Locally Differentially Private Decentralized Stochastic Bilevel Optimization with Guaranteed Convergence Accuracy

Ziqin Chen [1] **Yongqiang Wang** [1]

Abstract

Decentralized bilevel optimization based machine learning techniques are achieving remarkable success in a wide variety of domains. However, the intensive exchange of information (involving nested-loops of consensus or communication iterations) in existing decentralized bilevel-optimization algorithms leads to a great challenge to ensure rigorous differential privacy, which, however, is necessary to bring the benefits of machine learning to domains where involved data are sensitive. By proposing a new decentralized stochastic bilevel-optimization algorithm which avoids nested-loops of information-exchange iterations, we achieve, for the first time, both differential privacy and accurate convergence in decentralized bilevel optimization. This is significant since even for single-level decentralized optimization and learning, existing differential-privacy solutions have to sacrifice convergence accuracy for privacy. Besides characterizing the convergence rate under nonconvex/convex/strongly convex conditions, we also rigorously quantify the price of differential privacy in the convergence rate. Experimental results on machine learning models confirm the efficacy of our algorithm.

1. Introduction

Bilevel stochastic optimization is evolving as an effective tool for solving many machine learning problems having a nested structure, with typical examples including meta-learning (Bertinetto et al., 2019; Rajeswaran et al., 2019), hyperparameter optimization (Franceschi et al., 2018), imitation learning (Arora et al., 2020), and neural architecture search (Liu et al., 2018). So far, numerous centralized stochastic bilevel-optimization algorithms have been proposed (Ghadimi & Wang, 2018; Khanduri et al., 2021; Ji et al., 2021; Hong et al., 2023). Recently, with the increasingly pressing need to parallelize learning algorithms in order to handle the enormous growth in data and model sizes, the following decentralized stochastic bilevel-optimization (DSBO) problem is gaining increased traction (Lu et al., 2022; Yang et al., 2022; Gao et al., 2023; Chen et al., 2023; Zhang et al., 2023; Dong et al., 2023; Kong et al., 2024):

$$\begin{aligned} &\min_{x\in\mathbb{R}^p} F(x), \quad F(x)=\frac{1}{m}\sum_{i=1}^{m} f_i(x,y^*(x)),\\ &\text{s.t. } y^*(x)=\text{argmin}_{y\in\mathbb{R}^q} g(x,y):=\frac{1}{m}\sum_{i=1}^{m} g_i(x,y), \end{aligned} \tag{1}$$

where $x\in\mathbb{R}^p$ and $y\in\mathbb{R}^q$ represent the optimization parameters and m denotes the number of agents. Each agent i only has access to its local upper-level objective function f_i and lower-level objective function g_i, which, in machine learning applications, are usually given by

$$\begin{aligned} f_i(x,y)&=\mathbb{E}_{\varphi_i}[h(x,y;\varphi_i)],\\ g_i(x,y)&=\mathbb{E}_{\xi_i}[l(x,y;\xi_i)]. \end{aligned} \tag{2}$$

In (2), φ_i and ξ_i represent random data samples which usually follow unknown and heterogeneous distributions across different agents.

All above DSBO algorithms require participating agents to explicitly share model updates in every iteration, which raises severe privacy concerns when involved data are sensitive. In fact, recent studies (Zhu et al., 2019; Triastcyn & Faltings, 2020) have shown that even though raw data are not shared, exploiting information shared in decentralized optimization, external adversaries can still precisely recover the raw data used for training (pixel-wise accurate for images and token-wise matching for texts). As differential privacy is evolving as the de facto standard for privacy preservation due to its rigorous mathematical foundations yet implementation simplicity and post-processing immunity (Dwork et al., 2010; 2014), it is of great interest to achieve differential privacy in DSBO. However, given that existing DSBO algorithms all involve nested-loops of com-

[1]Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, 29634, United States. Correspondence to: Yongqiang Wang <yongqiw@clemson.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

munication (consensus) iterations[1], directly incorporating the standard differential-privacy noise injection mechanism in existing DSBO algorithms will inevitably result in an exploding cumulative privacy budget as the iteration proceeds, leading to diminishing privacy protection in the long run. Another challenge is to maintain the accuracy of DSBO under the constraint of differential privacy. In fact, even for the simpler single-level decentralized optimization problem, existing differential-privacy solutions have to trade optimization accuracy for privacy (Bellet et al., 2018; Cyffers et al., 2022; Bietti et al., 2022), which is undesirable in accuracy-sensitive applications.

1.1. Our Contributions

1. We propose a differentially private DSBO algorithm that can ensure both accurate convergence and rigorous differential privacy, with the cumulative privacy budget bounded even when the number of iterations tends to infinity. To the best of our knowledge, no such results have been reported before. Moreover, by employing the local differential privacy framework, our results can be applied to the fully decentralized setting where no data aggregator or mediator exists to gather data or assist privacy design.

2. A key enabler for our approach to achieving both differential privacy and accurate convergence is a novel algorithm for DSBO. Except limited recent works (Zhang et al., 2023; Dong et al., 2023; Kong et al., 2024) which use single-loop consensus, most existing DSBO algorithms employ nested-loops of consensus iterations. Our new algorithm successfully circumvents nested-loops of consensus, which makes it possible to alleviate the growth of the cumulative privacy budget as the number of iterations increases. In fact, given that using intensive (nested-loops of) consensus or communication rounds is the only approach to ensuring accurate convergence when the objective functions are heterogeneous across the agents (note that the results in Zhang et al. (2023), Dong et al. (2023), and Kong et al. (2024) are subject to an optimization error that is on the order of the constant stepsize therein), our algorithm is of independent interest even if privacy is not of concern.

3. We establish the convergence rate of our algorithm for nonconvex/convex/strongly convex objective functions f_i, which is different from existing DSBO results (Lu et al., 2022; Gao et al., 2023; Chen et al., 2023) that focus solely on the nonconvex case. Moreover, our convergence analysis relaxes the assumption that g_i is Lipschitz continuous with respect to y, which is widely used in existing DSBO literature (see, e.g., Chen et al. (2022) and Yang et al. (2022)).

[1]Note that the algorithm in Gao et al. (2023) assumes identical data distributions for ξ_i and hence $g_1 = g_2 = \cdots = g_m$ (see equations (2) and (3) in Gao et al. (2023) or Appendix C.2 in Chen et al. (2023)), and thus does not apply to our general setting here.

4. Despite retaining accurate convergence, our algorithm does pay a price for obtained differential privacy in convergence rate. We systematically quantify the tradeoff between privacy and convergence rate. It is worth noting that by avoiding estimating the full Hessian or Jacobian matrix, our algorithm still achieves improved computational complexity compared with the result for DSBO in Chen et al. (2022), which does not consider privacy protection.

5. We conduct experiment evaluation using several machine learning problems. The results confirm the efficiency of our algorithm on both the synthetic and the real-world datasets.

1.2. Related Work

1.2.1. Bilevel Optimization

Bilevel optimization was first discussed in Bracken & McGill (1973) for solving resource allocation problems. Historically, it was treated by viewing the lower-level optimality condition as constraints to the upper-level problem (Hansen et al., 1992). More recently, Couellan & Wang (2016) proposed a gradient-based algorithm providing asymptotic convergence and Ghadimi & Wang (2018) developed a nested-loop stochastic approximated algorithm establishing non-asymptotic convergence. Following these developments, various centralized approaches have been introduced, trying to improve the efficiency in solving bilevel-optimization problems (Khanduri et al., 2021; Ji et al., 2021; Hong et al., 2023).

Driven by the need for parallelized learning algorithms to handle the enormous growth in data and model sizes in machine learning, plenty of DSBO algorithms have been proposed recently (Lu et al., 2022; Chen et al., 2022; Yang et al., 2022; Gao et al., 2023; Chen et al., 2023). For example, Lu et al. (2022) and Gao et al. (2023) considered the DSBO problem where the lower-level objective function is fully accessible to every agent. Chen et al. (2022), Yang et al. (2022), and Chen et al. (2023) considered the case where neither the upper-level function nor the lower-level function is fully accessible to every local agent. In addition, the approaches in Chen et al. (2022) and Yang et al. (2022) require computing the full Jacobian and/or Hessian matrix, entailing a computational complexity of the order $\mathcal{O}(pq)$ or $\mathcal{O}(q^2)$ in every iteration. To reduce the computational complexity, Chen et al. (2023) proposed to estimate the Hessian-vector and Jacobian-vector products, which reduces the per-iteration complexity from $\mathcal{O}(pq)$ (or $\mathcal{O}(q^2)$) to $\mathcal{O}(\max\{p, q\})$. However, none of the existing results have addressed differential privacy for DSBO. In fact, as discussed in Section 1, to ensure accurate enough local estimation of the hypergradient, all of these algorithms employ nested-loops of consensus (communication) iterations, which will result in an exploding cumulative privacy budget if we incorporate these algorithms with the standard

Laplace-noise mechanism in Dwork et al. (2014) to achieve differential privacy. In Table 1, we summarize the difference between our algorithm and existing results.

1.2.2. Differential Privacy

Widely regarded as the "gold standard" for privacy protection (Cummings et al., 2021), differential privacy has found numerous applications in distributed computation scenarios, including distributed deep learning (Papernot et al., 2018; Ghazi et al., 2021; Kairouz et al., 2021), distributed stochastic optimization (Bassily et al., 2019; Asi et al., 2021; Altschuler & Talwar, 2022), and federated learning (Geyer et al., 2017; Zhang et al., 2022). Note that the commonly used differential-privacy framework assumes the presence of a data aggregator/curator to collect the raw data and inject noise. In the decentralized scenario, to ensure agent-level privacy, we employ the local differential privacy (LDP) framework (Kasiviswanathan et al., 2011), in which random perturbations are performed locally by each agent, thereby protecting individual data against external adversaries and neighboring agents. LDP has been implemented in decentralized stochastic optimization and federated learning algorithms (Bellet et al., 2018; Cyffers et al., 2022; Bietti et al., 2022); however, these algorithms often face a fundamental tradeoff between optimization accuracy and privacy. It is worth noting that although using the information-theoretic approach, Kasiviswanathan et al. (2011) and Dwork et al. (2014) have proven the possibility to retain accurate convergence in differentially private learning by trading convergence rate for privacy, it is only recently that Wang & Nedić (2023) and Chen & Wang (2023) proposed concrete implementable algorithms that actually achieve this goal in decentralized optimization and learning. Nevertheless, these results are for the conventional single-level decentralized optimization and they cannot be combined with existing bilevel-optimization algorithms to ensure both differential privacy and accurate convergence. In fact, due to the existence of nested-loops of consensus iterations in existing DSBO algorithms, directly applying the differential-privacy mechanisms in Wang & Nedić (2023) and Chen & Wang (2023) will result in both loss of convergence accuracy and explosion of the cumulative privacy budget.

Notations: We denote $\nabla F(x) \in \mathbb{R}^p$ as the gradient of $F(x)$. We use $\nabla_x g(x,y)$ and $\nabla_y g(x,y)$ to represent the gradients of g with respect to x and y, respectively. We write $\nabla^2_{xy} g(x,y) \in \mathbb{R}^{p\times q}$ for the Jacobian matrix of g and $\nabla^2_{yy} g(x,y) \in \mathbb{R}^{q\times q}$ for the Hessian matrix of g with respect to y. We denote $\|\cdot\|_1$ and $\|\cdot\|$ as the l_1-norm and the l_2-norm of vectors, respectively. We use $\mathbf{1}_p$ to denote the all-ones vector in $\mathbb{R}^p$. We add an overbar to a letter to denote the average of all agents, e.g., $\bar{x}_t = \frac{1}{m}\sum_{i=1}^m x_{i,t}$. We use bold font to represent stacked vectors of all agents, e.g., $\boldsymbol{x}_t = \mathrm{col}(x_{1,t}, \cdots, x_{m,t})$. We write $\mathbb{P}[\mathcal{A}]$ for the probability of an event $\mathcal{A}$. We use $\mathrm{Lap}(\nu)$ to denote the Laplace distribution with a parameter $\nu > 0$, featuring a probability density function $f(x|\nu) = \frac{1}{2\nu} e^{\frac{-|x|}{\nu}}$. $\mathrm{Lap}(\nu)$ has a mean of zero and a variance of $2\nu^2$. We denote the set of m agents as $[m]$ and the neighboring set of agent i as $\mathcal{N}_i$. We denote the coupling weight matrix as $W = \{w_{ij}\} \in \mathbb{R}^{m\times m}$, in which $w_{ij} > 0$ if agent j interacts with agent i, and $w_{ij} = 0$ otherwise.

2. Preliminaries

2.1. Hypergradient Estimation

The major challenge in solving DSBO lies in the absence of explicit knowledge of $y^*(x)$, which makes it impossible for individual agents to evaluate the hypergradient $\nabla F(x, y^*(x))$. By leveraging the results for centralized stochastic bilevel optimization (Ghadimi & Wang, 2018), recently, Chen et al. (2022) proposed to calculate the hypergradient using the following relation:

$$\begin{aligned}\nabla F(x) = \frac{1}{m}\sum_{i=1}^m \nabla_x f_i(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x)) \\ \times \left[\nabla^2_{yy} g(x, y^*(x))\right]^{-1} \frac{1}{m}\sum_{i=1}^m \nabla_y f_i(x, y^*(x)).\end{aligned} \tag{3}$$

It is evident that computing $\nabla F(x)$ requires global information about $g(x,y) = \frac{1}{m}\sum_{i=1}^m g_i(x,y)$, which is inaccessible to agent i in a decentralized setting. A natural approach is to use ∇g_i as a surrogate; however, due to data heterogeneity across the agents, this approach results in steady-state errors. Therefore, every agent has to maintain local estimates of the global hypergradient. Instead of estimating the entire Hessian/Jacobian matrix, Chen et al. (2023) proposed to estimate the Hessian-inverse-vector product:

$$z^* = \left(\sum_{i=1}^m \nabla_{yy} g_i(x, y^*(x))\right)^{-1} \left(\sum_{i=1}^m \nabla_y f_i(x, y^*(x))\right). \tag{4}$$

According to (3), the global hypergradient is given by

$$\nabla F(x) = \frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_i(x, y^*(x)) - \nabla^2_{xy} g_i(x, y^*(x)) z^*\right), \tag{5}$$

where $\nabla^2_{xy} g_i(x, y^*(x)) z^*$ will be referred to as the Jacobian-vector product.

From (5), we know that if each agent i can have an accurate enough estimation of $\nabla_x f_i(x, y^*(x))$, z^*, and $\nabla^2_{xy} g_i(x, y^*(x)) z^*$, then every agent can have a good estimate of the global hypergradient. Notably, estimating the vector-valued z^* and $\nabla^2_{xy} g_i(x, y^*(x)) z^*$ circumvents the need for estimating the full Hessian and Jacobian matrices, which substantially reduces the per-iteration computational complexity.

Table 1. We compare our Algorithm 2 (LDP-DSBO) with existing algorithms, including the centralized bilevel-optimization algorithm BSA (Ghadimi & Wang, 2018), personalized DSBO algorithms SPDB (Lu et al., 2022) and VRDSBO (Gao et al., 2023), and DSBO algorithms DSBO-JHIP (Chen et al., 2022), GBDSBO (Yang et al., 2022), and MA-DSBO (Chen et al., 2023). In the table, we use δ to denote the optimization error. We use "Jacobian" to represent whether the algorithm requires computing the full Hessian or Jacobian matrix. We use "DP" to represent whether the algorithm considers differential privacy. We also use "Privacy Budget" to refer to the cumulative privacy budget of the algorithm when it is combined with the Laplace noise used in our algorithm to enable differential privacy. The detailed cumulative privacy budget calculation is provided in Appendix H.2).

ALGORITHM	DECENTRALIZED?	COMPUTATIONAL COMPLEXITY	JACOBIAN	DP	PRIVACY BUDGET
BSA	NO	$\mathcal{O}(\delta^{-3}+(q^2\log(\delta^{-1})+pq)\delta^{-2})$	YES	NO	$\mathcal{O}(\delta^{-3})$
SPDB	YES	$\mathcal{O}(\max\{p,q\}\log(\delta^{-1})\delta^{-2})$	NO	NO	$\mathcal{O}(\delta^{-2})$
VRDSBO	YES	$\mathcal{O}((pq+q^2)\delta^{-\frac{3}{2}})$	YES	NO	$\mathcal{O}(\delta^{-\frac{3}{2}})$
DSBO-JHIP	YES	$\mathcal{O}(pq\log(\delta^{-1})\delta^{-3})$	YES	NO	$\mathcal{O}(\delta^{-3})$
GBDSBO	YES	$\mathcal{O}((q^2\log(\delta^{-1})+pq)\delta^{-2})$	YES	NO	$\mathcal{O}(\delta^{-2})$
MA-DSBO	YES	$\mathcal{O}(\max\{p,q\}\log(\delta^{-1})\delta^{-2})$	NO	NO	$\mathcal{O}(\delta^{-2})$
LDP-DSBO	YES	$\mathcal{O}(\max\{p,q\}\delta^{-2.6})$	NO	YES	$\mathcal{O}(1)$

2.2. Assumptions

Assumption 2.1. The weight matrix $W=\{w_{ij}\}\in\mathbb{R}^{m\times m}$ is symmetric and satisfies $\mathbf{1}^TW=\mathbf{0}^T$ and $W\mathbf{1}=\mathbf{0}$. The eigenvalues of $I+W$ (after arranged in an increasing order) satisfy $0=\delta_1<\delta_2\leq\cdots\leq\delta_m<1$.

Assumption 2.2. For any $i\in[m]$, functions f_i, ∇f_i, ∇g_i, and $\nabla^2 g_i$ are $L_{f,0}$, $L_{f,1}$, $L_{g,1}$, and $L_{g,2}$ Lipschitz continuous, respectively. Moreover, each function g_i is μ_g-strongly convex in y.

Assumption 2.3. The stochastic oracles $\nabla h(x,y;\varphi)$, $\nabla^2 h(x,y;\varphi)$, $\nabla l(x,y;\xi)$, $\nabla^2 l(x,y;\xi)$, and $\nabla^3 l(x,y;\xi)$ are unbiased with bounded variances, which are represented as $\sigma_{f,1}^2$, $\sigma_{f,2}^2$, $\sigma_{g,1}^2$, $\sigma_{g,2}^2$, and $\sigma_{g,3}^2$, respectively.

Assumptions 2.2 and 2.3 are standard in the DSBO literature (Lu et al., 2022; Chen et al., 2022; Yang et al., 2022; Chen et al., 2023; Gao et al., 2023). They allow f_i and g_i to be heterogeneous across the agents, which are more general than the homogeneous-function assumption in Lu et al. (2022) and Gao et al. (2023). In addition, we relax the assumption that lower-level objective functions g_i are Lipschitz continuous with respect to y, which is used in Chen et al. (2022) and Yang et al. (2022).

2.3. Local Differential Privacy

In this paper, we consider the case where data arrive sequentially in a serial manner, and only one data point is acquired by each agent at each time instant, i.e., at time instant T, the dataset $\mathcal{D}_i$ accessible to agent i is given by $\mathcal{D}_i=\{\xi_{i,1},\cdots,\xi_{i,T}\}$. Then, we can introduce the following definitions for differential privacy:

Definition 2.4. (Adjacency) Given two local datasets $\mathcal{D}_i=\{\xi_{i,1},\cdots,\xi_{i,T}\}$ and $\mathcal{D}_i'=\{\xi_{i,1}',\cdots,\xi_{i,T}'\}$ for any $i\in[m]$ and any time $T\in\mathbb{N}$, $\mathcal{D}_i$ and $\mathcal{D}_i'$ are adjacent if there exists a time instant $k\in\{1,\cdots,T\}$ such that $\xi_{i,k}\neq\xi_{i,k}'$ while $\xi_{i,t}=\xi_{i,t}'$ for all $t\neq k,\ t\in\{1,\cdots,T\}$.

Definition 2.5. (Local Differential Privacy) Denote a DSBO algorithm as a mapping $\mathcal{A}_i(\mathcal{D}_i,x_{-i})\mapsto\mathcal{O}_i$, where x_{-i} denotes all messages received by agent i and $\mathcal{O}_i$ represents the set of all possible observations on agent i. Then, for any given $\epsilon_i>0$, we say that $\mathcal{A}_i$ is ϵ_i-locally differentially private if for any adjacent datasets $\mathcal{D}_i$ and $\mathcal{D}_i'$, the following inequality holds:

$$\mathbb{P}[\mathcal{A}_i(\mathcal{D}_i,x_{-i})\in\mathcal{O}_i]\leq e^{\epsilon_i}\mathbb{P}[\mathcal{A}_i(\mathcal{D}_i',x_{-i})\in\mathcal{O}_i]. \quad (6)$$

The parameter ϵ_i is referred to as the cumulative privacy budget for iterations $1,2,\cdots,T$. A smaller ϵ_i indicates closer distributions of observations under adjacent datasets, thereby a higher level of privacy protection. Clearly, if ϵ_i grows to infinity as the iteration number T tends to infinity, privacy will be lost eventually in the infinite-time horizon.

Different from the commonly used "centralized" differential-privacy framework where a data curator is needed to gather data and inject noise, in LDP, each agent acts as its own data curator and designs its noise independently of other agents (Bellet et al., 2018). Therefore, adjacent datasets in LDP allow variations in all agents' data at a single time instant. This is different from most existing differential privacy solutions for (decentralized) stochastic convex optimization (e.g., Zhang et al. (2018); Bassily et al. (2019); Lü et al. (2020); Asi et al. (2021); Altschuler & Talwar (2022); Huang et al. (2024)), which, at any time instant, only allow difference in one agent's dataset in adjacency definition.

3. The Proposed Algorithm

In this section, we first introduce an approach for individual agents to locally estimate Hessian-inverse-vector product under the constraint of differential privacy, which is necessary for individual agents to locally estimate the global

Algorithm 1 Subroutine for Estimating Hessian-Inverse-Vector Product for Agent i, $i \in [m]$

Input: Parameters $x_{i,t}$, $y_{i,t}$, and $z_{i,t}$; Data samples $\{\varphi_{i,k}\}_{k\in[0,t]}$ and $\{\xi_{i,t}\}_{k\in[0,t]}$; Stepsize $\lambda_{z,t} = \frac{\lambda_{z,0}}{(t+1)^{v_z}}$ with $\lambda_{z,0} > 0$ and $v_z \in (0,1)$; DP-noise $\vartheta_{i,t}$ satisfying Assumption 3.1.
$H_{i,t}z_{i,t} = \nabla^2_{yy} g_{i,t}(x_{i,t}, y_{i,t}) z_{i,t}$.
$b_{i,t} = \nabla_y f_{i,t}(x_{i,t}, y_{i,t})$.
$\nabla_z \phi_{i,t}(z_{i,t}) = H_{i,t}z_{i,t} - b_{i,t}$.
$z_{i,t+1} = z_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(z_{j,t} + \vartheta_{j,t} - z_{i,t}) - \lambda_{z,t}\nabla_z \phi_{i,t}(z_{i,t})$.
Output: $z_{i,t+1}$ on agent i.

hypergradient according to (5). Using it as a subroutine, we will then propose our differentially private DSBO algorithm.

Approximating z^* in (4) amounts to letting each agent solve for the following equation:

$$\sum_{i=1}^m H_i z^* = \sum_{i=1}^m b_i \text{ or } z^* \triangleq \left(\sum_{i=1}^m H_i\right)^{-1} \left(\sum_{i=1}^m b_i\right), \quad (7)$$

where H_i and b_i are given by $H_i = \nabla^2_{yy} g_i(x, y^*(x))$ and $b_i = \nabla_y f_i(x, y^*(x))$, respectively. Equality (7) is essentially the optimality condition of the following optimization problem:

$$\min_{z\in\mathbb{R}^q} \frac{1}{m}\sum_{i=1}^m \phi_i(z), \quad \phi_i(z) = \frac{1}{2} z^T H_i z - b_i^T z. \quad (8)$$

We present Algorithm 1 that enables all agents to collaboratively find the optimal solution z^* to problem (8).

Since objective functions f_i and g_i in problem (8) are expectations over unknown distributions (see the equations in (2)), they are inaccessible and can only be approximated from sampled data in practical implementations. Therefore, under our setting of serially arriving data, we use $f_{i,t}(x,y) = \frac{1}{t+1}\sum_{k=0}^t h(x,y;\varphi_{i,k})$ and $g_{i,t}(x,y) = \frac{1}{t+1}\sum_{k=0}^t l(x,y;\xi_{i,k})$.

Building on Algorithm 1, agent i can estimate the hypergradient $\nabla F(x)$ in (5) locally by using the following equality:

$$u_{i,t} = \nabla_x f_{i,t}(x_{i,t}, y_{i,t}) - \nabla^2_{xy} g_{i,t}(x_{i,t}, y_{i,t}) z_{i,t}. \quad (9)$$

With the hypergradient estimation (9), we propose a locally differentially private algorithm to solve the DSBO problem (1) in Algorithm 2. The injected DP noises satisfy the following assumption:

Assumption 3.1. For every $i \in [m]$ and $t \geq 0$, each element of DP-noise vectors $\chi_{i,t}$, $\zeta_{i,t}$, and $\vartheta_{i,t}$ follows Laplace distributions $\text{Lap}\left(\frac{\sigma_{i,x}}{\sqrt{2}(t+1)^{\varsigma_{i,x}}}\right)$, $\text{Lap}\left(\frac{\sigma_{i,y}}{\sqrt{2}(t+1)^{\varsigma_{i,y}}}\right)$, and $\text{Lap}\left(\frac{\sigma_{i,z}}{\sqrt{2}(t+1)^{\varsigma_{i,z}}}\right)$, respectively, where $\sigma_{i,x}$, $\sigma_{i,y}$, and $\sigma_{i,z}$ are positive constants and the rates of noise variances satisfy

$$\max_{i\in[m]}\{\varsigma_{i,x}\} < v_x, \ \max_{i\in[m]}\{\varsigma_{i,y}\} < v_y, \text{ and } \max_{i\in[m]}\{\varsigma_{i,z}\} < v_z,$$

where $v_x, v_y, v_z \in (0,1)$ are the decaying rates of the stepsizes $\lambda_{x,t}$, $\lambda_{y,t}$, and $\lambda_{z,t}$, respectively, in Algorithm 2.

Algorithm 2 LDP Design for DSBO Algorithm for Agent i, $i \in [m]$

Input: Random initialization $x_{i,0} \in \mathbb{R}^p$, $y_{i,0} \in \mathbb{R}^q$, and $z_{i,0} \in \mathbb{R}^q$ for each agent $i \in [m]$. Stepsizes $\lambda_{x,t} = \frac{\lambda_{x,0}}{(t+1)^{v_x}}$ and $\lambda_{y,t} = \frac{\lambda_{y,0}}{(t+1)^{v_y}}$ with $\lambda_{x,0} > 0$, $\lambda_{y,0} > 0$, and $v_x, v_y \in (0,1)$; DP-noises $\chi_{i,t}$ and $\zeta_{i,t}$ satisfying Assumption 3.1.
for $t = 0, 1, \cdots, T-1$ **do**
  Acquire current data $\varphi_{i,t}$ and $\xi_{i,t}$.
  $y_{i,t+1} = y_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(y_{j,t} + \zeta_{j,t} - y_{i,t}) - \lambda_{y,t}\nabla_y g_{i,t}(x_{i,t}, y_{i,t})$.
  Run Algorithm 1 and obtain the output $z_{i,t+1}$.
  Estimate hypergradient $u_{i,t}$ by using (9).
  $x_{i,t+1} = x_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(x_{j,t} + \chi_{j,t} - x_{i,t}) - \lambda_{x,t}u_{i,t}$.
end for
Output: $x_{i,T}$ on agent i.

It is worth noting that different from existing DSBO algorithms in Chen et al. (2022), Yang et al. (2022), and Gao et al. (2023) which estimate the full Hessian matrix or Jacobian matrix, Algorithm 2 only estimates a vector of dimension $\max\{p, q\}$, and hence has reduced computational complexity. In addition, different from existing DSBO algorithms in Chen et al. (2022) and Chen et al. (2023) which use a nested communication (consensus) loop to estimate z^*, Algorithm 2 avoids any nested-loops of consensus operations. The avoidance of nested consensus loops is significant in that under nested-loops of consensus iterations, the cumulative privacy budget will grow quickly as iteration proceeds, making it impossible to ensure a finite cumulative privacy budget in the infinite-time horizon (see detailed explanations in Appendix H.1).

4. Main Results

4.1. Convergence Rate of Algorithm 2

Theorem 4.1. *Denote the lowest decaying rates of DP-noise variances as* $\varsigma_x = \min_{i\in[m]}\{\varsigma_{i,x}\}$, $\varsigma_y = \min_{i\in[m]}\{\varsigma_{i,y}\}$, *and* $\varsigma_z = \min_{i\in[m]}\{\varsigma_{i,z}\}$. *Under Assumptions 2.1-2.3, and 3.1, if the stepsize rates satisfy* $0 < v_z < v_y < v_x < 1$, *then we have the following results for the iterates* $\{x_i\}$ *generated by Algorithm 2:*

(i) If $F(x)$ *is strongly convex and the rates of DP-noise*

variances satisfy $2\varsigma_x > v_x$, $2\varsigma_x > v_z + v_y$, $2\varsigma_y > v_z + v_y$, *and* $2\varsigma_z > v_y$, *then we have*

$$\mathbb{E}\left[\|x_{i,T} - x^*\|^2\right] \leq \mathcal{O}\left(T^{-\beta_1}\right), \tag{10}$$

where the rate β_1 *is given by* $\beta_1 = \min\{2\varsigma_x - v_x, 2\varsigma_x - 2v_z, 2\varsigma_y - 2v_z, 2\varsigma_z - v_z, 2\varsigma_y - v_y, 2 - 2v_y\}$.

(ii) If F *is convex and the rates of DP-noise variances satisfy* $\varsigma_x > \frac{1}{2}$, $2\varsigma_x > v_z + v_y$, $2\varsigma_x > 2v_z + 2 - 2v_x$, $2\varsigma_y > v_z + v_y$, $2\varsigma_y > 2v_z + 2 - 2v_x$, $2\varsigma_y > v_y + 2 - 2v_x$, $2\varsigma_z > v_z + 2 - 2v_x$, *and* $2\varsigma_z > v_y$, *then we have*

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[F(x_{i,t}) - F(x^*)\right] \leq \mathcal{O}\left(T^{-(1-v_x)}\right). \tag{11}$$

(iii) If F *is nonconvex and the rates of DP-noise variances satisfy* $\varsigma_x > \frac{1}{2}$, $2\varsigma_x > v_z + v_y$, $2\varsigma_x > 2v_z + 1 - v_x$, $2\varsigma_y > 2v_z + 1 - v_x$, $2\varsigma_y > v_y + 1 - v_x$, $2\varsigma_y > v_z + v_y$, $2\varsigma_z > v_z + 1 - v_x$, *and* $2\varsigma_z > v_y$, *then we have*

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right] \leq \mathcal{O}\left(T^{-(1-v_x)}\right). \tag{12}$$

Theorem 4.1 proves that the optimization errors for strongly convex, convex, and nonconvex $F(x)$ decrease with iterations at rates $\mathcal{O}\left(T^{-\beta_1}\right)$, $\mathcal{O}\left(T^{-(1-v_x)}\right)$, and $\mathcal{O}\left(T^{-(1-v_x)}\right)$, respectively.

Moreover, to give a more intuitive description of the computational complexity, we define a δ-solution to problem (1):

Definition 4.2. (Lian et al., 2017) For any $i \in [m]$ and some positive integer T, if $\mathbb{E}\left[\|x_{i,T} - x^*\|^2\right] \leq \delta$ holds when F is strongly convex, or $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[F(x_{i,t}) - F(x^*)\right] \leq \delta$ holds when F is convex, or $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right] \leq \delta$ holds when F is nonconvex, then we say that the sequence $\{x_{i,t}\}_{t=0}^{T}$ can reach a δ-solution to problem (1).

Definition 4.2 provides a direct quantitative measure of the optimization error with respect to the optimal solution x^* under strongly convex F. This measure is stronger than the metrics in Ghadimi & Wang (2018) and Yang et al. (2022) that characterize the distance between $F(\bar{x}_T)$ and $F(x^*)$. Moreover, in the nonconvex case, compared with Chen et al. (2023), which uses the minimum hypergradient over all iterations (i.e., $\min_{0<t<T}\mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right] \leq \delta$), Definition 4.2 is much more stringent.

Corollary 4.3. *(i) For a strongly convex* $F(x)$, *if we choose* $T = \mathcal{O}(\delta^{-\frac{1}{\beta_1}})$, *then the computational complexity of Algorithm 2 is* $\mathcal{O}(\max\{p,q\}\delta^{-\frac{1}{\beta_1}})$ *in finding a* δ*-solution. For example, setting* $v_x = 0.66$, $v_y = 0.64$, $v_z = 0.43$, $\varsigma_x = 0.65$, $\varsigma_y = 0.63$, *and* $\varsigma_z = 0.42$ *yields* $\beta_1 = 0.4$ *and a computational complexity of* $\mathcal{O}(\max\{p,q\}\delta^{-2.5})$.
(ii) For a convex $F(x)$, *if we set* $T = \mathcal{O}\big(\delta^{-\frac{1}{1-v_x}}\big)$, *then the computational complexity of Algorithm 2 is* $\mathcal{O}\big(\max\{p,q\}\delta^{-\frac{1}{1-v_x}}\big)$ *in finding a* δ*-solution. For example, with* $v_x = 0.77$, $v_y = 0.75$, $v_z = 0.5$, $\varsigma_x = 0.76$, $\varsigma_y = 0.74$, *and* $\varsigma_z = 0.49$, *we have* $1 - v_x = 0.23$ *and hence a computational complexity of* $\mathcal{O}(\max\{p,q\}\delta^{-4.35})$.
(iii) For a nonconvex $F(x)$, *if we choose* $T = \mathcal{O}\big(\delta^{-\frac{1}{1-v_x}}\big)$, *then the computational complexity of Algorithm 2 is* $\mathcal{O}\big(\max\{p,q\}\delta^{-\frac{1}{1-v_x}}\big)$ *in finding a* δ*-solution. For example, using* $v_x = 0.615$, $v_y = 0.60375$, $v_z = 0.4$, $\varsigma_x = 0.61125$, $\varsigma_y = 0.6$, *and* $\varsigma_z = 0.398125$ *yields* $1 - v_x = 0.385$ *and a computational complexity of* $\mathcal{O}(\max\{p,q\}\delta^{-2.6})$.

Corollary 4.3 provides computational complexities under different convexity assumptions. It is more comprehensive than existing DSBO results (Chen et al., 2022; Gao et al., 2023; Chen et al., 2023), which only focus on a nonconvex function $F(x)$. Moreover, it is worth noting that compared with the computational complexity of $\mathcal{O}(pq\log(\delta^{-1})\delta^{-3})$ in Chen et al. (2022), our Algorithm 2 ensures an improved computational complexity of $\mathcal{O}(\max\{p,q\}\delta^{-2.6})$, even under the additional constraint of differential privacy.

4.2. Differential Privacy Analysis for Algorithm 2

In this subsection, we prove that besides accurate convergence, Algorithm 2 can simultaneously ensure rigorous ϵ_i-LDP for each agent, with a finite cumulative privacy budget even when the number of iterations tends to infinity.

Assumption 4.4. Functions ∇h, ∇l, and $\nabla^2 l$ are $L_{h,1}$, $L_{l,1}$, and $L_{l,2}$ Lipschitz continuous, respectively. Moreover, there exist some positive constants c_{h0} and c_{l0} such that $\|\nabla_y h(x,y;\varphi_i)\|_1 \leq c_{h0}$ and $\|\nabla_y l(x,y;\xi_i)\|_1 \leq c_{l0}$ hold for all $i \in [m]$.

Assumption 4.4 is commonly used in differential-privacy design for decentralized learning/optimization (Huang et al., 2015; Bellet et al., 2018; Cyffers et al., 2022; Bietti et al., 2022). Although it is stricter than Assumption 2.2 (which assumes Lipschitz continuity of the gradients of expected functions f_i and g_i), it is not required in our convergence analysis. In fact, existing DSBO results (Chen et al., 2022; Yang et al., 2022; Gao et al., 2023; Chen et al., 2023) often do not clearly differentiate between Assumption 2.2 and Assumption 4.4, and usually assume Lipschitz continuity of loss functions h and l and their first-and second-order moments, similar to Assumption 4.4 (see e.g., Assumptions 3.3 and 3.4 in Yang et al. (2022) and Assumption 2.1 in Chen et al. (2023)).

Theorem 4.5. *Under Assumptions 2.1 and 4.4, if each element of* $\chi_{i,t}$, $\zeta_{i,t}$, *and* $\vartheta_{i,t}$ *follows the Laplace distributions given in Assumption 3.1, then* $x_{i,t}$ *(resp.* $F(x_{i,t})$ *and* $\nabla F(x_{i,t})$ *in the general convex case and nonconvex case, respectively) in Algorithm 2 converges in mean square to the optimal solution* x^* *to problem* (1) *(resp. in mean to* $F(x^*)$

and in mean square to zero, respectively). Furthermore,

(i) For any $T \in \mathbb{N}^+$, agent i's implementation of Algorithm 2 is locally differentially private with a cumulative privacy budget bounded by $\epsilon_i = \epsilon_{i,x} + \epsilon_{i,y} + \epsilon_{i,z}$, where $\epsilon_{i,x}$, $\epsilon_{i,y}$, and $\epsilon_{i,z}$ are given by $\epsilon_{i,x} \leq \sum_{t=1}^{T} \frac{\sqrt{2}C_{\epsilon x}}{\sigma_{i,x}(t+1)^{1+v_x-\varsigma_x}}$, $\epsilon_{i,y} \leq \sum_{t=1}^{T} \frac{\sqrt{2}C_{\epsilon y}}{\sigma_{i,y}(t+1)^{1+v_y-\varsigma_y}}$, and $\epsilon_{i,z} \leq \sum_{t=1}^{T} \frac{\sqrt{2}C_{\epsilon z}}{\sigma_{i,z}(t+1)^{1+v_z-\varsigma_z}}$ with $\bar{w} = \min_{i\in[m]}\{|w_{ii}|\}$, $C_{\epsilon x} = \frac{4}{\bar{w}}\left(\frac{4(1+v_x)}{e\ln\left(\frac{4}{2-2\bar{w}}\right)}\right)^{1+v_x}$, $C_{\epsilon y} = \frac{4}{\bar{w}}\left(\frac{4(1+v_y)}{e\ln\left(\frac{4}{2-2\bar{w}}\right)}\right)^{1+v_y}$, $C_{\epsilon z} = \frac{4}{\bar{w}}\left(\frac{4(1+v_z)}{e\ln\left(\frac{4}{2-2\bar{w}}\right)}\right)^{1+v_z}$.

