Optimization Letters
https://doi.org/10.1007/s11590-024-02101-4

ORIGINAL PAPER

Decentralized bilevel optimization

Xuxing Chen[1] · **Minhui Huang**[2] · **Shiqian Ma**[3]

Received: 5 March 2023 / Accepted: 1 February 2024

Abstract

Bilevel optimization has been successfully applied to many important machine learning problems. Algorithms for solving bilevel optimization have been studied under various settings. In this paper, we study the nonconvex-strongly-convex bilevel optimization under a decentralized setting. We design decentralized algorithms for both deterministic and stochastic bilevel optimization problems. Moreover, we analyze the convergence rates of the proposed algorithms in difference scenarios including the case where data heterogeneity is observed across agents. Numerical experiments on both synthetic and real data demonstrate that the proposed methods are efficient.

Keywords Decentralized optimization · Bilevel optimization · Hypergradient estimation

Here we provide a brief history of this paper. This paper appeared on arxiv on 06/12/2022 (arxiv ID 2206.05670). To the best of our knowledge, this is the first paper discussing decentralized algorithms for bilevel optimization. All other papers on the same topic appeared later than ours, including [1] which appeared on arxiv on 06/22/2022 (arxiv ID 2206.10870), Gao et al. [2] which appeared on arxiv on 06/30/2022 (arxiv ID 2206.15025), Terashita and Hara [3] which appeared on 10/05/2022 (arxiv ID 2210.02129), Chen et al. [4] which appeared on 10/23/2022 (arxiv ID 2210.12839), and [5] which appeared on arxiv on 12/20/2022 (arxiv ID 2212.10048). This paper was first submitted to NeurIPS 2022 and was rejected, although the reviewers did not raise any questions about the novelty and correctness.

✉ Shiqian Ma
sqma@rice.edu

Xuxing Chen
xuxchen@ucdavis.edu

Minhui Huang
mhhuang@ucdavis.edu

[1] Department of Mathematics, University of California, Davis, Davis, USA

[2] Department of Electrical and Computer Engineering, University of California, Davis, Davis, USA

[3] Department of Computational Applied Mathematics and Operations Research, Rice University, Houston, USA

Published online: 26 March 2024

1 Introduction

Bilevel optimization provides a framework for solving problems arising from meta learning [6–8], hyperparameter optimization [9, 10], reinforcement learning [10, 11], etc. It aims at minimizing an objective in the upper level under a constraint given by another optimization problem in the lower level, and has been studied intensively in recent years [7, 10–14]. Mathematically, it can be formulated as:

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \Phi(x) = f(x, y^*(x)), \quad \text{(upper level)} \\ \text{s.t.} \quad & y^*(x) = \arg\min_{y\in\mathbb{R}^q} g(x, y), \quad \text{(lower level)} \end{aligned} \tag{1}$$

where g is the lower level function which is usually assumed to be strongly convex with respect to y for all x, and f is the upper level function which is possibly non-convex. Designing a bilevel optimization algorithm requires estimation of the hypergradient $\nabla\Phi(x)$, which by chain rule and optimality condition of the lower level problem is:

$$\nabla\Phi(x) = \nabla_x f(x, y^*(x)) - \nabla_{xy} g(x, y^*(x))\Big(\nabla_y^2 g(x, y^*(x))\Big)^{-1} \nabla_y f(x, y^*(x)), \tag{2}$$

where $\nabla_{xy} g$ and $\nabla_y^2 g$ represent Jacobian matrix of $\nabla_y g$ and Hessian matrix of g respectively. Decentralized optimization aims at solving the finite-sum problem:

$$\min_{x\in\mathbb{R}^p} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{3}$$

where the ith agent only has access to the information related to f_i, and each agent communicates with neighbors to cooperatively solve the original problem. There is no central server collecting local updates. Decentralized algorithms are better choices in certain scenarios [15, 16]. Since decentralized training has been proved to be efficient, it is natural to ask:

Can we design an algorithm to solve bilevel optimization problems in a decentralized regime?

We will see the answer is affirmative. Our contributions can be summarized as follows.

