\\
&+\frac{3\alpha}{2rm}L_f^2+\frac{1}{1-\lambda}L_f^2\beta]+\frac{1}{1-\lambda}\alpha^2\frac{1}{\sqrt{\beta}}\\
\leq& (\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2 L_\ell}{2m}-\frac{\alpha\tau}{m})+4\alpha^2[\frac{1}{\beta}\frac{(1-\lambda)}{32}\sqrt{\beta}+\frac{(1-\lambda)}{160L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2}\\
&+\frac{3\alpha}{2rm}L_f^2+\frac{1}{1-\lambda}L_f^2\beta]+\frac{1}{1-\lambda}\alpha^2\frac{1}{\sqrt{\beta}}\\
\leq& (\frac{\alpha\tau}{12m}+\frac{\alpha\tau}{12m})+\frac{\alpha\tau}{6m}-\frac{\alpha\tau}{m}+\frac{\alpha\tau}{12m}+\frac{\alpha\tau}{12m}+\frac{\alpha\tau}{12m}+\frac{\alpha\tau}{6m}+\frac{\alpha\tau}{12m}=-\frac{\alpha\tau}{6m},
\end{aligned}\tag{67}$$

where $\alpha \leq \min\{\frac{\tau}{3L_\ell}, \frac{8\sqrt{\beta}\tau}{12m(1-\lambda)}, \frac{20L_y^2\tau}{27(1-\lambda)\beta^{1.5}L_f^2 m}, \frac{\tau(1-\lambda)}{24mL_f^2\beta}, \frac{\tau\sqrt{\beta}(1-\lambda)}{12m}\}, r=\frac{\tau}{6+3\tau}$.

To ensure $C_3' \leq 0$, we have

$$\begin{aligned}
C_3 =& L^2\frac{3\alpha}{2rm}-\frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\\
\leq& \frac{\mu_g^2\beta}{8}\frac{(1-\lambda)}{240L_y^2}\sqrt{\beta}-\frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\\
\leq& \frac{\mu_g\beta}{8}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}-\frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\\
\leq& -\frac{\mu_g\beta}{8}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\leq-\frac{\beta^{1.5}}{8}\frac{\mu_g^2(1-\lambda)}{240L_y^2},
\end{aligned}\tag{68}$$

where $\alpha \leq \frac{\mu_g^2\beta^{1.5}}{8}\frac{(1-\lambda)}{2880L_y^2L^2}$.

To ensure $C_4' \leq 0$, we have

$$C_4' = \frac{3\alpha}{2rm}+(\lambda-1)\beta \leq 0,\tag{69}$$

where $\alpha \leq (1-\lambda)\beta\frac{2m}{3}\frac{\tau}{6+3\tau}, r=\frac{\tau}{6+3\tau}$.

To ensure $C_5' \leq 0$, we have

$$\begin{aligned}
C_5' =& -\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2}+\beta^3\frac{1}{1-\lambda}L_f^2+\beta^2(\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2\mu}+\frac{3\alpha}{2rm}L_f^2)\\
\leq& -\frac{\mu_g(1-\lambda)}{240L_y^2}\frac{\beta^{2.5}}{2}+\beta^3\frac{1}{1-\lambda}L_f^2+\beta^2(\frac{(1-\lambda)}{160L_y^2}\sqrt{\beta}L_f^2\frac{9\beta}{2}+\frac{3\alpha}{2rm}L_f^2)\\
\leq& -\frac{\mu_g(1-\lambda)}{240L_y^2}\frac{\beta^{2.5}}{2}+\frac{\mu_g(1-\lambda)}{240L_y^2}\frac{\beta^{2.5}}{6}+\frac{\mu_g(1-\lambda)}{240L_y^2}\frac{\beta^{2.5}}{6}+\frac{\mu_g(1-\lambda)}{240L_y^2}\frac{\beta^{2.5}}{6}\leq 0,
\end{aligned} \tag{70}$$

where $\beta \leq \min\{(\frac{\mu_g(1-\lambda)^2}{1440L_y^2L_f^2})^2, \frac{2\mu_g}{81L_f^2}\}, \alpha \leq \frac{\mu_g(1-\lambda)}{240L_y^2}\frac{\beta^{2.5}}{9L_f^2}m\frac{\tau}{6+3\tau}, r=\frac{\tau}{6+3\tau}$.

With the above conditions, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T}\left(\mathbb{E}\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+\mathbb{E}\|\mathbf{y}_t-\mathbf{y}_t^*\|^2\right)\\
&\leq \frac{\mathbb{E}\left[\mathfrak{p}_0-\mathfrak{p}^*\right]}{T\min\left\{\frac{1-\lambda}{4}\frac{1}{\sqrt{\beta}},\frac{\alpha\tau}{6m},\frac{\beta^{1.5}}{8}\frac{\mu_g^2(1-\lambda)}{240L_y^2}\right\}}+\bar{C}_\sigma' = \mathcal{O}(1/T),
\end{aligned} \tag{71}$$

where $\mathfrak{p}_t = \ell(\bar{\mathbf{x}}_t)+h(\bar{\mathbf{x}}_t)+\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2+\beta\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2+\frac{1}{\sqrt{\beta}}\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2), \bar{C}_\sigma' = \frac{\frac{3\alpha}{2rm}C_{g_{xy}}^2C_{f_y^2}\left(1-\frac{\mu_g}{L_g}\right)^{2K}}{\min\left\{\frac{1-\lambda}{4}\frac{1}{\sqrt{\beta}},\frac{\alpha\tau}{6m},\frac{\beta^{1.5}}{8}\frac{\mu_g^2(1-\lambda)}{240L_y^2}\right\}}$.

The exact expressions of the constants shown in Theorem 1 are: $C_{1,1}=\frac{(1-\lambda)m}{2\sqrt{\beta}(L_\ell+\tau)}\frac{\tau}{6+3\tau}, C_{1,2}=\frac{(1-\lambda)m}{8\sqrt{\beta}L_f^2}\frac{\tau}{6+3\tau}, C_{1,3}=\frac{\tau}{3L_\ell}, C_{1,4}=\frac{(1-\lambda)\mu_g^2\beta^{1.5}}{23040L_y^2L^2}, C_{1,5}=\frac{8\sqrt{\beta}\tau}{12m(1-\lambda)}, C_{1,6}=\frac{20L_y^2\tau}{27(1-\lambda)\beta^{1.5}L_f^2m}, C_{1,6}=\frac{\tau(1-\lambda)}{24mL_f^2\beta}, C_{1,7}=(1-\lambda)\beta\frac{2m}{3}\frac{\tau}{6+3\tau}, C_{1,8}=\frac{\mu_g(1-\lambda)}{240L_y^2}\frac{\beta^{2.5}}{9L_f^2}m\frac{\tau}{6+3\tau}, C_{1,9}=\frac{\tau\sqrt{\beta}(1-\lambda)}{12m}, C_{2,1}=\frac{\sqrt{40}L_y}{3L_f}, C_{2,2}=\frac{1-\lambda}{16L_f}, C_{2,3}=(\frac{\mu_g(1-\lambda)^2}{1440L_y^2L_f^2})^2, C_{2,4}=\frac{2\mu_g}{81L_f^2}$.