(ii) The cumulative privacy budget ϵ_i is finite even when the number of iterations T tends to infinity.

Theorem 4.5 shows that Algorithm 2 can ensure rigorous ϵ_i-LDP and accurate convergence simultaneously. This differs from most existing differential-privacy solutions for decentralized single-level optimization (e.g., Bellet et al. (2018); Cyffers et al. (2022); Bietti et al. (2022)), which have to trade convergence accuracy for differential privacy.

A key reason for Algorithm 2 to ensure non-diminishing privacy protection using diminishing noise variances is that our algorithm design leads to a diminishing sensitivity. Specifically, Lemma 2 in Huang et al. (2015) proves that when $\sum_{t=1}^{\infty} \frac{\Delta}{\nu_t} \leq \epsilon$ (where Δ is the sensitivity and ν_t is the parameter of Laplacian distribution $\text{Lap}(\nu_t)$) is satisfied, the iterative algorithm is ϵ-differentially private. According to (182)-(184) in the supplemental material, the sensitivities $\Delta_{i,t,x} \leq \mathcal{O}(t^{-(1+v_x)})$, $\Delta_{i,t,y} \leq \mathcal{O}(t^{-(1+v_y)})$, and $\Delta_{i,t,z} \leq \mathcal{O}(t^{-(1+v_z)})$ of Algorithm 2 decay faster than the Laplacian variances $\nu_{i,t,x} \leq \mathcal{O}(t^{-\varsigma_x})$, $\nu_{i,t,y} \leq \mathcal{O}(t^{-\varsigma_y})$ and $\nu_{i,t,z} \leq \mathcal{O}(t^{-\varsigma_z})$, which ensures $\sum_{t=1}^{\infty}\left(\frac{\Delta_{i,t,x}}{\nu_{i,t,x}} + \frac{\Delta_{i,t,y}}{\nu_{i,t,y}} + \frac{\Delta_{i,t,z}}{\nu_{i,t,z}}\right) \leq \epsilon_i < \infty$.

Remark 4.6. We would like to point out that the accurate convergence of Algorithm 2 does not conflict with the constraints of differential privacy. More specifically, according to the differential-privacy theory, conventional query mechanisms on a dataset can only achieve differential privacy by sacrificing query accuracies, but the considered stochastic optimization algorithm does not correspond to a simple query mechanism on the optimal solution. Instead, what are queried in stochastic optimization (machine learning) are input data, and revealing the precise optimal solution is not equivalent to revealing accurate input data (which is the actual query target).

In fact, the achievement of rigorous LDP and accurate convergence of Algorithm 2 comes at the expense of a reduced convergence rate. We use the convergence rate and cumulative privacy budget under a nonconvex $F(x)$ as an example to quantify this tradeoff:

Corollary 4.7. *For any given cumulative privacy budget $\epsilon_i > 0$, $i \in [m]$, the convergence rate of Algorithm 2 is $\mathcal{O}\left(\frac{T^{-(1-v_x)}}{\min_{i\in[m]}\{\epsilon_i^2\}}\right)$ with $v_x \in (0.6, 1)$.*

Corollary 4.7 indicates that a higher level of differential privacy, i.e., a smaller cumulative privacy budget ϵ_i, corresponds to a reduced convergence rate.

Although the result of convergence rate in Corollary 4.7 appears inferior to the one achieved in Bassily et al. (2019) for centralized single-level stochastic optimization under the constraint of differential privacy, we would like to emphasize that this difference is caused by the increased complexity of decentralized bilevel optimization over centralized single-level optimization. In fact, when we only consider the lower-level optimization part in our algorithm, where our bilevel optimization problem reduces to single-level optimization, we can prove that our algorithm has exactly the same order of convergence rate as that in Bassily et al. (2019) (we summarize the result for this special lower-level only case of our Algorithm 2 in Appendix G.3).

5. Experiments

In this section, we study the applications of Algorithm 2 in both hyperparameter optimization and meta learning. In each experiment, we compared Algorithm 2 with state-of-the-art DSBO algorithms, including MA-DSBO (Chen et al., 2023) and GBDSBO (Yang et al., 2022). The interaction pattern contains 10 agents connected in a circle, where each agent can only communicate with its two immediate neighbors. For the weight matrix W, we set $w_{ij} = 0.3$ if agents i and j are neighbors, and $w_{ij} = 0$ otherwise. Due to space limitation, we leave detailed experimental setups in Appendix A.1.

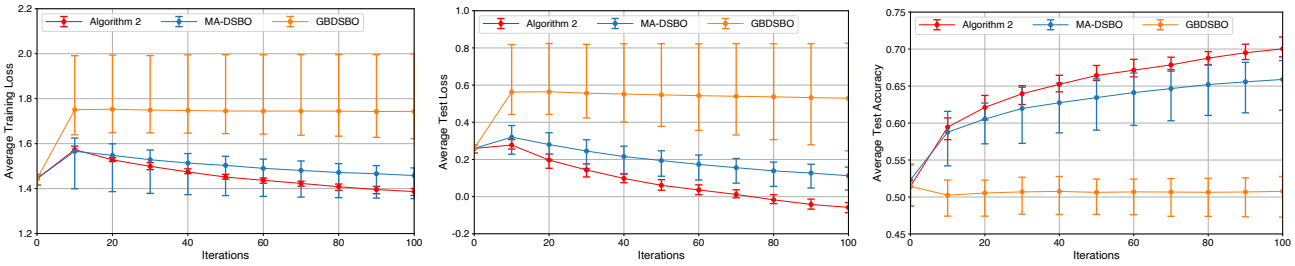
To evaluate the convergence performance of Algorithm 2 in the absence of differential-privacy constraints, we also conducted experiments without Laplacian noises, with the results given in Appendix A.2.1. Furthermore, we provided comparison results with VRDSBO in Gao et al. (2023) (which only addresses the special case of $g_1 = \cdots = g_m$) in Appendix A.2.2. In addition, we tested the efficacy of our Algorithm 2 on various network topologies and with diverse heterogeneous data distributions, with the respective results given in Appendix A.2.3 and Appendix A.2.4.

5.1. Hyperparameter Optimization

The objective of a hyperparameter optimization problem can be formulated as follows:

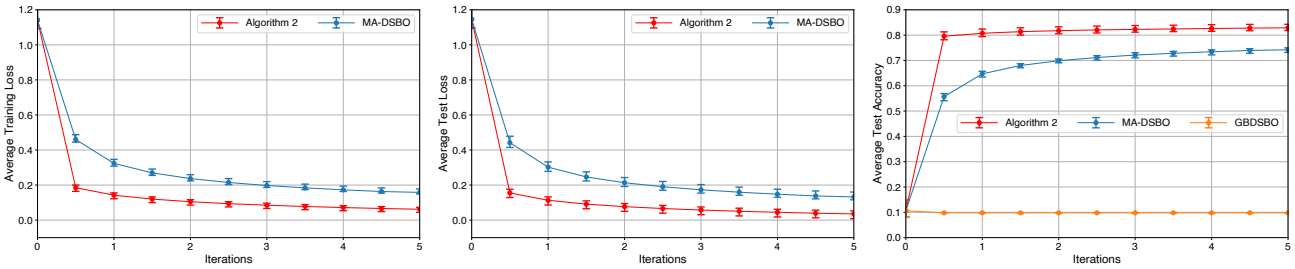
$$\min_{\lambda\in\mathbb{R}^p} \quad \frac{1}{m}\sum_{i=1}^{m} f_i(\lambda, \omega^*(\lambda)),$$

$$\text{s.t.} \quad \omega^*(\lambda) = \text{argmin}_{\omega\in\mathbb{R}^q} \frac{1}{m}\sum_{i=1}^{m} g_i(\lambda, \omega),$$



(a) Training loss (b) Test loss (c) Test accuracy

Figure 1. Comparison by using the synthetic dataset under differential-privacy constraints.



(a) Training loss (b) Test loss (c) Test accuracy

Figure 2. Comparison by using the "MNIST" dataset under differential-privacy constraints

in which we aim to find an optimal hyperparameter λ under the constraint that $\omega^*(\lambda)$ is the optimal model parameter with a given λ. We conducted experiments on both synthetic and real-world datasets. The detailed experimental setup is given in Appendix A.1.

Synthetic Data Following Chen et al. (2023), we define loss functions for each agent i as follows:

$$h(\lambda, \omega; \varphi_i) = \sum_{(x_{i,e}, y_{i,e}) \in \mathcal{D}_{i,t}^h} L(y_{i,e} x_{i,e}^T \omega),$$

$$l(\lambda, \omega; \xi_i) = \sum_{(x_{i,e}, y_{i,e}) \in \mathcal{D}_{i,t}^l} L(y_{i,e} x_{i,e}^T \omega) + \frac{1}{2} \sum_{s=1}^{200} e^{\lambda_s} \omega_s^2,$$

where λ_s and ω_s represent the s-th element of $\lambda \in \mathbb{R}^{200}$ and $\omega \in \mathbb{R}^{200}$, respectively. The function $L(\cdot)$ is given by $L(x) = \log(1 + e^{-x})$. $\mathcal{D}_{i,t}^l$ and $\mathcal{D}_{i,t}^h$ represent the training dataset and the validation dataset for agent i, at time t, respectively. For each agent i, the data distribution of $x_{i,e}$ was drawn from a normal distribution $\mathcal{N}(0, i^2)$, which is heterogeneous due to the difference in variances. The label y_e was generated by $y_{i,e} = x_{i,e}^T \omega + 0.1\varepsilon$, where $\varepsilon \in \mathbb{R}^{200}$ denotes the noise vector sampled from the standard normal distribution. The algorithm was executed for 100 iterations, with each agent randomly selecting 50 training samples in every iteration. The test dataset contains $20,000$ samples, with $1,000$ samples randomly selected for each iteration.

The resulting training loss, test loss, and test accuracy are shown in Figures 1(a), 1(b), and 1(c), respectively. It is clear that the proposed algorithm has much lower training loss and higher test accuracy under differential-privacy constraints.

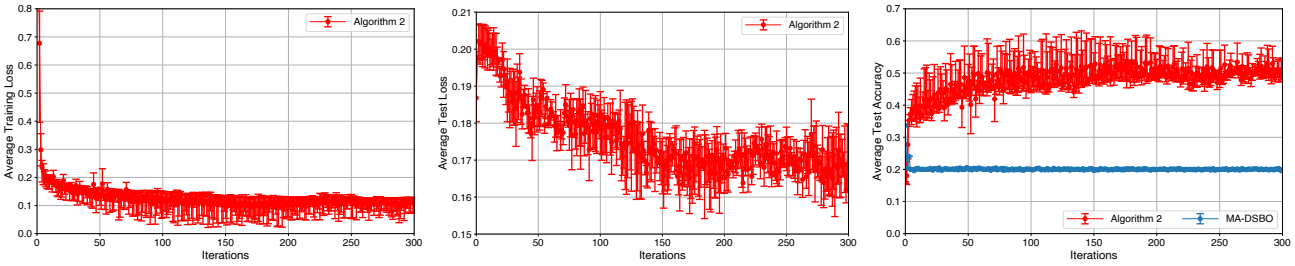
MNIST We evaluated the performance of Algorithm 2 by using the "MNIST" dataset (Grazzi et al., 2020). The loss functions are defined as follows:

$$h(\lambda, \omega; \varphi_{i,t}) = \frac{1}{|\mathcal{D}_{i,t}^h|} \sum_{(x_{i,e}, y_{i,e}) \in \mathcal{D}_{i,t}^h} L(x_{i,e}^T \omega, y_{i,e}),$$

$$l(\lambda, \omega; \xi_{i,t}) = \frac{1}{|\mathcal{D}_{i,t}^l|} \sum_{(x_{i,e}, y_{i,e}) \in \mathcal{D}_{i,t}^l} L(x_{i,e}^T \omega, y_{i,e}) + \frac{1}{cp} \sum_{r=1}^{c} \sum_{s=1}^{p} e^{\lambda_s} \omega_{rs},$$

where $c = 10$ and $p = 784$ represent the number of classes and features, respectively. λ_s denotes the s-th element of the hyperparameter $\lambda \in \mathbb{R}^p$ and ω_{rs} denotes the element in the r-row and s-column of the model parameter $\omega \in \mathbb{R}^{c \times p}$. The function $L(\cdot)$ is used to calculate the cross entropy loss.

Figure 2 shows the results of our Algorithm 2, GBDSBO and MA-DSBO. It is worth noting that GBDSBO failed to train in this case (both its training and testing losses tended to infinity), and hence we omitted its values in Figure 2(a) and Figure 2(b). The results once again confirm the effectiveness



(a) Training loss (b) Test loss (c) Test accuracy

Figure 3. Comparison by using the "CIFAR-10" dataset under differential-privacy constraints

of our algorithm under differential-privacy constraints.

5.2. Decentralized Meta Learning

Following Finn et al. (2017), we consider a meta-learning problem with N tasks $\{\mathcal{T}_q, q = 1, \cdots, N\}$. Each task $\mathcal{T}_q$ has a loss function $L(x, y_q; \xi)$ over each data sample ξ, where x is the parameter of an embedding model shared by all tasks, and y_q is the task-specific parameter. The goal of meta learning is to find an optimal parameter x^* that benefits all tasks, and building on this parameter x^*, the model can quickly adapt its own parameter y_q to any new task $\mathcal{T}_q$ using only a few data points and training iterations. The meta learning problem is essentially a bilevel optimization problem. In the lower level, given a parameter x, the base-learner of task $\mathcal{T}_q$ searches y_q^* as the minimizer of its loss over a training dataset $\mathcal{D}_q^{\text{tra}}$. In the upper level, the meta-learner evaluates the minimizers y_q^* on a validation dataset $\mathcal{D}_q^{\text{val}}$. Let $\boldsymbol{y}^* = \text{col}\{y_1^*, \cdots, y_N^*\}$ be all task-specific optimal parameters.

Different from the conventional setting of centralized meta learning, in decentralized training for a meta-learning task $\mathcal{T}_q$, training and validation data are distributed across various devices (called agents). Therefore, we allocate each agent $i \in [m]$ with its own local datasets for task $\mathcal{T}_q$, including both a local training dataset $\mathcal{D}_{i,q}^{\text{tra}}$ and a validation dataset $\mathcal{D}_{i,q}^{\text{val}}$. Then, in the lower-level optimization, building on a given parameter x, m base-learners associated with task $\mathcal{T}_q$ cooperatively search for y_q^*, while in the upper-level optimization, m meta-learners cooperatively find an optimal parameter x^*. The decentralized meta learning problem can be formulated as follows:

$$\begin{aligned} \min_{x} \quad & F(x) := \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{q=1}^{N} f_{i,q}(x, y_q^*(x)), \\ \text{s.t.} \quad & \boldsymbol{y}^*(x) := \text{argmin}_{y_q} \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{q=1}^{N} g_{i,q}(x, y_q), \end{aligned} \tag{13}$$

where $f_{i,q}(x, y_q^*(x)) = \frac{1}{|\mathcal{D}_{i,q}^{\text{val}}|}\sum_{\xi_{i,q}\in\mathcal{D}_{i,q}^{\text{val}}} L(x, y_q^*(x); \xi_{i,q})$ and $g_{i,q}(x, y) = \frac{1}{|\mathcal{D}_{i,q}^{\text{tra}}|}\sum_{\xi_{i,q}\in\mathcal{D}_{i,q}^{\text{tra}}} L(x, y_q; \xi_{i,q}) + R_{i,x}(y_q)$ with $R_{i,x}(y)$ denoting a strongly-convex regularizer w.r.t y.

In this experiment, we evaluated the performance of Algorithm 2 on the "CIFAR-10" dataset (Krizhevsky et al., 2010). The detailed experimental setup is given in Appendix A.1.

From Figure 3, we can see that MA-DSBO fails to train in the meta learning task, whereas our Algorithm 2 has much better training and test accuracies. Moreover, since the training loss and test loss of MA-DSBO tended to infinity, they were not plotted in Figures 3-(a) and 3-(b). Note that we did not show the results for the GBDSBO algorithm in Yang et al. (2022) since its large overhead in computing and communicating the full Hessian and Jacobian matrices prohibited training under the large model and dataset.

6. Conclusions

In this paper, we proposed a decentralized stochastic bilevel-optimization algorithm that can simultaneously ensure both accurate convergence and rigorous differential privacy. This is significant because even for the simpler problem of single-level decentralized optimization/learning, existing differential-privacy solutions have to sacrifice convergence accuracy for privacy. Lying at the core of our approach is a new algorithm for decentralized stochastic bilevel optimization that avoids any nested-loops of consensus (communication) iterations. This is important since all existing decentralized algorithms for bilevel optimization rely on nested-loops of consensus iterations, which, unfortunately, constitutes an obstacle for achieving differential privacy because the intensive consensus operations lead to an exploding cumulative privacy budget. We systematically characterized the convergence performance of our algorithm under both nonconvex and convex objective functions, and quantified the price and tradeoff in the convergence rate. Experimental results on practical machine learning models confirm the efficacy of our algorithm.

Acknowledgements

We greatly appreciate the Area Chair's and Reviewers' time and effort in handling/reviewing our paper. We also thank the National Science Foundation for supporting this work under Grants ECCS-1912702, CCF-2106293, CCF-2215088, CNS-2219487, and CCF 2334449.

Impact Statement

This paper presents work whose goal is to advance the field of differentially private optimization and learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Altschuler, J. and Talwar, K. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *Advances in Neural Information Processing Systems*, 35:3788–3800, 2022.

Arora, S., Du, S., Kakade, S., Luo, Y., and Saunshi, N. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pp. 367–376. PMLR, 2020.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32: 11282–11291, 2019.

Bellet, A., Guerraoui, R., Taziki, M., and Tommasi, M. Personalized and private peer-to-peer machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 473–481. PMLR, 2018.

Bertinetto, L., Henriques, J., Torr, P., and Vedaldi, A. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Bietti, A., Wei, C.-Y., Dudik, M., Langford, J., and Wu, S. Personalization improves privacy-accuracy tradeoffs in federated learning. In *International Conference on Machine Learning*, pp. 1945–1962. PMLR, 2022.

Bollobás, B. *Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability*. Cambridge University Press, 1986.

Bracken, J. and McGill, J. T. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

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

Chen, X., Huang, M., Ma, S., and Balasubramanian, K. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International Conference on Machine Learning*, pp. 4641–4671. PMLR, 2023.

Chen, Z. and Wang, Y. Locally differentially private gradient tracking for distributed online learning over directed graphs. *arXiv preprint arXiv:2310.16105*, 2023.

Couellan, N. and Wang, W. On the convergence of stochastic bi-level gradient methods. *Optimization*, pp. 13833, 2016.

Cummings, R., Kaptchuk, G., and Redmiles, E. M. "I need a better description": an investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3037–3052, 2021.

Cyffers, E., Even, M., Bellet, A., and Massoulié, L. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. *Advances in Neural Information Processing Systems*, 35:15889–15902, 2022.

Dong, Y., Ma, S., Yang, J., and Yin, C. A single-loop algorithm for decentralized bilevel optimization. *arXiv preprint arXiv:2311.08945*, 2023.

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Differential privacy under continual observation. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, pp. 715–724, 2010.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

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. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *International Conference on Artificial Intelligence and Statistics*, pp. 9238–9281. PMLR, 2023.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

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

Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., and Zhang, C. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34: 27131–27145, 2021.

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.

Hansen, P., Jaumard, B., and Savard, G. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing*, 13(5):1194–1217, 1992.

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

Huang, L., Wu, J., Shi, D., Dey, S., and Shi, L. Differential privacy in distributed optimization with gradient tracking. *IEEE Transactions on Automatic Control*, 2024.

Huang, Z., Mitra, S., and Vaidya, N. Differentially private distributed optimization. In *Proceedings of the 16th International Conference on Distributed Computing and Networking*, pp. 1–10, 2015.

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.

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR, 2021.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

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*, 34:30271–30283, 2021.

Kong, B., Zhu, S., Lu, S., Huang, X., and Yuan, K. Decentralized bilevel optimization over graphs: Loopless algorithmic update and transient iteration complexity. *arXiv preprint arXiv:2402.03167*, 2024.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research), 2010.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30:5330–5340, 2017.

Liu, H., Simonyan, K., and Yang, Y. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2018.

Lü, Q., Liao, X., Xiang, T., Li, H., and Huang, T. Privacy masking stochastic subgradient-push algorithm for distributed online optimization. *IEEE Transactions on Cybernetics*, 51(6):3224–3237, 2020.

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*, pp. 5543–5547. IEEE, 2022.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, U. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*, 32:113–124, 2019.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Triastcyn, A. and Faltings, B. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*, pp. 9583–9592. PMLR, 2020.

Wang, Y. and Nedić, A. Tailoring gradient methods for differentially-private distributed optimization. *IEEE Transactions on Automatic Control (Early Access)*, 2023.

Yang, S., Zhang, X., and Wang, M. Decentralized gossip-based stochastic bilevel optimization over communication networks. *Advances in Neural Information Processing Systems*, 35:238–252, 2022.

Zhang, X., Khalili, M. M., and Liu, M. Improving the privacy and accuracy of ADMM-based distributed algorithms. In *International Conference on Machine Learning*, pp. 5796–5805. PMLR, 2018.

Zhang, X., Chen, X., Hong, M., Wu, Z. S., and Yi, J. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning*, pp. 26048–26067. PMLR, 2022.

Zhang, Y., Thai, M. T., Wu, J., and Gao, H. On the communication complexity of decentralized bilevel optimization. *arXiv preprint arXiv:2311.11342*, 2023.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32: 14774–14784, 2019.

Outline

- Section A: Experimental setups and additional experimental results
  - A.1 Experimental setups
  - A.2 Additional experiment results
- Section B: Notations and auxiliary lemmas
  - B.1 Additional notations
  - B.2 Auxiliary lemmas
- Section C: Empirical risk minimization problems and useful properties of empirical functions
  - C.1 ERM problem with respect to problem (1)
  - C.2 ERM problem with respect to problem (8)
- Section D: Results of Algorithm 2
  - D.1-D.10 Technical lemmas for consensus errors
  - D.11 Technical lemmas for the estimation error on the hypergradient
- Section E: Proof of Theorem 4.1
  - E.1 Proof for a strongly convex upper-level function
  - E.2 Proof for a convex upper-level function
  - E.3 Proof for a nonconvex upper-level function
- Section F: Proof of Theorem 4.5
- Section G: Proofs of Corollary 4.3 and Corollary 4.7, as well as further discussion
- Section H: The reason why existing DSBO algorithms cannot ensure rigorous ϵ_i-local differential privacy
  - H.1 The limitation of existing DSBO algorithms under differential-privacy constraints
  - H.2 The calculations of the cumulative privacy budget for the algorithms listed in Table 1

The structure of main proofs is given in Figure 4. Given that auxiliary lemmas from Sections B and C are utilized in several lemmas, they are omitted from this figure.

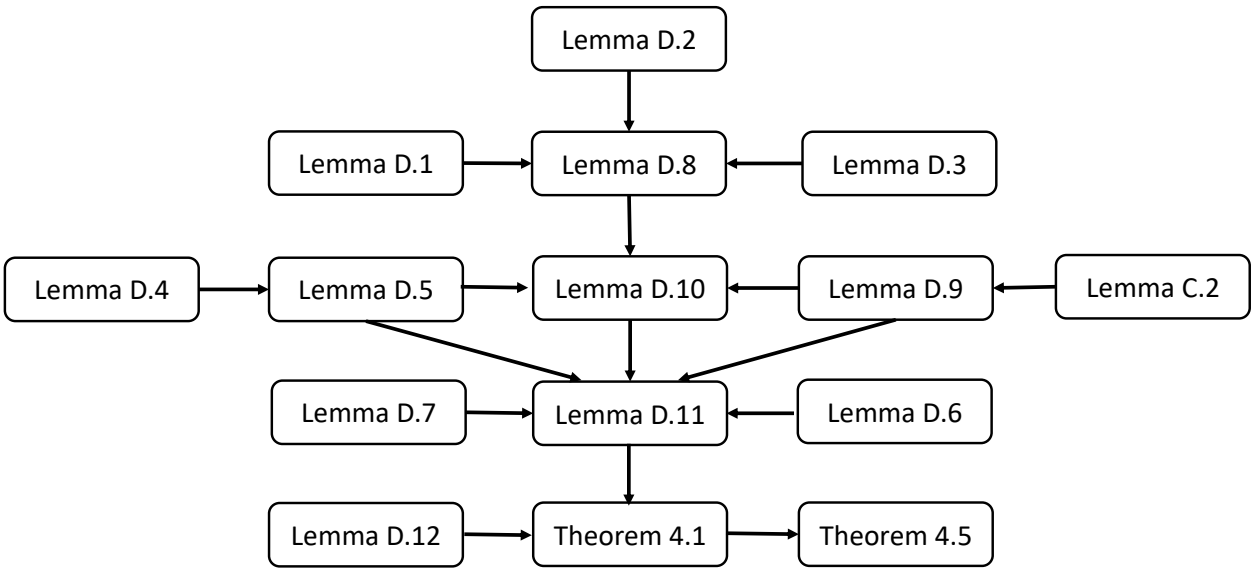


Figure 4. Structure of proofs.

A. Experimental Setups and Additional Experiments

A.1. Experimental Setups

Synthetic data In the synthetic-data experiment of Section 5.1, the stepsizes for Algorithm 2 were set to $\lambda_{x,t} = \frac{0.05}{(t+1)^{0.95}}$, $\lambda_{y,t} = \frac{0.05}{(t+1)^{0.87}}$, and $\lambda_{z,t} = \frac{0.02}{(t+1)^{0.75}}$. Each element of DP-noise vectors $\chi_{i,t}$, $\zeta_{i,t}$, and $\vartheta_{i,t}$ for agent i follows Laplace distributions $\text{Lap}\big(\frac{1}{\sqrt{2}(t+1)^{0.8+0.01i}}\big)$, $\text{Lap}\big(\frac{1}{\sqrt{2}(t+1)^{0.76+0.01i}}\big)$, and $\text{Lap}\big(\frac{1}{\sqrt{2}(t+1)^{0.6+0.01i}}\big)$, respectively. In our comparison, near-optimal stepsizes were selected for MA-DSBO and GBDSBO, ensuring that doubling these stepsizes would lead to non-converging behaviors. The number of nested-loops for MA-DSBO and GBDSBO was set to 10. We applied the fastest decaying DP-noise variance $\text{Lap}\big(\frac{1}{\sqrt{2}(t+1)^{0.8+0.01i}}\big)$ to MA-DSBO and GBDSBO, as using a slower decaying DP noise to make their privacy budget the same as ours results in divergence of both algorithms (this gives them an edge in accuracy comparison).

MNIST In the "MNIST" experiments of Section 5.1, both the training and validation datasets consist of $60,000$ images, and the test dataset contains $10,000$ images. We assigned 50% of the data from the i-th class to agent i, while the remaining 50% of the data is split evenly among the other agents. In each iteration, 50 images (forming a mini-batch of size 50) are randomly sampled from each agent's local training dataset. The training is conducted over 6 epochs. For Algorithm 2, the stepsizes were set to $\lambda_{x,t} = \frac{1.2}{(t+1)^{0.95}}$, $\lambda_{y,t} = \frac{1.2}{(t+1)^{0.87}}$, and $\lambda_{z,t} = \frac{1.2}{(t+1)^{0.75}}$. All other parameters were the same as those employed in the previous synthetic-data experiment.

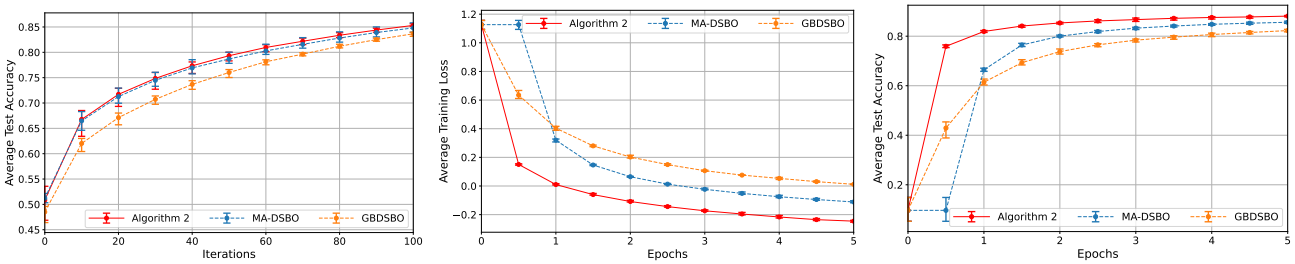
Meta learning on the "CIFAR-10" dataset In the decentralized meta learning experiment of Section 5.2, all agents were equipped with the same convolutional neural network architecture given in the MAML framework (Finn et al., 2017). In this specific application, y corresponds to the parameter of the last linear layer of the neural network and x corresponds to the parameter of the remaining layers. This setup ensures that the lower-level objective function $g_{i,q}(x,y)$ is strongly-convex w.r.t y while the upper-level objective function $F(x)$ is generally nonconvex w.r.t x. The algorithm was executed over 32 batches of tasks over 1000 iterations, with each task involving a training dataset $\mathcal{D}_q^{\text{tra}}$ and a validation dataset $\mathcal{D}_q^{\text{val}}$, both designed for 5-way classification with 50-shot for each class. Different from conventional centralized meta learning, the training data and validation data were distributed among different agents for cooperative learning in each task. We considered heterogeneous distribution, in which 30% of the data from the i-th class was assigned to agent i, while the remaining 70% was evenly distributed among the other agents. Note that heterogeneous distribution is particularly likely to happen in the decentralized learning setting since the data are collected by multiple agents from multiple sources. In this experiment, we also compared our Algorithm 2 with the nested-loop-based decentralized bilevel optimization algorithm (MA-DSBO) in (Chen et al., 2023). For Algorithm 2, the stepsizes were set to $\lambda_{x,t} = \frac{1}{(k+1)^{0.53}}$, $\lambda_{y,t} = \frac{1}{(k+1)^{0.52}}$, and $\lambda_{z,t} = \frac{1}{(k+1)^{0.50}}$. Each element of DP-noise vectors $\chi_{i,t}$, $\zeta_{i,t}$ and $\vartheta_{i,t}$ for agent i follows Laplace distribution $\text{Lap}(\frac{1}{(t+1)^{0.42+0.01i}})$, $\text{Lap}(\frac{1}{(t+1)^{0.41+0.01i}})$, and $\text{Lap}(\frac{1}{(t+1)^{0.4+0.01i}})$, respectively. In our comparison, the same stepsizes and DP-noises were applied to the MA-DSBO algorithm.

A.2. Additional Experimental Results

A.2.1. Comparison between Algorithm 2 with MA-DSBO and GBDSBO in the absence of DP-Noise

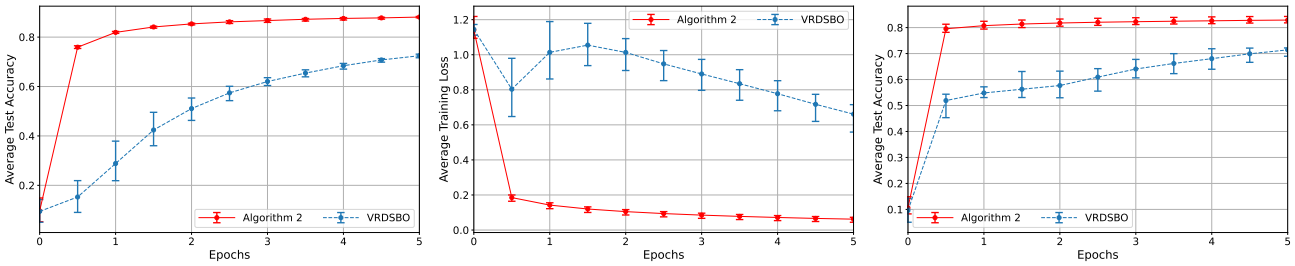
To further assess the performance of our Algorithm 2 in the absence of DP-noise, we conducted additional experiments to compare Algorithm 2 with MA-DSBO and GBDSBO using both synthetic dataset and the "MNIST" dataset. In the synthetic-data experiment, we chose the stepsizes for our Algorithm 2 as $\lambda_{x,t} = \frac{0.05}{(t+1)^{0.55}}$, $\lambda_{y,t} = \frac{0.05}{(t+1)^{0.5}}$, and $\lambda_{z,t} = \frac{0.02}{(t+1)^{0.45}}$. The stepsizes for MA-DSBO (Chen et al., 2023) were set to $\alpha = \beta = 0.03$ and $\gamma = 0.01$, and the stepsizes for GBDSBO (Yang et al., 2022) were set to $\alpha = \beta = 0.05$ and $\gamma = 0.02$. Those stepsizes were set in accordance with the guidelines provided in these works. In the "MNIST" experiment, the stepsizes for our Algorithm 2 were set to $\lambda_{x,t} = \frac{1.2}{(t+1)^{0.55}}$, $\lambda_{y,t} = \frac{1.2}{(t+1)^{0.5}}$, and $\lambda_{z,t} = \frac{1.2}{(t+1)^{0.45}}$. The stepsizes for MA-DSBO and GBDSBO were all set to 0.1. For all experiments, the number of nested-loops for both MA-DSBO and GBDSBO was set to 10. This setup corresponds to 10 outer iterations, which is equivalent to 100 iterations used in our algorithm, ensuring a fair comparison.

Figure 5-(a) shows that Algorithm 2 achieves similar test accuracy to MA-DSBO and higher test accuracy than GBDSBO in the synthetic-data experiment. Figures 5-(b) and 5-(c) confirm the advantage of our proposed algorithm in both training loss and test accuracy.



(a) Test accuracy, the synthetic data (b) Training loss, the "MNIST" dataset (c) Test accuracy, the "MNIST" dataset

Figure 5. Comparison by using the synthetic dataset and the "MNIST" dataset in the absence of DP noises.



(a) Test accuracy, DP noises free (b) Training loss, DP noises (c) Test accuracy, DP noises

Figure 6. Comparison Algorithm 2 with VRDSBO by using the "MNIST" dataset in the presence and the absence of DP noises.

A.2.2. Comparison between Algorithm 2 with VRDSBO

In this subsection, we compared our algorithm with the single-loop algorithm VRDSBO in Gao et al. (2023). While VRDSBO eliminates the need for nested-loops of communication (consensus) iterations, it is not applicable to general DSBO problems because it implicitly assumes homogeneous lower-level functions (a detailed illustration was provided in Appendix C.2 in Chen et al. (2023)). Therefore, we did not include this comparative experiment in the main text.

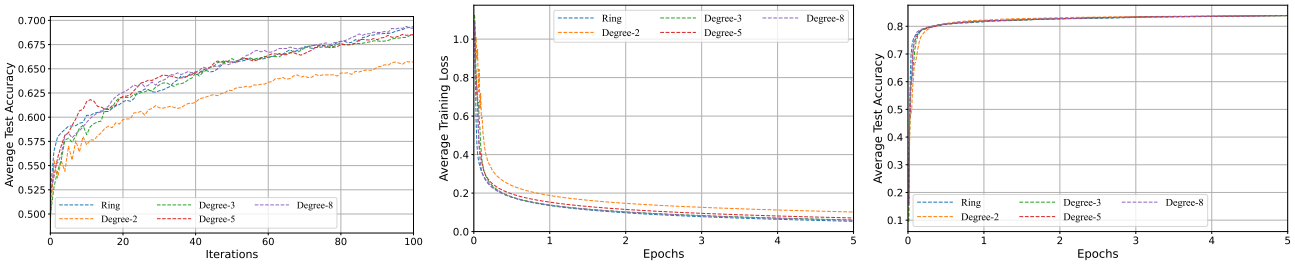
In the absence of DP-noise, the stepsizes for our Algorithm 2 were set to $\lambda_{x,t} = \frac{1.2}{(t+1)^{0.55}}$, $\lambda_{y,t} = \frac{1.2}{(t+1)^{0.5}}$, and $\lambda_{z,t} = \frac{1.2}{(t+1)^{0.45}}$. When considering DP-noise, the stepsizes of our Algorithm 2 were set to $\lambda_{x,t} = \frac{1.2}{(t+1)^{0.95}}$, $\lambda_{y,t} = \frac{1.2}{(t+1)^{0.87}}$, and $\lambda_{z,t} = \frac{1.2}{(t+1)^{0.75}}$. The stepsizes for VRDSBO were set to $\alpha_1 = \alpha_2 = 3$, $\beta_1 = \beta_2 = 1$, and $\eta = \frac{1.2}{(t+1)^{0.95}}$ (with η specifically designed to avoid divergent behaviors). The DP-noise variances were the same as those employed in Section 5.1. Figure 6 shows that under heterogeneous lower-level objective functions, our Algorithm 2 outperforms VRDSBO both in the presence and the absence of differential-privacy constraints.

A.2.3. Experimental results on various network topologies

We have conducted additional experimental results to evaluate the efficacy of our Algorithm 2 under different network topologies. We considered a network of $m = 10$ agents, with the interaction graph being a ring network and random r-regular graph (Bollobás, 1986) with r set to 2, 3, 5, and 8. We used the same parameters as those employed in Section 5.1. The results in Figure 7 show that the performance of our algorithm is insensitive to changes in topologies.

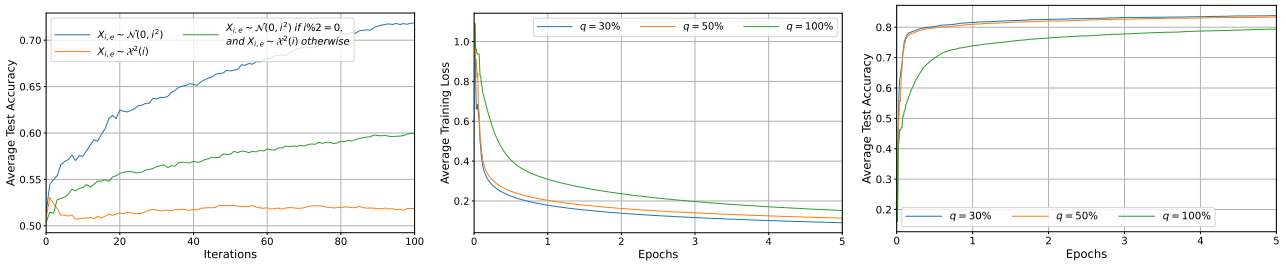
A.2.4. Experimental results on various heterogeneous data distributions

We have also conducted new experimental results to evaluate the convergence performance of our Algorithm 2 under different degrees of heterogeneity in data distributions using both the synthetic dataset and the "MNIST" dataset. More specifically, for the synthetic dataset experiment, we considered three heterogeneous data distributions for each agent i: (i) data $x_{i,e}$ followed a normal distribution $\mathcal{N}(0, i^2)$; (ii) data $x_{i,e}$ followed a Chi-squared distribution $\mathcal{X}^2(i)$; (iii) if agents $i \mod (2) = 0$ (i.e., i is an even number), data $x_{i,e}$ followed a normal distribution $\mathcal{N}(0, i^2)$, otherwise, data $x_{i,e}$ following a Chi-squared distribution $\mathcal{X}^2(i)$. In the "MNIST" dataset experiment, the training dataset contains $60,000$ images, while the



(a) Test accuracy, the synthetic dataset (b) Training loss, the "MNIST" dataset (c) Test accuracy, the "MNIST" dataset

Figure 7. Convergence performance of Algorithm 2 for different network topologies under LDP constraints.