- We propose a novel algorithm to estimate the hypergradient in different cases.
- We design a decentralized bilevel optimization (DBO) algorithm and analyze its convergence rate. We also analyze the convergence results for the stochastic version of DBO. To the best of our knowledge, our paper is the first work proposing provably convergent decentralized bilevel optimization algorithms in the presence of data heterogeneity.
- We study the effect of gradient tracking in the deterministic decentralized bilevel optimization and analyze the convergence rates.
- We conduct numerical experiments on several hyperparameter optimization problems. The results demonstrate the efficiency of our algorithms.

1.1 Related work

Bilevel optimization can be dated back to [17, 18]. Due to its great success in solving problems in meta learning [6–8], hyperparameter optimization [9, 10] and many others [19–21], there is a flurry of work proposing and analyzing bilevel optimization algorithms. The major challenge in bilevel optimization is the estimation of the hypergradient in (2). Computing each hypergradient requires access to $y^*(x)$, which is often intractable. Even if $y^*(x)$ is available, the nonlinearity in $\left[\nabla_y^2 g\right]^{-1}$ still requires careful consideration. There are several strategies to overcome this: approximate implicit differentiation (AID) [9, 12, 13, 22–24], iterative differentiation (ITD) [10, 13, 22, 24, 25] and Neumann series-based approach [11–14]. All of them only require first order information, Jacobian-vector and Hessian-vector products. Based on different algorithm designs, bilevel problems can be solved via single-loop [11, 26] or double-loop algorithms [12–14]. It is worth noting that variance reduction and momentum methods have also been introduced to bilevel optimization recently [27–29].

Decentralized optimization plays a key role in distributed optimization. It is gaining popularity in recent years due to its superior scalability for handling large language models and heterogeneous environments (i.e., different bandwidth, latency, data distribution, etc.) [30, 31]. Under a decentralized setting, the data is distributed to different agents, and each agent communicates with neighbors to solve a finite-sum minimization problem. As opposed to centralized optimization, decentralized optimization aims at solving the problem without a central server that collects iterates from local agents. The main challenge is the data heterogeneity across agents, which should be mitigated by communications. It has been proved that decentralized algorithms have their own advantages such as faster convergence, data privacy preservation and robustness to low network bandwidth compared to the centralized setting and single-agent training [15]. For example, low network bandwidth will greatly hinder the communication with the central server if the algorithm is designed to be centralized.

An important approach to accelerate the decentralized algorithms is gradient tracking, which has been proved to be efficient [32–35]. We refer the interested readers to [36], which provides a comprehensive review of decentralized optimization in a unified variance reduction framework.

Distributed bilevel optimization can be directly applied to solve problems like hyperparameter optimization, min-max optimization, meta learning, etc, in a distributed manner. For example, meta learning, which aims at training a model on some learning tasks so that it can solve new learning tasks using only a few samples, has been studied in the context of medical data analysis [37, 38]. In this bilevel optimization model, the lower level problem targets to minimize the loss function using the training tasks, and the upper level problem targets to choose the shared model parameters using the testing tasks. Extending meta learning to the decentralized setting has also been studied [39], and one important reason to apply decentralized meta learning in medical data analysis is protecting patients' privacy. Different hospitals, as agents in decentralized training, can collaborate to train a model, but they

should not share patients' data during the training process, and decentralized meta learning can help achieve this. Motivated by such applications, there exist recent works considering bilevel optimization under distributed setting. Bilevel optimization under a federated setting has received some attention recently [40, 41], and so does min-max optimization under various distributed settings [42–44]. However, none of these papers considers bilevel optimization under the decentralized setting. There is a concurrent work [45] also studying decentralized bilevel optimization. However, it aims at solving decentralized bilevel optimization problems under a personalized setting, in a sense that the lower level problems are different among agents. In Sect. 3 we will see that our problem is substantially different. To the best of our knowledge, our paper is the first work on non-personalized decentralized bilevel optimization.