We would like to note that the term $\bar{C}_\sigma'$ decays exponentially fast with respect to K. To show the sample complexity, we note that the number of sample complexity per agent in the outer loops can be calculated as: $\lceil\frac{T}{q}\rceil\cdot n$. Also, the number of samples using in the inner loop can be calculated as TS. Hence, the total sample complexity can be calculated as:

$$\lceil\frac{T}{q}\rceil n+T\cdot S\leq\frac{T+q}{q}n+T\cdot K\sqrt{n}=T\sqrt{n}+n+T\cdot K\sqrt{n}=O(\sqrt{n}K\epsilon^{-1}+n).$$

Thus, the overall SFO complexity is $\mathcal{O}(\sqrt{n}K\epsilon^{-1}+n)$. This completes the proof.

Proof of Proposition 3

Based on stochastic gradient estimator, we have

$$\begin{aligned}
&\mathbb{E}_\xi\|\frac{1}{m}\sum_{i=1}^{m}\nabla f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\bar{\mathbf{u}}_t\|^2=\mathbb{E}_\xi\|\frac{1}{m}\sum_{i=1}^{m}\nabla f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\frac{1}{m}\sum_{i=1}^{m}\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})\|^2\\
=&\mathbb{E}_\xi\|\frac{1}{m}\sum_{i=1}^{m}\nabla f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\frac{1}{m}\sum_{i=1}^{m}\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})\|^2\\
=&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_\xi\|\nabla f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})\|^2\\
=&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_\xi\|\nabla f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbb{E}_\xi[\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})]+\mathbb{E}_\xi[\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})]-\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})\|^2\\
\overset{(a)}{\leq}&\frac{2}{m}\sum_{i=1}^{m}\mathbb{E}_\xi\|\nabla f(\mathbf{x}_{i,t},\mathbf{y}_{i,t})-\mathbb{E}_\xi[\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})]\|^2
\end{aligned}$$

$$+\frac{2}{m}\sum_{i=1}^{m}\|\mathbb{E}_{\xi}[\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})]-\bar{\nabla}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\bar{\xi}_{i0})\|^2$$
$$\overset{(b)}{\leq}2(\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2+2\sigma_f^2, \tag{72}$$

where (a) follows from the triangle inequality and (b) is from Assumption 3(b) and Lemma 3.

Besides, with the results from Step 3 and the update rule of $\mathbf{p}_i(\mathbf{x}_{i,t},\mathbf{y}_{i,t})$, we have

$$\begin{aligned}
&\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\leq(1+c_1)\lambda^2\|\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1}\|^2+(1+\frac{1}{c_1})\alpha^2\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}_{t-1}\|^2,\\
&\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2-\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2\leq((1+c_2)\lambda^2-1)\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2\\
&\quad+(1+\frac{1}{c_2})L_f^2(\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2)\\
&\overset{(a)}{\leq}((1+c_2)\lambda^2-1)\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2+(1+\frac{1}{c_2})L_f^2(8\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2\\
&\quad+4\alpha^2\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+4\alpha^2\|(\tilde{\mathbf{x}}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2+\beta^2\sum_{i=1}^{m}\|\mathbf{v}_{i,t-1}\|^2),
\end{aligned}\tag{73}$$

where (a) follows from Eq. (40).

Additionally, we have

$$\mathbb{E}_{\zeta}\|\nabla g(\mathbf{x}_i,\mathbf{y}_i)-\mathbf{v}_{i,t}\|^2=\|\nabla g(\mathbf{x}_i,\mathbf{y}_i)-g(\mathbf{x}_i,\mathbf{y}_i;\zeta_{i0})\|^2=\sigma_g^2. \tag{74}$$

Thus, combing (44)-(73), we can conclude that

$$\begin{aligned}
&\mathbb{E}\{\ell(\bar{\mathbf{x}}_{T+1})-\ell(\bar{\mathbf{x}}_0)+h(\bar{\mathbf{x}}_{T+1})-h(\bar{\mathbf{x}}_0)+\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}\big[\|\mathbf{y}_{T+1}-\mathbf{y}_{T+1}^*\|^2-\|\mathbf{y}_0^*-\mathbf{y}_0\|^2\big]\\
&\quad+\frac{1}{\sqrt{\beta}}[\|\mathbf{x}_{T+1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{T+1}\|^2-\|\mathbf{x}_0-\mathbf{1}\otimes\bar{\mathbf{x}}_0\|^2]+\beta[\|\mathbf{u}_{T+1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{T+1}\|^2-\|\mathbf{u}_0-\mathbf{1}\otimes\bar{\mathbf{u}}_0\|^2]\}\\
&\leq(\frac{\alpha L_\ell}{2mr}+\frac{\alpha\tau}{2rm})\sum_{t=0}^{T}\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+(\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2L_\ell}{2m}-\frac{\alpha\tau}{m})\sum_{t=0}^{T}\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\\
&\quad+L^2\frac{3\alpha}{2rm}\sum_{i=1}^{m}\sum_{t=0}^{T}\|\mathbf{y}_t^*-\mathbf{y}_t\|^2+\frac{3\alpha}{2rm}\sum_{t=0}^{T}\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2\\
&\quad+\frac{3\alpha}{2rm}\sum_{t=0}^{T}\|\frac{1}{m}\sum_{i=1}^{m}\bar{\nabla}f(\mathbf{x}_t,\mathbf{y}_t)-\bar{\mathbf{u}}_t\|^2-\frac{\mu_g\beta^{3/2}}{4}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sum_{t=0}^{T}\|\mathbf{y}_t^*-\mathbf{y}_t\|^2\\
&\quad+\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}[\frac{9\beta}{2\mu_g}\sum_{t=0}^{T}\|\nabla_{\mathbf{y}}g(\mathbf{x}_t,\mathbf{y}_t)-\mathbf{v}_t\|^2-(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2}\sum_{t=0}^{T}\|\mathbf{v}_t\|^2\\
&\quad+\frac{5L_y^2}{\mu_g\beta}\sum_{t=0}^{T}(8\|(\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2+4\alpha^2\|(\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2+4\alpha^2\|(\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t)\|^2)]\\
&\quad+\frac{1}{\sqrt{\beta}}((1+c_1)\lambda^2-1)\sum_{t=1}^{T+1}\mathbb{E}\|\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1}\|^2+\frac{1}{\sqrt{\beta}}(1+\frac{1}{c_1})\alpha^2\sum_{t=1}^{T+1}\mathbb{E}\|\tilde{\mathbf{x}}_{t-1}-\mathbf{x}_{t-1}\|^2\\
&\quad+\beta((1+c_2)\lambda^2-1)\sum_{t=1}^{T+1}\mathbb{E}\|\mathbf{u}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{u}}_{t-1}\|^2\\
&\quad+(1+\frac{1}{c_2})\beta L_f^2\sum_{t=1}^{T+1}(8\mathbb{E}\|(\mathbf{x}_{t-1}-\mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2
\end{aligned}$$

$$+ 4\alpha^2\mathbb{E}\|(\mathbf{x}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + 4\alpha^2\mathbb{E}\|(\tilde{\mathbf{x}}_{t-1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{t-1})\|^2 + \beta^2\mathbb{E}\|\mathbf{v}_{t-1}\|^2). \tag{75}$$