(a) Test accuracy, the sythetic dataset (b) Training loss, the "MNIST" dataset (c) Test accuracy, the "MNIST" dataset

Figure 8. Convergence performance of Algorithm 2 for different data distributions under LDP constraints.

validation dataset contains $10,000$ images. We assigned q, $q = 30\%, 50\%, 100\%$ of the data from the i-th class to agent i, while the remaining $(100\% - q)$ of the data is split evenly among the other agents. Note that q measures the heterogeneity in data distribution among all agents.

Figure 8-(a) shows that the degree of data heterogeneity does affect convergence performance of our algorithm. Figure 8-(b) and 8-(c) imply that our algorithm is moderately affected by the degree of data heterogeneity in the "MNIST" dataset.

B. Notations and Auxiliary Lemmas

B.1. Additional Notations

Throughout this paper, we add a bar over a letter to denote the average of all agents and use bold font to represent stacked vectors of m agents. For further notational simplicity, we introduce the following notations:

$$\hat{\boldsymbol{H}}_t = \boldsymbol{H}_t - \mathbf{1}_m \otimes \bar{H}_t, \qquad \hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \mathbf{1}_m \otimes \bar{x}_t, \qquad \hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \mathbf{1}_m \otimes \bar{y}_t,$$

$$g_t(x,y) = \frac{1}{m}\sum_{i=1}^m g_{i,t}(x,y), \quad F_t(x,y) = \frac{1}{m}\sum_{i=1}^m f_{i,t}(x,y), \quad F_t(x) = \frac{1}{m}\sum_{i=1}^m f_{i,t}(x, y^*(x)),$$

$$\hat{\boldsymbol{z}}_t = \boldsymbol{z}_t - \mathbf{1}_m \otimes \bar{z}_t, \qquad \breve{z}_t = (\nabla^2_{yy} g_t(\bar{x}_t, \bar{y}_t))^{-1} \nabla F_t(\bar{x}_t, \bar{y}_t), \qquad \hat{\boldsymbol{u}}_t = \boldsymbol{u}_t - \mathbf{1}_m \otimes \bar{u}_t,$$

$$\chi_{wi,t} = \sum_{j\in\mathcal{N}_i} w_{ij}\chi_{i,t}, \qquad \zeta_{wi,t} = \sum_{j\in\mathcal{N}_i} w_{ij}\zeta_{i,t}, \qquad \vartheta_{wi,t} = \sum_{j\in\mathcal{N}_i} w_{ij}\vartheta_{i,t},$$

$$\hat{\boldsymbol{\chi}}_t = \boldsymbol{\chi}_t - \mathbf{1}_m \otimes \bar{\chi}_t, \qquad \hat{\boldsymbol{\zeta}}_t = \boldsymbol{\zeta}_t - \mathbf{1}_m \otimes \bar{\zeta}_t, \qquad \hat{\boldsymbol{\vartheta}}_t = \boldsymbol{\vartheta}_t - \mathbf{1}_m \otimes \bar{\vartheta}_t,$$

$$\sigma_x^+ = \max_{i\in[m]}\{\sigma_{i,x}\}, \qquad \sigma_y^+ = \max_{i\in[m]}\{\sigma_{i,y}\}, \qquad \sigma_z^+ = \max_{i\in[m]}\{\sigma_{i,z}\},$$

$$\varsigma_x = \min_{i\in[m]}\{\varsigma_{i,x}\}, \qquad \varsigma_y = \min_{i\in[m]}\{\varsigma_{i,y}\}, \qquad \varsigma_z = \min_{i\in[m]}\{\varsigma_{i,z}\},$$

$$\sigma_{x,t} = \frac{\sigma_x^+}{(t+1)^{\varsigma_x}}, \qquad \sigma_{y,t} = \frac{\sigma_y^+}{(t+1)^{\varsigma_y}}, \qquad \sigma_{z,t} = \frac{\sigma_z^+}{(t+1)^{\varsigma_z}}.$$

B.2. Auxiliary Lemmas

In this subsection, we introduce some well-known results from the existing literature, along with auxiliary lemmas that will be used in our subsequent convergence analysis.

Lemma B.1. *(Ghadimi & Wang, 2018; Chen et al., 2023) Under Assumption 2.2, $\nabla F(x)$ defined in (1) is L_F-Lipschitz continuous, i.e., for any given $x_1, x_2 \in \mathbb{R}^p$, we have*

$$\|\nabla F(x_1) - \nabla F(x_2)\| \leq L_F \|x_1 - x_2\|, \tag{14}$$

where the Lipschitz constant L_F is given by $L_F = L_{f,1} + \frac{2L_{f,1}L_{g,1}+L_{g,2}L_{f,0}^2}{\mu_g} + \frac{2L_{g,1}L_{f,0}L_{g,2}+L_{g,1}^2 L_{f,1}}{\mu_g^2} + \frac{L_{g,2}L_{g,1}^2 L_{f,0}}{\mu_g^3}$.

Lemma B.2. *(Wang & Nedić, 2023) Let $\{v_t\}$ be a nonnegative sequence, and $\{a_t\}$ and $\{b_t\}$ be positive sequences satisfying $a_0 < 1$, $\lim_{t\to\infty} a_t = 0$, $\sum_{t=0}^{\infty} a_t = \infty$, and $\lim_{t\to\infty} \frac{b_t}{a_t} = 0$. If $v_{t+1} \leq (1 - a_t)v_t + b_t$ holds for all $t > 0$, then we always have $v_t \leq C\frac{b_t}{a_t}$ for all $t > 0$, where C is some positive constant.*

Lemma B.3. *For any given pairs $(x, y) \in \mathbb{R}^p \times \mathbb{R}^q$, we introduce an auxiliary function $l(x, y; \xi) : \mathbb{R}^p \times \mathbb{R}^q \mapsto \mathbb{R}$ with a random variable ξ. If $\mathbb{E}_\xi [l(x, y; \xi)]$ is L-Lipschitz continuous and $\nabla l(x, y; \xi)$ is unbiased with a bounded variance σ^2, then for any given pairs (x_1, y_1) and $(x_2, y_2) \in \mathbb{R}^p \times \mathbb{R}^q$, the following inequality always holds:*

$$\mathbb{E}_\xi \left[\|l(x_1, y_1; \xi) - l(x_2, y_2; \xi)\|^2\right] \leq 2(L^2 + \sigma^2)(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2). \tag{15}$$

Proof. The mean value theorem implies that there must exist some constant $r \in (0, 1)$ such that for any $x_r = rx_1 + (1-r)x_2$ and $y_r = ry_1 + (1 - r)y_2$, the following inequality holds:

$$\begin{aligned}
&\mathbb{E}\left[\|l(x_1, y_1; \xi) - l(x_2, y_2; \xi)\|^2\right] = \mathbb{E}\left[\left(\langle \nabla_x l(x_r, y_r; \xi), x_1 - x_2\rangle + \langle \nabla_y l(x_r, y_r; \xi), y_1 - y_2\rangle\right)^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla_x l(x_r, y_r; \xi)\|^2\right] \|x_1 - x_2\|^2 + 2\mathbb{E}\left[\|\nabla_y l(x_r, y_r; \xi)\|^2\right] \|y_1 - y_2\|^2.
\end{aligned}$$

Since both terms $\mathbb{E}[\|\nabla_x l(x_r, y_r; \xi)\|^2]$ and $\mathbb{E}[\|\nabla_y l(x_r, y_r; \xi)\|^2]$ are no larger than $\mathbb{E}[\|\nabla l(x_r, y_r; \xi)\|^2]$, we can arrive at (15) based on the relationship $\mathbb{E}[\|\nabla l(x_r, y_r; \xi)\|^2] \leq L^2 + \sigma^2$. □

C. Empirical Risk Minimization Problems and Useful Properties of Empirical Functions

C.1. Empirical Risk Minimization Problem with respect to Problem (1)

We introduce the following ERM problem to approximate problem (1) under sequentially arriving data:

$$\begin{aligned}
&\min_{x\in\mathbb{R}^p} F_t(x), \quad F_t(x) = \frac{1}{m}\sum_{i=1}^{m} f_{i,t}(x, y_t^*(x)), \\
&\text{s.t.} \quad y_t^*(x) = \text{argmin}_{y\in\mathbb{R}^q} g_t(x, y) := \frac{1}{m}\sum_{i=1}^{m} g_{i,t}(x, y),
\end{aligned} \tag{16}$$

for any $t \geq 0$, where empirical functions $f_{i,t}$ and $g_{i,t}$ are given by $f_{i,t}(x, y) = \frac{1}{t+1}\sum_{k=0}^{t} h(x, y; \varphi_{i,k})$ and $g_{i,t}(x, y) = \frac{1}{t+1}\sum_{k=0}^{t} l(x, y; \xi_{i,k})$, respectively.

In the following lemmas, we present some useful properties of empirical functions $F_t(x)$ and $g_t(x, y)$. To this end, we define an auxiliary function $F_t(x, y) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_{i,t}(x, y)$ for any given pair $(x, y) \in \mathbb{R}^p \times \mathbb{R}^q$.

Lemma C.1 proves the boundedness properties of $F_t(x, y)$ and $g_t(x, y)$.

Lemma C.1. *Under Assumptions 2.2 and 2.3, for any given pair $(x, y) \in \mathbb{R}^p \times \mathbb{R}^q$, the following inequalities hold:*

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla_y F_t(x, y)\|^2\right] \leq 2\sigma_{f,1}^2 + 2L_{f,0}^2, \quad \mathbb{E}[\|\nabla_{yy}^2 g_t(x, y)\|^2] \leq 2\sigma_{g,2}^2 + 2L_{g,1}^2, \\
&\mathbb{E}[\|\nabla_{xy}^2 g_t(x, y)\|^2] \leq 2\sigma_{g,2}^2 + 2L_{g,1}^2, \quad \mathbb{E}[\|\nabla_{yy}^2 g_t(x, y)\|^2] \geq \mu_g^2.
\end{aligned} \tag{17}$$

Proof. By using the definition of $F_t(x,y)$, Assumption 2.2, and Assumption 2.3, we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla_y F_t(x,y)\|^2\right] &\le \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left\|\frac{1}{t+1}\sum_{k=0}^t \nabla_y h(x,y;\varphi_{i,k}) - \nabla_y f_i(x,y) + \nabla_y f_i(x,y)\right\|^2\right]\\ &\le \frac{2\sigma_{f,1}^2}{t+1} + \frac{2}{m}\sum_{i=1}^m \|\nabla_y f_i(x,y)\|^2 \le \frac{2\sigma_{f,1}^2}{t+1} + 2L_{f,0}^2 \le 2\sigma_{f,1}^2 + 2L_{f,0}^2.\end{aligned}$$

Similarly, based on the definition of $g_t(x,y)$, Assumption 2.2, and Assumption 2.3, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\nabla_{yy}^2 g_t(x,y)\|^2\right] &\le \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left\|\frac{1}{t+1}\sum_{k=0}^t \nabla_{yy}^2 l(x,y;\xi_{i,k}) - \nabla_{yy}^2 g_i(x,y) + \nabla_{yy}^2 g_i(x,y)\right\|^2\right]\\ &\le \frac{2\sigma_{g,2}^2}{t+1} + \frac{2}{m}\sum_{i=1}^m \|\nabla_{yy}^2 g_i(x,y)\|^2 \le \frac{2\sigma_{g,2}^2}{t+1} + 2L_{g,1}^2 \le 2\sigma_{g,2}^2 + 2L_{g,1}^2,\end{aligned}$$

and the following inequality:

$$\mathbb{E}\left[\|\nabla_{xy}^2 g_t(x,y)\|^2\right] \le \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left\|\frac{1}{t+1}\sum_{k=0}^t \nabla_{xy}^2 l(x,y;\xi_{i,k}) - \nabla_{xy}^2 g_i(x,y) + \nabla_{xy}^2 g_i(x,y)\right\|^2\right] \le 2\sigma_{g,2}^2 + 2L_{g,1}^2.$$

The μ_g-strongly convexity of lower-level functions g_i in Assumption 2.2 implies

$$\mathbb{E}\left[\nabla_{yy}^2 g_t(x,y)\right] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\frac{1}{t+1}\sum_{k=0}^t \nabla_{yy}^2 l(x,y;\xi_{i,k}) - \nabla_{yy}^2 g_i(x,y) + \nabla_{yy}^2 g_i(x,y)\right] = \nabla_{yy}^2 g(x,y) \ge \mu_g I_q,$$

which implies the last inequality in (17). □

By using Lemma B.3, we establish Lemma C.2 for Lipschitz continuity of functions $F_t(x,y)$ and $g_t(x,y)$.

Lemma C.2. *Under Assumptions 2.2 and 2.3, we have the following statements:*

(i) For any given pairs $(x_1,y_1) \in \mathbb{R}^p \times \mathbb{R}^q$ and $(x_2,y_2) \in \mathbb{R}^p \times \mathbb{R}^q$ and any $t > 0$, we have

$$\mathbb{E}\left[\|\nabla_y F_t(x_2,y_2) - \nabla_y F_t(x_1,y_1)\|^2\right] \le 2(L_{f,1}^2 + \sigma_{f,2}^2)\left(\|x_2 - x_1\|^2 + \|y_2 - y_1\|^2\right). \tag{18}$$

(ii) For any given pairs $(x_1,y_1) \in \mathbb{R}^p \times \mathbb{R}^q$ and $(x_2,y_2) \in \mathbb{R}^p \times \mathbb{R}^q$ and any $t > 0$, we obtain

$$\mathbb{E}\left[\left\|\left(\nabla_{yy}^2 g_t(x_2,y_2)\right)^{-1} - \left(\nabla_{yy}^2 g_t(x_1,y_1)\right)^{-1}\right\|^2\right] \le \frac{2(L_{g,2}^2 + \sigma_{g,3}^2)}{\mu_g^4}\left(\|x_2 - x_1\|^2 + \|y_2 - y_1\|^2\right), \tag{19}$$

$$\mathbb{E}\left[\left\|\nabla_y^2 g_t(x_2,y_2) - \nabla_y^2 g_t(x_1,y_1)\right\|^2\right] \le 2(L_{g,1}^2 + \sigma_{g,2}^2)\left(\|x_2 - x_1\|^2 + \|y_2 - y_1\|^2\right). \tag{20}$$

Proof. (i) By using the definition of $F_t(x,y)$ and Lemma B.3, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\nabla_y F_t(x_2,y_2) - \nabla_y F_t(x_1,y_1)\|^2\right] &\le \frac{1}{m}\sum_{i=1}^m \frac{1}{t+1}\sum_{k=0}^t \mathbb{E}\left[\|\nabla_y h(x_2,y_2;\varphi_{i,k}) - \nabla_y h(x_1,y_1;\varphi_{i,k})\|^2\right]\\ &\le 2(L_{f,1}^2 + \sigma_{f,2}^2)\left(\|x_2 - x_1\|^2 + \|y_2 - y_1\|^2\right),\end{aligned}$$

where we have used $\nabla_y f_i(x,y) = \mathbb{E}\left[\nabla_y h(x,y;\varphi_{i,k})\right]$, $L_{f,1}$-Lipschitz continuity of $\nabla_y f_i(x,y)$, and the bounded variance $\sigma_{f,2}^2$ of $\nabla^2 h(x,y;\varphi_{i,k})$ in the last inequality.

(ii) According to the definition of $g_t(x,y)$, we use Lemma B.3 to obtain

$$\mathbb{E}\left[\left\|\left(\nabla^2_{yy}g_t(x_2,y_2)\right)^{-1}-\left(\nabla^2_{yy}g_t(x_1,y_1)\right)^{-1}\right\|^2\right]\leq\frac{\mathbb{E}\left[\left\|\nabla^2_{yy}g_t(x_2,y_2)-\nabla^2_{yy}g_t(x_1,y_1)\right\|^2\right]}{\mu_g^4}$$
$$\leq\frac{2(L^2_{g,2}+\sigma^2_{g,3})}{\mu_g^4}\left(\|x_2-x_1\|^2+\|y_2-y_1\|^2\right),$$

where in the derivation we have used the following inequality from the proof of Lemma 2.2 in Ghadimi & Wang (2018) for any symmetrical matrices $A_1\in\mathbb{R}^{q\times q}$ and $A_2\in\mathbb{R}^{q\times q}$ satisfying $A_1\geq\mu_g I$ and $A_2\geq\mu_g I$:

$$\|A_1^{-1}-A_2^{-1}\|=\|A_1^{-1}(A_2-A_1)A_2^{-1}\|\leq\|A_1^{-1}\|\|A_2^{-1}\|\|A_2-A_1\|\leq\frac{\|A_2-A_1\|}{\mu_g^2}. \tag{21}$$

Additionally, using an argument similar to the derivation of (18), we arrive at (20). □

Lemma C.3 establishes the variations of functions $\nabla_y F_{t+1}(x,y)$ and $\nabla_{yy}g_t(x,y)$ over iterations.

Lemma C.3. *Under Assumptions 2.2 and 2.3, for any given pairs (x,y) and any $t>0$, the following inequalities hold:*

$$\mathbb{E}\left[\|\nabla_y F_{t+1}(x,y)-\nabla_y F_t(x,y)\|^2\right]\leq\frac{8(\sigma^2_{f,1}+L^2_{f,0})}{(t+2)^2}\quad\text{and}\quad\mathbb{E}\left[\|\nabla_{yy}g_{t+1}(x,y)-\nabla_{yy}g_t(x,y)\|^2\right]\leq\frac{8(\sigma^2_{g,2}+L^2_{g,1})}{(t+2)^2}. \tag{22}$$

Proof. We estimate an upper bound on $\mathbb{E}\left[\|\nabla_y F_{t+1}(x,y)-\nabla_y F_t(x,y)\|^2\right]$ by using the definition of $F_t(x,y)$:

$$\begin{aligned}&\mathbb{E}\left[\|\nabla_y F_{t+1}(x,y)-\nabla_y F_t(x,y)\|^2\right]\\&\leq\frac{1}{m}\sum_{i=1}^m\mathbb{E}\left[\left\|\frac{1}{t+2}\nabla_y h(x,y;\varphi_{i,t+1})+\frac{1}{t+2}\sum_{k=0}^t\nabla_y h(x,y;\varphi_{i,k})-\frac{1}{t+1}\sum_{k=0}^t\nabla_y h(x,y;\varphi_{i,k})\right\|^2\right]\\&\leq\frac{2}{m(t+2)^2}\sum_{i=1}^m\mathbb{E}\left[\|\nabla_y h(x,y;\varphi_{i,t+1})\|^2\right]+\frac{2}{m}\sum_{i=1}^m\left(\frac{1}{(t+2)(t+1)}\right)^2\mathbb{E}\left[\left\|\sum_{k=0}^t\nabla_y h(x,y;\varphi_{i,k})\right\|^2\right].\end{aligned} \tag{23}$$

The first term on the right hand side of (23) satisfies

$$\mathbb{E}\left[\|\nabla_y h(x,y;\varphi_{i,t+1})\|^2\right]\leq\mathbb{E}\left[2\|\nabla_y h(x,y;\varphi_{i,t+1})-\nabla_y f_i(x,y)\|^2+2\|\nabla_y f_i(x,y)\|^2\right]\leq 2\sigma^2_{f,1}+2L^2_{f,0}. \tag{24}$$

The second term on the right hand side of (23) satisfies

$$\mathbb{E}\left[\left\|\sum_{k=0}^t\nabla_y h(x,y;\varphi_{i,k})\right\|^2\right]\leq(t+1)\sum_{k=0}^t\mathbb{E}\left[\|\nabla_y h(x,y;\varphi_{i,k})\|^2\right]\leq 2(t+1)^2(\sigma^2_{f,1}+L^2_{f,0}), \tag{25}$$

where we have used $(a_1+\cdots+a_n)^2\leq n(a_1^2+\cdots+a_n^2)$ in the first inequality and (24) in the last inequality.

After substituting (24) and (25) into (23), we arrive at the first term in (22). Furthermore, by employing an argument similar to the derivation of the first term in (22), we can obtain the second term in (22). □

Lemma C.4 quantifies the distance between the optimal solution $y_t^*(x)$ to the lower-level ERM problem in (16) and the true optimal solution $y^*(x)$ to the lower-level optimization problem in (1):

Lemma C.4. *Under Assumptions 2.2 and 2.3, for any given $x\in\mathbb{R}^p$ and any $t>0$, we have*

$$\mathbb{E}\left[\|y_t^*(x)-y^*(x)\|^2\right]\leq\frac{4\sigma^2_{g,1}}{\mu_g^2(t+1)}. \tag{26}$$

Proof. We introduce the auxiliary functions $\bar{g}_{x,t}(y) = g_t(x,y)$ and $\bar{g}_x(y) = g(x,y)$, each with its optimal solution denoted as $y_t^* = \text{argmin}_{y\in\mathbb{R}^q}\bar{g}_{x,t}(y)$ and $y^* = \text{argmin}_{y\in\mathbb{R}^q}\bar{g}_x(y)$, respectively. For any given $x \in \mathbb{R}^p$, at time t, it follows that $y_t^* = y_t^*(x)$ and $y^* = y^*(x)$.

Given the definition of y_t^*, we obtain $\bar{g}_{x,t}(y_t^*) \le \bar{g}_{x,t}(y^*)$, which further implies

$$\bar{g}_x(y_t^*) - \bar{g}_x(y^*) \le (\bar{g}_x(y_t^*) - \bar{g}_{x,t}(y_t^*)) - (\bar{g}_x(y^*) - \bar{g}_{x,t}(y^*)). \tag{27}$$

By applying the mean value theorem to (27), we have

$$\bar{g}_x(y_t^*) - \bar{g}_x(y^*) \le \langle \nabla_y\bar{g}_x(\theta) - \nabla_y\bar{g}_{x,t}(\theta), y_t^* - y^*\rangle \le \|\nabla_y\bar{g}_x(\theta) - \nabla_y\bar{g}_{x,t}(\theta)\|\|y_t^* - y^*\|, \tag{28}$$

where the variable θ is given by $\theta = ry_t^* + (1-r)y^*$ with some constant $r \in (0,1)$.

The definition $\nabla_y\bar{g}_x(\theta) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\nabla_y l(x,\theta;\xi_i)]$ implies

$$\begin{aligned}\mathbb{E}\left[\|\nabla_y\bar{g}_{x,t}(\theta) - \nabla_y\bar{g}_x(\theta)\|\right] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \nabla_y g_{i,t}(x,\theta) - \nabla_y g(x,\theta)\right\|\right] \\ &\le \frac{1}{m}\sum_{i=1}^m \frac{1}{t+1}\sum_{k=0}^t \mathbb{E}\left[\|\nabla_y l(x,\theta;\xi_{i,k}) - \mathbb{E}\left[\nabla_y l(x,\theta;\xi_{i,k})\right]\|\right].\end{aligned} \tag{29}$$

Considering that the data points $\xi_{i,k}$ are independently and identically distributed across iterations, we use Assumption 2.3 and the Lyapunov inequality $E[\|X\|] \le (E[\|X\|^p])^{\frac{1}{p}},\ \forall p \ge 1$ to obtain

$$\begin{aligned}\sum_{k=0}^t \mathbb{E}\left[\|\nabla_y l(x,\theta;\xi_{i,k}) - \mathbb{E}\left[\nabla_y l(x,\theta;\xi_{i,k})\right]\|\right] &\le \sqrt{\mathbb{E}\left[\left(\sum_{k=0}^t \|\nabla_y l(x,\theta;\xi_{i,k}) - \mathbb{E}\left[\nabla_y l(x,\theta;\xi_{i,k})\right]\|\right)^2\right]} \\ &\le \sqrt{\mathbb{E}\left[\sum_{k=0}^t \|\nabla_y l(x,\theta;\xi_{i,k}) - \nabla_y g_i(x,\theta)\|^2\right]} \le \sigma_{g,1}\sqrt{t+1}.\end{aligned} \tag{30}$$

Substituting (30) into (29) yields $\mathbb{E}\left[\|\nabla_y\bar{g}_{x,t}(\theta) - \nabla_y\bar{g}_x(\theta)\|\right] \le \frac{\sigma_{g,1}}{\sqrt{t+1}}$. Further combing this relation with (28) leads to

$$\mathbb{E}\left[\|\bar{g}_x(y_t^*) - \bar{g}_x(y^*)\|\right] \le \frac{\sigma_{g,1}}{\sqrt{t+1}}\mathbb{E}\left[\|y_t^* - y^*\|\right]. \tag{31}$$

The μ_g-strongly convex of g_i implies $\frac{\mu_g}{2}\|y_t^* - y^*\|^2 \le \bar{g}_x(y_t^*) - \bar{g}_x(y^*)$. By combing this relation with (31), we have

$$\frac{\mu_g}{2}\mathbb{E}\left[\|y_t^* - y^*\|^2\right] \le \frac{\sigma_{g,1}}{\sqrt{t+1}}\mathbb{E}\left[\|y_t^* - y^*\|\right], \tag{32}$$

which implies $\mathbb{E}[\|y_t^* - y^*\|] \le \frac{2\sigma_{g,1}}{\mu_g\sqrt{t+1}}$. Substituting this inequality into (32), we obtain $\mathbb{E}[\|y_t^* - y^*\|^2] \le \frac{4\sigma_{g,1}^2}{\mu_g^2(t+1)}$. Recalling relationships $y_t^* = y_t^*(x)$ and $y^* = y^*(x)$ for any given $x \in \mathbb{R}^p$, at time t, we arrive at (26). □

Remark C.5. Since $\nabla_y g(x, y^*(x)) = 0$ is valid for any given $x \in \mathbb{R}^p$, it follows from Lemma C.4 that

$$\mathbb{E}\left[\|\nabla_y g(x,y_t^*(x))\|^2\right] = \mathbb{E}\left[\|\nabla_y g(x,y_t^*(x)) - \nabla_y g(x,y^*(x))\|^2\right] \le L_{g,1}^2\mathbb{E}\left[\|y_t^*(x) - y^*(x)\|^2\right] \le \frac{4L_{g,1}^2\sigma_{g,1}^2}{\mu_g^2(t+1)}. \tag{33}$$

We would like to point out that the relation (33) is a key to circumventing the assumption of Lipschitz continuity of the lower-level objective function $g(x,y)$ with respect to y, which is used in existing DSBO results (see Assumption 2.1 in Chen et al. (2022) and Assumption 3.4(iv) in Yang et al. (2022).)

Furthermore, we define $y_i^*(x) = \text{argmin}_{y\in\mathbb{R}^q} g_i(x,y)$ for any given $x \in \mathbb{R}^p$. By using an argument similar to the derivation of (26), we can obtain

$$\mathbb{E}\left[\|\nabla_y g_i(x,y_t^*(x))\|^2\right] = \mathbb{E}\left[\|\nabla_y g_i(x,y_t^*(x)) - \nabla_y g_i(x,y_i^*(x))\|^2\right] \le L_{g,1}^2\mathbb{E}\left[\|y_t^*(x) - y_i^*(x)\|^2\right] \le \frac{4L_{g,1}^2\sigma_{g,1}^2}{\mu_g^2(t+1)}. \tag{34}$$

In Lemma C.6, we quantify the variation of $y_t^*(x)$ over iteration t.

Lemma C.6. *Under Assumptions 2.2 and 2.3, for any given $x \in \mathbb{R}^p$, the following inequality always holds:*

$$\mathbb{E}\left[\|y_{t+1}^*(x) - y_t^*(x)\|^2\right] \leq \frac{2\sigma_{g,1}^2(\mu_g^2 + 4L_{g,1}^2)}{\mu_g^4(t+1)^2}. \tag{35}$$

Proof. For any given $x \in \mathbb{R}^p$, the definition of $y_t^*(x)$ implies $\nabla_y g_t(x, y_t^*(x)) = 0$, which further implies

$$\nabla_{yx}^2 g_t(x, y_t^*(x)) + \nabla_{yy}^2 g_t(x, y_t^*(x))\nabla_x y_t^*(x) = 0 \quad \text{or} \quad \nabla_x y_t^*(x) = -\left(\nabla_{yy}^2 g_t(x, y_t^*(x))\right)^{-1} \nabla_{yx}^2 g_t(x, y_t^*(x)). \tag{36}$$

Taking the squared norm and expectation on both sides of (36), we obtain the following inequality based on Lemma C.1:

$$\mathbb{E}\left[\|\nabla_x y_t^*(x)\|^2\right] \leq \frac{2\sigma_{g,2}^2 + 2L_{g,1}^2}{\mu_g^2}. \tag{37}$$

The differential mean value theorem implies Lipschitz continuity of $y_t^*(x)$:

$$\mathbb{E}\left[\|y_t^*(x_2) - y_t^*(x_1)\|^2\right] \leq \frac{2\sigma_{g,2}^2 + 2L_{g,1}^2}{\mu_g^2}\|x_2 - x_1\|^2. \tag{38}$$

We proceed to estimate an upper bound on $\mathbb{E}\left[\|y_{t+1}^*(x) - y_t^*(x)\|\right]$.

For any given $x \in \mathbb{R}^p$, we define an auxiliary function $g_{x,t}(y) \triangleq \frac{1}{m}\sum_{i=1}^m l(x, y; \xi_{i,t})$. Considering the definition of $g_t(x, y)$, we obtain the relation $g_t(x, y) = \frac{1}{t+1}\sum_{k=0}^t g_{x,k}(y)$, which further implies the following two inequalities based on $y_t^*(x) = \operatorname{argmin}_{y\in\mathbb{R}^q} g_t(x, y)$:

$$\sum_{k=0}^{t} \nabla_y g_{x,k}(y_t^*(x)) = 0 \quad \text{and} \quad \sum_{k=0}^{t+1} \nabla_y g_{x,k}(y_{t+1}^*(x)) = 0. \tag{39}$$

Given $\sum_{k=0}^{t+1} \nabla_y g_{x,k}(y_{t+1}^*(x)) = \sum_{k=0}^{t} \nabla_y g_{x,k}(y_{t+1}^*(x)) + \nabla_y g_{x,t+1}(y_{t+1}^*(x))$, we use (39) to obtain

$$\begin{aligned}
&\sum_{k=0}^{t} \left\langle y_{t+1}^*(x) - y_t^*(x), \nabla_y g_{x,k}(y_{t+1}^*(x)) - \nabla_y g_{x,k}(y_t^*(x))\right\rangle \\
&= \left\langle y_{t+1}^*(x) - y_t^*(x), \sum_{k=0}^{t+1} \nabla_y g_{x,k}(y_{t+1}^*(x)) - \nabla_y g_{x,t+1}(y_{t+1}^*(x)) - \sum_{k=0}^{t} \nabla_y g_{x,k}(y_t^*(x))\right\rangle \\
&= -\left\langle y_{t+1}^*(x) - y_t^*(x), \nabla_y g_{x,t+1}(y_{t+1}^*(x))\right\rangle.
\end{aligned} \tag{40}$$

Recalling the definition $g_t(x, y) = \frac{1}{t+1}\sum_{k=0}^t g_{x,k}(y)$, Assumptions 2.2, and 2.3, for any given $x \in \mathbb{R}^p$, $y_1 \in \mathbb{R}^q$, and $y_2 \in \mathbb{R}^q$, the following inequality always holds:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{t} \langle y_1 - y_2, \nabla_y g_{x,k}(y_1) - \nabla_y g_{x,k}(y_2)\rangle\right] = (t+1)\mathbb{E}\left[\langle y_1 - y_2, \nabla_y g_t(x, y_1) - \nabla_y g_t(x, y_2)\rangle\right] \\
&= (t+1)\langle y_1 - y_2, \nabla_y g(x, y_1) - \nabla_y g(x, y_2)\rangle \geq \mu_g(t+1)\|y_1 - y_2\|^2,
\end{aligned}$$

which further implies

$$\mathbb{E}\left[\sum_{k=0}^{t} \left\langle y_{t+1}^*(x) - y_t^*(x), \nabla_y g_{x,k}(y_{t+1}^*(x)) - \nabla_y g_{x,k}(y_t^*(x))\right\rangle\right] \geq \mu_g(t+1)\mathbb{E}\left[\|y_{t+1}^*(x) - y_t^*(x)\|^2\right]. \tag{41}$$

Combing (40) and (41) leads to

$$-\mathbb{E}\left[\left\langle y_{t+1}^*(x) - y_t^*(x), \nabla_y g_{x,t+1}(y_{t+1}^*(x))\right\rangle\right] \geq (t+1)\mu_g\mathbb{E}\left[\|y_{t+1}^*(x) - y_t^*(x)\|^2\right]. \tag{42}$$

By using Assumption 2.2, Assumption 2.3, and Lemma C.4, we have

$$\mathbb{E}\left[\|\nabla_y g_{x,t+1}(y^*_{t+1}(x))\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \nabla_y l(x, y^*_{t+1}(x); \xi_{i,t+1}) - \nabla_y g(x, y^*_{t+1}(x)) + \nabla_y g(x, y^*_{t+1}(x))\right\|^2\right]$$
$$\leq 2\sigma^2_{g,1} + 2\mathbb{E}\left[\left\|\nabla_y g(x, y^*_{t+1}(x)) - \nabla_y g(x, y^*(x))\right\|^2\right] \leq 2\sigma^2_{g,1} + \frac{8L^2_{g,1}\sigma^2_{g,1}}{\mu^2_g(t+1)},$$

which implies $\mathbb{E}\left[\|\nabla_y g_{x,t+1}(y^*_{t+1}(x))\|\right] \leq \sigma_{g,1}\sqrt{2 + \frac{8L^2_{g,1}}{\mu^2_g}}$. Further combing this inequality and (42), we arrive at

$$\sigma_{g,1}\sqrt{2 + \frac{8L^2_{g,1}}{\mu^2_g}}\mathbb{E}\left[\|y^*_{t+1}(x) - y^*_t(x)\|\right] \geq (t+1)\mu_g\mathbb{E}\left[\|y^*_{t+1}(x) - y^*_t(x)\|^2\right], \tag{43}$$

which implies (35) in Lemma C.6. □

C.2. Empirical Risk Minimization Problem with respect to Problem (8)

We introduce the following ERM problem to approximate problem (8) under sequentially arriving data:

$$\min_{z\in\mathbb{R}^q} \frac{1}{m}\sum_{i=1}^m \phi_{i,t}(z), \quad \phi_{i,t}(z) = \frac{1}{2}z^T H^*_{i,t} z - (b^*_{i,t})^T z, \tag{44}$$

Here, $H^*_{i,t}$ and $b^*_{i,t}$ are given by $H^*_{i,t} = \nabla^2_{yy} g_{i,t}(x, y^*(x))$ and $b^*_{i,t} = \nabla_y f_{i,t}(x, y^*(x))$.

D. Results of Algorithm 2

This section is devoted to analyzing the consensus error of the iterative variables generated by Algorithm 2. To this end, several technical lemmas are presented in Subsections D.1-D.10, with their interrelationships depicted in Figure 4.

D.1. Estimation of $\mathbb{E}\left[\|\bar{x}_{t+1} - \bar{x}_t\|^2\right]$ in Lemma D.1 and Its Proof

Recalling Algorithm 2 Step 7: $x_{i,t+1} = x_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(x_{j,t} + \chi_{j,t} - x_{i,t}) - \lambda_{x,t}u_{i,t}$, we express the update rule of $\bar{x}_{t+1}$ as follows:

$$\bar{x}_{t+1} = \bar{x}_t + \bar{\chi}_t - \lambda_{x,t}\bar{u}_t \quad \text{with} \quad \bar{u}_t = \frac{1}{m}\sum_{i=1}^m \left(\nabla_x f_{i,t}(x_{i,t}, y_{i,t}) - \nabla^2_{xy} g_{i,t}(x_{i,t}, y_{i,t}) z_{i,t}\right). \tag{45}$$

Lemma D.1. *Under Assumptions 2.1-2.3 and 3.1, for any $t > 0$, we have*

$$\mathbb{E}\left[\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] \leq \sigma^2_{x,t} + c_{\bar{x}1}\lambda^2_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + c_{\bar{x}2}\lambda^2_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + c_{\bar{x}3}\lambda^2_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + c_{\bar{x}4}\lambda^2_{x,t}\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right]$$
$$+ c_{\bar{x}5}\lambda^2_{x,t}\mathbb{E}\left[\|\bar{y}_t - y^*_t(\bar{x}_t)\|^2\right] + c_{\bar{x}6}\lambda^2_{x,t}, \tag{46}$$

where the constants $c_{\bar{x}1}$ to $c_{\bar{x}6}$ are given by $c_{\bar{x}1} = \frac{36L^2_{f,0}(\sigma^2_{g,2}+L^2_{g,1})}{m\mu^2_g}$, $c_{\bar{x}2} = \frac{18L^2_{f,0}}{m}$, $c_{\bar{x}3} = \frac{12(\sigma^2_{g,2}+L^2_{g,1})}{m}$, $c_{\bar{x}4} = c_{\bar{x}3}m$, $c_{\bar{x}5} = c_{\bar{x}2}m$, and $c_{\bar{x}6} = 6(\sigma^2_{f,1} + L^2_{f,0}) + \frac{24\left(\sigma^2_{g,2}+L^2_{g,1}\right)(\sigma^2_{f,1}+L^2_{f,0})}{\mu^2_g}$.