2 Preliminaries

In this paper we consider the following decentralized optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \Phi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x, y^*(x)), \quad \text{(upper level)} \\ \text{s.t.} \quad & y^*(x) = \arg\min_{y\in\mathbb{R}^q} g(x,y) := \frac{1}{n}\sum_{i=1}^{n} g_i(x,y), \quad \text{(lower level)} \end{aligned} \tag{4}$$

where $x \in \mathbb{R}^p, y \in \mathbb{R}^q$. f_i is possibly nonconvex and g_i is strongly convex in y. Here n denotes the number of agents. The local objectives f_i and g_i are defined as:

$$f_i(x,y) = \mathbb{E}_{\phi\sim\mathcal{D}_{f_i}}\left[F(x,y;\phi)\right], \quad g_i(x,y) = \mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}\left[G(x,y;\xi)\right].$$

$\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$ represent the data distributions used to generate the objectives for agent i, and each agent only has access to f_i and g_i. In practice we can replace the expectation by empirical loss,

$$f_i(x,y) = \frac{1}{n_{f_i}}\sum_{j=1}^{n_{f_i}} F(x,y;\phi_{ij}), \quad g_i(x,y) = \frac{1}{n_{g_i}}\sum_{j=1}^{n_{g_i}} G(x,y;\xi_{ij}),$$

and then use mini-batch or full batch gradient descent in the updates. When we use mini-batch gradient descent, we call it *"stochastic case"*, and when we use full batch gradient descent, we call it *"deterministic case"*. We will study the convergence rates under these two cases in Sect. 3.

Notation We denote by $\nabla f(x,y)$ and $\nabla^2 f(x,y)$ the gradient and Hessian matrix of f, respectively. We use $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ to represent the gradients of f with respect to x and y, respectively. Denote by $\nabla_{xy} f(x,y) = \nabla_x \nabla_y f(x,y) \in \mathbb{R}^{p\times q}$ the Jacobian matrix of $\nabla_y f(x,y)$ and $\nabla_y^2 f(x,y)$ the Hessian matrix of f with respect to y. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and Frobenius norm for matrices, and $\|\cdot\|_2$ denotes the spectral norm for matrices. $\mathbf{1}_n$ is the all one vector in $\mathbb{R}^n$.

The following assumptions will be used, which are standard in bilevel optimization [11–14] and decentralized optimization literature [15, 16, 34, 35].

Assumption 2.1 (*Smoothness and convexity*) For any i, functions f_i, ∇f_i, ∇g_i, $\nabla^2 g_i$ are $L_{f,0}, L_{f,1}, L_{g,1}, L_{g,2}$ Lipschitz continuous respectively, i.e.,

$$|f_i(z) - f_i(z')| \le L_{f,0}\|z - z'\|,\ \|\nabla f_i(z) - \nabla f_i(z')\| \le L_{f,1}\|z - z'\|,$$
$$\|\nabla g_i(z) - \nabla g_i(z')\| \le L_{g,1}\|z - z'\|,\ \|\nabla^2 g_i(z) - \nabla^2 g_i(z')\| \le L_{g,2}\|z - z'\|,$$

for any $z = (x, y)$ and $z' = (x', y')$. Function g_i is μ-strongly convex in y for all i, i.e., $\nabla_y^2 g_i(x, y) \succeq \mu I$. Moreover, we define $L = \max\left(L_{f,1}, L_{g,1}\right)$, and $\kappa = \frac{L}{\mu}$.

Assumption 2.2 (*Network topology*) Suppose the communication network is represented by a weight matrix $W = (w_{ij}) \in \mathbb{R}^{n\times n}$, i.e., $w_{ij} \ge 0$ and is non-zero if and only if node i is a neighbor of j. W is symmetric and doubly stochastic, i.e.,

$$W = W^{\top}, \quad W\mathbf{1_n} = \mathbf{1_n}, \quad w_{ij} \ge 0, \forall i, j,$$

and its eigenvalues satisfy $1 = \lambda_1 > \lambda_2 \ge \cdots \ge \lambda_n > -1$ and $\rho := \max\left(|\lambda_2|, |\lambda_n|\right) < 1$.

Assumption 2.3 (*Data homogeneity on g*) Assume the data associated with g_i is independent and identically distributed, i.e., $\mathcal{D}_{g_i} = \mathcal{D}_g$. (We do not require data homogeneity in the upper level.)

Assumption 2.4 (*Bounded variance*) The stochastic derivatives $\nabla f_i(x, y; \phi)$, $\nabla g_i(x, y; \xi)$, $\nabla^2 g_i(x, y; \xi)$ are unbiased with bounded variances σ_f^2, $\sigma_{g,1}^2$, $\sigma_{g,2}^2$, respectively.