Choosing $c_1 = c_2 = \frac{1}{\lambda} - 1$, we have

$$\begin{aligned}
&\mathbb{E}\{\ell(\bar{\mathbf{x}}_{T+1}) - \ell(\bar{\mathbf{x}}_0) + h(\bar{\mathbf{x}}_{T+1}) - h(\bar{\mathbf{x}}_0)\\
&+ \frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}[\|\bar{\mathbf{y}}_{T+1} - \mathbf{y}^*_{T+1}\|^2 - \|\mathbf{y}^*_0 - \bar{\mathbf{y}}_0\|^2] + \frac{1}{\sqrt{\beta}}[\|\mathbf{x}_{T+1} - \mathbf{1}\otimes\bar{\mathbf{x}}_{T+1}\|^2\\
&- \|\mathbf{x}_0 - \mathbf{1}\otimes\bar{\mathbf{x}}_0\|^2] + \beta[\|\mathbf{u}_{T+1} - \mathbf{1}\otimes\bar{\mathbf{u}}_{T+1}\|^2 - \|\mathbf{u}_0 - \mathbf{1}\otimes\bar{\mathbf{u}}_0\|^2]\}\\
\le& C_1\sum_{t=0}^{T}\mathbb{E}\|\mathbf{x}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + C_2\sum_{t=0}^{T}\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2 + C_3\sum_{t=0}^{T}\mathbb{E}\|\mathbf{y}_t - \mathbf{y}^*_t\|^2\\
&+ C_4\sum_{t=0}^{T}\mathbb{E}\|\mathbf{u}_t - \mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2 + C_5\sum_{t=0}^{T}\mathbb{E}\|\mathbf{v}_t\|^2\\
&+ [\frac{3\alpha}{2rm}(2(\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2 + 2\sigma_f^2) + \frac{9\beta}{2\mu_g}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}\sigma_g^2]\cdot T.
\end{aligned} \tag{76}$$

where the constants are

$$C_1 = (\frac{\alpha L_\ell}{2mr} + \frac{\alpha\tau}{2rm}) + (8+4\alpha^2)[\frac{5L_y^2\sqrt{\beta}}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2} + \frac{1}{1-\lambda}L_f^2\beta] + (\lambda-1)\frac{1}{\sqrt{\beta}}, \tag{77}$$

$$\begin{aligned}
C_2 =& (\frac{\alpha r}{m} + \frac{\alpha\tau r}{2m} + \frac{\alpha^2 L_\ell}{2m} - \frac{\alpha\tau}{m})\\
&+4\alpha^2[\frac{5L_y^2}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta} + \frac{1}{1-\lambda}L_f^2\beta] + \frac{1}{1-\lambda}\alpha^2\frac{1}{\sqrt{\beta}},
\end{aligned} \tag{78}$$

$$C_3 = L^2\frac{3\alpha}{2rm} - \frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}, \tag{79}$$

$$C_4 = \frac{3\alpha}{2rm} + (\lambda-1)\beta, \tag{80}$$

$$C_5 = -\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2} + \beta^3\frac{1}{1-\lambda}L_f^2. \tag{81}$$

To ensure $C_1 \le \frac{1-\lambda}{4}$, we have

$$\begin{aligned}
C_1 =& (\frac{\alpha L_\ell}{2mr} + \frac{\alpha\tau}{2rm}) + (8+4\alpha^2)[\frac{5L_y^2\sqrt{\beta}}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2} + \frac{1}{1-\lambda}L_f^2\beta] + (\lambda-1)\frac{1}{\sqrt{\beta}}\\
=&(\frac{\alpha L_\ell}{2mr} + \frac{\alpha\tau}{2rm})\\
&+ (8+4\alpha^2)[\frac{5L_y^2\sqrt{\beta}}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}] + (8+4\alpha^2)[\frac{1}{1-\lambda}L_f^2\beta] + (\lambda-1)\frac{1}{\sqrt{\beta}}\\
\le&[\frac{1-\lambda}{4\sqrt{\beta}} + \frac{1-\lambda}{4\sqrt{\beta}m} + (\frac{1-\lambda}{8\beta} + \frac{1-\lambda}{16\beta}) + (\lambda-1)\frac{1}{\sqrt{\beta}}]\\
\le&[\frac{1-\lambda}{4\sqrt{\beta}} + \frac{1-\lambda}{4\sqrt{\beta}} + (\frac{1-\lambda}{8\sqrt{\beta}} + \frac{1-\lambda}{8\sqrt{\beta}}) + (\lambda-1)\frac{1}{\sqrt{\beta}}] = -\frac{1-\lambda}{4}\frac{1}{\sqrt{\beta}}
\end{aligned} \tag{82}$$

where $\alpha \le \min\{\frac{(1-\lambda)m}{2\sqrt{\beta}(L_\ell+\tau)}\frac{\tau}{6+3\tau}, \frac{1-\lambda}{8\beta L_f}\}, \beta \le \frac{1-\lambda}{8L_f}, r = \frac{\tau}{6+3\tau}$.

To ensure $C_2 \le 0$, we have

$$C_2 = (\frac{\alpha r}{m} + \frac{\alpha\tau r}{2m} + \frac{\alpha^2 L_\ell}{2m} - \frac{\alpha\tau}{m})$$

$$
\begin{aligned}
&+4\alpha^2[\frac{5L_y^2}{\mu_g\beta}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}+\frac{1}{1-\lambda}L_f^2\beta]+\frac{1}{1-\lambda}\alpha^2\frac{1}{\sqrt{\beta}}\\
&\leq(\frac{\alpha r}{m}+\frac{\alpha\tau r}{2m}+\frac{\alpha^2L_\ell}{2m}-\frac{\alpha\tau}{m})+\alpha^2[\frac{1}{\sqrt{\beta}}\frac{(1-\lambda)}{12m}+\frac{4}{1-\lambda}L_f^2\beta]+\frac{1}{1-\lambda}\alpha^2\frac{1}{\sqrt{\beta}}\\
&\leq\frac{\alpha\tau}{12m}+\frac{\alpha\tau}{12m}+\frac{\alpha\tau}{6m}-\frac{\alpha\tau}{m}+(\frac{\alpha\tau}{72m}+\frac{\alpha\tau}{12m})+\frac{\alpha\tau}{6m}\leq-\frac{\alpha\tau}{3m},
\end{aligned}\tag{83}
$$

where $\alpha\leq\min\{\frac{\tau}{3L_\ell},\frac{\tau\sqrt{\beta}}{6(1-\lambda)},\frac{\tau(1-\lambda)}{48mL_f^2\beta}\},r=\frac{\tau}{6+3\tau}$.