Proof. Considering the definition of $\bar{u}_t$ in (45), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{u}_t\|^2\right] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla_x f_{i,t}(x_{i,t},y_{i,t}) - \nabla^2_{xy} g_{i,t}(x_{i,t},y_{i,t}) z_{i,t}\|^2\right] \\
&\le \frac{2}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla_x f_{i,t}(x_{i,t},y_{i,t})\|^2\right] + \frac{2}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla^2_{xy} g_{i,t}(x_{i,t},y_{i,t})\|^2 \|\bar{z}_t\|^2\right], \\
&\le \frac{2}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla_x f_{i,t}(x_{i,t},y_{i,t}) - \nabla_x f_i(x_{i,t},y_{i,t}) + \nabla_x f_i(x_{i,t},y_{i,t}) - \nabla_x f_i(x_{i,t},y_t^*(x_{i,t})) + \nabla_x f_i(x_{i,t},y_t^*(x_{i,t}))\|^2\right] \\
&\quad + \frac{2}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla^2_{xy} g_{i,t}(x_{i,t},y_{i,t})\|^2 \|\bar{z}_t\|^2\right] \\
&\le \frac{6\sigma_{f,1}^2}{t+1} + \frac{6}{m}\sum_{i=1}^m \mathbb{E}\left[\|\nabla_x f_i(x_{i,t},y_{i,t}) - \nabla_x f_i(x_{i,t},y_t^*(x_{i,t}))\|^2\right] + 6L_{f,0}^2 + \frac{4}{m}\left(\sigma_{g,2}^2 + L_{g,1}^2\right)\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right] \\
&\le \frac{6\sigma_{f,1}^2}{t+1} + \frac{6L_{f,0}^2}{m}\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + 6L_{f,0}^2 + \frac{4}{m}\left(\sigma_{g,2}^2 + L_{g,1}^2\right)\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right],
\end{aligned} \tag{47}$$

where $\boldsymbol{y}_t$ and $\boldsymbol{y}_t^*(\boldsymbol{x})$ are given by $\boldsymbol{y}_t = \mathrm{col}(y_{1,t},\cdots,y_{m,t})$ and $\boldsymbol{y}_t^*(\boldsymbol{x}) = \mathrm{col}(y_t^*(x_{1,t}),\cdots,y_t^*(x_{m,t}))$.

To further analyze the term $\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right]$ in (47), we use the following decomposition:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] \le \mathbb{E}\left[\|\boldsymbol{y}_t - \mathbf{1}_m\otimes\bar{y}_t + \mathbf{1}_m\otimes\bar{y}_t - \mathbf{1}_m\otimes y_t^*(\bar{x}_t) + \mathbf{1}_m\otimes y_t^*(\bar{x}_t) - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] \\
&\le 3\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + 3m\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + 3\sum_{i=1}^m \mathbb{E}\left[\|y_t^*(\bar{x}_t) - y_t^*(x_{i,t})\|^2\right] \\
&\le 3\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + 3m\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + \frac{6(\sigma_{g,2}^2 + L_{g,1}^2)}{\mu_g^2}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right],
\end{aligned} \tag{48}$$

with $\hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \mathbf{1}_m\otimes\bar{y}_t$ and $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \mathbf{1}_m\otimes\bar{x}_t$. In the last inequality, we have used (38).

We now focus on characterizing the term $\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right]$ in (47). Considering that both the first term in (17) from Lemma C.1 and Assumption 2.2 lead to $\mathbb{E}\left[\|\breve{z}_t\|^2\right] = \mathbb{E}\left[\|(\nabla^2_{yy} g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_t,\bar{y}_t)\|^2\right] \le \frac{2\sigma_{f,1}^2 + 2L_{f,0}^2}{\mu_g^2}$, we subsequently obtain

$$\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right] = \mathbb{E}\left[\left\|\hat{\boldsymbol{z}}_t + \mathbf{1}_m\otimes(\bar{z}_t - \breve{z}_t) + \mathbf{1}_m\otimes\breve{z}_t\right\|^2\right] \le 3\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 3m\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \frac{6m(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}, \tag{49}$$

where $\hat{\boldsymbol{z}}_t$ is defined as $\hat{\boldsymbol{z}}_t = \boldsymbol{z}_t - \mathbf{1}_m\otimes\bar{z}_t$.

Substituting (48) and (49) into (47), we arrive at

$$\begin{aligned}
\mathbb{E}\left[\|\bar{u}_t\|^2\right] \le{}& \frac{6\sigma_{f,1}^2}{t+1} + \frac{36L_{f,0}^2(\sigma_{g,2}^2 + L_{g,1}^2)}{m\mu_g^2}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \frac{18L_{f,0}^2}{m}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + \frac{12\left(\sigma_{g,2}^2 + L_{g,1}^2\right)}{m}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] \\
&+ 18L_{f,0}^2\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + 12\left(\sigma_{g,2}^2 + L_{g,1}^2\right)\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \frac{24\left(\sigma_{g,2}^2 + L_{g,1}^2\right)\left(\sigma_{f,1}^2 + L_{f,0}^2\right)}{\mu_g^2} + 6L_{f,0}^2.
\end{aligned} \tag{50}$$

Taking the squared norm and the expectation on both sides of (45) and then substituting (50) into (45), we arrive at (46). □

D.2. Estimation of $\mathbb{E}\left[\|\bar{y}_{t+1} - \bar{y}_t\|^2\right]$ in Lemma D.2 and Its Proof

Recalling Algorithm 2 Step 4: $y_{i,t+1} = y_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(y_{j,t} + \zeta_{j,t} - y_{i,t}) - \lambda_{y,t}\nabla_y g_{i,t}(x_{i,t},y_{i,t})$, we express the update rule of $\bar{y}_{t+1}$ as follows:

$$\bar{y}_{t+1} = \bar{y}_t + \bar{\zeta}_t - \lambda_{y,t}\frac{1}{m}\sum_{i=1}^m \nabla_y g_{i,t}(x_{i,t},y_{i,t}). \tag{51}$$

Lemma D.2. *Under Assumptions 2.1-2.3 and 3.1, for any $t > 0$, we have*

$$\mathbb{E}\left[\|\bar{y}_{t+1} - \bar{y}_t\|^2\right] \leq \sigma_{y,t}^2 + c_{\bar{y}1}\lambda_{y,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + c_{\bar{y}2}\lambda_{y,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + c_{\bar{y}3}\lambda_{y,t}^2\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + c_{\bar{y}4}\frac{\lambda_{y,t}^2}{t+1}, \tag{52}$$

with $c_{\bar{y}1} = \frac{24L_{g,1}^2(\sigma_{g,2}^2+L_{g,1}^2)}{m\mu_g^2}$, $c_{\bar{y}2} = \frac{12L_{g,1}^2}{m}$, $c_{\bar{y}3} = c_{\bar{y}2}m$, *and* $c_{\bar{y}4} = 2\sigma_{g,1}^2\left(1 + \frac{8L_{g,1}^2}{\mu_g^2}\right)$.

Proof. By taking the squared norm and expectation on both sides of (51), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{y}_{t+1} - \bar{y}_t\|^2\right] \leq \mathbb{E}\left[\|\bar{\zeta}_t\|^2\right] + \lambda_{y,t}^2\mathbb{E}\left[\frac{2}{m}\sum_{i=1}^m\|\nabla_y g_{i,t}(x_{i,t},y_{i,t}) - \nabla_y g_i(x_{i,t},y_{i,t})\|^2 + 2\left\|\frac{1}{m}\sum_{i=1}^m\nabla_y g_i(x_{i,t},y_{i,t})\right\|^2\right]\\
&\leq \sigma_{y,t}^2 + \frac{2\sigma_{g,1}^2\lambda_{y,t}^2}{t+1} + 2\lambda_{y,t}^2\mathbb{E}\left[2\frac{1}{m}\sum_{i=1}^m\|\nabla_y g_i(x_{i,t},y_{i,t}) - \nabla_y g_i(x_{i,t},y_t^*(x_{i,t}))\|^2 + 2\left\|\frac{1}{m}\sum_{i=1}^m\nabla_y g_i(x_{i,t},y_t^*(x_{i,t}))\right\|^2\right]\\
&\leq \sigma_{y,t}^2 + \frac{2\sigma_{g,1}^2\lambda_{y,t}^2}{t+1} + \frac{4L_{g,1}^2}{m}\lambda_{y,t}^2\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + 4\lambda_{y,t}^2\mathbb{E}\left[\|\nabla_y g(x_{i,t},y_t^*(x_{i,t}))\|^2\right]\\
&\leq \sigma_{y,t}^2 + \frac{4L_{g,1}^2}{m}\lambda_{y,t}^2\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + 2\sigma_{g,1}^2\left(1 + \frac{8L_{g,1}^2}{\mu_g^2}\right)\frac{\lambda_{y,t}^2}{t+1},
\end{aligned} \tag{53}$$

where we have used (33) in the last inequality. Further substituting (48) into (53) yields (52). □

D.3. Estimation of $\mathbb{E}\left[\|\breve{z}_{t+1} - \breve{z}_t\|^2\right]$ in Lemma D.3 and Its Proof

Recalling the definition $\breve{z}_t = (\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_t,\bar{y}_t)$ with $\nabla_y F_t(\bar{x}_t,\bar{y}_t) \triangleq \frac{1}{m}\sum_{i=1}^m\nabla f_{i,t}(\bar{x}_t,\bar{y}_t)$, we express $\breve{z}_{t+1} - \breve{z}_t$ as follows:

$$\breve{z}_{t+1} - \breve{z}_t = (\nabla_{yy}^2 g_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_t,\bar{y}_t). \tag{54}$$

Lemma D.3. *Under Assumptions 2.2 and 2.3, for any $t > 0$, we have*

$$\mathbb{E}\left[\|\breve{z}_{t+1} - \breve{z}_t\|^2\right] < c_{\breve{z}1}\mathbb{E}\left[\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] + c_{\breve{z}1}\mathbb{E}\left[\|\bar{y}_{t+1} - \bar{y}_t\|^2\right] + \frac{c_{\breve{z}2}}{(t+2)^2}, \tag{55}$$

with $c_{\breve{z}1} = \frac{8(L_{f,1}^2+\sigma_{f,2}^2)}{\mu_g^2} + \frac{16(L_{g,2}^2+\sigma_{g,3}^2)(\sigma_{f,1}^2+L_{f,0}^2)}{\mu_g^4}$ *and* $c_{\breve{z}2} = \frac{32(\sigma_{f,1}^2+L_{f,0}^2)}{\mu_g^2}\left(1 + \frac{2(\sigma_{g,2}^2+L_{g,1}^2)}{\mu_g^2}\right)$.

Proof. By taking the squared norm and the expectation on both sides of (54), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\breve{z}_{t+1} - \breve{z}_t\|^2\right] = \mathbb{E}\left[\left\|(\nabla_{yy}^2 g_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_t,\bar{y}_t)\right\|^2\right]\\
&\leq 4\mathbb{E}\left[\left\|(\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_t,\bar{y}_t)\right\|^2\right]\\
&\quad + 4\mathbb{E}\left[\left\|(\nabla_{yy}^2 g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right]\\
&\quad + 4\mathbb{E}\left[\left\|(\nabla_{yy}^2 g_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla_{yy}^2 g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right]\\
&\quad + 4\mathbb{E}\left[\left\|(\nabla_{yy}^2 g_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla_{yy}^2 g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right].
\end{aligned} \tag{56}$$

Using both (17) in Lemma C.1 and (18) in Lemma C.2, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_t,\bar{y}_t)\right\|^2\right]\\
&\leq \mathbb{E}\left[\left\|(\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\right\|^2\right]\mathbb{E}\left[\|\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1}) - \nabla_y F_t(\bar{x}_t,\bar{y}_t)\|^2\right]\\
&\leq \frac{2(L_{f,1}^2+\sigma_{f,2}^2)}{\mu_g^2}\left(\mathbb{E}\left[\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] + \mathbb{E}\left[\|\bar{y}_{t+1} - \bar{y}_t\|^2\right]\right).
\end{aligned} \tag{57}$$

Similarly, using (17) in Lemma C.1 and (19) in Lemma C.2, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\nabla^2_{yy} g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla^2_{yy} g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right] \\
&\le \mathbb{E}\left[\left\|(\nabla^2_{yy} g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1} - (\nabla^2_{yy} g_t(\bar{x}_t,\bar{y}_t))^{-1}\right\|^2\right]\mathbb{E}\left[\left\|\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right] \\
&\le \frac{4(L^2_{g,2}+\sigma^2_{g,3})(\sigma^2_{f,1}+L^2_{f,0})}{\mu_g^4}\left(\mathbb{E}\left[\|\bar{x}_{t+1}-\bar{x}_t\|^2\right] + \mathbb{E}\left[\|\bar{y}_{t+1}-\bar{y}_t\|^2\right]\right).
\end{aligned} \tag{58}$$

Using (17) in Lemma C.1 and the first term in (22) of Lemma C.3, one yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\nabla^2_{yy} g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla^2_{yy} g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right] \\
&\le \frac{1}{\mu_g^2}\mathbb{E}\left[\left\|\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - \nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right] \le \frac{8(\sigma^2_{f,1}+L^2_{f,0})}{\mu_g^2(t+2)^2}.
\end{aligned} \tag{59}$$

Utilizing (21), the results in (17) from Lemma C.1 and the second term in (22) of Lemma C.3, we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\nabla^2_{yy} g_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - (\nabla^2_{yy} g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right] \\
&\le \mathbb{E}\left[\left\|(\nabla^2_{yy} g_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1} - (\nabla^2_{yy} g_t(\bar{x}_{t+1},\bar{y}_{t+1}))^{-1}\right\|^2\right]\mathbb{E}\left[\left\|\nabla_y F_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1})\right\|^2\right] \\
&\le \frac{1}{\mu_g^4}\mathbb{E}\left[\|\nabla^2_{yy} g_{t+1}(\bar{x}_{t+1},\bar{y}_{t+1}) - \nabla^2_{yy} g_t(\bar{x}_{t+1},\bar{y}_{t+1})\|^2\right]\mathbb{E}\left[\|\nabla_y F_t(\bar{x}_{t+1},\bar{y}_{t+1})\|\right] \\
&\le \frac{16(\sigma^2_{g,2}+L^2_{g,1})(\sigma^2_{f,1}+L^2_{f,0})}{\mu_g^4(t+2)^2}.
\end{aligned} \tag{60}$$

Substituting (57) to (60) into (56), we arrive at (55). □

In the following Subsections D.4-D.7, we quantify the distance between the iterative variables generated by Algorithm 2 and their corresponding average values.

D.4. Estimation of $\mathbb{E}\left[\|\hat{\boldsymbol{u}}_t\|^2\right]$ in Lemma D.4 and Its Proof

Here, we use the definitions $\hat{\boldsymbol{u}}_t = \boldsymbol{u}_t - \mathbf{1}_m \otimes \bar{u}_t$, $\boldsymbol{u}_t = \mathrm{col}(u_{1,t},\cdots,u_{m,t})$, and $\bar{u}_t = \frac{1}{m}\sum_{i=1}^m u_{i,t}$ with $u_{i,t}$ given by

$$u_{i,t} = \nabla_x f_{i,t}(x_{i,t},y_{i,t}) - \nabla^2_{xy} g_{i,t}(x_{i,t},y_{i,t}) z_{i,t}. \tag{61}$$

Lemma D.4. *Under Assumptions 2.2 and 2.3, for any $t>0$, the following inequality always holds:*

$$\mathbb{E}\left[\|\hat{\boldsymbol{u}}_t\|^2\right] \le c_{\hat{u}1}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + c_{\hat{u}2}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + c_{\hat{u}3}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + c_{\hat{u}4}\mathbb{E}\left[\|\bar{z}_t-\check{z}_t\|^2\right] + c_{\hat{u}5}\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + c_{\hat{u}6}, \tag{62}$$

where the constants $c_{\hat{u}1}$ to $c_{\hat{u}6}$ are given by $c_{\hat{u}1} = \frac{144L^2_{f,0}(\sigma^2_{g,2}+L^2_{g,1})}{\mu_g^2}$, $c_{\hat{u}2} = 72L^2_{f,0}$, $c_{\hat{u}3} = 48(\sigma^2_{g,2}+L^2_{g,1})$, $c_{\hat{u}4} = c_{\hat{u}3}m$, $c_{\hat{u}5} = c_{\hat{u}2}m$, and $c_{\hat{u}6} = 24m\sigma^2_{f,1} + 24mL^2_{f,0} + \frac{2c_{\hat{u}4}(\sigma^2_{f,1}+L^2_{f,0})}{\mu_g^2}$.

Proof. We first determine an upper bound on $\mathbb{E}\left[\|\boldsymbol{u}_t\|^2\right]$. Based on (61) and Lemma C.1, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{u}_t\|^2\right] \le 2\sum_{i=1}^m \mathbb{E}\left[\|\nabla_x f_{i,t}(x_{i,t},y_{i,t})\|^2 + \|\nabla^2_{xy} g_{i,t}(x_{i,t},y_{i,t})\|^2\|z_{i,t}\|^2\right] \\
&\le 2\sum_{i=1}^m \mathbb{E}\left[\|\nabla_x f_{i,t}(x_{i,t},y_{i,t}) - \nabla_x f_i(x_{i,t},y_{i,t}) + \nabla_x f_i(x_{i,t},y_{i,t})\|^2\right] + 2\sum_{i=1}^m \mathbb{E}\left[\|\nabla^2_{xy} g_{i,t}(x_{i,t},y_{i,t})\|^2\|z_{i,t}\|^2\right] \\
&\le \frac{6m\sigma^2_{f,1}}{t+1} + 6\sum_{i=1}^m \mathbb{E}\left[\|\nabla_x f_i(x_{i,t},y_{i,t}) - \nabla_x f_i(x_{i,t},y_t^*(x_{i,t}))\|^2\right] + 6mL^2_{f,0} + 4\left(\frac{\sigma^2_{g,2}}{t+1} + L^2_{g,1}\right)\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right] \\
&\le \frac{6m\sigma^2_{f,1}}{t+1} + 6L^2_{f,0}\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + 6mL^2_{f,0} + 4\left(\sigma^2_{g,2}+L^2_{g,1}\right)\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right].
\end{aligned} \tag{63}$$

Then, we characterize the term $\mathbb{E}\left[\|\mathbf{1}_m \otimes \bar{u}_t\|^2\right]$. By using (47), we have

$$\mathbb{E}\left[\|\mathbf{1}_m \otimes \bar{u}_t\|^2\right] \leq \frac{6m\sigma_{f,1}^2}{t+1} + 6L_{f,0}^2\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + 6mL_{f,0}^2 + 4\left(\sigma_{g,2}^2 + L_{g,1}^2\right)\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right]. \tag{64}$$

Based on the relation $\|\hat{\boldsymbol{u}}_t\|^2 = 2\|\boldsymbol{u}_t\|^2 + 2\|\mathbf{1}_m \otimes \bar{u}_t\|^2$, by summing up the corresponding sides of (63) and (64), we obtain

$$\mathbb{E}\left[\|\hat{\boldsymbol{u}}_t\|^2\right] \leq \frac{24m\sigma_{f,1}^2}{t+1} + 24L_{f,0}^2\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + 24mL_{f,0}^2 + 16\left(\sigma_{g,2}^2 + L_{g,1}^2\right)\mathbb{E}\left[\|\boldsymbol{z}_t\|^2\right]. \tag{65}$$

Substituting (48) and (49) into (65), we can arrive at (62). □

D.5. Estimation of $\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]$ in Lemma D.5 and Its Proof

Recalling the definitions $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \mathbf{1}_m \otimes \bar{x}_t$, $\boldsymbol{x}_t = \mathrm{col}(x_{1,t}, \cdots, x_{m,t})$, and $\bar{x}_t = \frac{1}{m}\sum_{i=1}^m x_{i,t}$ with $x_{i,t+1} = x_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(x_{j,t} + \chi_{j,t} - x_{i,t}) - \lambda_{x,t}u_{i,t}$ in Algorithm 2 Step 7, we have

$$\hat{\boldsymbol{x}}_{t+1} = (I + W \otimes I_p)\hat{\boldsymbol{x}}_t + \hat{\boldsymbol{\chi}}_t - \lambda_{x,t}\hat{\boldsymbol{u}}_t. \tag{66}$$

Lemma D.5. *Under Assumptions 2.1-2.3 and 3.1, for any $t > 0$, we have*

$$\begin{aligned}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{t+1}\|^2\right] \leq& \left(1 - \frac{\delta_2}{2} + c_{\hat{x}1}\lambda_{x,t}^2\right)\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + 4m\sigma_{x,t}^2 + c_{\hat{x}2}\lambda_{x,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + c_{\hat{x}3}\lambda_{x,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] \\ &+ c_{\hat{x}4}\lambda_{x,t}^2\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + c_{\hat{x}5}\lambda_{x,t}^2\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + c_{\hat{x}6}\lambda_{x,t}^2,\end{aligned} \tag{67}$$

where $c_{\hat{x}1}$ to $c_{\hat{x}6}$ are given by $c_{\hat{x}i} = \left(1 + \frac{2}{\delta_2}\right)c_{\hat{u}i},\ i = \{1, \cdots, 6\}$ with $c_{\hat{u}i}$ given in the statement of Lemma D.4.

Proof. By taking the squared norm and the expectation on both sides of (66), we obtain

$$\begin{aligned}&\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{t+1}\|^2\right] = \|I + W \otimes I_q\|^2\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + 4m\sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{u}}_t\|^2\right] - 2\mathbb{E}\left[\langle(I + W \otimes I_q)\hat{\boldsymbol{x}}_t, \lambda_{x,t}\hat{\boldsymbol{u}}_t\rangle\right] \\ &\leq \left(1 - \frac{\delta_2}{2}\right)\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + 4m\sigma_{x,t}^2 + \left(1 + \frac{2}{\delta_2}\right)\lambda_{x,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{u}}_t\|^2\right],\end{aligned} \tag{68}$$

where in the derivation we have used Assumptions 2.1, Assumption 3.1, and the following inequality:

$$-2\mathbb{E}\left[\langle(I + W \otimes I_q)\hat{\boldsymbol{x}}_t, \lambda_{x,t}\hat{\boldsymbol{u}}_t\rangle\right] \leq \frac{\delta_2}{2}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \frac{2\lambda_{x,t}^2}{\delta_2}\mathbb{E}\left[\|\hat{\boldsymbol{u}}_t\|^2\right].$$

Substituting (62) from Lemma D.4 into (68), we arrive at (67). □

D.6. Estimation of $\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]$ in Lemma D.6 and Its Proof

Recalling the definitions $\hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \mathbf{1}_m \otimes \bar{y}_t$, $\boldsymbol{y}_t = \mathrm{col}(y_{1,t}, \cdots, y_{m,t})$, and $\bar{y}_t = \frac{1}{m}\sum_{i=1}^m y_{i,t}$ with $y_{i,t+1} = y_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(x_{j,t} + \zeta_{j,t} - y_{i,t}) - \lambda_{y,t}\nabla_y g_{i,t}(x_{i,t}, y_{i,t})$ given in Algorithm 2 Step 4, we have

$$\hat{\boldsymbol{y}}_{t+1} = (I + W \otimes I_q)\hat{\boldsymbol{y}}_t + \hat{\boldsymbol{\zeta}}_t - \lambda_{y,t}\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t, \boldsymbol{y}_t), \tag{69}$$

with $\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t, \boldsymbol{y}_t) = \mathrm{col}(\nabla_y\hat{g}_{1,t}, \cdots, \hat{g}_{m,t})$ and $\nabla_y\hat{g}_{i,t} = \nabla_y g_{i,t}(x_{i,t}, y_{i,t}) - \frac{1}{m}\sum_{i=1}^m \nabla_y g_{i,t}(x_{i,t}, y_{i,t})$.

Lemma D.6. *Under Assumptions 2.1- 2.3 and 3.1, for any $t > 0$, the following inequality always holds:*

$$\mathbb{E}\left[\|\hat{\boldsymbol{y}}_{t+1}\|^2\right] \leq \left(1 - \frac{\delta_2}{2} + c_{\hat{y}1}\lambda_{y,t}^2\right)\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + 4m\sigma_{y,t}^2 + c_{\hat{y}2}\lambda_{y,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + c_{\hat{y}3}\lambda_{y,t}^2\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + c_{\hat{y}4}\frac{\lambda_{y,t}^2}{t+1}, \tag{70}$$

where the constants $c_{\hat{y}1}$ to $c_{\hat{y}4}$ are given by $c_{\hat{y}1} = 48L_{g,1}^2\left(1 + \frac{2}{\delta_2}\right)$, $c_{\hat{y}2} = \left(1 + \frac{2}{\delta_2}\right)\frac{96(\sigma_{g,2}^2 + L_{g,1}^2)L_{g,1}^2}{\mu_g^2}$, $c_{\hat{y}3} = c_{\hat{y}1}m$, and $c_{\hat{y}4} = 8\sigma_{g,1}^2 m\left(1 + \frac{2}{\delta_2}\right)\left(1 + \frac{8L_{g,1}^2}{\mu_g^2}\right)$.

Proof. By taking the squared norm and expectation on both sides of (69), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_{t+1}\|^2\right] &= \|I+W\otimes I_q\|^2\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + 4m\sigma_{y,t}^2 + \lambda_{y,t}^2\mathbb{E}\left[\|\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t,\boldsymbol{y}_t)\|^2\right] \\ &\quad - 2\mathbb{E}\left[\langle (I+W\otimes I_q)\hat{\boldsymbol{y}}_t, \lambda_{y,t}\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle\right] \\ &\le \left(1-\frac{\delta_2}{2}\right)\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + 4m\sigma_{y,t}^2 + \left(1+\frac{2}{\delta_2}\right)\lambda_{y,t}^2\mathbb{E}\left[\|\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t,\boldsymbol{y}_t)\|^2\right].\end{aligned} \tag{71}$$

We proceed to characterize the term $\mathbb{E}\left[\|\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t,\boldsymbol{y}_t)\|^2\right]$ in (71). Considering the definition of $\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t,\boldsymbol{y}_t)$, we have

$$\mathbb{E}\left[\|\nabla_y\hat{\boldsymbol{g}}_t(\boldsymbol{x}_t,\boldsymbol{y}_t)\|^2\right] \le 2\sum_{i=1}^m\mathbb{E}\left[\|\nabla_y g_{i,t}(x_{i,t},y_{i,t})\|^2\right] + 2\sum_{i=1}^m\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})\right\|^2\right]. \tag{72}$$

We first analyze the first term on the right hand side of (72):

$$\begin{aligned}&\sum_{i=1}^m\mathbb{E}\left[\|\nabla_y g_{i,t}(x_{i,t},y_{i,t})\|^2\right] \le \frac{2m\sigma_{g,1}^2}{t+1} + 2\sum_{i=1}^m\mathbb{E}\left[\|\nabla_y g_i(x_{i,t},y_{i,t})\|^2\right] \\ &\le \frac{2m\sigma_{g,1}^2}{t+1} + 2\sum_{i=1}^m\mathbb{E}\left[2\|\nabla_y g_i(x_{i,t},y_{i,t}) - \nabla_y g_i(x_{i,t},y_t^*(x_{i,t}))\|^2 + 2\|\nabla_y g_i(x_{i,t},y_t^*(x_{i,t}))\|^2\right] \\ &\le 4L_{g,1}^2\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + \frac{2\sigma_{g,1}^2 m}{t+1}\left(1+\frac{8L_{g,1}^2}{\mu_g^2}\right),\end{aligned} \tag{73}$$

where we have used (34) in the last inequality. Similarly, the second term on the right hand side of (72) satisfies

$$\sum_{i=1}^m\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})\right\|^2\right] \le 4L_{g,1}^2\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*(\boldsymbol{x})\|^2\right] + \frac{2\sigma_{g,1}^2 m}{t+1}\left(1+\frac{8L_{g,1}^2}{\mu_g^2}\right). \tag{74}$$

Substituting (73) and (74) into (72) and subsequently substituting (72) and (48) into (71), we arrive at (70). □

D.7. Estimation of $\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right]$ in Lemma D.7 and Its Proof

Using $\bar{z}_t = \frac{1}{m}\sum_{i=1}^m z_{i,t}$, $z_{i,t+1} = z_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(x_{j,t}+\vartheta_{j,t}-z_{i,t}) - \lambda_{z,t}\nabla_z\phi_{i,t}(z_{i,t})$ from Algorithm 1 Step 5, and $\nabla_z\phi_{i,t}(z_{i,t}) = H_{i,t}z_{i,t} - b_{i,t}$ from Algorithm 1 Step 4, we have

$$\bar{z}_{t+1} = \bar{z}_t + \bar{\vartheta}_t - \lambda_{z,t}\frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t} + \lambda_{z,t}\bar{b}_t, \tag{75}$$

with $\bar{b}_t = \frac{1}{m}\sum_{i=1}^m b_{i,t}$ and $b_{i,t} = \nabla_y f_{i,t}(x_{i,t},y_{i,t})$.

Recalling definitions $\hat{z}_{i,t} = z_{i,t} - \bar{z}_t$, $H_{i,t} = \nabla_{yy}^2 g_{i,t}(x_{i,t},y_{i,t})$, and $\bar{H}_t = \frac{1}{m}\sum_{i=1}^m H_{i,t}$, we obtain

$$\begin{aligned}H_{i,t}z_{i,t} - \frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t} &= H_{i,t}z_{i,t} - \frac{1}{m}\sum_{i=1}^m H_{i,t}(\hat{z}_{i,t}+\bar{z}_t) = H_{i,t}z_{i,t} - \frac{1}{m}\sum_{i=1}^m H_{i,t}\hat{z}_{i,t} - \bar{H}_t\bar{z}_t \\ &= H_{i,t}\hat{z}_{i,t} - \frac{1}{m}\sum_{i=1}^m H_{i,t}\hat{z}_{i,t} + \left(H_{i,t}-\bar{H}_t\right)\bar{z}_t.\end{aligned} \tag{76}$$

We define auxiliary variables $\tilde{\boldsymbol{H}}_t = \breve{\boldsymbol{H}}_t - \frac{1}{m}(\mathbf{1}_m\otimes I_q)(\boldsymbol{H}_t)^T \in \mathbb{R}^{mq\times mq}$ with $\breve{\boldsymbol{H}}_t = \mathrm{diag}(H_{1,t},\cdots,H_{m,t}) \in \mathbb{R}^{mq\times mq}$ and $\boldsymbol{H}_t = \mathrm{col}(H_{1,t},\cdots,H_{m,t})$. Further using the definitions $\hat{\boldsymbol{z}}_t = \boldsymbol{z}_t - \mathbf{1}_m\otimes\bar{z}_t \in \mathbb{R}^{mq}$, $\hat{\boldsymbol{b}}_t = \boldsymbol{b}_t - \mathbf{1}_m\otimes\bar{b}_t \in \mathbb{R}^{mq}$, and $\hat{\boldsymbol{H}}_t = \boldsymbol{H}_t - \mathbf{1}_m\otimes\bar{H}_t \in \mathbb{R}^{mq\times q}$, and then combining (75) and (76), we obtain the following equality:

$$\hat{\boldsymbol{z}}_{t+1} = (I+W\otimes I_q)\,\hat{\boldsymbol{z}}_t + \hat{\boldsymbol{\vartheta}}_t - \lambda_{z,t}\tilde{\boldsymbol{H}}_t\hat{\boldsymbol{z}}_t - \lambda_{z,t}\hat{\boldsymbol{H}}_t\bar{z}_t + \lambda_{z,t}\hat{\boldsymbol{b}}_t. \tag{77}$$

Lemma D.7. *Under Assumptions 2.1-2.3 and 3.1, for any $t > 0$, the following inequality always holds:*

$$\mathbb{E}\left[\|\hat{\boldsymbol{z}}_{t+1}\|^2\right] \leq \left(1 - \frac{\delta_2}{2}\right)\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 4m\sigma_{z,t}^2 + c_{\hat{z}1}\lambda_{z,t}^2\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + c_{\hat{z}2}\lambda_{z,t}^2, \tag{78}$$

where $c_{\hat{z}1}$ and $c_{\hat{z}2}$ are given by $c_{\hat{z}1} = 8mL_{g,1}^2\left(3 + \frac{8(1-\delta_2)^2}{\delta_2}\right)$ and $c_{\hat{z}2} = \frac{c_{\hat{z}1}}{2L_{g,1}^2}\left(\frac{4L_{g,1}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2} + L_{f,0}^2\right)$.

Proof. By taking the squared norm and expectation on both sides of (77), and then using inequality $(a+b+c+d)^2 \leq a^2+b^2+c^2+d^2+2ab+2ac+2ad+2bc+2bd+2cd$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{z}}_{t+1}\|^2\right] &= \mathbb{E}\left[\|\left(I + W \otimes I_q\right)\hat{\boldsymbol{z}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{\vartheta}}_t\|^2\right] + \lambda_{z,t}^2\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|^2\|\hat{\boldsymbol{z}}_t\|^2\right] + \lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{H}}_t\|^2\|\bar{z}_t\|^2\right] + \lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{b}}_t\|^2\right] \\
&\quad - 2\mathbb{E}\left[\left\langle\left(I + W \otimes I_q\right)\hat{\boldsymbol{z}}_t, \lambda_{z,t}\tilde{\boldsymbol{H}}_t\hat{\boldsymbol{z}}_t\right\rangle\right] - 2\mathbb{E}\left[\left\langle\left(I + W \otimes I_q\right)\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{H}}_t\bar{z}_t\right\rangle\right] + 2\mathbb{E}\left[\left\langle\left(I + W \otimes I_q\right)\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{b}}_t\right\rangle\right] \\
&\quad + 2\mathbb{E}\left[\left\langle\lambda_{z,t}\tilde{\boldsymbol{H}}_t\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{H}}_t\bar{z}_t\right\rangle\right] - 2\mathbb{E}\left[\left\langle\lambda_{z,t}\tilde{\boldsymbol{H}}_t\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{b}}_t\right\rangle\right] - 2\mathbb{E}\left[\left\langle\lambda_{z,t}\hat{\boldsymbol{H}}_t\bar{z}_t, \lambda_{z,t}\hat{\boldsymbol{b}}_t\right\rangle\right],
\end{aligned} \tag{79}$$

where in the derivation we have used Assumption 3.1, which implies $\mathbb{E}[\langle\cdot, \hat{\boldsymbol{\vartheta}}_t\rangle] = 0$.

By using the relationships $2ab \leq a^2 + b^2$ and $2\langle a, \lambda_{z,t}b\rangle \leq \kappa_1 a^2 + \frac{1}{\kappa_1}\lambda_{z,t}^2 b^2$ holding for all $\kappa_1 > 0$, we can obtain

$$\begin{cases}
-2\mathbb{E}\left[\left\langle\left(I + W \otimes I_q\right)\hat{\boldsymbol{z}}_t, \lambda_{z,t}\tilde{\boldsymbol{H}}_t\hat{\boldsymbol{z}}_t\right\rangle\right] \leq 2\lambda_{z,t}\mathbb{E}\left[\|I + W \otimes I_q\|\,\|\tilde{\boldsymbol{H}}_t\|\,\|\hat{\boldsymbol{z}}_t\|^2\right], \\
-2\mathbb{E}\left[\left\langle\left(I + W \otimes I_q\right)\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{H}}_t\bar{z}_t\right\rangle\right] \leq \kappa_1\|I + W \otimes I_q\|^2\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \frac{\lambda_{z,t}^2}{\kappa_1}\mathbb{E}\left[\|\hat{\boldsymbol{H}}_t\|^2\|\bar{z}_t\|^2\right], \\
2\mathbb{E}\left[\left\langle\left(I + W \otimes I_q\right)\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{b}}_t\right\rangle\right] \leq \kappa_1\|I + W \otimes I_q\|^2\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \frac{\lambda_{z,t}^2}{\kappa_1}\mathbb{E}\left[\|\hat{\boldsymbol{b}}_t\|^2\right], \\
2\mathbb{E}\left[\left\langle\lambda_{z,t}\tilde{\boldsymbol{H}}_t\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{H}}_t\bar{z}_t\right\rangle\right] \leq \lambda_{z,t}^2\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|^2\|\hat{\boldsymbol{z}}_t\|^2\right] + \lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{H}}_t\|^2\|\bar{z}_t\|^2\right], \\
2\mathbb{E}\left[\left\langle\lambda_{z,t}\tilde{\boldsymbol{H}}_t\hat{\boldsymbol{z}}_t, \lambda_{z,t}\hat{\boldsymbol{b}}_t\right\rangle\right] \leq \lambda_{z,t}^2\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|^2\|\hat{\boldsymbol{z}}_t\|^2\right] + \lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{b}}_t\|^2\right], \\
-2\mathbb{E}\left[\left\langle\lambda_{z,t}\hat{\boldsymbol{H}}_t\bar{z}_t, \lambda_{z,t}\hat{\boldsymbol{b}}_t\right\rangle\right] \leq \lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{H}}_t\|^2\|\bar{z}_t\|^2\right] + \lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{b}}_t\|^2\right].
\end{cases} \tag{80}$$

Substituting (80) into (79), we arrive at

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{z}}_{t+1}\|^2\right] &= \|I + W \otimes I_q\|^2\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{\vartheta}}_t\|^2\right] + \left(3\lambda_{z,t}^2 + \frac{\lambda_{z,t}^2}{\kappa_1}\right)\mathbb{E}\left[\|\hat{\boldsymbol{H}}_t\|^2\|\bar{z}_t\|^2\right] + (3\lambda_{z,t}^2 + \frac{\lambda_{z,t}^2}{\kappa_1})\mathbb{E}\left[\|\hat{\boldsymbol{b}}_t\|^2\right] \\
&\quad + 3\lambda_{z,t}^2\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|^2\right]\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 2\kappa_1\|I + W \otimes I_q\|^2\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 2\lambda_{z,t}\|I + W \otimes I_q\|\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|\right]\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right].
\end{aligned} \tag{81}$$

By using the definition of $\tilde{\boldsymbol{H}}_t$ and Assumption 2.2, we obtain

$$\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|^2\right] \leq 2\mathbb{E}\left[\|\breve{\boldsymbol{H}}_t\|^2\right] + 2\mathbb{E}\left[\left\|\frac{1}{m}(\mathbf{1}_m \otimes I_q)(\boldsymbol{H}_t)^T\right\|^2\right] \leq 4mL_{g,1}^2. \tag{82}$$

We choose $\kappa_1 \leq \frac{\delta_2}{8(1-\delta_2)^2}$, leading to $2\kappa_1(1-\delta_2)^2 \leq \frac{\delta_2}{4}$. Additionally, since the stepsize $\lambda_{z,t}$ decays with time, the inequality $12mL_{g,1}^2\lambda_{z,t}^2 + 4\sqrt{m}L_{g,1}\lambda_{z,t}(1-\delta_2) \leq \frac{\delta_2}{4}$ always holds for a sufficiently large iteration T. Without loss of generality, we can set $\lambda_{z,0}$ as a small constant, ensuring that above inequality is satisfied. This strategy is commonly used in the DSBO result, such as Yang et al. (2022). Then, the summation of the last three terms on the right hand side of (81) can be simplified as follows:

$$3\lambda_{z,t}^2\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|^2\right]\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 2\lambda_{z,t}\|I + W \otimes I_q\|\mathbb{E}\left[\|\tilde{\boldsymbol{H}}_t\|\right]\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 2\kappa_1\|I + W \otimes I_q\|^2\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] \leq \frac{\delta_2}{2}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right], \tag{83}$$

where in the derivation we have used (82) and $\|I + W \otimes I_q\| \leq 1 - \delta_2$ from Assumption 2.1.