3 Our algorithms

If it is a single-agent system, i.e., $n = 1$, a natural idea to solve bilevel optimization (4) is to apply gradient descent for the upper level problem, which leads to the following updating scheme:

$$x^{k+1} = x^k - \eta_k \nabla\Phi(x^k),$$

where $\eta_k > 0$ is a step size, and $\nabla\Phi(x^k)$ is the hypergradient at x^k. However, computing $\nabla\Phi(x^k)$ requires $y^*(x^k)$. To obtain an approximation to $y^*(x^k)$, we can apply gradient descent to solve the lower-level problem. Therefore, a prototype of the gradient descent method for solving bilevel optimization (4) can be described as:

$$\begin{aligned}
&\text{for } k = 0, 1, \ldots, K \\
&\text{for } t = 0, 1, \ldots, T-1 \\
&y^{t+1} = y^t - \eta_y \nabla_y g(x^k, y^t) \\
&x^{k+1} = x^k - \eta_x \widetilde{\nabla\Phi}(x^k),
\end{aligned}$$

where $\widetilde{\nabla\Phi}(x^k)$ is an approximation of the hypeprgradient $\nabla\Phi(x^k)$ and is defined as

$$\widetilde{\nabla\Phi}(x^k) = \nabla_x f(x^k, y^T) - \nabla_{xy} g(x^k, y^T)\Big(\nabla_y^2 g(x^k, y^T)\Big)^{-1} \nabla_y f(x^k, y^T).$$

Clearly, there are two loops involved. We call the one updating x^k the outer loop, and the one updating y^t the inner loop.

When it comes to the decentralized setting in a multi-agent system, there are a few new challenges. Here we first discuss the main challenge when there is data heterogeneity, i.e., when Assumption 2.3 does not hold. In the outer loop of bilevel optimization algorithms [11–14], we typically focus on estimating the hypergradient so that we can perform gradient descent according to the hypergradient estimate. In the decentralized setting, where each agent has their own hypergradient given by:

$$\nabla\Phi_i(x) = \nabla_x f_i(x, y^*(x)) - \nabla_{xy} g(x, y^*(x))\Big(\nabla_y^2 g(x, y^*(x))\Big)^{-1} \nabla_y f_i(x, y^*(x)). \tag{5}$$

Note that node i does not have access to $\nabla_{xy} g(x, y^*(x))\Big(\nabla_y^2 g(x, y^*(x))\Big)^{-1}$ and $y^*(x)$ which both require global information about g. One natural idea is to use the following function as a local surrogate (here $y_i^*(x) := \arg\min_y g_i(x, y)$):

$$\nabla f_i(x, y_i^*(x)) = \nabla_x f_i(x, y_i^*(x)) - \nabla_{xy} g_i(x, y_i^*(x))\Big(\nabla_y^2 g_i(x, y_i^*(x))\Big)^{-1} \nabla_y f_i(x, y_i^*(x)). \tag{6}$$

Unfortunately, the hypergradient estimation error (i.e., $\Big\|\nabla\Phi_i(x) - \nabla f_i(x, y_i^*(x))\Big\|$) may not be diminishing. For example, when $f_i(x, y) = \frac{1}{2} y^\top y$, and $g_i(x, y) = \frac{i}{2} y^\top y - x^\top y$, we have $\nabla f_i(x, y_i^*(x)) = \frac{x}{i^2}$, $\nabla\Phi_i(x) = \frac{2x}{(n+1)i}$, which implies

$$\Big\|\nabla\Phi_i(x) - \nabla f_i(x, y_i^*(x))\Big\| = \left(\frac{n+1-2i}{i^2(n+1)}\right)\|x\|$$

which cannot be diminishing no matter how the algorithm proceeds if $x \neq 0$. Thus we cannot directly apply (6) in our problem when Assumption 2.3 does not hold. Note that the difference between our work and [45] can be viewed as the difference between (6) and (5). Their problem formulation ((1a) and (1b) in [45]) is essentially $\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n f_i(x, y_i^*(x))$, which means using (6) is sufficient for computing the global hypergradient. Mathematically, in our setting we would like to compute $Z \in \mathbb{R}^{q\times p}$ such that

$$Z^\top = \left(\sum_{i=1}^n \nabla_{xy} g_i(x,y)\right)\left(\sum_{i=1}^n \nabla_y^2 g_i(x,y)\right)^{-1} \tag{7}$$

on node i for any given (x, y). In the next section we design a novel oracle to solve this subproblem with heterogeneous data at the price of the computation of Jacobian matrices.