To ensure $C_3\leq 0$, we have

$$
\begin{aligned}
C_3&=L^2\frac{3\alpha}{2rm}-\frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}\\
&\leq\frac{\mu_g^2\beta}{8}\frac{(1-\lambda)}{240L_y^2}\sqrt{\beta}-\frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}\\
&\leq\frac{\mu_g\beta}{8}\frac{\mu_g(1-\lambda)}{20(8+4\alpha^2)L_y^2}\sqrt{\beta}-\frac{\mu_g\beta}{4}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}\\
&\leq-\frac{\mu_g\beta}{8}\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}\leq-\frac{\beta^{1.5}}{8}\frac{\mu_g^2(1-\lambda)}{240mL_y^2}<0,
\end{aligned}\tag{84}
$$

where $\alpha\leq\frac{\mu_g^2\beta^{1.5}}{8}\frac{(1-\lambda)}{2880L_y^2L^2}$,

To ensure $C_4\leq 0$, we have

$$
C_4=\frac{3\alpha}{2rm}+(\lambda-1)\beta\leq 0,\tag{85}
$$

where $\alpha\leq(1-\lambda)\beta\frac{2rm}{3}$.

To ensure $C_5\leq 0$, we have

$$
\begin{aligned}
C_5=&-\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}(1+\frac{\mu_g\beta}{4})\frac{\beta^2}{2}+\beta^3\frac{1}{1-\lambda}L_f^2\\
\leq&-\frac{\mu_g(1-\lambda)}{240mL_y^2}\sqrt{\beta}\frac{\beta^2}{2}+\beta^3\frac{1}{1-\lambda}L_f^2\leq 0,
\end{aligned}\tag{86}
$$

where $\beta\leq\frac{(1-\lambda)^4\mu_g^2}{480^2mL_y^2L_f^2}$.

With the above conditions, we have

$$
\frac{1}{T}\sum_{t=0}^{T}\left(\mathbb{E}\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\right)\leq\frac{\mathbb{E}\left[\mathfrak{p}_0-\mathfrak{p}^*\right]+C_\sigma\cdot T}{T\min\left\{\frac{1-\lambda}{4}\frac{1}{\sqrt{\beta}},\frac{\alpha\tau}{3m}\right\}},\tag{87}
$$

where $\mathfrak{p}_t=\ell(\bar{\mathbf{x}}_t)+h(\bar{\mathbf{x}}_t)+\frac{\mu_g(1-\lambda)}{20m(8+4\alpha^2)L_y^2}\sqrt{\beta}\|\bar{\mathbf{y}}_t-\mathbf{y}_t^*\|^2+\beta\|\mathbf{u}_t-\mathbf{1}\otimes\bar{\mathbf{u}}_t\|^2+\frac{1}{\sqrt{\beta}}\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2,C_\sigma=[\frac{3\alpha}{2rm}(2(\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2+2\sigma_f^2)+\frac{9\beta^{1.5}(1-\lambda)}{40m(8+4\alpha^2)L_y^2}\sigma_g^2]$.

Let $\alpha=\mathcal{O}(T^{-\frac{1}{2}}),\alpha\leq\frac{(1-\lambda)m}{4\sqrt{\beta}\tau},\beta=\mathcal{O}(T^{-\frac{1}{3}}),r=\frac{\tau}{6+3\tau}$

Thus, we have

$$
C'_\sigma=\frac{\frac{3\alpha}{rm}((\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2+\sigma_f^2)+\frac{9\beta^{1.5}(1-\lambda)}{40m(8+4\alpha^2)L_y^2}\sigma_g^2}{\min\left\{\frac{1-\lambda}{4}\frac{1}{\sqrt{\beta}},\frac{\alpha\tau}{3m}\right\}}
$$

$$
\begin{aligned}
=&\frac{\frac{3\alpha}{rm}((\frac{C_{g_{xy}}C_{f_y}}{\mu_g}\left(1-\frac{\mu_g}{L_g}\right)^K)^2+\sigma_f^2)+\frac{9\beta^{1.5}(1-\lambda)}{40m(8+4\alpha^2)L_y^2}\sigma_g^2}{\frac{\alpha\tau}{3m}}\\
=&\frac{9(6+3\tau)}{\tau^2}((\frac{C_{g_{xy}}C_{f_y}}{\mu_g}(1-\frac{\mu_g}{L_g})^K)^2+\sigma_f^2)+\frac{27(1-\lambda)}{40(8+4\alpha^2)L_y^2}\frac{\beta^{1.5}}{\alpha\tau}\sigma_g^2. \qquad (88)
\end{aligned}
$$

Then, we can conclude that:

$$
\frac{1}{T}\sum_{t=0}^{T}\left(\mathbb{E}\|\mathbf{x}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2+\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{1}\otimes\bar{\mathbf{x}}_t\|^2\right)\leq\frac{\mathbb{E}\left[\mathfrak{p}_0-\mathfrak{p}^*\right]+C_\sigma\cdot T}{T\cdot\frac{\alpha\tau}{3m}}=\mathcal{O}(1/\sqrt{T})+C'_\sigma. \qquad (89)
$$

C. Supporting Lemmas

C.1. Proof of Lemma 1

$$
\begin{aligned}
&\left\|\nabla f\left(\mathbf{x}_1,\mathbf{y};\bar{\xi}\right)-\nabla f\left(\mathbf{x}_2,\mathbf{y};\bar{\xi}\right)\right\|^2\\
&\overset{(a)}{\leq}2\left\|\nabla_{\mathbf{x}}f\left(\mathbf{x}_1,\mathbf{y};\xi_i^0\right)-\nabla_{\mathbf{x}}f\left(\mathbf{x}_2,\mathbf{y};\xi_i^0\right)\right\|^2+2\|\nabla^2_{\mathbf{xy}}g\left(\mathbf{x}_1,\mathbf{y};\zeta_i^0\right)\left[\frac{1}{L_g}\mathbf{H}_{i,K}\right]\nabla_{\mathbf{y}}f\left(\mathbf{x}_1,\mathbf{y};\xi_i^0\right)\\
&\quad-\nabla^2_{\mathbf{xy}}g\left(\mathbf{x}_2,\mathbf{y};\zeta_i^0\right)\left[\frac{1}{L_g}\mathbf{H}_{i,K}\right]\nabla_{\mathbf{y}}f\left(\mathbf{x}_2,\mathbf{y};\xi_i^0\right)\|^2\\
&\overset{(b)}{\leq}2L_{f_x}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2+2\|\nabla^2_{\mathbf{xy}}g\left(\mathbf{x}_1,\mathbf{y};\zeta_i^0\right)\left[\frac{1}{L_g}\mathbf{H}_{i,K}\right]\nabla_{\mathbf{y}}f\left(\mathbf{x}_1,\mathbf{y};\xi_i^0\right)\\
&-\nabla^2_{\mathbf{xy}}g\left(\mathbf{x}_2,\mathbf{y};\zeta_i^0\right)\left[\frac{1}{L_g}\mathbf{H}_{i,K}\right]\nabla_{\mathbf{y}}f\left(\mathbf{x}_2,\mathbf{y};\xi_i^0\right)\|^2, \qquad (90)
\end{aligned}
$$

where (a) follows from triangle inequality and the definition of $\nabla f\left(\mathbf{x},\mathbf{y};\bar{\xi}\right)$, (b) follows from the gradient Liptichz assumption.