Substituting (83) into (81) and using $(1-\delta_2)^2 < 1-\delta_2$ based on $\delta_2 < 1$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{z}}_{t+1}\|^2\right] &\le \left(1-\frac{\delta_2}{2}\right)\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \left(3+\frac{1}{\kappa_1}\right)\lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{H}}_t\|^2\|\bar{z}_t\|^2\right] + \left(3+\frac{1}{\kappa_1}\right)\lambda_{z,t}^2\mathbb{E}\left[\|\hat{\boldsymbol{b}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{\vartheta}}_t\|^2\right] \\
&\le \left(1-\frac{\delta_2}{2}\right)\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 4mL_{g,1}^2\left(3+\frac{1}{\kappa_1}\right)\lambda_{z,t}^2\left(2\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right] + 2\mathbb{E}\left[\|\breve{z}_t\|^2\right]\right) + 4mL_{f,0}^2\left(3+\frac{1}{\kappa_1}\right)\lambda_{z,t}^2 + 4m\sigma_{z,t}^2 \\
&\le \left(1-\frac{\delta_2}{2}\right)\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 4m\sigma_{z,t}^2 + c_{\hat{z}1}\lambda_{z,t}^2\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right] + c_{\hat{z}2}\lambda_{z,t}^2,
\end{aligned}$$

where we have used $\mathbb{E}[\|\hat{\boldsymbol{H}}_t\|^2] \le 4mL_{g,1}^2$ and $\mathbb{E}[\|\hat{\boldsymbol{b}}_t\|^2] \le 4mL_{f,0}^2$ from Assumption 2.2, as well as $\mathbb{E}[\|\hat{\boldsymbol{\vartheta}}_t\|^2] \le 4m\sigma_{z,t}^2$ from Assumption 3.1 in the second inequality. Moreover, we have utilized $\mathbb{E}[\|\breve{z}_t\|^2] \le \frac{2\sigma_{f,1}^2+2L_{f,0}^2}{\mu_g^2}$ from Lemma C.1 in the last inequality. □

D.8. Estimation of $\mathbb{E}\left[\|\bar{z}_{t+1}-\breve{z}_{t+1}\|^2\right]$ in Lemma D.8 and Its Proof

Here, we use definitions $\bar{z}_t = \frac{1}{m}\sum_{i=1}^m z_{i,t}$ and $\breve{z}_t = (\nabla_{yy}^2 g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_y F_t(\bar{x}_t,\bar{y}_t)$. The update of $\bar{z}_{t+1}$ satisfies

$$\bar{z}_{t+1} = \bar{z}_t + \bar{\vartheta}_t - \lambda_{z,t}\frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t} + \lambda_{z,t}\bar{b}_t. \tag{84}$$

Lemma D.8. *Under Assumptions 2.1-2.3 and 3.1, for any $t > 0$, we have*

$$\begin{aligned}
\mathbb{E}\left[\|\bar{z}_{t+1}-\breve{z}_{t+1}\|^2\right] \le{}& \left(1-\frac{\lambda_{z,t}\mu_g}{4}+c_{z1}\lambda_{z,t}^2+c_{z2}\frac{\lambda_{x,t}^2}{\lambda_{z,t}}\right)\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right] \\
&+\left(c_{z3}\lambda_{z,t}+c_{z4}\kappa_2+c_{z5}\frac{\lambda_{x,t}^2}{\lambda_{z,t}}+c_{z6}\frac{\lambda_{y,t}^2}{\lambda_{z,t}}\right)\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \left(c_{z3}\lambda_{z,t}+c_{z4}\kappa_2+c_{z7}\frac{\lambda_{x,t}^2}{\lambda_{z,t}}+c_{z8}\frac{\lambda_{y,t}^2}{\lambda_{z,t}}\right)\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] \\
&+\left(c_{z9}\lambda_{z,t}+c_{z10}\frac{\lambda_{x,t}^2}{\lambda_{z,t}}\right)\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \left(c_{z11}\frac{\lambda_{x,t}^2}{\lambda_{z,t}}+c_{z12}\frac{\lambda_{y,t}^2}{\lambda_{z,t}}\right)\mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right] \\
&+c_{z13}\sigma_{z,t}^2+c_{z14}\frac{\sigma_{x,t}^2}{\lambda_{z,t}}+c_{z14}\frac{\sigma_{y,t}^2}{\lambda_{z,t}}+c_{z15}(\lambda_{z,t})^2+c_{z16}\frac{\lambda_{x,t}^2}{\lambda_{z,t}}+c_{z17}\frac{\lambda_{y,t}^2}{\lambda_{z,t}(t+1)}+c_{z18}\frac{1}{\lambda_{z,t}(t+2)^2},
\end{aligned} \tag{85}$$

where the constants c_{z1} to c_{z18} are given by $c_{z1} = c_{\bar{z}1}\left(1+\frac{\lambda_{z,0}\mu_g}{4}\right)$, $c_{z2} = c_{\breve{z}1}c_{\bar{x}4}\left(\lambda_{z,0}+\frac{4}{\mu_g}\right)$, $c_{z3} = \frac{c_{z1}c_{\bar{z}2}}{c_{\bar{z}1}}$, $c_{z4} = \frac{c_{z1}c_{\bar{z}3}}{c_{\bar{z}1}}$, $c_{z5} = \frac{c_{z2}c_{\bar{x}1}}{c_{\bar{x}4}}$, $c_{z6} = \frac{c_{z2}c_{\bar{y}1}}{c_{\bar{x}4}}$, $c_{z7} = \frac{c_{z2}c_{\bar{x}2}}{c_{\bar{x}4}}$, $c_{z8} = \frac{c_{z2}c_{\bar{y}2}}{c_{\bar{x}4}}$, $c_{z9} = \frac{c_{z1}c_{\bar{z}3}}{c_{\bar{z}1}}$, $c_{z10} = \frac{c_{z2}c_{\bar{x}3}}{c_{\bar{x}4}}$, $c_{z11} = \frac{c_{z2}c_{\bar{x}5}}{c_{\bar{x}4}}$, $c_{z12} = \frac{c_{z2}c_{\bar{y}3}}{c_{\bar{x}4}}$, $c_{z13} = \frac{c_{z1}}{c_{\bar{z}1}}$, $c_{z14} = \frac{c_{z2}}{c_{\bar{x}4}}$, $c_{z15} = \frac{c_{z1}c_{\bar{z}4}}{c_{\bar{z}1}}$, $c_{z15} = c_{z12}c_{\bar{z}4}$, $c_{z16} = \frac{c_{z2}c_{\bar{x}6}}{c_{\bar{x}4}}$, $c_{z17} = \frac{c_{z2}c_{\bar{y}4}}{c_{\bar{x}4}}$, and $c_{z18} = \frac{c_{z2}c_{\bar{z}2}}{c_{\bar{x}4}}$.

Proof. According to the update of $\bar{z}_{t+1}$ in (84) and the definition of $\breve{z}_t$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{z}_{t+1}-\breve{z}_t\|^2\right] ={}& \mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right] + \mathbb{E}\left[\|\bar{\vartheta}_t\|^2\right] + \lambda_{z,t}^2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t}-\bar{b}_t\right\|^2\right] \\
&-2\mathbb{E}\left[\left\langle \bar{z}_t-\breve{z}_t, \lambda_{z,t}\left(\frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t}-\bar{b}_t\right)\right\rangle\right].
\end{aligned} \tag{86}$$

The definition of $\hat{z}_{i,t}$ implies $z_{i,t} = \hat{z}_{i,t} + \bar{z}_t$, which further implies

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t}-\bar{H}_t\bar{z}_t\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m H_{i,t}\hat{z}_{i,t}\right\|^2\right] \le \frac{2(\sigma_{g,2}^2+L_{g,1}^2)}{m}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right], \tag{87}$$

where in the derivation we have used $\mathbb{E}[\|H_{i,t}\|^2] = \mathbb{E}[\|\nabla_{yy}^2 g_{i,t}(x_{i,t},y_{i,t})\|^2] \le 2(\sigma_{g,2}^2+L_{g,1}^2)$ from Lemma C.1.

Substituting (87) into the third term on the right hand side of (86) yields

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}H_{i,t}z_{i,t}-\bar{b}_t\right\|^2\right] &\leq 2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}H_{i,t}z_{i,t}-\bar{H}_t\bar{z}_t\right\|^2\right]+2\mathbb{E}\left[\left\|\bar{H}_t\bar{z}_t-\bar{b}_t\right\|^2\right]\\ &\leq \frac{4(\sigma_{g,2}^2+L_{g,1}^2)}{m}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right]+16\left(\sigma_{g,2}^2+L_{g,1}^2\right)\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right]+32\left(\sigma_{g,2}^2+L_{g,1}^2\right)\frac{\sigma_{f,1}^2+L_{f,0}^2}{\mu_g^2}+8\left(\sigma_{f,1}^2+L_{f,0}^2\right),\end{aligned} \tag{88}$$

where we have used the following inequality in the last inequality:

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{H}_t\bar{z}_t-\bar{b}_t\right\|^2\right] &\leq \mathbb{E}\left[2\left\|\bar{H}_t\right\|^2\left(2\left\|\bar{z}_t-\breve{z}_t\right\|^2+2\|\breve{z}_t\|^2\right)+2\left\|\bar{b}_t\right\|^2\right]\\ &\leq 8\left(\sigma_{g,2}^2+L_{g,1}^2\right)\mathbb{E}\left[\left\|\bar{z}_t-\breve{z}_t\right\|^2\right]+\frac{16\left(\sigma_{g,2}^2+L_{g,1}^2\right)\left(\sigma_{f,1}^2+L_{f,0}^2\right)}{\mu_g^2}+8\left(\sigma_{f,1}^2+L_{f,0}^2\right),\end{aligned} \tag{89}$$

and relations $\mathbb{E}[\|\bar{H}_t\|^2]\leq 2\sigma_{g,2}^2+2L_{g,1}^2$, $\mathbb{E}[\|\breve{z}_t\|^2]\leq\frac{2\sigma_{f,1}^2+2L_{f,0}^2}{\mu_g^2}$ and $\mathbb{E}[\|\bar{b}_t\|^2]\leq 2\sigma_{f,1}^2+2L_{f,0}^2$ from Lemma C.1.

To characterize the last term on the right hand side of (86), we define an auxiliary variable $\breve{z}_t'$ as follows:

$$\breve{z}_t'=(\bar{H}_t)^{-1}\bar{b}_t=\left(\frac{1}{m}\sum_{i=1}^{m}\nabla_{yy}^2 g_{i,t}(x_{i,t},y_{i,t})\right)^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}\nabla_y f_{i,t}(x_{i,t},y_{i,t})\right).$$

Then, we can obtain the following relationship:

$$\lambda_{z,t}\mathbb{E}\left[\left\langle\bar{z}_t-\breve{z}_t',(\bar{H}_t\bar{z}_t-\bar{b}_t)\right\rangle\right]=\lambda_{z,t}\mathbb{E}\left[\left\langle\bar{z}_t-\breve{z}_t',(\bar{H}_t\bar{z}_t-\bar{H}_t\breve{z}_t')\right\rangle\right]\geq\lambda_{z,t}\mu_g\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t'\|^2\right], \tag{90}$$

where we have used $\mathbb{E}_\xi\left[\nabla_{yy}g_t(x,y)\right]=\nabla_{yy}g(x,y)$ for any given (x,y) and Assumption 2.2 in the last inequality.

By using (90) and $2\langle a,\lambda_{z,t}b\rangle\leq\kappa_2 a^2+\frac{1}{\kappa_2}\lambda_{z,t}^2b^2$ for any $\kappa_2>0$, we obtain the following inequality:

$$\begin{aligned}&2\lambda_{z,t}\mathbb{E}\left[\left\langle\bar{z}_t-\breve{z}_t,\bar{H}_t\bar{z}_t-\bar{b}_t\right\rangle\right]\\ &=2\lambda_{z,t}\mathbb{E}\left[\left\langle\bar{z}_t-\breve{z}_t',\bar{H}_t\bar{z}_t-\bar{b}_t\right\rangle\right]+2\lambda_{z,t}\mathbb{E}\left[\left\langle\breve{z}_t'-\breve{z}_t,\bar{H}_t\bar{z}_t-\bar{b}_t\right\rangle\right]\\ &\geq 2\lambda_{z,t}\mu_g\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t'\|^2\right]+2\lambda_{z,t}\mathbb{E}\left[\left\langle\breve{z}_t'-\breve{z}_t,(\bar{H}_t\bar{z}_t-\bar{b}_t)\right\rangle\right]\\ &\geq\lambda_{z,t}\mu_g\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right]-2\lambda_{z,t}\mu_g\mathbb{E}\left[\|\breve{z}_t'-\breve{z}_t\|^2\right]-\left(\kappa_2\mathbb{E}\left[\|\breve{z}_t'-\breve{z}_t\|^2\right]+\frac{\lambda_{z,t}^2}{\kappa_2}\mathbb{E}\left[\|\bar{H}_t\bar{z}_t-\bar{b}_t\|^2\right]\right),\end{aligned} \tag{91}$$

where in the last inequality we have used the inequality $\|b\|^2\leq 2\|a\|^2+2\|b-a\|^2$ resulting in $\|a\|^2\geq\frac{\|b\|^2}{2}-\|b-a\|^2$ for any $a,b,c\in\mathbb{R}^q$.

According to definitions $\breve{z}_t=(\nabla_{yy}^2g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_yF_t(\bar{x}_t,\bar{y}_t)$ and $\breve{z}_t'=(\sum_{i=1}^m\nabla_{yy}^2g_{i,t}(x_{i,t},y_{i,t}))^{-1}\sum_{i=1}^m\nabla_yf_{i,t}(x_{i,t},y_{i,t})$, we estimate an upper bound on $\mathbb{E}\left[\|\breve{z}_t'-\breve{z}_t\|^2\right]$ as follows:

$$\begin{aligned}&\mathbb{E}\left[\|\breve{z}_t'-\breve{z}_t\|^2\right]\\ &=\mathbb{E}\left[\left\|(\sum_{i=1}^m\nabla_{yy}^2g_{i,t}(x_{i,t},y_{i,t}))^{-1}\sum_{i=1}^m\nabla_yf_{i,t}(x_{i,t},y_{i,t})-(\nabla_{yy}^2g_t(\bar{x}_t,\bar{y}_t))^{-1}\nabla_yF_t(\bar{x}_t,\bar{y}_t)\right\|^2\right]\\ &\leq 2\mathbb{E}\left[\left\|(\sum_{i=1}^m\nabla_{yy}^2g_{i,t}(x_{i,t},y_{i,t}))^{-1}-(\nabla_{yy}^2g_t(\bar{x}_t,\bar{y}_t))^{-1}\right\|^2\right]\mathbb{E}\left[\left\|\sum_{i=1}^m\nabla_yf_{i,t}(x_{i,t},y_{i,t})\right\|^2\right]\\ &\quad+2\mathbb{E}\left[\|(\nabla_{yy}^2g_t(\bar{x}_t,\bar{y}_t))^{-1}\|^2\right]\mathbb{E}\left[\left\|\nabla_yF_t(\bar{x}_t,\bar{y}_t)-\sum_{i=1}^m\nabla_yf_{i,t}(x_{i,t},y_{i,t})\right\|^2\right]\\ &\leq c_{\breve{z}3}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+c_{\breve{z}3}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right],\end{aligned} \tag{92}$$

where we have used Lemma C.1, as well as (18) and (19) from Lemma C.2 in the second inequality. The constants $c_{\breve{z}3}$ is given by $c_{\breve{z}3} = \frac{c_{\breve{z}1}}{2m}$ with $c_{\breve{z}1}$ given in the statement of Lemma D.3.

By using inequalities (87), (89), (91), and (92), the last term on the right hand side of (86) satisfies

$$\begin{aligned}
&-2\mathbb{E}\left[\left\langle \bar{z}_t - \breve{z}_t, \lambda_{z,t}\left(\frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t} - \bar{b}_t\right)\right\rangle\right] \\
&= -2\lambda_{z,t}\mathbb{E}\left[\left\langle \bar{z}_t - \breve{z}_t, \bar{H}_t\bar{z}_t - \bar{b}_t\right\rangle\right] + 2\lambda_{z,t}\mathbb{E}\left[\left\langle \bar{z}_t - \breve{z}_t, \bar{H}_t\bar{z}_t - \frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t}\right\rangle\right] \\
&\le -\lambda_{z,t}\mu_g\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + (2\lambda_{z,t}\mu_g + \kappa_2)\mathbb{E}\left[\|\breve{z}'_t - \breve{z}_t\|^2\right] + \frac{\lambda_{z,t}^2}{\kappa_2}\mathbb{E}\left[\|\bar{H}_t\bar{z}_t - \bar{b}_t\|^2\right] \\
&\quad + \left(\frac{\lambda_{z,t}\mu_g}{2}\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \frac{2\lambda_{z,t}}{\mu_g}\mathbb{E}\left[\left\|\bar{H}_t\bar{z}_t - \frac{1}{m}\sum_{i=1}^m H_{i,t}z_{i,t}\right\|^2\right]\right) \\
&\le -\frac{\lambda_{z,t}\mu_g}{2}\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + (2\lambda_{z,t}\mu_g + \kappa_2)c_{\breve{z}3}\left(\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]\right) + \frac{4(\sigma_{g,2}^2 + L_{g,1}^2)}{m\mu_g}\lambda_{z,t}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] \\
&\quad + \frac{8\left(\sigma_{g,2}^2 + L_{g,1}^2\right)}{\kappa_2}\lambda_{z,t}^2\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \frac{8\left(\sigma_{f,1}^2 + L_{f,0}^2\right)}{\kappa_2}\left(\frac{4\left(\sigma_{g,2}^2 + L_{g,1}^2\right)}{\mu_g^2} + 1\right)\lambda_{z,t}^2.
\end{aligned} \tag{93}$$

Substituting (88) and (93) into (86), we arrive at

$$\begin{aligned}
\mathbb{E}\left[\|\bar{z}_{t+1} - \breve{z}_t\|^2\right] &\le \left(1 - \frac{\lambda_{z,t}\mu_g}{2} + c_{\breve{z}1}\lambda_{z,t}^2\right)\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \sigma_{z,t}^2 \\
&\quad + (c_{\breve{z}2}\lambda_{z,t} + \kappa_2 c_{\breve{z}3})\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + (c_{\breve{z}2}\lambda_{z,t} + \kappa_2 c_{\breve{z}3})\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + c_{\breve{z}3}\lambda_{z,t}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + c_{\breve{z}4}\lambda_{z,t}^2,
\end{aligned} \tag{94}$$

where the constants $c_{\breve{z}1}$ to $c_{\breve{z}4}$ are given by $c_{\breve{z}1} = 8\left(\sigma_{g,2}^2 + L_{g,1}^2\right)\left(2 + \frac{1}{\kappa_2}\right)$, $c_{\breve{z}2} = 2\mu_g c_{\breve{z}3}$, $c_{\breve{z}3} = \frac{4(\sigma_{g,2}^2 + L_{g,1}^2)}{m}(\frac{1}{\mu_g} + \lambda_{z,0})$, and $c_{\breve{z}4} = \left(\frac{32(\sigma_{g,2}^2 + L_{g,1}^2)(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2} + 8(\sigma_{f,1}^2 + L_{f,0}^2)\right)(1 + \frac{1}{\kappa_2})$.

We proceed to use the following decomposition:

$$\|\bar{z}_{t+1} - \breve{z}_{t+1}\|^2 \le \left(1 + \frac{\lambda_{z,t}\mu_g}{4}\right)\|\bar{z}_{t+1} - \breve{z}_t\|^2 + \left(1 + \frac{4}{\lambda_{z,t}\mu_g}\right)\|\breve{z}_{t+1} - \breve{z}_t\|^2. \tag{95}$$

Substituting (55) in Lemma D.3 into (95) yields

$$\begin{aligned}
\mathbb{E}\left[\|\bar{z}_{t+1} - \breve{z}_{t+1}\|^2\right] &\le \left(1 + \frac{\lambda_{z,t}\mu_g}{4}\right)\mathbb{E}\left[\|\bar{z}_{t+1} - \breve{z}_t\|^2\right] \\
&\quad + \left(1 + \frac{4}{\lambda_{z,t}\mu_g}\right)\left(c_{\breve{z}1}\mathbb{E}\left[\|\bar{x}_{t+1} - \bar{x}_t\|^2\right] + c_{\breve{z}1}\mathbb{E}\left[\|\bar{y}_{t+1} - \bar{y}_t\|^2\right] + \frac{c_{\breve{z}2}}{(t+2)^2}\right).
\end{aligned} \tag{96}$$

Further substituting (46) in Lemma D.1, (52) in Lemma D.2, and (94) into (96), we arrive at (85). □

D.9. Estimation of $\mathbb{E}\left[\|\bar{y}_{t+1} - y_{t+1}^*(\bar{x}_{t+1})\|^2\right]$ in Lemma D.9 and Its Proof

Here, we use definitions $\bar{y}_t = \frac{1}{m}\sum_{i=1}^m y_{i,t}$ and $y_t^*(\bar{x}_t) := \operatorname{argmin}_{y\in\mathbb{R}^q} g_t(\bar{x}_t, y)$ with $\bar{x}_t = \frac{1}{m}\sum_{i=1}^m x_{i,t}$. We express the update rule of $\bar{y}_{t+1}$ as follows:

$$\bar{y}_{t+1} = \bar{y}_t + \bar{\zeta}_t - \lambda_{y,t}\frac{1}{m}\sum_{i=1}^m \nabla_y g_{i,t}(x_{i,t}, y_{i,t}). \tag{97}$$

Lemma D.9. *Under Assumptions 2.1-2.3 and 3.1, for any $t > 0$, we have*

$$\begin{aligned}
\mathbb{E}\left[\|\bar{y}_{t+1} - y_{t+1}^*(\bar{x}_{t+1})\|^2\right] &\le \left(1 - \frac{\lambda_{y,t}\mu_g}{4} + c_{y1}\lambda_{y,t}^2\right)\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + c_{y2}\sigma_{y,t}^2 + c_{y3}\frac{\lambda_{y,t}^2}{t+1} \\
&\quad + c_{y4}\lambda_{y,t}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + c_{y5}\lambda_{y,t}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + \frac{c_{y6}}{\lambda_{y,t}(t+1)^2},
\end{aligned} \tag{98}$$

where the constants c_{y1} *to* c_{y6} *are given by* $c_{y1} = \left(1+\frac{\lambda_{y,0}\mu_g}{4}\right)c_{\bar{y}3}$, $c_{y2} = \frac{c_{y1}}{c_{\bar{y}3}}$, $c_{y3} = c_{y2}c_{\bar{y}4}$, $c_{y4} = c_{y2}\left(\frac{8(L_{g,1}^2+\sigma_{g,2}^2)}{m\mu_g}+c_{\bar{y}1}\lambda_{y,0}\right)$, $c_{y5} = c_{y2}\left(\frac{8(L_{g,1}^2+\sigma_{g,2}^2)}{m\mu_g}+c_{\bar{y}2}\lambda_{y,0}\right)$, *and* $c_{y6} = \left(\lambda_{y,0}+\frac{4}{\mu_g}\right)\frac{2\sigma_{g,1}^2(\mu_g^2+4L_{g,1}^2)}{\mu_g^4}$.

Proof. Taking the squared norm and the expectation on both sides of (97), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\bar{y}_{t+1}-y_t^*(\bar{x}_t)\|^2\right] \le \mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right]+\sigma_{y,t}^2+\lambda_{y,t}^2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})\right\|^2\right]\\ -2\lambda_{y,t}\mathbb{E}\left[\left\langle\bar{y}_t-y_t^*(\bar{x}_t),\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})\right\rangle\right].\end{aligned} \tag{99}$$

By using an argument similar to the derivation of (52), we have

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})\right\|^2\right] \le c_{\bar{y}1}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+c_{\bar{y}2}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]+c_{\bar{y}3}\mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right]+\frac{c_{\bar{y}4}}{t+1}. \tag{100}$$

By using (20) in Lemma C.2, we obtain

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})-\nabla_y g_t(\bar{x}_t,\bar{y}_t)\right\|^2\right] \le \frac{4(L_{g,1}^2+\sigma_{g,2}^2)}{m}\left(\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]\right), \tag{101}$$

which further implies

$$\begin{aligned}&-2\lambda_{y,t}\mathbb{E}\left[\left\langle\bar{y}_t-y_t^*(\bar{x}_t),\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})\right\rangle\right] = -2\lambda_{y,t}\mathbb{E}\left[\langle\bar{y}_t-y_t^*(x),\nabla_y g_t(\bar{x}_t,\bar{y}_t)\rangle\right]\\ &\quad+2\lambda_{y,t}\mathbb{E}\left[\left\langle\bar{y}_t-y_t^*(\bar{x}_t),\nabla_y g_t(\bar{x}_t,\bar{y}_t)-\frac{1}{m}\sum_{i=1}^m\nabla_y g_{i,t}(x_{i,t},y_{i,t})\right\rangle\right]\\ &\le -\lambda_{y,t}\mu_g\mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right]+\frac{\lambda_{y,t}\mu_g}{2}\mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right]+\frac{8(L_{g,1}^2+\sigma_{g,2}^2)\lambda_{y,t}}{m\mu_g}\left(\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]\right),\end{aligned} \tag{102}$$

where we have used Assumption 2.2 and (101) in the last inequality.

Substituting (100) and (102) into (99), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\bar{y}_{t+1}-y_t^*(\bar{x}_t)\|^2\right] \le \left(1-\frac{\lambda_{y,t}\mu_g}{2}+c_{\bar{y}3}\lambda_{y,t}^2\right)\mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right]+\sigma_{y,t}^2+c_{\bar{y}4}\frac{\lambda_{y,t}^2}{t+1}\\ +\left(\frac{8(L_{g,1}^2+\sigma_{g,2}^2)}{m\mu_g}+c_{\bar{y}1}\lambda_{y,0}\right)\lambda_{y,t}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+\left(\frac{8(L_{g,1}^2+\sigma_{g,2}^2)}{m\mu_g}+c_{\bar{y}2}\lambda_{y,0}\right)\lambda_{y,t}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right].\end{aligned} \tag{103}$$

We proceed to use the following decomposition:

$$\mathbb{E}\left[\|\bar{y}_{t+1}-y_{t+1}^*(\bar{x}_{t+1})\|^2\right] = \left(1+\frac{\lambda_{y,t}\mu_g}{4}\right)\mathbb{E}\left[\|\bar{y}_{t+1}-y_t^*(\bar{x}_t)\|^2\right]+\left(1+\frac{4}{\lambda_{y,t}\mu_g}\right)\mathbb{E}\left[\|y_{t+1}^*(\bar{x}_{t+1})-y_t^*(\bar{x}_t)\|^2\right]. \tag{104}$$

By substituting (35) and (103) into (104), we arrive at (98). □

D.10. Consensus Errors of Algorithm 2

In this subsection, we summarize the consensus errors of the iterative variables generated by Algorithm 2. The analysis is based on the definitions: $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \mathbf{1}_m\otimes\bar{x}_t$, $\hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \mathbf{1}_m\otimes\bar{y}_t$, and $\hat{\boldsymbol{z}}_t = \boldsymbol{z}_t - \mathbf{1}_m\otimes\bar{z}_t$.

Lemma D.10. *Under Assumptions 2.1-2.3 and 3.1, if the stepsize rates satisfy $1 > v_x > v_y > v_z > 0$ and the rates of DP-noise variances satisfy $2\varsigma_x > v_z + v_y$, $2\varsigma_y > v_z + v_y$ and $2\varsigma_z > v_y$, then the following inequality always holds:*

$$\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] \leq \frac{C_0}{(t+1)^{\beta_0}}, \tag{105}$$

where the rate β_0 is given by $\beta_0 = \min\{2\varsigma_x - v_z - v_y, 2\varsigma_y - v_z - v_y, 2\varsigma_z - v_y, 2 - 2v_y\}$ and $C_0 > 0$ is some constant.

Proof. We sum up both sides of (67), (70), (78), (85), and (98) to obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{t+1}\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{y}}_{t+1}\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{z}}_{t+1}\|^2\right] + \mathbb{E}\left[\|\bar{z}_{t+1} - \breve{z}_{t+1}\|^2\right] + \mathbb{E}\left[\|\bar{y}_{t+1} - y_t^*(\bar{x}_{t+1})\|^2\right] \\
&\leq \left(1 - \frac{\delta_2}{2} + c_{\hat{x}1}\lambda_{x,t}^2 + c_{\hat{y}2}\lambda_{y,t}^2 + c_{z3}\lambda_{z,t} + c_{z4}\kappa_2 + c_{z5}\frac{\lambda_{x,t}^2}{\lambda_{z,t}} + c_{z6}\frac{\lambda_{y,t}^2}{\lambda_{z,t}} + c_{y4}\lambda_{y,t}\right)\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] \\
&\quad + \left(1 - \frac{\delta_2}{2} + c_{\hat{y}1}\lambda_{y,t}^2 + c_{\hat{x}2}\lambda_{x,t}^2 + c_{z3}\lambda_{z,t} + c_{z4}\kappa_2 + c_{z7}\frac{\lambda_{x,t}^2}{\lambda_{z,t}} + c_{z8}\frac{\lambda_{y,t}^2}{\lambda_{z,t}} + c_{y5}\lambda_{y,t}\right)\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] \\
&\quad + \left(1 - \frac{\delta_2}{2} + c_{\hat{x}3}\lambda_{x,t}^2 + c_{z9}\lambda_{z,t} + c_{z10}\frac{\lambda_{x,t}^2}{\lambda_{z,t}}\right)\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] \\
&\quad + \left(1 - \frac{\lambda_{z,t}\mu_g}{4} + c_{z1}\lambda_{z,t}^2 + c_{z2}\frac{\lambda_{x,t}^2}{\lambda_{z,t}} + c_{\hat{x}4}\lambda_{x,t}^2 + c_{\hat{z}1}\lambda_{z,t}^2\right)\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] \\
&\quad + \left(1 - \frac{\lambda_{y,t}\mu_g}{4} + c_{y1}\lambda_{y,t}^2 + c_{\hat{x}5}\lambda_{x,t}^2 + c_{\hat{y}3}\lambda_{y,t}^2 + c_{z11}\frac{\lambda_{x,t}^2}{\lambda_{z,t}} + c_{z12}\frac{\lambda_{y,t}^2}{\lambda_{z,t}}\right)\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] \\
&\quad + 4m\sigma_{x,t}^2 + (4m + c_{y2})\sigma_{y,t}^2 + (4m + c_{z13})\sigma_{z,t}^2 + c_{z14}\frac{\sigma_{x,t}^2}{\lambda_{z,t}} + c_{z14}\frac{\sigma_{y,t}^2}{\lambda_{z,t}} + c_{\hat{x}6}\lambda_{x,t}^2 + (c_{\hat{y}4} + c_{y3})\frac{\lambda_{y,t}^2}{t+1} \\
&\quad + (c_{\hat{z}2} + c_{z15})(\lambda_{z,t}^2) + c_{z16}\frac{\lambda_{x,t}^2}{\lambda_{z,t}} + c_{z17}\frac{\lambda_{y,t}^2}{\lambda_{z,t}(t+1)} + \frac{c_{y6}}{\lambda_{y,t}(t+1)^2} + \frac{c_{z18}}{\lambda_{z,t}(t+1)^2}.
\end{aligned} \tag{106}$$

To guarantee $c_{z4}\kappa_2 \leq \frac{\delta_2}{4}$, we select $\kappa_2 \leq \frac{\delta_2}{4c_{z4}}$. Furthermore, considering decaying stepsizes satisfying $\lambda_{x,t} \leq \lambda_{x,0}$, $\lambda_{y,t} \leq \lambda_{y,0}$, and $\lambda_{z,t} \leq \lambda_{z,0}$, we can choose the initial stepsizes $\lambda_{x,0}$, $\lambda_{y,0}$, and $\lambda_{z,0}$ to satisfy the following inequalities:

$$\begin{cases}
\frac{\delta_2}{4} \geq \frac{\lambda_{y,0}\mu_g}{8} + c_{\hat{x}1}\lambda_{x,0}^2 + c_{\hat{y}2}\lambda_{y,0}^2 + c_{z3}\lambda_{z,0} + c_{z5}\frac{\lambda_{x,0}^2}{\lambda_{z,0}} + c_{z6}\frac{\lambda_{y,0}^2}{\lambda_{z,0}} + c_{y4}\lambda_{y,0}, \\
\frac{\delta_2}{4} \geq \frac{\lambda_{y,0}\mu_g}{8} + c_{\hat{y}1}\lambda_{y,0}^2 + c_{\hat{x}2}\lambda_{x,0}^2 + c_{z3}\lambda_{z,0} + c_{z7}\frac{\lambda_{x,0}^2}{\lambda_{z,0}} + c_{z8}\frac{\lambda_{y,0}^2}{\lambda_{z,t}} + c_{y5}\lambda_{y,0}, \\
\frac{\delta_2}{2} \geq \frac{\lambda_{y,0}\mu_g}{8} + c_{\hat{x}3}\lambda_{x,0}^2 + c_{z9}\lambda_{z,0} + c_{z10}\frac{\lambda_{x,0}^2}{\lambda_{z,0}}, \\
\frac{\mu_g}{8} \geq c_{z1}\lambda_{z,0} + c_{z2}\frac{\lambda_{x,0}^2}{\lambda_{z,0}^2} + c_{\hat{x}4}\frac{\lambda_{x,0}^2}{\lambda_{z,0}} + c_{\hat{z}1}\lambda_{z,0}, \\
\frac{\mu_g}{8} \geq c_{y1}\lambda_{y,0} + c_{\hat{x}5}\frac{\lambda_{x,0}^2}{\lambda_{y,0}} + c_{\hat{y}3}\lambda_{y,0} + c_{z11}\frac{\lambda_{x,0}^2}{\lambda_{z,0}\lambda_{y,0}} + c_{z12}\frac{\lambda_{y,0}}{\lambda_{z,0}}.
\end{cases} \tag{107}$$

It should be noted that in practical applications, the initial stepsizes $\lambda_{x,0}$, $\lambda_{y,0}$, and $\lambda_{z,0}$ can be chosen as any positive constants, without strictly following (107). This flexibility is due to the decaying property of the terms on the right hand side of (107), which guarantees that there will be a time instant $T_0 > 0$ such that (107) is valid for all $t > T_0$.

Considering the relations in (107), inequality (106) can be rewritten as

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{t+1}\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{y}}_{t+1}\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{z}}_{t+1}\|^2\right] + \mathbb{E}\left[\|\bar{z}_{t+1} - \breve{z}_{t+1}\|^2\right] + \mathbb{E}\left[\|\bar{y}_{t+1} - y_t^*(\bar{x}_{t+1})\|^2\right] \\
&\leq \left(1 - \frac{\lambda_{y,0}\mu_g}{8(t+1)^{v_y}}\right)\left(\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right]\right) + \Phi_t,
\end{aligned} \tag{108}$$

where Φ_t is given by

$$\Phi_t = \frac{4m(\sigma_x^+)^2}{(t+1)^{2\varsigma_x}} + \frac{(4m+c_{y2})(\sigma_y^+)^2}{(t+1)^{2\varsigma_y}} + \frac{(4m+c_{z13})(\sigma_z^+)^2}{(t+1)^{2\varsigma_z}} + \frac{c_{z14}(\sigma_x^+)^2}{\lambda_{z,0}^2(t+1)^{2\varsigma_x-v_z}} + \frac{c_{z14}(\sigma_y^+)^2}{\lambda_{z,0}^2(t+1)^{2\varsigma_y-v_z}} + \frac{c_{\hat{x}6}\lambda_{x,0}^2}{(t+1)^{2v_x}} + \frac{(c_{\hat{y}4}+c_{y3})\lambda_{y,0}^2}{(t+1)^{2v_y+1}} + \frac{(c_{\hat{z}2}+c_{z15})\lambda_{z,0}^2}{(t+1)^{2v_z}} + \frac{c_{z16}\lambda_{x,0}^2}{\lambda_{z,0}(t+1)^{2v_x-v_z}} + \frac{c_{z17}\lambda_{y,0}^2}{\lambda_{z,0}(t+1)^{2v_y+1-v_z}} + \frac{c_{y6}}{\lambda_{y,0}(t+1)^{2-v_y}} + \frac{c_{z18}}{\lambda_{z,0}(t+1)^{2-v_z}} \leq \frac{c_1}{(t+1)^s}, \tag{109}$$

with $c_1 = 4m(\sigma_x^+)^2 + (4m+c_{y2})(\sigma_y^+)^2 + (4m+c_{z13})(\sigma_z^+)^2 + c_{z14}(\sigma_x^+)^2 + c_{z14}(\sigma_y^+)^2 + c_{\hat{x}6}\lambda_{x,0}^2 + (c_{\hat{y}4}+c_{y3})\lambda_{y,0}^2 + (c_{\hat{z}2}+c_{z15})(\lambda_{z,0})^2 + c_{z16}\lambda_{x,0}^2 + c_{z17}\lambda_{y,0}^2 + \frac{c_{y6}}{\lambda_{y,0}} + \frac{c_{z18}}{\lambda_{z,0}}$, and $s = \min\{2\varsigma_x - v_z, 2\varsigma_y - v_z, 2\varsigma_z, 2 - v_y\}$.

Recalling the conditions $1 > v_x > v_y > v_z > 0$, $2\varsigma_x > v_z + v_y$, $2\varsigma_y > v_z + v_y$, and $2\varsigma_z > v_y$ given in the lemma statement, we know that $s > v_y$ always holds. Hence, using Lemma B.2 leads to (105). □

To accurately characterize the consensus error of iterative variables generated by Algorithm 2, we present the following lemma, which is derived from Lemma D.10.

Lemma D.11. *Under the same assumptions given in Lemma D.10, we have*

$$\begin{aligned} &\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] \leq \frac{\hat{c}_x}{(t+1)^{2\varsigma_x}}, \qquad \mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] \leq \frac{\hat{c}_y}{(t+1)^{2\varsigma_y}}, \qquad \mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] \leq \frac{\bar{c}_y}{(t+1)^{\min\{2\varsigma_y - v_y, 2-2v_y\}}}, \\ &\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] \leq \frac{\hat{c}_z}{(t+1)^{2\varsigma_z}}, \quad \mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] \leq \frac{\bar{c}_z}{(t+1)^{\min\{2\varsigma_x - 2v_z, 2\varsigma_y - 2v_z, 2\varsigma_z - v_z\}}}, \end{aligned} \tag{110}$$

where the constants $\hat{c}_x$, $\hat{c}_y$, $\hat{c}_z$, $\bar{c}_y$, and $\bar{c}_z$ are given in (112), (113), (114), (116), and (118), respectively.