3.1 Jacobian–Hessian–Inverse Product oracle

We introduce the Jacobian–Hessian–Inverse Product (JHIP) oracle, which is essentially a decentralized subroutine. Denote by $H_i \in \mathbb{S}_{++}^{q\times q}$ and $J_i \in \mathbb{R}^{p\times q}$ the Hessian matrix of g_i and the Jacobian matrix of $\nabla_y g_i$. Every agent aims at finding $Z \in \mathbb{R}^{q\times p}$ (i.e., (7)) such that:

$$\sum_{i=1}^n H_i Z = \sum_{i=1}^n J_i^\top \quad \text{or equivalently, } Z^\top = \left(\sum_{i=1}^n J_i\right)\left(\sum_{i=1}^n H_i\right)^{-1}. \tag{8}$$

Notice that this is exactly the optimality condition of:

$$\min_{Z\in\mathbb{R}^{q\times p}} \frac{1}{n}\sum_{i=1}^n h_i(Z), \quad \text{where } h_i(Z) = \frac{1}{2}\operatorname{Tr}(Z^\top H_i Z) - \operatorname{Tr}(J_i Z). \tag{9}$$

The objective in (9) is strongly convex since each H_i is symmetric positive definite. Hence we can design a decentralized algorithm with gradient tracking so that all the agents can collaborate to solve for (8) without a central server. The algorithm is described in Algorithm 1, where we use the bold texts to highlight the different updates when the problem is deterministic and stochastic.

Algorithm 1 Jacobian-Hessian-Inverse Product oracle

Input: $Z_i^{(0)} \in \mathbb{R}^{q\times p}$, stepsizes $\{\gamma_t\}_{t=0}^\infty$, N, and initialization.
- **if deterministic,** $Y_i^{(0)} = H_i Z_i^{(0)} - J_i^\top, G_i^{(0)} = H_i Z_i^{(0)}$,
- **if stochastic,** $Y_i^{(0)} = \hat{H}_i^{(0)} Z_i^{(0)} - \left(\hat{J}_i^{(0)}\right)^\top, G_i^{(0)} = \hat{H}_i Z_i^{(0)} - \left(\hat{J}_i^{(0)}\right)^\top$.

Data: $H_i \in \mathbb{S}_{++}^{q\times q}, J_i \in \mathbb{R}^{p\times q}$ accessible only to agent i (deterministic). $(\hat{H}_i^{(t)}, \hat{J}_i^{(t)}), t \in \{0, 1, ..., N-1\}$ accessible only to agent i (stochastic).
for $t = 0, 1, ..., N-1$ **do**
  $Z_i^{(t+1)} = \sum_{j=1}^n w_{ij} Z_j^{(t)} - \gamma_t Y_i^{(t)}$,
  $G_i^{(t+1)} = H_i Z_i^{(t+1)}$ **if deterministic, else** $\hat{H}_i^{(t+1)} Z_i^{(t+1)} - \left(\hat{J}_i^{(t+1)}\right)^\top$,
  $Y_i^{(t+1)} = \sum_{i=1}^n w_{ij} Y_j^{(t)} + G_i^{(t+1)} - G_i^{(t)}$, for $i = 1, ..., n$.
end for

Note that for the deterministic case we can just maintain $G_i^{(t+1)} = H_i Z_i^{(t+1)}$ instead of the gradient $H_i Z_i^{(t+1)} - J_i^\top$ because we only use $G_i^{(t+1)}$ in line 6—the gradient tracking step, and the constant term $J_i^\top$ will be cancelled if we set $G_i^{(t+1)}$ as the gradient. We use $\hat{H}_i^{(t)} Z_i^{(t)} - \left(\hat{J}_i^{(t)}\right)^\top$ to represent the stochastic gradient of $h_i(Z)$ at $Z_i^{(t+1)}$. Each Hessian-matrix product $H_i Z_i^{(t+1)}$ in line 5 can be viewed as p Hessian-vector products, which is cheaper than computing the Hessian matrix when p is small. This oracle also requires computing the exact Jacobian matrix, which is more expensive than Jacobian-vector product. Moreover, the convergence rates have been well understood [34, 36, 46]. In general, one can also design other oracles (e.g., decentralized ADMM [47–51]) to solve (9). The convergence rates of this algorithm are summarized in Lemma 15.