For the last term, we have

$$
\begin{aligned}
&\|\nabla^2_{\mathbf{xy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^0\right)\left[\frac{1}{L_g}\mathbf{H}_{i,K}\right]\nabla_{\mathbf{y}}f\left(\mathbf{x}_1,\mathbf{y};\xi_i^0\right)-\nabla^2_{\mathbf{xy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^0\right)\left[\frac{1}{L_g}\mathbf{H}_{i,K}\right]\nabla_{\mathbf{y}}f\left(\mathbf{x}_1,\mathbf{y};\xi_i^0\right)\|^2\\
&\overset{(a)}{\leq}\quad 3C_{g_{xy}}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}\left\|\nabla_{\mathbf{y}}f\left(\mathbf{x}_1,\mathbf{y};\xi_i^0\right)-\nabla_{\mathbf{y}}f\left(\mathbf{x}_1,\mathbf{y};\xi_i^0\right)\right\|^2\\
&\quad+3C_{f_y}^2\frac{L^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}\left\|\nabla^2_{\mathbf{xy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^0\right)-\nabla^2_{\mathbf{xy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^0\right)\right\|^2\\
&\quad+3C_{g_{xy}}^2C_{f_y}^2\|\frac{1}{L_g}\sum_{j=1}^{K}\prod_{p=1}^{j}\left(I-\frac{1}{L_g}\nabla^2_{\mathbf{yy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^p\right)\right)-\frac{1}{L_g}\sum_{j=1}^{K}\prod_{p=1}^{j}\left(I-\frac{1}{L_g}\nabla^2_{\mathbf{yy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^p\right)\right)^2\\
&\overset{(b)}{\leq}3C_{g_{xy}}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}L_{f_y}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2+3C_{f_y}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}L_{g_{xy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\\
&\quad+3C_{g_{xy}}^2C_{f_y}^2\frac{K}{L_g^2}\sum_{j=1}^{K}\left\|\prod_{p=1}^{j}\left(I-\frac{1}{L_g}\nabla^2_{\mathbf{yy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^p\right)\right)-\prod_{p=1}^{j}\left(I-\frac{1}{L_g}\nabla^2_{\mathbf{yy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^p\right)\right)\right\|^2\\
&\overset{(c)}{\leq}3C_{gy}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}L_{f_y}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2+3C_{f_y}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}L_{g_{xy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\\
&\quad+3C_{g_{xy}}^2C_{f_y}^2\frac{K}{L_g^2}\sum_{j=1}^{K}j\sum_{p=1}^{j}\left(1-\frac{\mu_g}{L_g}\right)^{2(j-1)}\frac{1}{L_g^2}\left\|\nabla^2_{\mathbf{yy}}g_i\left(\mathbf{x}_1,\mathbf{y};\zeta_i^p\right)-\nabla^2_{\mathbf{yy}}g_i\left(x_2,\mathbf{y};\zeta_i^p\right)\right\|^2
\end{aligned}
$$

$$\overset{(d)}{\leq} 3C_{gy}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}L_{f_y}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2+3C_{f_y}^2\frac{K^2}{L_g^2}\left(1-\frac{\mu_g}{L_g}\right)^{2k}L_{g_{xy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2$$
$$+3C_{g_{xy}}^2C_{f_y}^2\frac{K}{L_g^2}\sum_{j=1}^{K}j^2\left(1-\frac{\mu_g}{L_g}\right)^{2(j-1)}\frac{1}{L_g^2}L_{g_{yy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2, \tag{91}$$

where (a) and (d) follow from Assumption 1-2 and the triangle inequality and the Lemma A.1 in (Khanduri et al., 2021), and (b) follows from L_{f_y}-Lipschitz continuity assumption and expanding j to k, (c) is because of the triangle inequality, and (d) follows from $L_{g_{yy}}$-Liptichz continuity assumption.

On both sides taking expectation w.r.t k, we have

$$\begin{aligned}
&\mathbb{E}_k\left\|\nabla f\left(\mathbf{x}_1,\mathbf{y};\bar{\xi}\right)-\nabla f\left(\mathbf{x}_2,\mathbf{y};\bar{\xi}\right)\right\|^2\\
\leq&2L_{f_x}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2+2(3C_{gy}^2\frac{K^2}{L_g^2}\mathbb{E}_k[\left(1-\frac{\mu_g}{L_g}\right)^{2k}]L_{f_y}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\\
&+3C_{f_y}^2\frac{K^2}{L_g^2}\mathbb{E}_k[\left(1-\frac{\mu_g}{L_g}\right)^{2k}]L_{g_{xy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\\
&+3C_{g_{xy}}^2C_{f_y}^2\frac{K}{L_g^2}\sum_{j=1}^{K}j^2\left(1-\frac{\mu_g}{L_g}\right)^{2(j-1)}\frac{1}{L_g^2}L_{g_{yy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2)\\
\leq&2L_{f_x}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2+2+2(3C_{gy}^2\frac{K^2}{L_g^2}\frac{1}{K}\left(\frac{L_g^2}{2\mu_gL_g-\mu_g^2}\right)L_{f_y}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\\
&+3C_{f_y}^2\frac{K^2}{L_g^2}\frac{1}{K}\left(\frac{L_g^2}{2\mu_gL_g-\mu_g^2}\right)L_{g_{xy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\\
&+3C_{g_{xy}}^2C_{f_y}^2\frac{K}{L_g^2}\sum_{j=1}^{K}j^2\left(1-\frac{\mu_g}{L_g}\right)^{2(j-1)}\frac{1}{L_g^2}L_{g_{yy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2).
\end{aligned}\tag{92}$$

Thus, we have

$$\begin{aligned}
L_f^2:=&2L_{f_x}^2++6C_{g_{xy}}^2L_{f_y}^2\left(\frac{K}{2\mu_gL_g-\mu_g^2}\right)+6C_{f_y}^2L_{g_{xy}}^2\left(\frac{K}{2\mu_gL_g-\mu_g^2}\right)\\
&+6C_{g_{xy}}^2C_{f_y}^2\frac{K}{L_g^2}\sum_{j=1}^{K}j^2\left(1-\frac{\mu_g}{L_g}\right)^{2(j-1)}\frac{1}{L_g^2}L_{g_{yy}}^2.
\end{aligned}\tag{93}$$

Further, $\mathbb{E}_k\left\|\nabla f\left(\mathbf{x},\mathbf{y}_1;\bar{\xi}\right)-\nabla f\left(\mathbf{x},\mathbf{y}_2;\bar{\xi}\right)\right\|^2\leq L_f\|\mathbf{y}_1-\mathbf{y}_2\|^2$ follows the same procedure.