Proof. Combing (105) in Lemma D.10 with (67) in Lemma D.5 yields

$$\begin{aligned} \mathbb{E}\left[\|\hat{\boldsymbol{x}}_{t+1}\|^2\right] &\leq \left(1 - \frac{\delta_2}{2} + c_{\hat{x}1}\lambda_{x,t}^2\right)\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \frac{4m(\sigma_x^+)^2}{(t+1)^{2\varsigma_x}} + \frac{\sum_{i=2}^5 c_{\hat{x}i}C_0\lambda_{x,0}^2}{(t+1)^{2v_x+\beta_0}} + \frac{c_{\hat{x}6}\lambda_{x,0}^2}{(t+1)^{2v_x}} \\ &\leq \left(1 - \frac{\delta_2}{4}\right)\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \frac{c_x}{(t+1)^{2\varsigma_x}}, \end{aligned} \tag{111}$$

where the constant c_x is given by $c_x = 4m(\sigma_x^+)^2 + \sum_{i=2}^5 c_{\hat{x}i}C_0\lambda_{x,0}^2 + c_{\hat{x}6}\lambda_{x,0}^2$.

By using Lemma 11 from Chen & Wang (2023), we can obtain the following inequality:

$$\mathbb{E}\left[\|\hat{x}_t\|^2\right] \leq \mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] \leq \frac{\hat{c}_x}{(t+1)^{2\varsigma_x}} \quad \text{with} \quad \hat{c}_x = c_x\left(\frac{8\varsigma_x}{e\ln(\frac{8}{8-\delta_2})}\right)^{2\varsigma_x}\left(\frac{\mathbb{E}\left[\|\hat{\boldsymbol{x}}_0\|^2\right](4-\delta_2)}{4c_x} + \frac{8}{\delta_2}\right). \tag{112}$$

By combining (105) in Lemma D.10 with (70) in Lemma D.6 and (78) in Lemma D.7, we use again Lemma 11 from Chen & Wang (2023) to obtain

$$\mathbb{E}\left[\|\hat{y}_t\|^2\right] \leq \mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] \leq \frac{\hat{c}_y}{(t+1)^{2\varsigma_y}} \quad \text{with} \quad \hat{c}_y = c_y\left(\frac{8\varsigma_y}{e\ln(\frac{8}{8-\delta_2})}\right)^{2\varsigma_y}\left(\frac{\mathbb{E}\left[\|\hat{\boldsymbol{y}}_0\|^2\right](4-\delta_2)}{4c_y} + \frac{8}{\delta_2}\right), \tag{113}$$

$$\mathbb{E}\left[\|\hat{z}_t\|^2\right] \leq \mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] \leq \frac{\hat{c}_z}{(t+1)^{2\varsigma_z}} \quad \text{with} \quad \hat{c}_z = c_z\left(\frac{8\varsigma_z}{e\ln(\frac{8}{8-\delta_2})}\right)^{2\varsigma_z}\left(\frac{\mathbb{E}\left[\|\hat{\boldsymbol{z}}_0\|^2\right](4-\delta_2)}{4c_z} + \frac{8}{\delta_2}\right), \tag{114}$$

where c_y and c_z are given by $c_y = 4m(\sigma_y^+)^2 + (c_{\hat{y}2} + c_{\hat{y}3})C_0\lambda_{y,0}^2 + c_{\hat{y}4}\lambda_{y,0}^2$ and $c_z = 4m(\sigma_z^+)^2 + c_{\hat{z}1}C_0\lambda_{z,0}^2 + c_{\hat{z}2}\lambda_{z,0}^2$.

Utilizing (105) in Lemma D.10, (107), (112), (113), and (98) in Lemma D.9, we obtain

$$\begin{aligned} &\mathbb{E}\left[\|\bar{y}_{t+1} - y_{t+1}^*(\bar{x}_{t+1})\|^2\right] \leq \left(1 - \frac{\lambda_{y,0}\mu_g}{8(t+1)^{v_y}}\right)\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + \frac{c_{y2}(\sigma_y^+)^2}{(t+1)^{2\varsigma_y}} + \frac{c_{y3}\lambda_{y,0}^2}{(t+1)^{2v_y+1}} + \frac{c_{y4}\lambda_{y,0}\hat{c}_x}{(t+1)^{2\varsigma_x+v_y}} \\ &+ \frac{c_{y5}\lambda_{y,0}\hat{c}_y}{(t+1)^{2\varsigma_y+v_y}} + \frac{c_{y6}}{\lambda_{y,0}(t+1)^{2-v_y}} \leq \left(1 - \frac{\lambda_{y,0}\mu_g}{8(t+1)^{v_y}}\right)\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + \frac{c_{\bar{y}*}}{(t+1)^{\min\{2\varsigma_y, 2-v_y\}}}, \end{aligned} \tag{115}$$

where the constant $c_{\bar{y}*}$ is given by $c_{\bar{y}*} = c_{y2}(\sigma_y^+)^2 + c_{y3}\lambda_{y,0}^2 + c_{y4}\lambda_{y,0}\hat{c}_x + c_{y5}\lambda_{y,0}\hat{c}_y + c_{y6}$.

Applying Lemma B.2 to (115), we have

$$\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] \leq \frac{\bar{c}_y}{(t+1)^{\beta_{\bar{y}}}}, \tag{116}$$

where the rate $\beta_{\bar{y}}$ is given by $\beta_{\bar{y}} = \min\{2\varsigma_y - v_y, 2 - 2v_y\}$ and $\bar{c}_y$ is some positive constant.

Furthermore, we use (105) in Lemma D.10, (107), (112), (113), (114), and (85) in Lemma D.8 to obtain

$$\begin{aligned}
\mathbb{E}\left[\|\bar{z}_{t+1} - \breve{z}_{t+1}\|^2\right] &\leq \left(1 - \frac{\lambda_{z,t}\mu_g}{8}\right)\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \left(c_{z3}\lambda_{z,0} + c_{z4}\kappa_2 + c_{z5}\frac{\lambda_{x,0}^2}{\lambda_{z,0}} + c_{z6}\frac{\lambda_{y,0}^2}{\lambda_{z,0}}\right)\frac{\hat{c}_x}{(t+1)^{2\varsigma_x}} \\
&+ \left(c_{z3}\lambda_{z,0} + c_{z4}\kappa_2 + c_{z7}\frac{\lambda_{x,0}^2}{\lambda_{z,0}} + c_{z8}\frac{\lambda_{y,0}^2}{\lambda_{z,0}}\right)\frac{\hat{c}_x}{(t+1)^{2\varsigma_y}} + \left(c_{z9} + c_{z10}\frac{\lambda_{x,0}^2}{\lambda_{z,0}^2}\right)\frac{\lambda_{z,0}\hat{c}_z}{(t+1)^{v_z+2\varsigma_z}} \\
&+ \left(c_{z11}\frac{\lambda_{x,0}^2}{\lambda_{y,0}^2} + c_{z12}\right)\frac{\lambda_{y,0}^2\bar{c}_y}{\lambda_{z,0}(t+1)^{2v_y - v_z + \beta_{\bar{y}}}} + \frac{c_{z13}(\sigma_z^+)^2}{(t+1)^{2\varsigma_z}} + \frac{c_{z14}(\sigma_x^+)^2}{\lambda_{z,0}(t+1)^{2\varsigma_x - v_z}} + \frac{c_{z14}(\sigma_y^+)^2}{\lambda_{z,0}(t+1)^{2\varsigma_y - v_z}} \\
&+ \frac{c_{z15}(\lambda_{z,0})^2}{(t+1)^{2v_z}} + \frac{c_{z16}(\lambda_{x,0})^2}{\lambda_{z,0}(t+1)^{2v_x - v_z}} + \frac{c_{z17}(\lambda_{y,0})^2}{\lambda_{z,0}(t+1)^{2v_y + 1 - v_z}} + \frac{c_{z18}}{\lambda_{z,0}(t+1)^{2 - v_z}} \\
&\leq \left(1 - \frac{\lambda_{z,0}\mu_g}{8(t+1)^{v_z}}\right)\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] + \frac{c_{\bar{z}\breve{z}}}{(t+1)^{\min\{2\varsigma_x - v_z, 2\varsigma_y - v_z, 2\varsigma_z\}}},
\end{aligned} \tag{117}$$

where the constant $c_{c_{\bar{z}\breve{z}}}$ is given by $c_{c_{\bar{z}\breve{z}}} = 2c_{z4}\kappa_2\hat{c}_x + (2c_{z3}\hat{c}_x + c_{z9}\hat{c}_z)\lambda_{z,0} + ((c_{z5} + c_{z7})\hat{c}_x + c_{z10}\hat{c}_z + c_{z11}\bar{c}_y + c_{z16})\frac{\lambda_{x,0}^2}{\lambda_{z,0}} + c_{z13}(\sigma_z^+)^2 + ((c_{z6} + c_{z8})\hat{c}_x + c_{z12}\bar{c}_y + c_{z17})\frac{\lambda_{y,0}^2}{\lambda_{z,0}} + (c_{z14}((\sigma_x^+)^2 + (\sigma_y^+)^2) + c_{z18})\frac{1}{\lambda_{z,0}} + c_{z15}\lambda_{z,0}^2$.

By applying Lemma B.2 to (117), we arrive at

$$\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right] \leq \frac{\bar{c}_z}{(t+1)^{\beta_{\bar{z}}}}, \tag{118}$$

where the rate $\beta_{\bar{z}}$ is given by $\beta_{\bar{z}} = \min\{2\varsigma_x - 2v_z, 2\varsigma_y - 2v_z, 2\varsigma_z - v_z\}$ and $\bar{c}_z$ is some positive constant. □

D.11. Estimation of $\mathbb{E}\left[\|\bar{u}_t - u_t^*\|^2\right]$ in Lemma D.12 and Its Proof

Here, we use the definitions $\bar{u}_t = \frac{1}{m}\sum_{i=1}^m u_{i,t}$ and $\breve{u}_t = \nabla_x F_t(\bar{x}_t, \bar{y}_t) + \nabla_{xy}^2 g_t(\bar{x}_t, \bar{y}_t)\breve{z}_t$. Moreover, we define the following auxiliary variables:

$$\begin{aligned}
\bar{z}_t^* &= \left(\nabla_{yy}^2 g_t(\bar{x}_t, y^*(\bar{x}_t))\right)^{-1}\nabla_y F_t(\bar{x}_t, y^*(\bar{x}_t)), \quad \bar{u}_t^* = \nabla_x F_t(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_{xy}^2 g_t(\bar{x}_t, y^*(\bar{x}_t))\bar{z}_t^*, \\
z_t^* &= \left(\nabla_{yy}^2 g(\bar{x}_t, y^*(\bar{x}_t))\right)^{-1}\nabla_y F(\bar{x}_t, y^*(\bar{x}_t)), \quad u_t^* = \nabla_x F(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_{xy}^2 g(\bar{x}_t, y^*(\bar{x}_t))z_t^*.
\end{aligned} \tag{119}$$

Lemma D.12. *Under Assumptions 2.1- 2.3 and 3.1, for any $t > 0$, the following inequality always holds:*

$$\begin{aligned}
\mathbb{E}\left[\|\bar{u}_t - u_t^*\|^2\right] \leq \frac{3(c_{\bar{u}_1^*} + c_{\bar{u}_2^*} + c_{\bar{u}_4^*})}{t+1} &+ 3c_{\bar{u}_3^*}\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + 3c_{\bar{u}_5^*}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + 3c_{\bar{u}_5^*}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] \\
&+ 3c_{\bar{u}_6^*}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + 3c_{\bar{u}_7^*}\mathbb{E}\left[\|\bar{z}_t - \breve{z}_t\|^2\right],
\end{aligned} \tag{120}$$

where the constants $c_{\bar{u}_1^*}$ to $c_{\bar{u}_7^*}$ are given by $c_{\bar{u}_1^*} = 2\sigma_{f,1}^2 + \frac{8\sigma_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2} + 4L_{g,1}^2 c_{\bar{z}^*}$ with $c_{\bar{z}^*} = \frac{2\sigma_{f,1}^2}{\mu_g^2} + \frac{2L_{f,0}^2\sigma_{g,2}^2}{\mu_g^4}$, $c_{\bar{u}_2^*} = 12\sigma_{f,1}^2\left(1 + \frac{2(\sigma_{g,2}^2 + L_{g,1}^2)}{\mu_g^2}\right) + \frac{48(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}\left(\sigma_{g,2}^2 + \frac{\sigma_{g,2}^2(\sigma_{g,2}^2 + L_{g,1}^2)}{\mu_g^2}\right)$, $c_{\bar{u}_3^*} = 12L_{f,1}^2\left(1 + \frac{2(\sigma_{g,2}^2 + L_{g,1}^2)}{\mu_g^2}\right) + \frac{48(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}\left(L_{g,2}^2 + \frac{L_{g,2}^2(\sigma_{g,2}^2 + L_{g,1}^2)}{\mu_g^2}\right)$, $c_{\bar{u}_4^*} = 12\sigma_{f,1}^2 + \frac{48\sigma_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}$, $c_{\bar{u}_5^*} = \frac{12L_{f,1}^2}{m} + \frac{48L_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{m\mu_g^2}$, $c_{\bar{u}_6^*} = \frac{16(\sigma_{g,2}^2 + L_{g,1}^2)}{m}$, and $c_{\bar{u}_7^*} = 16(\sigma_{g,2}^2 + L_{g,1}^2)$.

Proof. We use the following decomposition:

$$\mathbb{E}\left[\|u_t^* - \bar{u}_t\|^2\right] \le 3\mathbb{E}\left[\|u_t^* - \bar{u}_t^*\|^2\right] + 3\mathbb{E}\left[\|\bar{u}_t^* - \breve{u}_t\|^2\right] + 3\mathbb{E}\left[\|\breve{u}_t - \bar{u}_t\|^2\right]. \tag{121}$$

By using Assumption 3.1, the definitions of $\bar{z}_t^*$ and z_t^*, and Lemma C.1, we have

$$\begin{aligned}\mathbb{E}\left[\|\bar{z}_t^* - z_t^*\|^2\right] &\le 2\mathbb{E}\left[\left\|\left(\nabla^2_{yy} g_t(\bar{x}_t, y^*(\bar{x}_t))\right)^{-1}\right\|^2 \|\nabla_y F_t(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_y F(\bar{x}_t, y^*(\bar{x}_t))\|^2\right] \\ &+ 2\mathbb{E}\left[\left\|\left(\nabla^2_{yy} g_t(\bar{x}_t, y^*(\bar{x}_t))\right)^{-1} - \left(\nabla^2_{yy} g(\bar{x}_t, y^*(\bar{x}_t))\right)^{-1}\right\|^2 \|\nabla_y F(\bar{x}_t, y^*(\bar{x}_t))\|^2\right] \le \frac{c_{\bar{z}^*}}{t+1},\end{aligned} \tag{122}$$

where $c_{\bar{z}^*}$ is given by $c_{\bar{z}^*} = \frac{2\sigma_{f,1}^2}{\mu_g^2} + \frac{2L_{f,0}^2\sigma_{g,2}^2}{\mu_g^4}$. Using the definitions of $\bar{u}_t^*$ and u_t^* and inequality (122), we further obtain

$$\begin{aligned}&\mathbb{E}\left[\|\bar{u}_t^* - u_t^*\|^2\right] \le 2\mathbb{E}\left[\|\nabla_x F_t(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_x F(\bar{x}_t, y^*(\bar{x}_t))\|^2\right] + 2\mathbb{E}\left[\|\nabla^2_{xy} g_t(\bar{x}_t, y^*(\bar{x}_t))\bar{z}_t^* - \nabla^2_{xy} g(\bar{x}_t, y^*(\bar{x}_t))z_t^*\|^2\right] \\ &\le \frac{2\sigma_{f,1}^2}{t+1} + 4\mathbb{E}\left[\|\nabla^2_{xy} g_t(\bar{x}_t, y^*(\bar{x}_t)) - \nabla^2_{xy} g(\bar{x}_t, y^*(\bar{x}_t))\|^2\|\bar{z}_t^*\|^2\right] + 4\mathbb{E}\left[\|\nabla^2_{xy} g(\bar{x}_t, y^*(\bar{x}_t))\|^2\|\bar{z}_t^* - z_t^*\|^2\right] \\ &\le \frac{2\sigma_{f,1}^2}{t+1} + \frac{8\sigma_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2(t+1)} + \frac{4L_{g,1}^2 c_{\bar{z}^*}}{t+1} = \frac{c_{\bar{u}_1^*}}{t+1},\end{aligned} \tag{123}$$

where we have used the relationship $\mathbb{E}\left[\|\bar{z}_t^*\|^2\right] \le \frac{2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}$ from Lemma C.1 in the last inequality.

We proceed to estimate an upper bound on $\mathbb{E}\left[\|\bar{u}_t^* - \breve{u}_t\|^2\right]$ in (121) based on the definitions of $\bar{u}_t^*$ and $\breve{u}_t$:

$$\begin{aligned}&\mathbb{E}\left[\|\bar{u}_t^* - \breve{u}_t\|^2\right] \le 2\mathbb{E}\left[\|\nabla_x F_t(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_x F_t(\bar{x}_t, \bar{y}_t)\|^2\right] + 2\mathbb{E}\left[\|\nabla^2_{xy} g_t(\bar{x}_t, y^*(\bar{x}_t))\bar{z}_t^* - \nabla^2_{xy} g_t(\bar{x}_t, \bar{y}_t)\breve{z}_t\|^2\right] \\ &\le 2\left(\frac{6\sigma_{f,1}^2}{t+1} + 6L_{f,1}^2\mathbb{E}\left[\|y_t^*(\bar{x}_t) - \bar{y}_t\|^2\right]\right) \\ &\quad + 4\mathbb{E}\left[\|\nabla^2_{xy} g_t(\bar{x}_t, y^*(\bar{x}_t))\bar{z}_t^* - \nabla^2_{xy} g_t(\bar{x}_t, \bar{y}_t)\bar{z}_t^*\|^2\right] + 4\mathbb{E}\left[\|\nabla^2_{xy} g_t(\bar{x}_t, \bar{y}_t)\bar{z}_t^* - \nabla^2_{xy} g_t(\bar{x}_t, \bar{y}_t)\breve{z}_t\|^2\right] \\ &\le \left(12\sigma_{f,1}^2 + \frac{48\sigma_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}\right)\frac{1}{t+1} + \left(12L_{f,1}^2 + \frac{48L_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}\right)\mathbb{E}\left[\|y_t^*(\bar{x}_t) - \bar{y}_t\|^2\right] \\ &\quad + 8(\sigma_{g,2}^2 + L_{g,1}^2)\mathbb{E}\left[\|\bar{z}_t^* - \breve{z}_t\|^2\right],\end{aligned} \tag{124}$$

where in the derivation we have used the following inequalities:

$$\begin{aligned}\mathbb{E}\left[\|\nabla_x F_t(x_2, y_2) - \nabla_x F_t(x_1, y_1)\|^2\right] &\le \frac{6\sigma_{f,1}^2}{t+1} + 6L_{f,1}^2\left(\|x_2 - x_1\|^2 + \|y_2 - y_1\|^2\right), \\ \mathbb{E}\left[\left\|\nabla^2_{xy} g_t(x_2, y_2) - \nabla^2_{xy} g_t(x_1, y_1)\right\|^2\right] &\le \frac{6\sigma_{g,2}^2}{t+1} + 6L_{g,2}^2\left(\|x_2 - x_1\|^2 + \|y_2 - y_1\|^2\right),\end{aligned} \tag{125}$$

for any given pairs $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^p \times \mathbb{R}^q$ and any $t > 0$. Moreover, we have utilized $\mathbb{E}\left[\|\bar{z}_t^*\|^2\right] \le \frac{2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^2}$ and $\mathbb{E}\left[\|\nabla^2_{xy} g_t(\bar{x}_t, \bar{y}_t)\|^2\right] \le 2(\sigma_{g,2}^2 + L_{g,1}^2)$ in the last inequality.

Next, we characterize the term $\mathbb{E}\left[\|\bar{z}_t^* - \breve{z}_t\|^2\right]$ in (124) as follows:

$$\begin{aligned}&\mathbb{E}\left[\|\bar{z}_t^* - \breve{z}_t\|^2\right] \le 2\mathbb{E}\left[\|\left(\nabla^2_{yy} g_t(\bar{x}_t, y^*(\bar{x}_t))\right)^{-1}\|^2\|\nabla_y F_t(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_y F_t(\bar{x}_t, \bar{y}_t)\|^2\right] \\ &\quad + 2\mathbb{E}\left[\|\left(\nabla^2_{yy} g_t(\bar{x}_t, y^*(\bar{x}_t))\right)^{-1} - \left(\nabla^2_{yy} g_t(\bar{x}_t, \bar{y}_t)\right)^{-1}\|^2\|\nabla_y F_t(\bar{x}_t, \bar{y}_t)\|^2\right] \\ &\le \left(\frac{12\sigma_{f,1}^2}{\mu_g^2} + \frac{24\sigma_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^4}\right)\frac{1}{t+1} + \left(\frac{12L_{f,1}^2}{\mu_g^2} + \frac{24L_{g,2}^2(\sigma_{f,1}^2 + L_{f,0}^2)}{\mu_g^4}\right)\mathbb{E}\left[\|y_t^*(\bar{x}_t) - \bar{y}_t\|^2\right],\end{aligned} \tag{126}$$

where we have used the following relationship in the last inequality:

$$\mathbb{E}\left[\left\|\left(\nabla^2_{yy}g_t(x_2,y_2)\right)^{-1}-\left(\nabla^2_{yy}g_t(x_1,y_1)\right)^{-1}\right\|^2\right]\leq\frac{6\sigma^2_{g,2}}{\mu_g^4(t+1)}+\frac{6L^2_{g,2}}{\mu_g^4}\left(\|x_2-x_1\|^2+\|y_2-y_1\|^2\right), \tag{127}$$

for any given pairs $(x_1,y_1),(x_2,y_2)\in\mathbb{R}^p\times\mathbb{R}^q$ and any $t>0$.

Substituting (126) into (124), we arrive at

$$\mathbb{E}\left[\|\bar{u}_t^*-\breve{u}_t\|^2\right]\leq\frac{c_{\bar{u}_2^*}}{t+1}+c_{\bar{u}_3^*}\mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right]. \tag{128}$$

Now we estimate an upper bound on $\mathbb{E}\left[\|\breve{u}_t-\bar{u}_t\|^2\right]$ in (121):

$$\mathbb{E}\left[\|\bar{u}_t-\breve{u}_t\|^2\right]\|\leq\frac{2}{m}\sum_{i=1}^m\left(\mathbb{E}\left[\|\nabla_x f_{i,t}(x_{i,t},y_{i,t})-\nabla_x f_{i,t}(\bar{x}_t,\bar{y}_t)\|^2\right]+\mathbb{E}\left[\|\nabla^2_{xy}g_{i,t}(x_{i,t},y_{i,t})z_{i,t}-\nabla^2_{xy}g_{i,t}(\bar{x}_t,\bar{y}_t)\breve{z}_t\|^2\right]\right). \tag{129}$$

The last term on the right hand side of (129) satisfies

$$\begin{aligned}
&\frac{2}{m}\sum_{i=1}^m\mathbb{E}\left[\|\nabla^2_{xy}g_{i,t}(x_{i,t},y_{i,t})z_{i,t}-\nabla^2_{xy}g_{i,t}(\bar{x}_t,\bar{y}_t)\breve{z}_t\|^2\right]\\
&\leq\frac{4}{m}\sum_{i=1}^m\mathbb{E}\left[\|\nabla^2_{xy}g_{i,t}(x_{i,t},y_{i,t})z_{i,t}-\nabla^2_{xy}g_{i,t}(x_{i,t},y_{i,t})\breve{z}_t\|^2\right]+\frac{4}{m}\sum_{i=1}^m\mathbb{E}\left[\|\nabla^2_{xy}g_{i,t}(x_{i,t},y_{i,t})\breve{z}_t-\nabla^2_{xy}g_{i,t}(\bar{x}_t,\bar{y}_t)\breve{z}_t\|^2\right]\\
&\leq\frac{8(\sigma^2_{g,2}+L^2_{g,1})}{m}\mathbb{E}\left[\|\boldsymbol{z}_t-\mathbf{1}_m\otimes\breve{z}_t\|^2\right]+\left(\frac{24\sigma^2_{g,2}}{t+1}+\frac{24L^2_{g,2}}{m}\left(\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]\right)\right)\frac{2(\sigma^2_{f,1}+L^2_{f,0})}{\mu_g^2},\\
&\leq\frac{48\sigma^2_{g,2}(\sigma^2_{f,1}+L^2_{f,0})}{\mu_g^2(t+1)}+\frac{48L^2_{g,2}(\sigma^2_{f,1}+L^2_{f,0})}{m\mu_g^2}\left(\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]\right)+\frac{16(\sigma^2_{g,2}+L^2_{g,1})}{m}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right]\\
&\quad+16(\sigma^2_{g,2}+L^2_{g,1})\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right],
\end{aligned} \tag{130}$$

where we have used the relationship $\mathbb{E}\left[\|\breve{z}_t\|^2\right]\leq\frac{2(\sigma^2_{f,1}+L^2_{f,0})}{\mu_g^2}$ in the third inequality.

By using (125) and substituting (130) into (129), we obtain

$$\mathbb{E}\left[\|\bar{u}_t-\breve{u}_t\|^2\right]\|\leq\frac{c_{\bar{u}_4^*}}{t+1}+c_{\bar{u}_5^*}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]+c_{\bar{u}_5^*}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]+c_{\bar{u}_6^*}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right]+c_{\bar{u}_7^*}\mathbb{E}\left[\|\bar{z}_t-\breve{z}_t\|^2\right], \tag{131}$$

where the constants $c_{\bar{u}_5^*}$ to $c_{\bar{u}_7^*}$ are given in the lemma statement.

Substituting (123), (128), and (131) into (121), we arrive at (120). □

E. Proof of Theorem 4.1

In this section, we establish convergence rates of Algorithm 2 under different convexity assumptions on the upper-level objective function F. Specifically, the convergence rate for a strongly convex F is given in Theorem E.1, for a convex F is given in Theorem E.2, and for a nonconvex F is given in Theorem E.3.

E.1. Convergence Rate for a Strongly Convex Upper-Level Objective Function

Theorem E.1. *Under Assumptions 2.1-2.3 and 3.1, if the upper-level objective function $F(x)$ is μ_f-strongly convex, the stepsize rates satisfy $0<v_z<v_y<v_x<1$, and the rates of DP-noise variances satisfy $2\varsigma_x>v_z+v_y$, $2\varsigma_y>v_z+v_y$, $2\varsigma_z>v_y$ and $2\varsigma_x>v_x$, then the following inequality always holds:*

$$\mathbb{E}\left[\|x_{i,T}-x^*\|^2\right]\leq\mathcal{O}\left(T^{-\beta_1}\right), \tag{132}$$

for all $T>0$ and any $i\in[m]$, where β_1 is given by $\beta_1=\min\{2\varsigma_x-v_x,2\varsigma_x-2v_z,2\varsigma_y-2v_z,2\varsigma_z-v_z,2\varsigma_y-v_y,2-2v_y\}$.

Proof. We first characterize the distance between the average sequence $\bar{x}_{t+1}$ and the optimal solution x^* to problem (1).

Recalling the update of $x_{i,t}$ in Algorithm 2 Step 7, we have $\bar{x}_{t+1} = \bar{x}_t + \bar{\chi}_t - \lambda_{x,t}\bar{u}_t$, which further implies

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{t+1} - x^*\|^2\right] \le \mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right] - 2\lambda_{x,t}\mathbb{E}\left[\langle \bar{x}_t - x^*, \bar{u}_t\rangle\right] \\
&\le \mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right] - 2\lambda_{x,t}\mathbb{E}\left[\langle \bar{x}_t - x^*, u_t^*\rangle\right] + 2\lambda_{x,t}\mathbb{E}\left[\langle \bar{x}_t - x^*, u_t^* - \bar{u}_t\rangle\right] \\
&\le \mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right] - \lambda_{x,t}\mu_f\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \frac{\lambda_{x,t}\mu_f}{2}\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \frac{2\lambda_{x,t}}{\mu_f}\mathbb{E}\left[\|u_t^* - \bar{u}_t\|^2\right] \\
&\le \left(1 - \frac{\lambda_{x,t}\mu_f}{2}\right)\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right] + \frac{2\lambda_{x,t}}{\mu_f}\mathbb{E}\left[\|u_t^* - \bar{u}_t\|^2\right],
\end{aligned} \tag{133}$$

where we have used the μ_f-strong convexity of $F(x)$, i.e., $2\lambda_{x,t}\langle \bar{x}_t - x^*, u_t^*\rangle \ge \lambda_{x,t}\mu_f\|\bar{x}_t - x^*\|^2$.

By substituting (48) and (49) into (47), we can obtain an upper bound on $\mathbb{E}\left[\|\bar{u}_t\|^2\right]$:

$$\mathbb{E}\left[\|\bar{u}_t\|^2\right] \le c_{\bar{x}1}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + c_{\bar{x}2}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] + c_{\bar{x}3}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + c_{\bar{x}4}\mathbb{E}\left[\|\bar{z}_t - \check{z}_t\|^2\right] + c_{\bar{x}5}\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + c_{\bar{x}6}. \tag{134}$$

By further substituting (134) and (120) in Lemma D.12 into (133), inequality (133) can be rewritten as

$$\begin{aligned}
\mathbb{E}\left[\|\bar{x}_{t+1} - x^*\|^2\right] \le &\left(1 - \frac{\lambda_{x,t}\mu_f}{2}\right)\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \sigma_{x,t}^2 + c_{x1}\lambda_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + c_{x2}\lambda_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] \\
&+ c_{x3}\lambda_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + c_{x4}\lambda_{x,t}\mathbb{E}\left[\|\bar{z}_t - \check{z}_t\|^2\right] + c_{x5}\lambda_{x,t}\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + c_{x6}\lambda_{x,t}^2 + c_{x7}\frac{\lambda_{x,t}}{t+1},
\end{aligned} \tag{135}$$

where the constants c_{x1} to c_{x7} are given by $c_{x1} = c_{\bar{x}1}\lambda_{x,0} + \frac{6c_{\bar{u}_5^*}}{\mu_f}$, $c_{x2} = c_{\bar{x}2}\lambda_{x,0} + \frac{6c_{\bar{u}_5^*}}{\mu_f}$, $c_{x3} = c_{\bar{x}3}\lambda_{x,0} + \frac{6c_{\bar{u}_6^*}}{\mu_f}$, $c_{x4} = c_{\bar{x}4}\lambda_{x,0} + \frac{6c_{\bar{u}_7^*}}{\mu_f}$, $c_{x5} = c_{\bar{x}5}\lambda_{x,0} + \frac{6c_{\bar{u}_3^*}}{\mu_f}$, $c_{x6} = c_{\bar{x}6}$, and $c_{x7} = \frac{6(c_{\bar{u}_1^*} + c_{\bar{u}_2^*} + c_{\bar{u}_4^*})}{\mu_f}$.

Using the results in Lemma D.11, we rewrite inequality (135) as follows:

$$\begin{aligned}
\mathbb{E}\left[\|\bar{x}_{t+1} - x^*\|^2\right] &\le \left(1 - \frac{\lambda_{x,0}\mu_f}{2(t+1)^{v_x}}\right)\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \frac{(\sigma_x^+)^2}{(t+1)^{2\varsigma_x}} + \frac{c_{x1}\lambda_{x,0}\hat{c}_x}{(t+1)^{2\varsigma_x + v_x}} + \frac{c_{x2}\lambda_{x,0}\hat{c}_y}{(t+1)^{2\varsigma_y + v_x}} + \frac{c_{x3}\lambda_{x,0}\hat{c}_z}{(t+1)^{2\varsigma_z + v_x}} \\
&+ \frac{c_{x4}\lambda_{x,0}\bar{c}_z}{(t+1)^{\min\{2\varsigma_x - 2v_z + v_x, 2\varsigma_y - 2v_z + v_x, 2\varsigma_z - v_z + v_x\}}} + \frac{c_{x5}\lambda_{x,0}\bar{c}_y}{(t+1)^{\min\{2\varsigma_y - v_y + v_x, 2 - 2v_y + v_x\}}} + \frac{c_{x6}\lambda_{x,0}^2}{(t+1)^{2v_x}} + \frac{c_{x7}\lambda_{x,0}}{(t+1)^{1+v_x}} \\
&\le \left(1 - \frac{\lambda_{x,0}\mu_f}{2(t+1)^{v_x}}\right)\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \frac{c_2}{(t+1)^{s_1}},
\end{aligned} \tag{136}$$

with $c_2 = (\sigma_x^+)^2 + (c_{x1}\hat{c}_x + c_{x2}\hat{c}_y + c_{x3}\hat{c}_z + c_{x4}\bar{c}_z + c_{x5}\bar{c}_y)\lambda_{x,0} + c_{x7}\lambda_{x,0}$ and $s_1 = \min\{2\varsigma_x, 2\varsigma_x - 2v_z + v_x, 2\varsigma_y - 2v_z + v_x, 2\varsigma_z - v_z + v_x, 2\varsigma_y - v_y + v_x, 2 - 2v_y + v_x\}$.

According to the conditions given in the theorem statement (or given in the statement of Theorem 4.1-(i)), we know that $s_1 > v_x$ always holds. Therefore, by using Lemma B.2, we arrive at

$$\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] \le \frac{c_3}{(t+1)^{\beta_1}}. \tag{137}$$

where the rate β_1 is given by $\beta_1 = \min\{2\varsigma_x - v_x, 2\varsigma_x - 2v_z, 2\varsigma_y - 2v_z, 2\varsigma_z - v_z, 2\varsigma_y - v_y, 2 - 2v_y\}$ and c_3 is some positive constant.

By using the definition $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \mathbf{1}_m \otimes \bar{x}_t$ and the first term of inequality (110) in Lemma D.11, we obtain

$$\mathbb{E}\left[\|x_{i,t} - x^*\|^2\right] \le 2\mathbb{E}\left[\|x_{i,t} - \bar{x}_t\|^2\right] + 2\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] \le C_1(t+1)^{-\beta_1}, \tag{138}$$

where the constant C_1 is given by $C_1 = 2(\hat{c}_x + c_3)$ and the rate β_1 satisfies $\beta_1 = \min\{2\varsigma_x - v_x, 2\varsigma_x - 2v_z, 2\varsigma_y - 2v_z, 2\varsigma_z - v_z, 2\varsigma_y - v_y, 2 - 2v_y\}$. Inequality (138) directly implies (132) in Theorem E.1 and (10) in Theorem 4.1-(i). □

E.2. Convergence Rate for a Convex Upper-Level Objective Function

Theorem E.2. *Under Assumptions 2.1-2.3 and 3.1, if the upper-level objective function $F(x)$ is convex, the stepsize rates satisfy $0 < v_z < v_y < v_x < 1$, and the rates of DP-noise variances satisfy $\varsigma_x > \frac{1}{2}$, $2\varsigma_x > v_z + v_y$, $2\varsigma_x > 2v_z + 2 - 2v_x$, $2\varsigma_y > 2v_z + 2 - 2v_x$, $2\varsigma_y > v_y + 2 - 2v_x$, $2\varsigma_y > v_z + v_y$, $2\varsigma_z > v_z + 2 - 2v_x$, and $2\varsigma_z > v_y$, then the following inequalities always hold:*

$$\begin{aligned} \mathbb{E}\left[\|\boldsymbol{x}_T - \mathbf{1}_m \otimes \bar{x}_T\|^2\right] &\leq \mathcal{O}\left(T^{-2\varsigma_x}\right), \\ \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[F(\bar{x}_t) - F(x^*)\right] &\leq \mathcal{O}\left(T^{v_x-1}\right), \\ \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[F(x_{i,t}) - F(x^*)\right] &\leq \mathcal{O}\left(T^{v_x-1}\right), \end{aligned} \tag{139}$$

for all $T > 0$ and any $i \in [m]$, where v_x is the rate of stepsize $\lambda_{x,t}$ given in Algorithm 2 satisfying $v_x - 1 < 0$.

Proof. (i) Based on the definition $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \mathbf{1}_m \otimes \bar{x}_t$, the first inequality in (139) follows naturally from (110) in Lemma D.11.

(ii) We now proceed to prove the second inequality in (139). Taking the squared norm and expectation on both sides of equality $\bar{x}_{t+1} = \bar{x}_t + \bar{\chi}_t - \lambda_{x,t}\bar{u}_t$ yields

$$\mathbb{E}\left[\|\bar{x}_{t+1} - x^*\|^2\right] \leq \mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right] - 2\mathbb{E}\left[\langle \bar{x}_t - x^*, \lambda_{x,t}\bar{u}_t\rangle\right]. \tag{140}$$

According to the definition $u_t^* = \nabla_x F(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_{xy}^2 g(\bar{x}_t, y^*(\bar{x}_t))z_t^*$, we have $u_t^* = \nabla F(\bar{x}_t)$. Using this relation and the convexity of F, the last term on the right hand side of (140) satisfies

$$\begin{aligned} &-2\mathbb{E}\left[\langle \bar{x}_t - x^*, \lambda_{x,t}\bar{u}_t\rangle\right] = 2\mathbb{E}\left[\langle x^* - \bar{x}_t, \lambda_{x,t}u_t^*\rangle\right] - 2\mathbb{E}\left[\langle \bar{x}_t - x^*, \lambda_{x,t}(\bar{u}_t - u_t^*)\rangle\right] \\ &\leq -2\lambda_{x,t}\mathbb{E}\left[F(\bar{x}_t) - F(x^*)\right] + a_t\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \frac{\lambda_{x,t}^2}{a_t}\mathbb{E}\left[\|\bar{u}_t - u_t^*\|^2\right], \end{aligned} \tag{141}$$

where a_t is an auxiliary decaying sequence satisfying $a_t = \frac{1}{(t+1)^r}$ with $1 < r < 2v_x$.

Substituting (141) into (140) leads to

$$\mathbb{E}\left[\|\bar{x}_{t+1} - x^*\|^2\right] \leq -2\lambda_{x,t}\mathbb{E}\left[F(\bar{x}_t) - F(x^*)\right] + (1 + a_t)\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \Phi_t, \tag{142}$$

where the term Φ_t is given by

$$\Phi_t = \frac{\lambda_{x,t}^2}{a_t}\mathbb{E}\left[\|\bar{u}_t - u_t^*\|^2\right] + \sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right]. \tag{143}$$

Since the relation $F(\bar{x}_t) \geq F(x^*)$ always holds, we drop the negative term $-2\lambda_{x,t}\mathbb{E}\left[F(\bar{x}_t) - F(x^*)\right]$ in (142) to obtain

$$\mathbb{E}\left[\|\bar{x}_{t+1} - x^*\|^2\right] \leq (1 + a_t)\mathbb{E}\left[\|\bar{x}_t - x^*\|^2\right] + \Phi_t \leq \left(\prod_{t=0}^{T}(1 + a_t)\right)\left(\mathbb{E}\left[\|\bar{x}_0 - x^*\|^2\right] + \sum_{t=0}^{T}\Phi_t\right). \tag{144}$$

By using the relation $\ln(1 + u) \leq u$ holding for any $u > 0$ and the definition $a_t = \frac{1}{(t+1)^r}$ with $1 < r < 2v_x$, we have

$$\ln\left(\prod_{t=0}^{T}(1 + a_t)\right) = \sum_{t=0}^{T}\ln(1 + a_t) \leq \sum_{t=0}^{T}a_t \leq a_0 + \sum_{t=1}^{T}\frac{1}{(t+1)^r} \leq a_0 + \int_1^{\infty}\frac{1}{x^r}dx \leq \frac{a_0(r-1)}{r-1}, \tag{145}$$

which implies $\prod_{t=0}^{T}(1 + a_t) \leq e^{\frac{a_0(r-1)}{r-1}}$. Then, inequality (144) can be rewritten as follows:

$$\mathbb{E}\left[\|\bar{x}_{t+1} - x^*\|^2\right] \leq e^{\frac{a_0(r-1)}{r-1}}\left(\mathbb{E}\left[\|\bar{x}_0 - x^*\|^2\right] + \sum_{t=0}^{T}\Phi_t\right). \tag{146}$$

Next, we estimate an upper bound on $\sum_{t=0}^{T}\Phi_t$, where Φ_t is defined in (143).