3.2 Hypergradient estimate

Algorithm 2 Hypergradient estimate

1: **Input:** x, y, N, M, β
2: **if** Assumption 2.3 holds **then**
3: **if** Deterministic case **then**
4: Run N-step conjugate gradient method on $\nabla_y^2 g_i(x,y)v = \nabla_y f_i(x,y)$ to get v^N. Set $\hat{\nabla} f_i = \nabla_x f_i(x,y) - \nabla_{xy} g_i(x,y) v^N$.
5: **else**
6: Set $\hat{\nabla} f_i = \nabla_x f_i(x, y;\ \phi^{(0)}) - \nabla_{xy} g_i(x, y;\ \phi^{(1)}) \cdot H_M \cdot \nabla_y f_i(x, y;\ \phi^{(0)})$,
7: where $H_M = \beta \sum_{s=0}^{M-1} \prod_{n=1}^{s} (I - \beta \nabla_y^2 g_i(x, y;\ \phi^{(M+1-n)}))$
8: **end if**
9: **else**
10: **if** Deterministic case **then**
11: Run N-step deterministic Algorithm 1 with $\gamma_t = \mathcal{O}(1)$ and
12: $H_i = \nabla_y^2 g_i(x,y), J_i = \nabla_{xy} g_i(x,y)$ to get $Z_i^{(N)}$.
13: Set $\hat{\nabla} f_i = \nabla_x f_i(x,y) - \left(Z_i^{(N)}\right)^\top \nabla_y f_i(x,y)$.
14: **else**
15: Run N-step stochastic Algorithm 1 with $\gamma_t = \mathcal{O}(\frac{1}{t})$ and
16: $\hat{H}_i^{(t)} = \nabla_y^2 g_i(x, y;\ \phi_i^{(t)}), \hat{J}_i^{(t)} = \nabla_{xy} g_i(x, y;\ \phi_i^{(t)})$ to get $Z_i^{(N)}$.
17: Set $\hat{\nabla} f_i = \nabla_x f_i(x, y;\ \phi_i^{(0)}) - \left(Z_i^{(N)}\right)^\top \nabla_y f_i(x, y;\ \phi_i^{(0)})$.
18: **end if**
19: **end if**
20: **Output:** $\hat{\nabla} f_i$ on node i.

Before we propose our algorithms, we first introduce hypergradient estimates under different cases.

- *Case 1: Deterministic + homogeneous data* In this case the hypergradient is estimated based on the AID approach [13], which is essentially utilizing conjugate gradient method, and only requires Hessian-vector product oracles instead of explicit Hessian matrix computation. We adopt the approximation error (Lemma 3 in [13]) in Lemma 11.
- *Case 2: Stochastic + homogeneous data* In this case the hypergradient is estimated based on a slight modification of the Neumann series approach [12]. Gradients ∇f_i and ∇g_i are replaced by their corresponding first order stochastic oracles (i.e., stochastic gradients). We have the error estimation in Lemma 35.
- *Case 3: Heterogeneous data* In this case we compute the global Jacobian-Hessian-Inverse product by using the JHIP oracle (Algorithm 1). The error estimation results are given in Lemma 14.