C.2. Proof of Lemma 3

$$\begin{aligned}
&\|\nabla f(\mathbf{x},\mathbf{y})-\mathbb{E}[\bar{\nabla}f(\mathbf{x},\mathbf{y};\bar{\xi})]\|\\
=&\|\nabla f(\mathbf{x},\mathbf{y})-\mathbb{E}[\nabla_{\mathbf{x}}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\xi_i^0)-\frac{1}{L_g}\nabla_{\mathbf{xy}}^2g\left(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\zeta_i^0\right)\mathbf{H}_{i,k}\nabla_{\mathbf{y}}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\xi_i^0)]\|\\
=&\|\nabla f(\mathbf{x},\mathbf{y})-\mathbb{E}\nabla_{\mathbf{x}}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\xi_i^0)\\
&-\mathbb{E}[\frac{1}{L_g}\nabla_{\mathbf{xy}}^2g\left(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\zeta_i^0\right)\sum_{j'=1}^{k(K)}\prod_{p=1}^{j'}\left(\mathbf{I}-\frac{\nabla_{\mathbf{yy}}^2g\left(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\zeta_i^p\right)}{L_g}\right)\nabla_{\mathbf{y}}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\xi_i^0)]\|\\
=&\|\nabla f(\mathbf{x},\mathbf{y})-\mathbb{E}\nabla_{\mathbf{x}}f(\mathbf{x}_{i,t},\mathbf{y}_{i,t};\xi_i^0)
\end{aligned}$$

$$
\begin{aligned}
&- [\frac{1}{L_g}\nabla^2_{\mathbf{xy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0\right) \sum_{j'=1}^{k(K)} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\right)}{L_g}\right)^{j'} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)]\| \\
=&\|\nabla f(\mathbf{x}, \mathbf{y}) - \mathbb{E}\nabla_{\mathbf{x}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0) \\
&- [\frac{1}{L_g}\nabla^2_{\mathbf{xy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0\right) \sum_{j'=1}^{\infty} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\right)}{L_g}\right)^{j'} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)] \\
&+ [\frac{1}{L_g}\nabla^2_{\mathbf{xy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0\right) \sum_{j'=k(K)+1}^{\infty} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\right)}{L_g}\right)^{j'} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)]\| \\
=&\|\nabla f(\mathbf{x}, \mathbf{y}) - \mathbb{E}\nabla_{\mathbf{x}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0) \\
&- [\frac{1}{L_g}\nabla^2_{\mathbf{xy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0\right) \sum_{j'=1}^{\infty} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\right)}{L_g}\right)^{j'} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)] \\
&+ [\frac{1}{L_g}\nabla^2_{\mathbf{xy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0\right) \sum_{j'=k(K)+1}^{\infty} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\right)}{L_g}\right)^{j'} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)]\| \\
=&\|[\frac{1}{L_g}\nabla^2_{\mathbf{xy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0\right) \sum_{i'=k(K)+1}^{\infty} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\right)}{L_g}\right)^{j'} \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)]\| \\
\leq& \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \xi)\| \cdot \left\|\nabla^2_{\mathbf{xy}} g(\mathbf{x}, \mathbf{y}; \zeta)\right\| \cdot \| \sum_{i'=k(K)+1}^{\infty} \left(\mathbf{I} - \frac{\nabla^2_{\mathbf{yy}} g\left(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}\right)}{L_g}\right)^{K} \| \\
\leq& C_{g_{xy}} C_{f_y} \frac{1}{\mu_g} (1 - \frac{\mu_g}{L_g})^K. \qquad (94)
\end{aligned}
$$

This completes the proof.

C.3. Proof of Lemma 4

Similar to Eqs. (90)–(91), and with the conventional stochastic gradient estimator $\bar{\nabla} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \bar{\xi}_{ij}) = \nabla_{\mathbf{x}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0) - \frac{K}{L_g}\nabla^2_{\mathbf{xy}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^0) \cdot \prod_{p=1}^{k(K)} (I - \frac{\nabla^2_{\mathbf{yy}} g(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \zeta_i^p)}{L_g}) \nabla_{\mathbf{y}} f(\mathbf{x}_{i,t}, \mathbf{y}_{i,t}; \xi_i^0)$, we have

$$
\begin{aligned}
&\mathbb{E}_k \left\|\nabla f\left(\mathbf{x}_1, \mathbf{y}; \bar{\xi}\right) - \nabla f\left(\mathbf{x}_2, \mathbf{y}; \bar{\xi}\right)\right\|^2 \\
\leq& 2L^2_{f_x} \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2 + 2(3C^2_{gy} \frac{K^2}{L_g^2} \frac{1}{K} \left(\frac{L_g^2}{2\mu_g L_g - \mu_g^2}\right) L^2_{f_y} \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2 \\
&+ 3C^2_{f_y} \frac{K^2}{L_g^2} \frac{1}{K} \left(\frac{L_g^2}{2\mu_g L_g - \mu_g^2}\right) L^2_{g_{xy}} \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2 \\
&+ 3C^2_{g_{xy}} C^2_{f_y} \frac{K^2}{L_g^2} \left\| \prod_{p=1}^{k(K)} \left(I - \frac{1}{L_g}\nabla^2_{\mathbf{yy}} g_i\left(\mathbf{x}_1, \mathbf{y}; \zeta_i^p\right)\right) - \prod_{p=1}^{k(K)} \left(I - \frac{1}{L_g}\nabla^2_{\mathbf{yy}} g_i\left(\mathbf{x}_1, \mathbf{y}; \zeta_i^p\right)\right) \right\|^2) \\
\leq& 2L^2_{f_x} \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2 + 6C^2_{gy} \frac{K^2}{L_g^2} \frac{1}{K} \left(\frac{L_g^2}{2\mu_g L_g - \mu_g^2}\right) L^2_{f_y} \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2 \\
&+ 6C^2_{f_y} \frac{K^2}{L_g^2} \frac{1}{K} \left(\frac{L_g^2}{2\mu_g L_g - \mu_g^2}\right) L^2_{g_{xy}} \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2 \\
&+ 6C^2_{g_{xy}} C^2_{f_y} \frac{K^2}{L_g^2} k(K) \sum_{p=1}^{k(K)} \left(1 - \frac{\mu_g}{L_g}\right)^{2(k(K)-1)} \frac{1}{L_g^2} \left\|\nabla^2_{\mathbf{yy}} g_i\left(\mathbf{x}_1, \mathbf{y}; \zeta_i^p\right) - \nabla^2_{\mathbf{yy}} g_i\left(x_2, \mathbf{y}; \zeta_i^p\right)\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
\leq & 2L_{f_x}^2 \left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2+6C_{gy}^2 \frac{K^2}{L_g^2} \frac{1}{K}\left(\frac{L_g^2}{2\mu_g L_g-\mu_g^2}\right) L_{f_y}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2 \\
& +6C_{f_y}^2 \frac{K^2}{L_g^2} \frac{1}{K}\left(\frac{L_g^2}{2\mu_g L_g-\mu_g^2}\right) L_{g_{xy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2 \\
& +6C_{g_{xy}}^2 C_{f_y}^2 \frac{K^2}{L_g^2} k(K)^2\left(1-\frac{\mu_g}{L_g}\right)^{2(k(K)-1)} \frac{1}{L_g^2} L_{g_{yy}}^2\left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2 . \qquad (95)
\end{aligned}
$$

Since we are aiming at finding a constant L_{conv} which satisfied the Liptichz inequality for all k, Eq. (95) needs to hold with the maximum value of $k(K)^2\left(1-\frac{\mu_g}{L_g}\right)^{2(k(K)-1)}$.