Substituting (120) and (134) into (143) and subsequently using (110) and the relation $a_t \leq 1$, we obtain

$$\begin{aligned}
\sum_{t=0}^{T}\Phi_t \leq & \sum_{t=0}^{T}\Bigg(\frac{3(c_{\bar{u}_1^*}+c_{\bar{u}_2^*}+c_{\bar{u}_4^*})\lambda_{x,t}^2}{a_t(t+1)}+\left(3c_{\bar{u}_3^*}+c_{\bar{x}5}\right)\frac{\lambda_{x,t}^2}{a_t}\mathbb{E}\left[\|\bar{y}_t-y_t^*(\bar{x}_t)\|^2\right]+\left(3c_{\bar{u}_5^*}+c_{\bar{x}1}\right)\frac{\lambda_{x,t}^2}{a_t}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right]\\
&+\left(3c_{\bar{u}_5^*}+c_{\bar{x}2}\right)\frac{\lambda_{x,t}^2}{a_t}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right]+\left(3c_{\bar{u}_6^*}+c_{\bar{x}3}\right)\frac{\lambda_{x,t}^2}{a_t}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right]+\left(3c_{\bar{u}_1^*}+c_{\bar{x}4}\right)\frac{\lambda_{x,t}^2}{a_t}\mathbb{E}\left[\|\bar{z}_t-\check{z}_t\|^2\right]+\sigma_{x,t}^2+c_{\bar{x}6}\lambda_{x,t}^2\Bigg)\\
\leq & \sum_{t=0}^{T}\frac{3\lambda_{x,0}^2(c_{\bar{u}_1^*}+c_{\bar{u}_2^*}+c_{\bar{u}_4^*})}{(t+1)^{2v_x-r+1}}+\sum_{t=0}^{T}\frac{\left(3c_{\bar{u}_3^*}+c_{\bar{x}5}\right)\bar{c}_y\lambda_{x,0}^2}{(t+1)^{\min\{2v_x-r+2\varsigma_y-v_y,2v_x-r+2-2v_y\}}}+\sum_{t=0}^{T}\frac{\left(3c_{\bar{u}_5^*}+c_{\bar{x}1}\right)\hat{c}_x\lambda_{x,0}^2}{(t+1)^{2v_x-r+2\varsigma_x}}\\
&+\sum_{t=0}^{T}\frac{\left(3c_{\bar{u}_5^*}+c_{\bar{x}2}\right)\hat{c}_y\lambda_{x,0}^2}{(t+1)^{2v_x-r+2\varsigma_y}}+\sum_{t=0}^{T}\frac{\left(3c_{\bar{u}_6^*}+c_{\bar{x}3}\right)\hat{c}_z\lambda_{x,0}^2}{(t+1)^{2v_x-r+2\varsigma_z}}+\sum_{t=0}^{T}\frac{\left(3c_{\bar{u}_1^*}+c_{\bar{x}4}\right)\bar{c}_z\lambda_{x,0}^2}{(t+1)^{\min\{2v_x-r+2\varsigma_x-2v_z,2v_x-r+2\varsigma_y-2v_z,2v_x-r+2\varsigma_z-v_z\}}}\\
&+\sum_{t=0}^{T}\frac{(\sigma_x^+)^2}{(t+1)^{2\varsigma_x}}+\sum_{t=0}^{T}\frac{c_{\bar{x}6}\lambda_{x,0}^2}{(t+1)^{2v_x}}.
\end{aligned}\tag{147}$$

By using the following inequality:

$$\sum_{t=0}^{T}\frac{1}{(t+1)^r}=1+\sum_{t=2}^{T+1}\frac{1}{t^s}\leq 1+\int_1^{\infty}\frac{1}{x^r}dx\leq\frac{r}{r-1},\tag{148}$$

and the constant r satisfying $1<r<2v_x$, we can rewrite inequality (147) as follows:

$$\begin{aligned}
\sum_{t=0}^{T}\Phi_t \leq & \frac{3\lambda_{x,0}^2(c_{\bar{u}_1^*}+c_{\bar{u}_2^*}+c_{\bar{u}_4^*})(2v_x-r+1)}{2v_x-r}+\frac{2(\sigma_x^+)^2\varsigma_x}{2\varsigma_x-1}+\frac{2c_{\bar{x}6}\lambda_{x,0}^2v_x}{2v_x-1}\\
&+\left(3c_{\bar{u}_3^*}+c_{\bar{x}5}\right)\bar{c}_y\lambda_{x,0}^2\max\left\{\frac{2v_x-r+2\varsigma_y-v_y}{2v_x-r+2\varsigma_y-v_y-1},\frac{2v_x-r+2-2v_y}{2v_x-r+1-2v_y}\right\}+\frac{\left(3c_{\bar{u}_5^*}+c_{\bar{x}1}\right)\hat{c}_x\lambda_{x,0}^2(2v_x-r+2\varsigma_x)}{2v_x-r+2\varsigma_x-1}\\
&+\frac{\left(3c_{\bar{u}_5^*}+c_{\bar{x}2}\right)\hat{c}_y\lambda_{x,0}^2(2v_x-r+2\varsigma_y)}{2v_x-r+2\varsigma_y-1}+\frac{\left(3c_{\bar{u}_6^*}+c_{\bar{x}3}\right)\hat{c}_z\lambda_{x,0}^2(2v_x-r+2\varsigma_z)}{2v_x-r+2\varsigma_z-1}\\
&+\left(3c_{\bar{u}_1^*}+c_{\bar{x}4}\right)\bar{c}_z\lambda_{x,0}^2\max\left\{\frac{2v_x-r+2\varsigma_x-2v_z}{2v_x-r+2\varsigma_x-2v_z-1},\frac{2v_x-r+2\varsigma_y-2v_z}{2v_x-r+2\varsigma_y-2v_z-1},\frac{2v_x-r+2\varsigma_z-v_z}{2v_x-r+2\varsigma_z-v_z-1}\right\}\triangleq c_4.
\end{aligned}\tag{149}$$

Substituting (149) into (146), we can arrive at

$$\mathbb{E}\left[\|\bar{x}_{t+1}-x^*\|^2\right]\leq e^{\frac{a_0(r-1)}{r-1}}\left(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+c_4\right).\tag{150}$$

We proceed to sum up both sides of (142) from 0 to T (T can be any positive integer):

$$\sum_{t=0}^{T}2\lambda_{x,t}\mathbb{E}\left[F(\bar{x}_t)-F(x^*)\right]\leq-\sum_{t=0}^{T}\mathbb{E}\left[\|\bar{x}_{t+1}-x^*\|^2\right]+\sum_{t=0}^{T}(1+a_t)\mathbb{E}\left[\|\bar{x}_t-x^*\|^2\right]+\sum_{t=0}^{T}\Phi_t.\tag{151}$$

The first and second terms on the right hand side of (151) can be simplified as follows:

$$\begin{aligned}
&\sum_{t=0}^{T}(1+a_t)\mathbb{E}\left[\|\bar{x}_t-x^*\|^2\right]-\sum_{t=0}^{T}\mathbb{E}\left[\|\bar{x}_{t+1}-x^*\|^2\right]\\
&\leq a_0\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+\sum_{t=1}^{T}a_t\mathbb{E}\left[\|\bar{x}_t-x^*\|^2\right]+\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]-\mathbb{E}\left[\|\bar{x}_{t+1}-x^*\|^2\right]\\
&\leq\sum_{t=1}^{T}\frac{1}{(t+1)^r}\left(e^{\frac{a_0(r-1)}{r-1}}\left(\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+c_4\right)\right)+(1+a_0)\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]\\
&\leq\left(\frac{re^{\frac{a_0(r-1)}{r-1}}}{r-1}+(1+a_0)\right)\mathbb{E}\left[\|\bar{x}_0-x^*\|^2\right]+\frac{c_4r}{r-1}\triangleq c_5,
\end{aligned}\tag{152}$$

where we have used (150) in the second inequality and (148) in the last inequality.

Substituting (149) and (152) into (151) and using $\lambda_{x,T} \leq \lambda_{x,t}$ for any $t \in [0,T]$ yield $\sum_{t=0}^{T} 2\lambda_{x,t}\mathbb{E}\left[F(\bar{x}_t) - F(x^*)\right] \leq c_4 + c_5$, which further implies

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[F(\bar{x}_t) - F(x^*)\right] \leq \frac{c_4+c_5}{2\lambda_{x,0}(T+1)^{1-v_x}} = \frac{C_2'}{(T+1)^{1-v_x}}, \tag{153}$$

with $C_2' = \frac{c_4+c_5}{2\lambda_{x,0}}$. Inequality (153) directly implies the second inequality in (139).

(iii) We now prove the third inequality in (139).

Assumption 2.2 implies $\mathbb{E}\left[F(x_{i,t}) - F(\bar{x}_t)\right] \leq L_{f,0}(\mathbb{E}\left[\|\hat{x}_t\|\right] + \mathbb{E}\left[\|\hat{y}_t\|\right])$. By using Lemma D.11, we have

$$\mathbb{E}\left[F(x_{i,t}) - F(\bar{x}_t)\right] \leq L_{f,0}\left(\frac{\sqrt{\hat{c}_x}}{(t+1)^{\varsigma_x}} + \frac{\sqrt{\hat{c}_y}}{(t+1)^{\varsigma_y}}\right). \tag{154}$$

Since $\sum_{t=0}^{T}\frac{1}{(t+1)^p} \leq \int_{x=0}^{T+1}\frac{1}{x^p}dx \leq \frac{(T+1)^{1-p}}{1-p}$ always holds for any $p \in (0,1)$, we arrive at

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[F(x_{i,t}) - F(\bar{x}_t)\right] \leq L_{f,0}\left(\frac{\sqrt{\hat{c}_x}}{(T+1)^{\varsigma_x}} + \frac{\sqrt{\hat{c}_y}}{(T+1)^{\varsigma_y}}\right) = \frac{C_2}{(T+1)^{\min\{\varsigma_x,\varsigma_y\}}}, \tag{155}$$

where the constant C_2 is given by $C_2 = L_{f,0}(\sqrt{\hat{c}_x} + \sqrt{\hat{c}_y})$.

According to the conditions $2\varsigma_x > v_z + v_y + 2 - 2v_x$ and $2\varsigma_y > v_z + v_y + 2 - 2v_x$ given in the theorem statement (or given in the statement of Theorem 4.1-(ii)), we have $1 - v_x < \varsigma_x$ and $1 - v_x < \varsigma_y$. Hence, by using (153), we arrive at the third inequality in (139) and (11) in Theorem 4.1-(ii). □

E.3. Convergence Rate for a Nonconvex Upper-Level Objective Function

Theorem E.3. *Under Assumptions 2.1-2.3 and 3.1, if the upper-level objective function $F(x)$ is nonconvex, the stepsize rates satisfy $0 < v_z < v_y < v_x < 1$, and the rates of DP-noise variances satisfy $\varsigma_x > \frac{1}{2}$, $2\varsigma_x > v_z + v_y$, $2\varsigma_x > 2v_z + 1 - v_x$, $2\varsigma_y > 2v_z + 1 - v_x$, $2\varsigma_y > v_y + 1 - v_x$, $2\varsigma_y > v_z + v_y$, $2\varsigma_z > v_z + 1 - v_x$, and $2\varsigma_z > v_y$, then the following inequalities always hold:*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_T - \mathbf{1}_m \otimes \bar{x}_T\|^2\right] &\leq \mathcal{O}\left(T^{-2\varsigma_x}\right),\\ \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right] &\leq \mathcal{O}\left(T^{v_x-1}\right),\end{aligned} \tag{156}$$

for all $T > 0$ and any $i \in [m]$, where v_x is the rate of stepsize $\lambda_{x,t}$ given in Algorithm 2 satisfying $v_x - 1 < 0$.

Proof. The first inequality in (156) follows naturally from (110) in Lemma D.11.

We proceed to prove the second inequality in (156).

Assumption 2.2 implies

$$F(\bar{x}_{t+1}) \leq F(\bar{x}_t) + \langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{L_{f,1}}{2}\|\bar{x}_{t+1} - \bar{x}_t\|. \tag{157}$$

Taking expectation on both sides of (157) yields

$$\mathbb{E}\left[F(\bar{x}_{t+1}) - F(\bar{x}_t)\right] \leq \mathbb{E}\left[\langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle\right] + \frac{L_{f,1}}{2}\mathbb{E}\left[\|\bar{x}_{t+1} - \bar{x}_t\|^2\right]. \tag{158}$$

Substituting the relation $\bar{x}_{t+1} - \bar{x}_t = \bar{\chi}_t - \lambda_{x,t}\bar{u}_t$ into the terms on the right hand side of (158) yields

$$\begin{aligned}&\mathbb{E}\left[\langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle\right] + \frac{L_{f,1}}{2}\mathbb{E}\left[\|\bar{x}_{t+1} - \bar{x}_t\|^2\right]\\ &= -\mathbb{E}\left[\langle \nabla F(\bar{x}_t), \lambda_{x,t}\bar{u}_t\rangle\right] + \frac{L_{f,1}}{2}\mathbb{E}\left[\|\bar{\chi}_t - \lambda_{x,t}\bar{u}_t\|^2\right]\\ &\leq -\mathbb{E}\left[\langle \nabla F(\bar{x}_t), \lambda_{x,t}\bar{u}_t\rangle\right] + \frac{L_{f,1}}{2}\left(\sigma_{x,t}^2 + \lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right]\right).\end{aligned} \tag{159}$$

The definition of u_t^* implies $u_t^* = \nabla F(\bar{x}_t)$. Hence, the first term on the right hand side of (159) satisfies

$$\begin{aligned}-\mathbb{E}\left[\langle \nabla F(\bar{x}_t), \lambda_{x,t}\bar{u}_t\rangle\right] &= -\lambda_{x,t}\mathbb{E}\left[\langle \nabla F(\bar{x}_t), u_t^*\rangle\right] - \lambda_{x,t}\mathbb{E}\left[\langle \nabla F(\bar{x}_t), \bar{u}_t - u_t^*\rangle\right] \\ &\leq -\lambda_{x,t}\mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right] + \frac{\lambda_{x,t}}{2}\mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right] + \frac{\lambda_{x,t}}{2}\mathbb{E}\left[\bar{u}_t - u_t^*\|^2\right] \\ &\leq -\frac{\lambda_{x,t}}{2}\mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right] + \frac{\lambda_{x,t}}{2}\mathbb{E}\left[\bar{u}_t - u_t^*\|^2\right].\end{aligned} \tag{160}$$

By substituting (159) and (160) into (158), we have

$$\mathbb{E}\left[F(\bar{x}_{t+1}) - F(\bar{x}_t)\right] \leq -\frac{\lambda_{x,t}}{2}\mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right] + \frac{\lambda_{x,t}}{2}\mathbb{E}\left[\bar{u}_t - u_t^*\|^2\right] + \frac{L_{f,1}}{2}\sigma_{x,t}^2 + \frac{L_{f,1}}{2}\lambda_{x,t}^2\mathbb{E}\left[\|\bar{u}_t\|^2\right]. \tag{161}$$

Summing up both sides of (161) from 0 to T and using the relationship $F(x^*) \leq F(\bar{x}_{t+1})$, we obtain

$$\begin{aligned}&\sum_{t=0}^{T}\frac{\lambda_{x,t}}{2}\mathbb{E}\left[\|\nabla F(\bar{x}_t)\|^2\right] \\ &\leq \mathbb{E}\left[F(\bar{x}_0) - F(x^*)\right] + \sum_{t=0}^{T}\frac{\lambda_{x,t}}{2}\mathbb{E}\left[\bar{u}_t - u_t^*\|^2\right] + \sum_{t=0}^{T}\frac{L_{f,1}(\sigma_x^+)^2}{2(t+1)^{2\varsigma_x}} + \sum_{t=0}^{T}\frac{L_{f,1}\lambda_{x,t}^2}{2}\mathbb{E}\left[\|\bar{u}_t\|^2\right].\end{aligned} \tag{162}$$

Combining (162) and the relation $\lambda_{x,t}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right] \leq \frac{\lambda_{x,t}}{2}\mathbb{E}\left[\|\nabla F(x_{i,t}) - F(\bar{x}_t)\|^2\right] + \frac{\lambda_{x,t}}{2}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right]$ yields

$$\sum_{t=0}^{T}\lambda_{x,t}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right] \leq \mathbb{E}\left[F(\bar{x}_0) - F(x^*)\right] + \sum_{t=0}^{t}\Phi_t. \tag{163}$$

where the term Φ_t is given by

$$\Phi_t = \lambda_{x,t}\mathbb{E}\left[\|\nabla F(\bar{x}_t) - \nabla F(x_{i,t})\|^2\right] + \frac{\lambda_{x,t}}{2}\mathbb{E}\left[\bar{u}_t - u_t^*\|^2\right] + \frac{L_{f,1}(\sigma_x^+)^2}{2(t+1)^{2\varsigma_x}} + \frac{L_{f,1}\lambda_{x,t}^2}{2}\mathbb{E}\left[\|\bar{u}_t\|^2\right]. \tag{164}$$

We proceed to estimate an upper bound on $\sum_{t=0}^{T}\Phi_t$.

Substituting (120) and (134) into (164), and then using Lemma B.1 and Lemma D.11, we have

$$\begin{aligned}\sum_{t=0}^{T}\Phi_t \leq &\sum_{t=0}^{T}\Bigg[\left(\frac{L_F}{m} + \frac{3c_{\bar{u}_5^*} + L_{f,1}c_{\bar{x}1}\lambda_{x,0}}{2}\right)\lambda_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{x}}_t\|^2\right] + \left(\frac{3c_{\bar{u}_5^*} + L_{f,1}c_{\bar{x}2}\lambda_{x,0}}{2}\right)\lambda_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{y}}_t\|^2\right] \\ &+ \left(\frac{3c_{\bar{u}_6^*} + L_{f,1}c_{\bar{x}3}\lambda_{x,0}}{2}\right)\lambda_{x,t}\mathbb{E}\left[\|\hat{\boldsymbol{z}}_t\|^2\right] + \left(\frac{3c_{\bar{u}_7^*} + L_{f,1}c_{\bar{x}4}\lambda_{x,0}}{2}\right)\lambda_{x,t}\mathbb{E}\left[\|\bar{z}_t - \check{z}_t\|^2\right] \\ &+ \left(\frac{3c_{\bar{u}_3^*} + L_{f,1}c_{\bar{x}5}\lambda_{x,0}}{2}\right)\lambda_{x,t}\mathbb{E}\left[\|\bar{y}_t - y_t^*(\bar{x}_t)\|^2\right] + \frac{3(c_{\bar{u}_1^*} + c_{\bar{u}_2^*} + c_{\bar{u}_4^*})}{2}\frac{\lambda_{x,t}}{t+1} + \frac{L_{f,1}(\sigma_x^+)^2}{2(t+1)^{2\varsigma_x}} + \frac{L_{f,1}c_{\bar{x}6}}{2}\lambda_{x,t}^2\Bigg] \\ \leq &\sum_{t=0}^{T}\frac{3\lambda_{x,0}(c_{\bar{u}_1^*} + c_{\bar{u}_2^*} + c_{\bar{u}_4^*})}{2(t+1)^{1+v_x}} + \sum_{t=0}^{T}\frac{L_{f,1}(\sigma_x^+)^2}{2(t+1)^{2\varsigma_x}} + \sum_{t=0}^{T}\frac{L_{f,1}c_{\bar{x}6}(\lambda_{x,0})^2}{2(t+1)^{2v_x}} \\ &+ \sum_{t=0}^{T}\left(\frac{L_F}{m} + \frac{3c_{\bar{u}_5^*} + L_{f,1}c_{\bar{x}1}\lambda_{x,0}}{2}\right)\frac{\hat{c}_x\lambda_{x,0}}{(t+1)^{2\varsigma_x+v_x}} + \sum_{t=0}^{T}\left(\frac{3c_{\bar{u}_5^*} + L_{f,1}c_{\bar{x}2}\lambda_{x,0}}{2}\right)\frac{\hat{c}_y\lambda_{x,0}}{(t+1)^{2\varsigma_y+v_x}} \\ &+ \sum_{t=0}^{T}\left(\frac{3c_{\bar{u}_6^*} + L_{f,1}c_{\bar{x}3}\lambda_{x,0}}{2}\right)\frac{\hat{c}_z\lambda_{x,0}}{(t+1)^{2\varsigma_z+v_x}} \\ &+ \sum_{t=0}^{T}\left(\frac{3c_{\bar{u}_7^*} + L_{f,1}c_{\bar{x}4}\lambda_{x,0}}{2}\right)\frac{\bar{c}_z\lambda_{x,0}}{(t+1)^{\min\{2\varsigma_x-2v_z+v_x,2\varsigma_y-2v_z+v_x,2\varsigma_z-v_z+v_x\}}} \\ &+ \sum_{t=0}^{T}\left(\frac{3c_{\bar{u}_3^*} + L_{f,1}c_{\bar{x}5}\lambda_{x,0}}{2}\right)\frac{\bar{c}_y\lambda_{x,0}}{(t+1)^{\min\{2\varsigma_y-v_y+v_x,2-2v_y+v_x\}}}.\end{aligned} \tag{165}$$

Using inequality (148) yields

$$\begin{aligned}\sum_{t=0}^{T}\Phi_t \le\ & \frac{3\lambda_{x,0}(c_{\bar{u}_1^*}+c_{\bar{u}_2^*}+c_{\bar{u}_4^*})(1+v_x)}{2v_x}+\frac{L_{f,1}(\sigma_x^+)^2\varsigma_x}{2\varsigma_x-1}+\frac{L_{f,1}c_{\bar{x}6}(\lambda_{x,0})^2v_x}{2v_x-1}\\ &+\left(\frac{L_F}{m}+\frac{3c_{\bar{u}_5^*}+L_{f,1}c_{\bar{x}1}\lambda_{x,0}}{2}\right)\frac{\hat{c}_x\lambda_{x,0}(2\varsigma_x+v_x)}{2\varsigma_x+v_x-1}+\left(\frac{3c_{\bar{u}_5^*}+L_{f,1}c_{\bar{x}2}\lambda_{x,0}}{2}\right)\frac{\hat{c}_y\lambda_{x,0}(2\varsigma_y+v_x)}{2\varsigma_y+v_x-1}\\ &+\left(\frac{3c_{\bar{u}_6^*}+L_{f,1}c_{\bar{x}3}\lambda_{x,0}}{2}\right)\frac{\hat{c}_z\lambda_{x,0}(2\varsigma_z+v_x)}{2\varsigma_z+v_x-1}\\ &+\left(\frac{3c_{\bar{u}_7^*}+L_{f,1}c_{\bar{x}4}\lambda_{x,0}}{2}\right)\bar{c}_z\lambda_{x,0}\max\left\{\frac{2\varsigma_x-2v_z+v_x}{2\varsigma_x-2v_z+v_x-1},\frac{2\varsigma_y-2v_z+v_x}{2\varsigma_y-2v_z+v_x-1},\frac{2\varsigma_z-v_z+v_x}{2\varsigma_z-v_z+v_x-1}\right\}\\ &+\left(\frac{3c_{\bar{u}_3^*}+L_{f,1}c_{\bar{x}5}\lambda_{x,0}}{2}\right)\bar{c}_y\lambda_{x,0}\max\left\{\frac{2\varsigma_y-v_y+v_x}{2\varsigma_y-v_y+v_x-1},\frac{2-2v_y+v_x}{1-2v_y+v_x}\right\}\triangleq c_6.\end{aligned}\tag{166}$$

Substituting (166) into (163) and defining $c_7\triangleq\mathbb{E}\left[F(\bar{x}_0)-F(x^*)\right]$, we can obtain $\sum_{t=0}^{T}\lambda_{x,t}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right]\le c_6+c_7$, which implies

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla F(x_{i,t})\|^2\right]\le\frac{c_6+c_7}{\lambda_{x,0}(T+1)^{1-v_x}}=\frac{C_3}{(T+1)^{1-v_x}},\tag{167}$$

with $C_3=\frac{c_6+c_7}{2\lambda_{x,0}}$. Inequality (167) directly implies the second inequality in (156) and (12) in Theorem 4.1-(iii). □

F. Proof of Theorem 4.5

In this section, we prove that in addition to accurate convergence, Algorithm 2 can also simultaneously ensure rigorous ϵ_i-LDP for each agent, even when the number of iterations T tends to infinity. To this end, we first provide a definition for the sensitivity of agent i's implementation $\mathcal{A}_i$:

Definition F.1. (Sensitivity) The sensitivity of agent i's implementation $\mathcal{A}_i$ is

$$\Delta_{i,t}=\max_{\text{Adj}(\mathcal{D}_i,\mathcal{D}_i')}\left\|\mathcal{A}_i(\mathcal{D}_i,\theta_{-i,t})-\mathcal{A}_i(\mathcal{D}_i',\theta_{-i,t})\right\|_1,\tag{168}$$

where $\text{Adj}(\mathcal{D}_i,\mathcal{D}_i')$ represents the adjacent relationship between agent i's adjacent datasets $\mathcal{D}_i$ and $\mathcal{D}_i'$, and $\theta_{-i,t}$ represents all information agent i receives from its neighbors at time t.

According to Definition F.1, under Algorithm 2, each agent i's implementation involves three sensitivities: $\Delta_{i,t,x}$, $\Delta_{i,t,y}$, and $\Delta_{i,t,z}$, which correspond to $x_{i,t}$, $y_{i,t}$, and $z_{i,t}$, respectively. With this understanding, we have the following lemma:

Lemma F.2. *(Huang et al., 2015) At each time $t\ge 0$, if agent i injects into each of its shared variables $x_{i,t}$, $y_{i,t}$, and $z_{i,t}$ noise vectors $\chi_{i,t}$, $\zeta_{i,t}$, and $\vartheta_{i,t}$ consisting of p, q, and q independent Laplace noises with parameters $\nu_{i,t,x}$, $\nu_{i,t,y}$, and $\nu_{i,t,z}$, respectively, such that $\sum_{t=1}^{\infty}\left(\frac{\Delta_{i,t,x}}{\nu_{i,t,x}}+\frac{\Delta_{i,t,y}}{\nu_{i,t,y}}+\frac{\Delta_{i,t,z}}{\nu_{i,t,z}}\right)\le\epsilon_i$, then agent i's implementation $\mathcal{A}_i$ of Algorithm 2 is ϵ_i-LDP.*

For the convenience of privacy analysis, we represent the different data points between upper-level adjacent datasets $\mathcal{D}_{f_i}$ and $\mathcal{D}_{f_i}'$ (as well as between lower-level adjacent datasets $\mathcal{D}_{g_i}$ and $\mathcal{D}_{g_i}'$) as the k-th one, i.e., $\varphi_{i,k}$ in $\mathcal{D}_{f_i}$ and $\varphi_{i,k}'$ in $\mathcal{D}_{f_i}'$ ($\xi_{i,k}$ in $\mathcal{D}_{g_i}$ and $\xi_{i,k}'$ in $\mathcal{D}_{g_i}'$), without loss of generality. We further denote $x_{i,t}$, $y_{i,t}$, and $z_{i,t}$ as the parameters generated by Algorithm 2 based on $\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$. We also use $x_{i,t}'$, $y_{i,t}'$, and $z_{i,t}'$ to represent the parameters generated by Algorithm 2 based on $\mathcal{D}_{f_i}'$ and $\mathcal{D}_{g_i}'$.

Now, we are in position to prove Theorem 4.5.

Proof. The convergence results follow naturally from Theorem 4.1.

(1) To prove the statement on privacy, we first analyze the sensitivities of agent i's implementation under Algorithm 2.

According to the definition of sensitivity, we have $z_{j,t}+\vartheta_{j,t}=z_{j,t}'+\vartheta_{j,t}'$, $y_{j,t}+\zeta_{j,t}=y_{j,t}'+\zeta_{j,t}'$, and $x_{j,t}+\chi_{j,t}=x_{j,t}'+\chi_{j,t}'$ for all $t\ge 0$ and $j\in\mathcal{N}_i$. Since we assume that only the k-th data point is different between $\mathcal{D}_{f_i}$ and $\mathcal{D}_{f_i}'$, as well as between $\mathcal{D}_{g_i}$ and $\mathcal{D}_{g_i}'$, when $t<k$, we have $z_{i,t}=z_{i,t}'$, $y_{i,t}=y_{i,t}'$, and $x_{i,t}=x_{i,t}'$. However, when $t\ge k$, since the difference in

loss functions kicks in at iteration k, i.e., $h(x,y;\varphi_{i,k}) \neq h(x,y;\varphi'_{i,k})$ and $l(x,y;\xi_{i,k}) \neq l(x,y;\xi'_{i,k})$, we have $z_{i,t} \neq z'_{i,t}$, $y_{i,t} \neq y'_{i,t}$, and $x_{i,t} \neq x'_{i,t}$. Hence, for agent i's implementation of Algorithm 2, we have

$$\|y_{i,t+1} - y'_{i,t+1}\|_1 = \|(1+w_{ii})(y_{i,t} - y'_{i,t}) - \lambda_{y,t}(\nabla_y g_{i,t}(x_{i,t}, y_{i,t}) - \nabla_y g'_{i,t}(x'_{i,t}, y'_{i,t}))\|_1,$$

for all $t \geq 0$. Let $\bar{w} = \min\{|w_{ii}|\},\ i \in [m]$, the sensitivity $\Delta_{i,t,y}$ satisfies

$$\begin{aligned}\Delta_{i,t+1,y} &\leq (1-\bar{w})\Delta_{i,t,y} + \frac{\lambda_{y,t}}{t+1}\sum_{p=k}^{t}\|\nabla_y l(x_{i,t}, y_{i,t}; \xi_{i,p}) - \nabla_y l(x'_{i,t}, y'_{i,t}; \xi'_{i,p})\|_1 \\ &\leq (1-\bar{w})\Delta_{i,t,y} + \frac{\lambda_{y,t}}{t+1}\sum_{p=0}^{t}\|\nabla_y l(x_{i,t}, y_{i,t}; \xi_{i,p}) - \nabla_y l(x'_{i,t}, y'_{i,t}; \xi'_{i,p})\|_1.\end{aligned} \tag{169}$$

Given that the difference in loss functions kicks in at iteration k, we have $\sum_{p=0}^{k-1}\nabla_y l(x_{i,t}, y_{i,t}; \xi_{i,p}) = \sum_{p=0}^{k-1}\nabla_y l(x'_{i,t}, y'_{i,t}; \xi'_{i,p})$, which are used in the last inequality.

By leveraging inequality (169) and the relation $\xi_{i,p} = \xi'_{i,p}$ for $p \neq k$, we have

$$\begin{aligned}\Delta_{i,t+1,y} \leq (1-\bar{w})\Delta_{i,t,y} &+ \frac{\lambda_{y,t}}{t+1}\sum_{p=0,p\neq k}^{t}\|\nabla_y l(x_{i,t}, y_{i,t}; \xi_{i,p}) - \nabla_y l(x'_{i,t}, y'_{i,t}; \xi_{i,p})\|_1 \\ &+ \frac{\lambda_{y,t}}{t+1}\|\nabla_y l(x_{i,t}, y_{i,t}; \xi_{i,k}) - \nabla_y l(x'_{i,t}, y'_{i,t}; \xi'_{i,k})\|_1.\end{aligned} \tag{170}$$

Assumption 4.4 implies that for the same data $\xi_{i,p}$, we can rewrite (170) as follows:

$$\Delta_{i,t+1,y} \leq \left(1 - \bar{w} + \frac{L_{l,1}\lambda_{y,t}t}{t+1}\right)\Delta_{i,t,y} + \frac{L_{l,1}\lambda_{y,t}t}{t+1}\Delta_{i,t,x} + \frac{2c_{l0}\lambda_{y,t}}{t+1}. \tag{171}$$

Similarly, by using the update of $x_{i,t}$ in the Step 7 of Algorithm 2, we have

$$\|x_{i,t+1} - x'_{i,t+1}\|_1 = \|(1+w_{ii})(x_{i,t} - x'_{i,t}) - \lambda_{x,t}(u_{i,t} - u'_{i,t})\|_1, \tag{172}$$

for all $t \geq 0$. Then, the sensitivity $\Delta_{i,t,x}$ satisfies

$$\begin{aligned}\Delta_{i,t+1,x} &\leq (1-\bar{w})\Delta_{i,t,x} + \frac{\lambda_{x,t}}{t+1}\sum_{p=k}^{t}\|\nabla_y h(x_{i,t}, y_{i,t}; \varphi_{i,p}) - \nabla_y h(x'_{i,t}, y'_{i,t}; \varphi'_{i,p})\|_1 \\ &\quad + \frac{\lambda_{x,t}}{t+1}\sum_{p=k}^{t}\|\nabla^2_{xy} l(x_{i,t}, y_{i,t}; \xi_{i,p})z_{i,t} - \nabla^2_{xy} l(x'_{i,t}, y'_{i,t}; \xi'_{i,p})z'_{i,t}\|_1 \\ &\leq (1-\bar{w})\Delta_{i,t,x} + \frac{\lambda_{x,t}}{t+1}\sum_{p=0}^{t}\|\nabla_y h(x_{i,t}, y_{i,t}; \varphi_{i,p}) - \nabla_y h(x'_{i,t}, y'_{i,t}; \varphi'_{i,p})\|_1 \\ &\quad + \frac{\lambda_{x,t}}{t+1}\sum_{p=0}^{t}\|\nabla^2_{xy} l(x_{i,t}, y_{i,t}; \xi_{i,p})z_{i,t} - \nabla^2_{xy} l(x'_{i,t}, y'_{i,t}; \xi'_{i,p})z'_{i,t}\|_1.\end{aligned} \tag{173}$$

By using the relation $\varphi_{i,p} = \varphi'_{i,p}$ for all $p \neq k$, the second term on the right hand side of (173) satisfies

$$\begin{aligned}&\frac{\lambda_{x,t}}{t+1}\sum_{p=0}^{t}\|\nabla_y h(x_{i,t}, y_{i,t}; \varphi_{i,p}) - \nabla_y h(x'_{i,t}, y'_{i,t}; \varphi'_{i,p})\|_1 \\ &\leq \frac{\lambda_{x,t}}{t+1}\sum_{p=0,p\neq k}^{t}\|\nabla_y h(x_{i,t}, y_{i,t}; \varphi_{i,p}) - \nabla_y h(x'_{i,t}, y'_{i,t}; \varphi_{i,p})\|_1 \\ &\quad + \frac{\lambda_{x,t}}{t+1}\|\nabla_y h(x_{i,t}, y_{i,t}; \varphi_{i,k}) - \nabla_y h(x'_{i,t}, y'_{i,t}; \varphi'_{i,k})\|_1 \\ &\leq \frac{L_{h,1}\lambda_{x,t}t}{t+1}(\Delta_{i,t,x} + \Delta_{i,t,y}) + \frac{2c_{h0}\lambda_{x,t}}{t+1}.\end{aligned} \tag{174}$$

Using an argument similar to the derivation of (174), the third term on the right hand side of (173) satisfies

$$\begin{aligned}
&\frac{\lambda_{x,t}}{t+1}\sum_{p=0}^{t}\|\nabla^2_{xy}l(x_{i,t},y_{i,t};\xi_{i,p})z_{i,t}-\nabla^2_{xy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})z'_{i,t}\|_1\\
&\le\frac{\lambda_{x,t}}{t+1}\sum_{p=0}^{t}\big(\|\nabla^2_{xy}l(x_{i,t},y_{i,t};\xi_{i,p})z_{i,t}-\nabla^2_{xy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})z_{i,t}\|_1\\
&\quad+\|\nabla^2_{xy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})z_{i,t}-\nabla^2_{xy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})z'_{i,t}\|_1\big)\\
&\le\frac{\lambda_{x,t}}{t+1}\sum_{p=0,p\neq k}^{t}\|\nabla^2_{xy}l(x_{i,t},y_{i,t};\xi_{i,p})-\nabla^2_{xy}l(x'_{i,t},y'_{i,t};\xi_{i,p})\|_1\|z_{i,t}\|_1\\
&\quad+\frac{\lambda_{x,t}}{t+1}\|\nabla^2_{xy}l(x_{i,t},y_{i,t};\xi_{i,k})-\nabla^2_{xy}l(x'_{i,t},y'_{i,t};\xi'_{i,k})\|_1\|z_{i,t}\|_1+\frac{\lambda_{x,t}}{t+1}\sum_{p=0}^{t}\|\nabla^2_{xy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})\|_1\|z_{i,t}-z'_{i,t}\|_1\\
&\le\frac{c_zL_{l,2}\lambda_{x,t}t}{t+1}(\Delta_{i,t,x}+\Delta_{i,t,y})+\frac{2c_zL_{l,1}\lambda_{x,t}}{t+1}+L_{l,1}\lambda_{x,t}\Delta_{i,t,z}.
\end{aligned}\tag{175}$$

Substituting (174) and (175) into (173) yields

$$\begin{aligned}
\Delta_{i,t+1,x}\le&\left(1-\bar{w}+\frac{(L_{h,1}+c_zL_{l,2})\lambda_{x,t}t}{t+1}\right)\Delta_{i,t,x}+\frac{(L_{h,1}+c_zL_{l,2})\lambda_{x,t}t}{t+1}\Delta_{i,t,y}\\
&+\frac{2(c_{h0}+c_zL_{l,1})\lambda_{x,t}}{t+1}+L_{l,1}\lambda_{x,t}\Delta_{i,t,z}.
\end{aligned}\tag{176}$$

Furthermore, by using the update of $z_{i,t}$ in the Step 5 of Algorithm 1, we have

$$\|z_{i,t+1}-z'_{i,t+1}\|_1=\|(1+w_{ii})(z_{i,t}-z'_{i,t})-\lambda_{z,t}(H_{i,t}z_{i,t}-H'_{i,t}z'_{i,t})+\lambda_{z,t}(b_{i,t}-b'_{i,t})\|_1,$$

for all $t\ge 0$. Then, the sensitivity $\Delta_{i,t,z}$ satisfies

$$\begin{aligned}
\Delta_{i,t+1,z}\le&(1-\bar{w})\Delta_{i,t,z}+\frac{\lambda_{z,t}}{t+1}\sum_{p=k}^{t}\|\nabla^2_{yy}l(x_{i,t},y_{i,t};\xi_{i,p})z_{i,t}-\nabla^2_{yy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})z'_{i,t}\|_1\\
&+\frac{\lambda_{z,t}}{t+1}\sum_{p=k}^{t}\|\nabla_yh(x_{i,t},y_{i,t};\varphi_{i,p})-\nabla_yh(x'_{i,t},y'_{i,t};\varphi'_{i,p})\|_1\\
\le&(1-\bar{w})\Delta_{i,t,z}+\frac{\lambda_{z,t}}{t+1}\sum_{p=0}^{t}\|\nabla_yh(x_{i,t},y_{i,t};\varphi_{i,p})-\nabla_yh(x'_{i,t},y'_{i,t};\varphi'_{i,p})\|_1\\
&+\frac{\lambda_{z,t}}{t+1}\sum_{p=0}^{t}\|\nabla^2_{yy}l(x_{i,t},y_{i,t};\xi_{i,p})z_{i,t}-\nabla^2_{yy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})z'_{i,t}\|_1.
\end{aligned}\tag{177}$$

Given that the difference in loss functions kicks in at iteration k, we have $\sum_{p=0}^{k-1}\nabla^2_{yy}l(x_{i,t},y_{i,t};\xi_{i,p})z_{i,t}=\sum_{p=0}^{k-1}\nabla^2_{yy}l(x'_{i,t},y'_{i,t};\xi'_{i,p})z'_{i,t}$, and $\sum_{p=0}^{k-1}\nabla_yh(x_{i,t},y_{i,t};\varphi_{i,p})=\sum_{p=0}^{k-1}\nabla_yh(x'_{i,t},y'_{i,t};\varphi'_{i,p})$, which are used in the last inequality.