3.3 Deterministic decentralized bilevel optimization

Algorithm 3 (Deterministic) Decentralized Bilevel Optimization

1: **Input:** $W, N, K, T, \eta_x, \eta_y, x_{i,0}, y_i^{(0)}$.
2: **for** $k = 0, 1, ..., K$ **do**
3: $\quad y_{i,k}^{(0)} = y_{i,k-1}^{(T)}$ if $k > 0$ otherwise $y_{i,k}^{(0)} = y_i^{(0)}$.
4: $\quad$ **for** $t = 0, 1, ..., T-1$ **do**
5: $\quad\quad$ **if** Assumption 2.3 holds **then**
6: $\quad\quad\quad y_{i,k}^{(t+1)} = y_{i,k}^{(t)} - \eta_y \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)})$, for $i = 1, ..., n$.
7: $\quad\quad$ **else**
8: $\quad\quad\quad v_{i,k}^{(t)} = \sum_{j=1}^n w_{ij} v_{j,k}^{(t-1)} + \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)}) - \nabla_y g_i(x_{i,k}, y_{i,k}^{(t-1)})$.
9: $\quad\quad\quad y_{i,k}^{(t+1)} = \sum_{j=1}^n w_{ij} y_{j,k}^{(t)} - \eta_y v_{i,k}^{(t)}$, for $i = 1, ..., n$.
10: $\quad\quad$ **end if**
11: $\quad$ **end for**
12: $\quad$ Run Algorithm 2 (with "deterministic case") to get $\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)})$.
13: $\quad x_{i,k+1} = \sum_{j=1}^n w_{ij} x_{j,k} - \eta_x \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)})$, for $i = 1, ..., n$.
14: **end for**
15: **Output:** $\bar{x}_K = \frac{1}{n} \sum_{i=1}^n x_{i,K}$.

We propose the decentralized bilevel optimization algorithm (DBO) in Algorithm 3. In the inner loop (lines 4–11) each agent performs local gradient descent updates for variable y in parallel. When Assumption 2.3 holds, we can simply run local gradient descent without communication because in the lower level local distribution already captures the global function information. When Assumption 2.3 does not hold, then the data distribution is substantially heterogeneous across agents, so we

add weighted averaging steps (line 9) to reach consensus and gradient tracking steps (line 8) to reduce the complexity. In the outer loop (lines 12–13) each agent communicates with neighbors and then performs gradient descent for variable x.[1] We have the following sublinear convergence result.

Theorem 3.1 *In Algorithm 3, suppose Assumptions 2.1 and 2.2 hold. Set* $\eta_x = \Theta(K^{-\frac{1}{3}}\kappa^{-\frac{8}{3}})$, $\eta_y = \frac{1}{\mu+L}$. *If Assumption 2.3 holds, we set* $T = \Theta(\kappa \log \kappa), N = \Theta(\sqrt{\kappa}\log\kappa)$. *If Assumption 2.3 does not hold, we set* $T = N = \Theta(\log K), \gamma_t = \Theta(1)$. *In both cases, we have*:

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2 = \mathcal{O}\left(\frac{\kappa^{\frac{8}{3}}}{K^{\frac{2}{3}}}\right).$$

3.4 Deterministic decentralized bilevel optimization with gradient tracking

In this section we study the effect of gradient tracking in decentralized bilevel optimization. We propose the Decentralized Bilevel Optimization with Gradient Tracking (DBOGT) Algorithm 4. We introduce u and v to serve as the update directions. For Algorithm 4 we have the following theorem.

Algorithm 4 (Deterministic) Decentralized Bilevel Optimization with Gradient Tracking

1: **Input:** $W, N, K, T, \eta_x, \eta_y, x_i^{(0)}, y_i^{(0)}$.
2: **for** $k = 0, 1, ..., K-1$ **do**
3: $\quad y_{i,k}^{(0)} = y_{i,k-1}^{(T)}$ if $k > 0$ otherwise $y_{i,k}^{(0)} = y_i^{(0)}$.
4: $\quad$ **for** $t = 0, 1, ..., T-1$ **do**
5: $\quad\quad$ **if** Assumption 2.3 holds **then**
6: $\quad\quad\quad y_{i,k}^{(t+1)} = y_{i,k}^{(t)} - \eta_y \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)})$, for $i = 1, ..., n$.
7: $\quad\quad$ **else**
8: $\quad\quad\quad v_{i,k}^{(t)} = \sum_{j=1}^n w_{ij} v_{j,k}^{(t-1)} + \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)}) - \nabla_y g_i(x_{i,k}, y_{i,k}^{(t-1)})$,
9: $\quad\quad\quad y_{i,k}^{(t+1)} = \sum_{j=1}^n w_{ij} y_{j,k}^{(t)} - \eta_y v_{i,k}^{(t)}$, for $i = 1, ..., n$.
10: $\quad\quad$ **end if**
11: $\quad$ **end for**
12: $\quad$ Run Algorithm 2 (with "deterministic case") to get $\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)})$.
13: $\quad u_{i,k} = \sum_{j=1}^n w_{ij} u_{j,k-1} + \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \hat{\nabla} f_i(x_{i,k-1}, y_{i,k-1}^{(T)})$,
14: $\quad x_{i,k+1} = \sum_{j=1}^n w_{ij} x_{j,k} - \eta_x u_{i,k}$, for $i = 1, ..., n$.
15: **end for**
16: **Output:** $\bar{x}_K = \frac{1}{n}\sum_{i=1}^n x_{i,K}$.