Thus, we have

$$
\begin{aligned}
L_{conv}^2 := & 2L_{f_x}^2+ +6C_{g_{xy}}^2 L_{f_y}^2\left(\frac{K}{2\mu_g L_g-\mu_g^2}\right)+6C_{f_y}^2 L_{g_{xy}}^2\left(\frac{K}{2\mu_g L_g-\mu_g^2}\right) \\
& +6C_{g_{xy}}^2 C_{f_y}^2 \frac{K^2}{L_g^2} \max_{k(K)}\{k(K)^2\left(1-\frac{\mu_g}{L_g}\right)^{2(k(K)-1)}\} \frac{1}{L_g^2} L_{g_{yy}}^2 . \qquad (96)
\end{aligned}
$$

Since $\max_{k(K)}\{k(K)^2\left(1-\frac{\mu_g}{L_g}\right)^{2(k(K)-1)}\} \geq \frac{1}{K}\sum_{j=1}^{K} j^2\left(1-\frac{\mu_g}{L_g}\right)^{2(j-1)}$.

Thus, we can conclude that $L_{conv} \geq L_f$.

Further, $\mathbb{E}_k\left\|\nabla f\left(\mathbf{x}, \mathbf{y}_1 ; \bar{\xi}\right)-\nabla f\left(\mathbf{x}, \mathbf{y}_2 ; \bar{\xi}\right)\right\|^2 \leq L_{conv}\|\mathbf{y}_1-\mathbf{y}_2\|^2$ follows the same procesure.

D. Further Experiments and Additional Results

D.1. Topology setting

We test three different topologies on a 10-agent system. The datasize for each agent is $n = 100$. We set the constant learning rate $\alpha = 0.5$, $\beta = 0.5$ and mini-batch size $q = \lceil\sqrt{n}\rceil = 10$, pre-defined parameter $K = 10$. As shown in Fig. 5, we can observe that Prometheus is insensitive to the network topology, but the convergence metric $\mathfrak{M}$ slightly increases as p_c decreases.

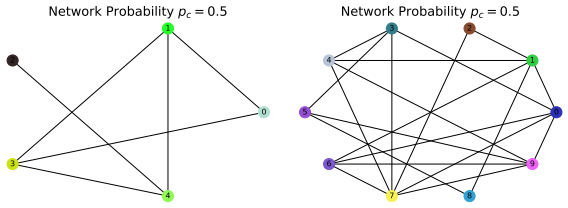

Figure 4. Network topology.

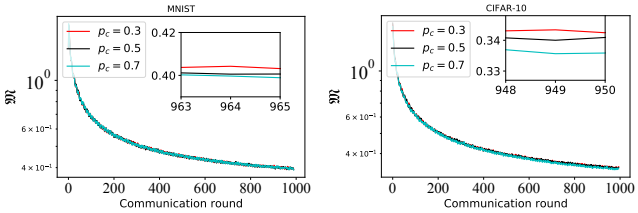

Figure 5. Network probability comparison.

D.2. Learning rate setting

We use a 10-agent system with a generated topology as shown in Fig. 4. In this experiment, the dataset size for each agent is $n = 100$, mini-batch size $q = \lceil\sqrt{n}\rceil = 10$, pre-defined parameter $K = 10$. Fig. 6 illustrates the convergence metric $\mathfrak{M}$ of Prometheus with different learning rates α and β. We fix a relatively small learning rate $\beta = 0.5$ while comparing α; and set $\alpha = 0.5$ while comparing β. In this experiment, we observe that methods with a smaller learning rate have a smaller slope in the figure, which implies a slower convergence.

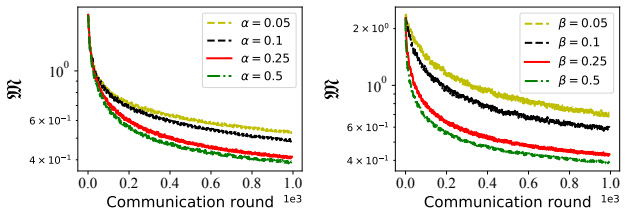

Figure 6. MNIST dataset.

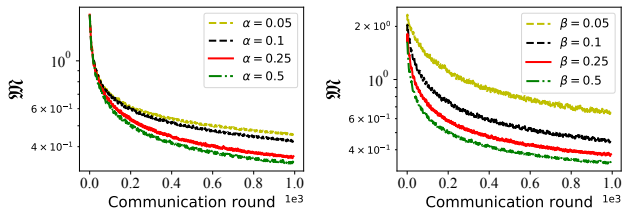

Figure 7. CIFAR-10 dataset.

D.3. Additional experiments on our new stochatic estimator

Recall that the conventional estimator to estimate the $\mathbf{A}^{-1}$ can be denoted as $\tilde{\mathbf{A}}_{conv}^{-1} = K\prod_{p=1}^{k(K)}(\mathbf{I} - \mathbf{A}_s)$, while the new estimator can be denoted as $\tilde{\mathbf{A}}^{-1} = \sum_{j'=1}^{k(K)}\prod_{p=1}^{j'}(\mathbf{I} - \mathbf{A}_s)$. Here we consider a 4-dimension matrix example $\mathbf{A} = 0.25 * \mathbf{I}_4$ and 10-dimension matrix example $\mathbf{A} = 0.1 * \mathbf{I}_{10}$. Let $\mathbf{A}_s$ be a random matrix obtained from $\mathbf{A}$ plus Gaussian noise. We use $\tilde{\mathbf{A}}_{conv}^{-1}$ and $\tilde{\mathbf{A}}^{-1}$ to estimate $\mathbf{A}^{-1}$, respectively. We run 10000 independent trials and the results are shown in Fig. 8 and Fig. 9. We can see from Fig. 8 and Fig. 9 that the new Hessian inverse estimator has a much smaller variance than the conventional one.