Furthermore, by leveraging (177) and using an argument similar to the derivation of (176), we have

$$\Delta_{i,t+1,z}\le(1-\bar{w}+c_{l1}\lambda_{z,t})\Delta_{i,t,z}+(L_{h,1}+c_zL_{l,2})\frac{\lambda_{z,t}t}{t+1}(\Delta_{i,t,x}+\Delta_{i,t,y})+\frac{2(c_{h0}+c_zL_{l,1})\lambda_{z,t}}{t+1}.\tag{178}$$

Summing up both sides of (171), (176), and (178), we obtain

$$\begin{aligned}\Delta_{i,t+1,x}+\Delta_{i,t+1,y}+\Delta_{i,t+1,z} &\leq \left(1-\bar{w}+\frac{L_{l,1}\lambda_{y,t}t}{t+1}+(L_{h,1}+c_zL_{l,2})\frac{\lambda_{x,t}t}{t+1}+(L_{h,1}+c_zL_{l,2})\frac{\lambda_{z,t}t}{t+1}\right)\Delta_{i,t,x}\\ &+\left(1-\bar{w}+\frac{L_{l,1}\lambda_{y,t}t}{t+1}+(L_{h,1}+c_zL_{l,2})\frac{\lambda_{x,t}t}{t+1}+(L_{h,1}+c_zL_{l,2})\frac{\lambda_{z,t}t}{t+1}\right)\Delta_{i,t,y}\\ &+(1-\bar{w}+L_{l,1}\lambda_{x,t}+c_{l1}\lambda_{z,t})\,\Delta_{i,t,z}+\frac{2c_{l0}\lambda_{y,t}}{t+1}+\frac{2(c_{h0}+c_zL_{l,1})\lambda_{x,t}}{t+1}+\frac{2(c_{h0}+c_zL_{l,1})\lambda_{z,t}}{t+1}.\end{aligned} \tag{179}$$

Since stepsizes $\lambda_{x,t}$, $\lambda_{y,t}$, and $\lambda_{z,t}$ are decaying sequences, we can choose proper initial stepsizes such that the following inequality always holds:

$$\begin{aligned}&\Delta_{i,t+1,x}+\Delta_{i,t+1,y}+\Delta_{i,t+1,z}\\ &\leq\left(1-\frac{\bar{w}}{2}\right)(\Delta_{i,t,x}+\Delta_{i,t,y}+\Delta_{i,t,z})+\frac{2c_{l0}\lambda_{y,t}}{t+1}+\frac{2(c_{h0}+c_zL_{l,1})\lambda_{x,t}}{t+1}+\frac{2(c_{h0}+c_zL_{l,1})\lambda_{z,t}}{t+1}\\ &\leq\left(1-\frac{\bar{w}}{2}\right)(\Delta_{i,t,x}+\Delta_{i,t,y}+\Delta_{i,t,z})+\frac{M_1}{(t+1)^{\beta_\epsilon}},\end{aligned} \tag{180}$$

with $M_1=2c_{l0}\lambda_{y,0}+2(c_{h0}+c_zL_{l,1})\lambda_{x,0}+2(c_{h0}+c_zL_{l,1})\lambda_{z,0}$ and $\beta_\epsilon=\min\{1+v_x,1+v_y,1+v_z\}$.

According to Lemma 11 in (Chen & Wang, 2023), the following inequality holds:

$$\Delta_{i,t,x}+\Delta_{i,t,y}+\Delta_{i,t,z}\leq M_2t^{-\beta_\epsilon}, \tag{181}$$

where the constant M_2 is given by $M_2=\frac{4}{\bar{w}}\left(\frac{4\beta_\epsilon}{e\ln(\frac{4}{2-2\bar{w}})}\right)^{\beta_\epsilon}$ with given $\Delta_{i,0,x}=\Delta_{i,0,y}=\Delta_{i,0,z}=0$.

According to (181), we have $\Delta_{i,t,x}\leq M_2$, $\Delta_{i,t,y}\leq M_2$, and $\Delta_{i,t,z}\leq M_2$ for all $t>0$. Substituting $\Delta_{i,t,x}\leq M_2$ into (171) and using again Lemma 11 in (Chen & Wang, 2023), we have

$$\Delta_{i,t,y}\leq\frac{C_{\epsilon y}}{(t+1)^{1+v_y}},\quad\text{with}\quad C_{\epsilon y}=\left(\frac{4(1+v_y)}{e\ln(\frac{4}{2-2\bar{w}})}\right)^{1+v_y}\left(\frac{\Delta_{i,0,y}(1-\frac{\bar{w}}{2})}{L_{l,1}\lambda_{y,0}M_2+2c_{l0}\lambda_{y,0}}+\frac{4}{\bar{w}}\right). \tag{182}$$

Similarly, substituting $\Delta_{i,t,x}\leq M_2$ and $\Delta_{i,t,y}\leq M_2$ into (178), we have

$$\Delta_{i,t,z}\leq\frac{C_{\epsilon z}}{(t+1)^{1+v_z}},\quad\text{with}\quad C_{\epsilon z}=\left(\frac{4(1+v_z)}{e\ln(\frac{4}{2-2\bar{w}})}\right)^{1+v_z}\left(\frac{\Delta_{i,0,z}(1-\frac{\bar{w}}{2})}{(2M_2(L_{h,1}+c_zL_{l,2})+2(c_{h0}+c_zL_{l,1}))\,\lambda_{z,0}}+\frac{4}{\bar{w}}\right). \tag{183}$$

Furthermore, substituting $\Delta_{i,t,y}\leq M_2$ and $\Delta_{i,t,z}\leq\frac{C_{\epsilon z}}{(t+1)^{1+v_z}}$ into (176) yields

$$\Delta_{i,t,x}\leq\frac{C_{\epsilon z}}{(t+1)^{1+v_x}},\text{ with }C_{\epsilon x}=\left(\frac{4(1+v_x)}{e\ln(\frac{4}{2-2\bar{w}})}\right)^{1+v_x}\left(\frac{\Delta_{i,0,x}(1-\frac{\bar{w}}{2})}{((L_{h,1}+c_zL_{l,2})M_2+2(c_{h0}+c_zL_{l,1})+L_{l,1}C_{\epsilon z})\,\lambda_{x,0}}+\frac{4}{\bar{w}}\right). \tag{184}$$

By using (182)-(184) and Lemma F.2, we arrive at

$$\sum_{t=1}^{T}\frac{\Delta_{i,t,x}}{\nu_{i,x}}\leq\sum_{t=1}^{T}\frac{\sqrt{2}C_{\epsilon x}}{\sigma_{i,x}(t+1)^{1+v_x-\varsigma_x}},\quad\sum_{t=1}^{T}\frac{\Delta_{i,t,y}}{\nu_{i,y}}\leq\sum_{t=1}^{T}\frac{\sqrt{2}C_{\epsilon y}}{\sigma_{i,y}(t+1)^{1+v_y-\varsigma_y}},\quad\sum_{t=1}^{T}\frac{\Delta_{i,t,z}}{\nu_{i,z}}\leq\sum_{t=1}^{T}\frac{\sqrt{2}C_{\epsilon z}}{\sigma_{i,z}(t+1)^{1+v_z-\varsigma_z}}, \tag{185}$$

where $\nu_{i,x}$, $\nu_{i,y}$, and $\nu_{i,z}$ are given by $\nu_{i,x}=\frac{\sigma_{i,x}}{\sqrt{2}(t+1)^{\varsigma_x}}$, $\nu_{i,y}=\frac{\sigma_{i,y}}{\sqrt{2}(t+1)^{\varsigma_y}}$, and $\nu_{i,z}=\frac{\sigma_{i,z}}{\sqrt{2}(t+1)^{\varsigma_z}}$ from Assumption 2.3. By incorporating $\Delta_{i,0,x}=\Delta_{i,0,y}=\Delta_{i,0,z}=0$ into $C_{\epsilon x}$, $C_{\epsilon y}$, and $C_{\epsilon z}$ in (182)-(184), we arrive at the result in Theorem 4.5-(i).

(2) The inequalities in (185) implies that $\epsilon_i=\epsilon_{i,z}+\epsilon_{i,y}+\epsilon_{i,x}$ is finite even when T tends to infinity since $v_x>\varsigma_x$, $v_y>\varsigma_y$, and $v_z>\varsigma_z$. □

G. Proofs of Corollary 4.3 and Corollary 4.7, as well as Further Discussion

G.1. Proof of Corollary 4.3

Proof. (i) For a strongly convex $F(x)$, the convergence rate of Algorithm 2 is $\mathcal{O}(T^{-\beta_1})$ based on (10). Therefore, setting $T^{-\beta_1}=\delta$ yields that the iteration complexity of Algorithm 2 is $\mathcal{O}(\delta^{-\frac{1}{\beta_1}})$ in finding a δ-solution. Furthermore, since the per-iteration complexity of Algorithm 2 is $\max\{p,q\}$, the computational complexity of Algorithm 2 is $\mathcal{O}(\max\{p,q\}\delta^{-\frac{1}{\beta_1}})$ in finding a δ-solution.

According to the conditions $0<v_z<v_y<v_z<1$, $2\varsigma_x>v_x$, $2\varsigma_x>v_z+v_y$, $2\varsigma_y>v_z+v_y$, and $2\varsigma_z>v_y$ given in Theorem 4.1-(i), we can choose $v_x=0.66$, $v_y=0.64$, $v_z=0.43$, $\varsigma_x=0.65$, $\varsigma_y=0.63$, and $\varsigma_z=0.42$. Under these parameters, we have $\beta_1=\min\{0.64,0.44,0.4,0.43,0.62,0.72\}=0.4$ and hence a computational complexity of $\mathcal{O}(\max\{p,q\}\delta^{-2.5})$.

(ii) Similarly, for a convex $F(x)$, the convergence rate of Algorithm 2 is $\mathcal{O}(T^{-(1-v_x)})$ based on (11). Therefore, the computational complexity of Algorithm 2 is $\mathcal{O}(\max\{p,q\}\delta^{-\frac{1}{1-v_x}})$ in finding a δ-solution. Recalling the conditions $0<v_z<v_y<v_z<1$, $\varsigma_x>\frac{1}{2}$, $2\varsigma_x>v_z+v_y$, $2\varsigma_x>2v_z+2-2v_x$, $2\varsigma_y>v_z+v_y$, $2\varsigma_y>2v_z+2-2v_x$, $2\varsigma_y>v_y+2-2v_x$, $2\varsigma_z>v_z+2-2v_x$, and $2\varsigma_z>v_y$ given in Theorem 4.1-(ii), we can select $v_x=0.77$, $v_y=0.75$, $v_z=0.5$, $\varsigma_x=0.76$, $\varsigma_y=0.74$, and $\varsigma_z=0.49$ yielding $1-v_x=0.23$ and a computational complexity of $\mathcal{O}(\max\{p,q\}\delta^{-4.35})$.

(iii) For a nonconvex $F(x)$, the convergence rate of Algorithm 2 is $\mathcal{O}(T^{-(1-v_x)})$ based on (12). Therefore, the computational complexity of Algorithm 2 is $\mathcal{O}(\max\{p,q\}\delta^{-\frac{1}{1-v_x}})$ in finding a δ-solution. We use $v_x=0.615$, $v_y=0.60375$, $v_z=0.4$, $\varsigma_x=0.61125$, $\varsigma_y=0.6$, and $\varsigma_z=0.398125$ to satisfy the conditions $0<v_z<v_y<v_x<1$, $\varsigma_x>\frac{1}{2}$, $2\varsigma_x>v_z+v_y$, $2\varsigma_x>2v_z+1-v_x$, $2\varsigma_y>2v_z+1-v_x$, $2\varsigma_y>v_y+1-v_x$, $2\varsigma_y>v_z+v_y$, $2\varsigma_z>v_z+1-v_x$, and $2\varsigma_z>v_y$ given in Theorem 4.1-(iii). Under these parameters, we have $1-v_x=0.385$ and hence a computational complexity of $\mathcal{O}(\max\{p,q\}\delta^{-2.6})$. □

G.2. Proof of Corollary 4.7

Proof. We select $v_x=\frac{3}{5}+\kappa$, $v_y=\frac{3}{5}+\frac{\kappa}{4}$, $v_z=\frac{2}{5}$, $\varsigma_x=v_x-\frac{\kappa}{4}$, $\varsigma_y=v_y-\frac{\kappa}{4}$, and $\varsigma_z=v_z-\frac{\kappa}{8}$ with $\kappa\in(0,\frac{2}{5})$ that satisfy the conditions given in Theorem 4.1-(iii). Therefore, the accurate convergence of Algorithm 2 remains attainable. Next, we quantify the tradeoff between the convergence rate of Algorithm 2 and the given cumulative privacy budget.

According to (185) in the proof of Theorem 4.5-(i), it has proven that the privacy budget ϵ_i is bounded by

$$\epsilon_i\leq\sum_{i=1}^{T}\frac{\sqrt{2}C_{\epsilon x}}{\sigma_{i,x}(t+1)^{1+v_x-\varsigma_x}}+\sum_{i=1}^{T}\frac{\sqrt{2}C_{\epsilon y}}{\sigma_{i,y}(t+1)^{1+v_y-\varsigma_y}}+\sum_{i=1}^{T}\frac{\sqrt{2}C_{\epsilon z}}{\sigma_{i,z}(t+1)^{1+v_z-\varsigma_z}},\tag{186}$$

where $C_{\epsilon x}$, $C_{\epsilon y}$, and $C_{\epsilon z}$ are positive constants, as given in the statement of Theorem 4.5-(i).

We proceed to characterize the inequality (186). By applying the following relationship

$$\sum_{t=1}^{T}\frac{1}{(t+1)^r}\leq\int_0^T\frac{1}{(x+1)^r}dx=\frac{1}{1-r}((T+1)^{1-r}-1),$$

to (186), we can derive

$$\begin{aligned}\epsilon_i=\epsilon_{i,x}+\epsilon_{i,y}+\epsilon_{i,y}&\leq\frac{\sqrt{2}C_{\epsilon x}}{\sigma_{i,x}(v_x-\varsigma_x)}\left(1-(T+1)^{-(v_x-\varsigma_x)}\right)+\frac{\sqrt{2}C_{\epsilon y}}{\sigma_{i,y}(v_y-\varsigma_y)}\left(1-(T+1)^{-(v_y-\varsigma_y)}\right)\\&\quad+\frac{\sqrt{2}C_{\epsilon z}}{\sigma_{i,z}(v_z-\varsigma_z)}\left(1-(T+1)^{-(v_z-\varsigma_z)}\right),\\&\leq\frac{\sqrt{2}C_{\epsilon x}}{\sigma_{i,x}(v_x-\varsigma_x)}+\frac{\sqrt{2}C_{\epsilon y}}{\sigma_{i,y}(v_y-\varsigma_y)}+\frac{\sqrt{2}C_{\epsilon z}}{\sigma_{i,z}(v_z-\varsigma_z)}.\end{aligned}\tag{187}$$

We denote the given cumulative privacy budget as $\epsilon_i'>0$. According to the inequality (187), we can choose DP-noise parameters as $\sigma_{i,x}=\frac{3\sqrt{2}C_{\epsilon x}}{(v_x-\varsigma_x)\epsilon_i'}$, $\sigma_{i,y}=\frac{3\sqrt{2}C_{\epsilon y}}{(v_y-\varsigma_y)\epsilon_i'}$, and $\sigma_{i,z}=\frac{3\sqrt{2}C_{\epsilon z}}{(v_z-\varsigma_z)\epsilon_i'}$, in which the parameters v_x, v_y, v_z, ς_x, ς_y and ς_z

are fixed constants (predetermined parameters) that satisfy the conditions given in the statement of Theorem 4.1-(iii). It is clear that a smaller ϵ_i' will result in larger values of $\sigma_{i,x}$, $\sigma_{i,y}$, and $\sigma_{i,z}$.

Next, we analyze the convergence rate of Algorithm 2 under a nonconvex $F(x)$, with setting DP-noise parameters $\sigma_{i,x} = \frac{3\sqrt{2}C_{\epsilon x}}{(v_x-\varsigma_x)\epsilon_i'}$, $\sigma_{i,y} = \frac{3\sqrt{2}C_{\epsilon y}}{(v_y-\varsigma_y)\epsilon_i'}$, and $\sigma_{i,z} = \frac{3\sqrt{2}C_{\epsilon z}}{(v_z-\varsigma_z)\epsilon_i'}$.

Based on (167) in the proof of Theorem 4.1-(iii), we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F\left(x_{i,t}\right)\right\|^2\right] \leq \frac{C_3}{(T+1)^{1-v_x}}. \tag{188}$$

Given that v_x is independent of $\sigma_{i,x}$, $\sigma_{i,y}$ and $\sigma_{i,z}$, the accurate convergence of Algorithm 2 remains attainable even if ϵ' tends to zero. However, C_3 is a positive constant that is positively correlated with DP-noise parameters $\sigma_{i,x}^2$, $\sigma_{i,y}^2$ and $\sigma_{i,z}^2$. This correlation can be evidenced by the definition $C_3 = \frac{c_6+c_7}{2\lambda_{x,0}}$ with c_6 given in (166) and c_7 given by $c_7 \triangleq \mathbb{E}\left[F(\bar{x}_0) - F(x^*)\right]$:

$$\begin{aligned}
C_3 \triangleq \frac{c_6+c_7}{2\lambda_{x,0}} &= \frac{3(c_{\bar{u}_1^*}+c_{\bar{u}_2^*}+c_{\bar{u}_4^*})(1+v_x)}{4v_x} + \frac{L_{f,1}(\sigma_x^+)^2\varsigma_x}{2\lambda_{x,0}(2\varsigma_x-1)} + \frac{L_{f,1}c_{\bar{x}6}\lambda_{x,0}v_x}{2(2v_x-1)} \\
&+ \left(\frac{L_F}{m} + \frac{3c_{\bar{u}_5^*}+L_{f,1}c_{\bar{x}1}\lambda_{x,0}}{2}\right)\frac{\hat{c}_x(2\varsigma_x+v_x)}{2(2\varsigma_x+v_x-1)} + \left(\frac{3c_{\bar{u}_5^*}+L_{f,1}c_{\bar{x}2}\lambda_{x,0}}{2}\right)\frac{\hat{c}_y(2\varsigma_y+v_x)}{2(2\varsigma_y+v_x-1)} \\
&+ \left(\frac{3c_{\bar{u}_6^*}+L_{f,1}c_{\bar{x}3}\lambda_{x,0}}{4}\right)\frac{\hat{c}_z(2\varsigma_z+v_x)}{2\varsigma_z+v_x-1} + \frac{\mathbb{E}\left[F(\bar{x}_0)-F(x^*)\right]}{2\lambda_{x,0}} \\
&+ \left(\frac{3c_{\bar{u}_7^*}+L_{f,1}c_{\bar{x}4}\lambda_{x,0}}{4}\right)\bar{c}_z\max\left\{\frac{2\varsigma_x-2v_z+v_x}{2\varsigma_x-2v_z+v_x-1}, \frac{2\varsigma_y-2v_z+v_x}{2\varsigma_y-2v_z+v_x-1}, \frac{2\varsigma_z-v_z+v_x}{2\varsigma_z-v_z+v_x-1}\right\} \\
&+ \left(\frac{3c_{\bar{u}_3^*}+L_{f,1}c_{\bar{x}5}\lambda_{x,0}}{4}\right)\bar{c}_y\max\left\{\frac{2\varsigma_y-v_y+v_x}{2\varsigma_y-v_y+v_x-1}, \frac{2-2v_y+v_x}{1-2v_y+v_x}\right\},
\end{aligned}$$

in which $\hat{c}_y$ and $\hat{c}_z$ are given by

$$\begin{aligned}
\hat{c}_y &= \left(4m(\sigma_y^+)^2 + (c_{\hat{y}2}+c_{\hat{y}3})C_0\lambda_{y,0}^2 + c_{\hat{y}4}\lambda_{y,0}^2\right)\left(\frac{8\varsigma_y}{e\ln(\frac{8}{8-\delta_2})}\right)^{2\varsigma_y}\left(\frac{\mathbb{E}\left[\|\hat{\boldsymbol{y}}_0\|^2\right](4-\delta_2)}{4c_y}+\frac{8}{\delta_2}\right), \\
\hat{c}_z &= \left(4m(\sigma_z^+)^2 + c_{\hat{z}1}C_0\lambda_{z,0}^2 + c_{\hat{z}2}\lambda_{z,0}^2\right)\left(\frac{8\varsigma_z}{e\ln(\frac{8}{8-\delta_2})}\right)^{2\varsigma_z}\left(\frac{\mathbb{E}\left[\|\hat{\boldsymbol{z}}_0\|^2\right](4-\delta_2)}{4c_z}+\frac{8}{\delta_2}\right).
\end{aligned}$$

Based on the definition $\sigma_x^+ = \max_{i\in[m]}\{\sigma_{i,x}\}$, $\sigma_y^+ = \max_{i\in[m]}\{\sigma_{i,y}\}$, and $\sigma_z^+ = \max_{i\in[m]}\{\sigma_{i,z}\}$, we have that C_3 is directly proportional to $\sigma_{i,x}^2$, $\sigma_{i,y}^2$, and $\sigma_{i,z}^2$. Given that $\sigma_{i,x}^2$, $\sigma_{i,y}^2$, and $\sigma_{i,z}^2$ are all inversely proportional to $(\epsilon_i')^2$, this leads us to deduce the following inequality based on (188):

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F\left(x_{i,t}\right)\right\|^2\right] \leq \frac{C_3}{(T+1)^{1-v_x}} = \mathcal{O}\left(\frac{T^{-(1-v_x)}}{\min_{i\in[m]}\{(\epsilon_i')^2\}}\right).$$

Recalling that $\epsilon_i' > 0$ represents any given cumulative privacy budgets and considering that $v_x = \frac{3}{5}+\kappa$ with $\kappa \in (0, \frac{2}{5})$ satisfies $v_x \in (0.6, 1)$, we arrive at the result in Corollary 4.7. □

G.3. Discussion on Relation with Existing Results on Single-Level Stochastic Optimization under The Constraint of Differential Privacy

In this subsection, we briefly show that when we only consider the lower-level optimization part of our algorithm, i.e., when our bilevel optimization problem reduces to single-level optimization, our algorithm has exactly the same order of convergence rate as that in Bassily et al. (2019) for single-level stochastic optimization under the constraint of differential privacy.

Proposition G.1. *In the special case where the bilevel optimization problem in (1) does not have the upper-level optimization objective function, problem (1) reduces to decentralized single-level optimization with the objective function* $g(y) = \frac{1}{m}\sum_{i=1}^{m} g_i(y)$ *and* $g_i(y) = \mathbb{E}_{\xi_i}[l(y;\xi_i)]$. *If the data* ξ_i *satisfies an arbitrary* n*-dimension (unknown) distribution, the objective function* $g(y)$ *is convex, the optimization variable* y *is within a convex set* $\mathcal{Y} \subset \mathbb{R}^d$, *and the number of iterations* T *is given by* $T = \mathcal{O}(\frac{(\max_{i\in[m]}\{\epsilon_i\}n)^4}{M^4 d^2})$ *with* $M = \max_{y\in\mathcal{Y}} \|y\|$, *then the loss at* y_T^i *generated by Algorithm 2 satisfies*

$$\mathbb{E}[g(y_T^i) - g(y^*)] \leq M \cdot \mathcal{O}\left(\frac{\sqrt{d}}{\min_{i\in[m]}\{\epsilon_i\}n} + \frac{1}{\sqrt{n}}\right),$$

where $y^* = argmin_{y\in\mathcal{Y}} g(y)$, ϵ_i *is the privacy budget of agent* i, *and* d *is the dimension of the optimization variable* y.

Proof. Following Bassily et al. (2019), we denote the data distribution as $\mathcal{P}^n$, $\mathcal{D}_i = (\xi_i^1, \cdots, \xi_i^n)$ as a sequence of i.i.d. samples from $\mathcal{P}^n$, and $g(y_t^i;\mathcal{P}^n) - g(y^*;\mathcal{P}^n)$ as the excess population loss of y.

By leveraging an argument similar to that of Theorem 3 in Chen & Wang (2023), we can derive

$$\mathbb{E}[g(y_T^i;\mathcal{D}_i) - \min_{y\in\mathcal{Y}} g(y;\mathcal{D}_i)] \leq \mathcal{O}(T^{-\beta}) \approx \mathcal{O}(T^{-0.25}), \tag{189}$$

where $\mathcal{D}_i \sim \mathcal{P}^n$ and $\beta = \frac{1-v_y}{2} = \frac{1}{4}$ (v_y represents the decaying rate of the stepsize $\lambda_{y,t}$ given in Algorithm 2; here we set $v_y = 0.5$).

We define $y_n^* = \text{argmin}_{y\in\mathcal{Y}} g(y;\mathcal{D})$ with $\mathcal{D} \sim \mathcal{P}^n$. According to Section 5.1.2 in Shapiro et al. (2021), we can obtain the following relationship:

$$\mathbb{E}[g(y_n^*;\mathcal{P}) - g(y^*;\mathcal{P})] \leq \mathcal{O}\left(\frac{M}{\sqrt{n}}\right). \tag{190}$$

Given that $\mathcal{D}_i$ and $\mathcal{D}$ follow the same distribution, we set T in (189) as $T = \mathcal{O}\left(\frac{((\max_{i\in[m]}\{\epsilon_i\}n)^4}{M^4 d^2}\right)$ and then combine (189) and (190) to obtain

$$\mathbb{E}[g(y_T^i;\mathcal{P}^n) - g(y^*;\mathcal{P}^n)] \leq M \cdot \mathcal{O}\left(\frac{\sqrt{d}}{\min_{i\in[m]}\{\epsilon_i\}n} + \frac{1}{\sqrt{n}}\right), \tag{191}$$

which implies the result in Proposition G.1. □

H. The Reason why Existing DSBO Algorithms cannot Ensure a Finite Cumulative Privacy Budget ϵ_i

H.1. The Limitation of Existing DSBO Algorithms under Differential-Privacy Constraints

In this section, we explain the limitation of existing DSBO algorithms in Chen et al. (2022), Yang et al. (2022), and Chen et al. (2023) under LDP constraints. Specifically, to obtain good approximations of the hypergradient and/or the optimal solution y^* to the lower-level optimization problem in (1), these algorithms incorporate inner-loop iterations into the outer algorithmic iteration, which leads to a cumulative privacy budget that grows to infinity as the number of outer iterations tends to infinity.

We use the DSBO-HIGP algorithm in Chen et al. (2023) as an example to illustrate this idea. To ensure privacy, persistent DP-noises have to be added to messages transmitted in each iteration of the DSBO-HIGP algorithm. Then, the modified DSBO-HIGP algorithm with injected DP-noises is described in the following Algorithm 3. It can be seen that Algorithm 3 has double inner-loops: a K-step inner-loop (lines 4-8) for achieving a good approximation of y^* (the optimal solution to the lower-level optimization problem in (1)) and an N-step inner-loop (lines 9-15) for a good estimation of the hypergradient $\nabla F(x)$. DP-noises have been injected into all communication steps to enable privacy. According to Theorem 3.3 in Chen et al. (2023), the convergence of the original DSBO-HIGP can be guaranteed only when $K = \log(T)$, $N \geq 1$, $\alpha_t = \mathcal{O}(\frac{1}{\sqrt{T}})$, $\forall T > 0$, $\beta_t = \mathcal{O}(\frac{1}{\sqrt{T}})$, $\forall T > 0$, and $\gamma \in (c_1, c_2)$ with $0 < c_1 < c_2$. It is worth noting that when T tends to infinity, the number of iterations K also tends to infinity.

Algorithm 3 LDP design for DSBO-HIGP

1: **Input:** Stepsizes α_t, β_t, and γ; Iterations $T > 0$, $K > 0$, and $N = \log(T)$; Initialization $y_{i,k}^0 = 0$, $x_{i,0} = r_{i,0} = 0$, $d_{i,t}^0 = -b_{i,t}^0$, $s_{i,t}^0 = -b_{i,t}^0$, and $z_{i,t}^0 = 0$; DP-noises $\vartheta_{i,t}^k$, $\zeta_{i,t}^k$, and $\chi_{i,t}^k$ satisfying Assumption 3.1.
2: **for** $t = 0, 1, \cdots, T-1$ **do**
3: $\quad y_{i,t}^0 = y_{i,t-1}^K$.
4: $\quad$ **for** $k = 0, 1, \cdots, K-1$ **do**
5: $\quad\quad$ **for** $i = 0, 1, \cdots, m-1$ **do**
6: $\quad\quad\quad y_{i,t}^{k+1} = y_{i,t}^k + \sum_{j\in\mathcal{N}_i} w_{ij}(y_{j,t}^k + \zeta_{j,t}^k - y_{i,t}^k) - \beta_t v_{i,t}^k$ with $v_{i,t}^k = \nabla_y g_i(x_{i,t}, y_{i,t}^k; \xi_{i,t}^k)$.
7: $\quad\quad$ **end for**
8: $\quad$ **end for**
9: $\quad$ **for** $k = 0, 1, \cdots, N-1$ **do**
10: $\quad\quad$ **for** $i = 0, 1, \cdots, m-1$ **do**
11: $\quad\quad\quad z_{i,t}^{k+1} = z_{i,t}^k + \sum_{j\in\mathcal{N}_i} w_{ij}(z_{j,t}^k + \vartheta_{j,t}^k - z_{i,t}^k) - \gamma d_{i,t}^k$,
12: $\quad\quad\quad s_{i,t}^{k+1} = H_{i,t}^{k+1} z_{i,t}^{k+1} - b_{i,t}^{k+1}$,
13: $\quad\quad\quad d_{i,t}^{k+1} = d_{i,t}^k + \sum_{j\in\mathcal{N}_i} w_{ij}(d_{j,t}^k + \vartheta_{j,t}^k - d_{i,t}^k) + s_{i,t}^{k+1} - s_{i,t}^k$.
14: $\quad\quad$ **end for**
15: $\quad$ **end for**
16: $\quad u_{i,t} = \nabla_x f_i(x_{i,t}, y_{i,t}^K; \varphi_{i,0}) - \nabla_{xy}^2 g_i(x_{i,t}, y_{i,t}^K; \xi_{i,0}) z_{i,t}^N$.
17: $\quad$ **for** $i = 0, 1, \cdots, m-1$ **do**
18: $\quad\quad x_{i,t+1} = x_{i,t} + \sum_{j\in\mathcal{N}_i} w_{ij}(x_{j,t} + \chi_{j,t} - x_{i,t}) - \alpha_t r_{i,t}$,
19: $\quad\quad r_{i,t+1} = (1-\alpha_t) r_{i,t} + \alpha_t u_{i,t}$.
20: $\quad$ **end for**
21: **end for**
22: **Output:** $\bar{x}_T = \frac{1}{m}\sum_{i=1}^m x_{i,T}$.

With this understanding, we first analyze the cumulative privacy budget $\epsilon_{i,y}$ associated with $y_{i,t}$ in Algorithm 3. By leveraging (185), the cumulative privacy budget $\epsilon_{i,y}$ of Algorithm 3 satisfies

$$\epsilon_{i,y} \leq \sum_{t=1}^{T}\sum_{k=1}^{K} \mathcal{O}\left(\frac{\beta_t}{\sigma_{i,y,t}^k (t+1)}\right), \tag{192}$$

where $\sigma_{i,y,t}^k$ represents the variance of the DP-noise $\zeta_{i,t}^k$.

When the DP-noise variance decays over the outer-loop iteration t (in this case, a fixed DP-noise is injected into the consensus operation at Algorithm 3 Step 6 during each inner-loop iteration, which degrades the estimation performance of the global y^*), the convergence of Algorithm 3 is significantly affected. Therefore, we consider the following two designs for $\sigma_{i,y,t}^k$:

(1) The DP-noise variance decays over both inner-loop iterations k and outer-loop iterations t, i.e., $\sigma_{i,y,t}^k = \mathcal{O}(\frac{1}{(t+1)^{\varsigma_y}(k+1)^{\varsigma_y}})$,

(2) The DP-noise variance decays over inner-loop iterations k, i.e.,$\sigma_{i,y,t}^k = \mathcal{O}(\frac{1}{(k+1)^{\varsigma_y}})$.

By using the decaying stepsize $\beta_t = \mathcal{O}(\frac{1}{(t+1)^{v_y}})$ with $v_y \in (0,1)$, the cumulative privacy budget $\epsilon_{i,y}$ for the aforementioned two scenarios satisfy

$$(1)\quad \epsilon_{i,y} \leq \sum_{t=1}^{T} \mathcal{O}\left(\frac{1}{(t+1)^{1+v_y-\varsigma_y}}\right)\sum_{k=1}^{K}\mathcal{O}\left((k+1)^{\varsigma_y}\right), \quad (2)\quad \epsilon_{i,y} \leq \sum_{t=1}^{T} \mathcal{O}\left(\frac{1}{(t+1)^{1+v_y}}\right)\sum_{k=1}^{K}\mathcal{O}\left((k+1)^{\varsigma_y}\right),$$

which imply that the cumulative privacy budget $\epsilon_{i,y}$ in both scenarios will grow to infinity when the number of outer iterations T tends to infinity, thus violating rigorous ϵ_i-LDP privacy constraints. Of course, employing a constant stepsize γ in the N-step inner-loop (lines 9-15) of Algorithm 3 exacerbates this issue, leading to a significant increase in the cumulative privacy budget $\epsilon_{i,z}$ (see the following Section H.2 for details).

The above mentioned issue also exists in other inner-loop-based DSBO algorithms (Chen et al., 2022; Yang et al., 2022).

H.2. The Calculations of the Cumulative Privacy Budget for the Algorithms Listed in Table 1

First, we compute the computational complexity and the cumulative privacy budget of our Algorithm 2, i.e., LDP-DSBO. We select $\upsilon_x = \frac{3}{5} + \kappa$, $\upsilon_y = \frac{3}{5} + \frac{\kappa}{4}$, $\upsilon_z = \frac{2}{5}$, $\varsigma_x = \upsilon_x - \frac{\kappa}{4}$, $\varsigma_y = \upsilon_y - \frac{\kappa}{4}$, and $\varsigma_z = \upsilon_z - \frac{\kappa}{8}$ with $\kappa \in (0, \frac{2}{5})$ that satisfy the conditions given in Theorem 4.1-(iii) (Since all results in Table 1 are obtained for a nonconvex F). Under these settings, the iteration complexity of Algorithm 2 is $\mathcal{O}\big(\delta^{-\frac{5}{2-5\kappa}}\big)$ and the cumulative privacy budget is $\mathcal{O}\big(\frac{1}{\kappa}\big)$ (Detailed computations of the iteration complexity and the cumulative privacy budget have been given in the proof of Corollary 4.7 in Appendix G.2). In this case, we can choose $\kappa \approx 0.015$ such that the iteration complexity of Algorithm 2 is no more than $\mathcal{O}\big(\delta^{-2.6}\big)$ and the cumulative privacy budget is 66.67, which is a constant and hence has an order of $\mathcal{O}(1)$.

Then, we compute the cumulative privacy budget of the remaining algorithms (except LDP-DSBO) listed in Table 1. For these algorithms, we employ the same Laplace noise used in our algorithm.

Given that all remaining algorithms in Table 1 use a constant stepsize, we estimate their cumulative privacy budgets ϵ_i under a stepsize $\gamma > 0$ and the DP-noise variance $\mathcal{O}\left(\frac{1}{(t+1)^{\varsigma}}\right)$ for some $\varsigma \in (0, 1)$. Additionally, we do not include inner-loops in this estimation. As explained in Subsection H.1, inner-loops cannot ensure a finite cumulative privacy budget in the infinite-time horizon, and thus a relaxed condition is considered for these algorithms, which makes the results better than the actual case. Based on (185), we obtain

$$\epsilon_i \leq \sum_{t=1}^{T} \mathcal{O}\left(\frac{\gamma}{\sigma_t(t+1)}\right) \leq \sum_{t=1}^{T} \mathcal{O}\left(\frac{1}{(t+1)^{1-\varsigma}}\right) \leq \mathcal{O}\left((T+1)^{\varsigma}\right), \tag{193}$$

where σ_t is the DP-noise variance and we have used the following relation for the last inequality:

$$\sum_{t=1}^{T} \frac{1}{(t+1)^{1-\varsigma}} \leq \int_{1}^{T+1} x^{\varsigma-1} dx \leq \frac{1}{\varsigma}(T+1)^{\varsigma} - \frac{1}{\varsigma} \leq \frac{1}{\varsigma}(T+1)^{\varsigma}. \tag{194}$$

By substituting the respective complexities of the algorithms listed in Table 1 into (194), we can obtain the results given in the last column of Table 1.