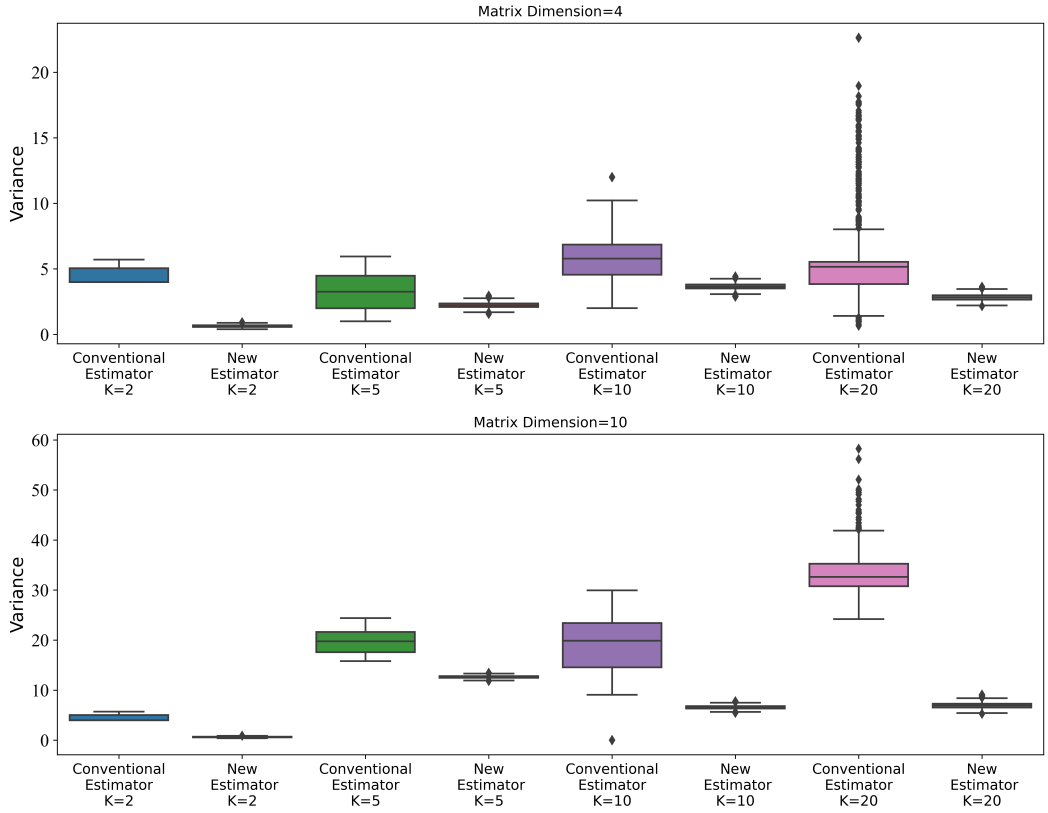

Figure 8. Variance comparisons on varying K.

D.4. Additional experiments on decentralized hyper-parameter

Next, we compare Prometheus with other baseline algorithms using the logistic regression problem (Grazzi et al., 2020; Ji et al., 2021) with the same formulation as in (1), where $f_i(\mathbf{x}, \mathbf{y}_i^*(\mathbf{x})) = \frac{1}{|\mathcal{D}_{\text{val},i}|}\sum_{(\mathbf{a}_j,\mathbf{c}_j)\in\mathcal{D}_{\text{val},i}} Q(\mathbf{a}_j^T\mathbf{y}_i^*, \mathbf{c}_j)$, $g_i(\mathbf{x}, \mathbf{y}_i) = \frac{1}{|\mathcal{D}_{\text{tr},i}|}\sum_{(\mathbf{a}_j,\mathbf{c}_j)\in\mathcal{D}_{\text{tr},i}} Q(\mathbf{a}_j^T\mathbf{y}_i, \mathbf{c}_j) + \frac{1}{q_1 p}\sum_{k=1}^{q_1}\sum_{r=1}^{p}\exp(\mathbf{x}_r)\mathbf{y}_{irk}^2$. $\mathcal{D}_{\text{tr},i}$ denotes the the training dataset and $\mathcal{D}_{\text{val},i}$ is the validation dataset for agent i, respectively, Q indicates the cross-entropy loss, q_1 denotes the number of classes, and p

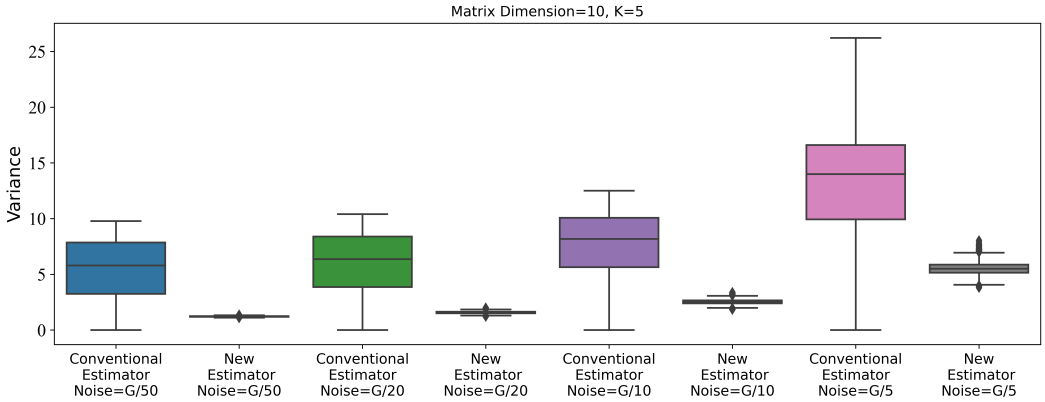

Figure 9. Variance comparisons on different level of noise G.

is the number of features. We use the "a9a" dataset from LIBSVM repository, which is publicly available at (Chang & Lin, 2011). We divide the a9a dataset into training, validation, and testing sets, which contain 40%, 40%, and 20% samples, respectively. We compare the proposed Prometheus algorithm in terms of test accuracy and loss, using ten-agent communication networks, with the network connection probability $p_c = 0.5$, step sizes $\alpha = \beta = 0.01$. As shown in Fig. 10, Prometheus performs better than all other algorithms in terms of the total number of communication rounds.

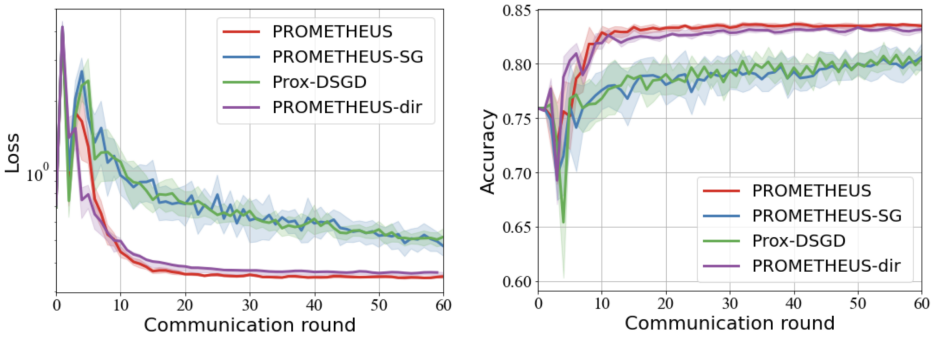

Figure 10. Hyper-parameter experiment on a ten-agent network.