[1] For simplicity we use constant stepsize η_x in the outer loop. Similar results can be obtained for diminishing stepsizes.

Theorem 3.2 *In Algorithm 4, suppose Assumptions 2.1 and 2.2 hold. Set* $\eta_x = \Theta(\kappa^{-3})$, $\eta_y = \Theta(1)$. *If Assumption 2.3 holds, we set* $T = \Theta(\kappa \log \kappa), N = \Theta(\sqrt{\kappa} \log \kappa)$. *If Assumption 2.3 does not hold, we set* $T = N = \Theta(\log K), \gamma_t = \Theta(1)$. *In both cases, we have*

$$\frac{1}{K+1} \sum_{j=0}^{K} \|\nabla \Phi(\bar{x}_j)\|^2 = \mathcal{O}\Big(\frac{1}{K}\Big).$$

Note that this result implies that in DBOGT we can set η_x as a constant that is independent of the total number of iterations K, which matches the results in gradient tracking literature [33–35, 52].

3.5 Decentralized stochastic bilevel optimization

Our stochastic version of the DBO algorithm: Decentralized Stochastic Bilevel Optimization (DSBO), is described in Algorithm 5. Its convergence rate is given in Theorem 3.3.

Algorithm 5 Decentralized Stochastic Bilevel Optimization

Input: $W, M, N, K, T, \eta_x, \eta_y, x_i^{(0)}, y_i^{(0)}$.
for $k = 0, 1, ..., K-1$ **do**
  $y_{i,k}^{(0)} = y_{i,k-1}^{(T)}$ if $k > 0$ otherwise $y_{i,k}^{(0)} = y_i^{(0)}$.
  for $t = 0, 1, ..., T-1$ **do**
    if Assumption 2.3 holds **then**
      $y_{i,k}^{(t+1)} = y_{i,k}^{(t)} - \eta_y \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)};\ \xi_{i,k}^{(t)})$, for $i = 1, ..., n$.
    else
      $y_{i,k}^{(t+1)} = \sum_{j=1}^{n} w_{ij}(y_{j,k}^{(t)} - \eta_y \nabla_y g_i(x_{i,k}, y_{i,k}^{(t)};\ \xi_{i,k}^{(t)}))$, for $i = 1, ..., n$.
    end if
  end for
  Run Algorithm 2 ("stochastic case" option) to get $\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)};\ \phi_{i,k})$.
  $x_{i,k+1} = \sum_{j=1}^{n} w_{ij} x_{j,k} - \eta_x \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)};\ \phi_{i,k})$, for $i = 1, ..., n$.
end for
Input: $\bar{x}_K = \frac{1}{n} \sum_{i=1}^{n} x_{i,K}$.

Theorem 3.3 *In Algorithm 5, suppose Assumptions 2.1 and 2.2 hold. Set* $\eta_x = \Theta(K^{-\frac{1}{2}}), T = \Theta(K^{\frac{1}{2}})$. *If Assumption 2.3 holds, we set* $M = \Theta(\log K), \eta_y^{(t)} = \Theta(K^{-\frac{1}{2}}), \beta = \min\Big(\frac{\mu}{\mu^2 + \sigma_{g,2}^2}, \frac{1}{L}\Big)$.

If Assumption 2.3 does not hold, we set $N = \Theta(\log K), \eta_y^{(t)} = \mathcal{O}(\frac{1}{t}), \gamma_t = \mathcal{O}(\frac{1}{t})$. *In both cases, we have*

$$\frac{1}{K+1}\sum_{j=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_j)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

4 Numerical experiments

In this section we conduct several experiments on hyperparameter optimization problems in the decentralized setting, which can be formulated as:

$$\begin{aligned} \min_{\lambda\in\mathbb{R}^p} \quad & \Phi(\lambda) = \frac{1}{n}\sum_{i=1}^{n} f_i(\lambda, \tau^*(\lambda)), \\ \text{s.t.} \quad & \tau^*(\lambda) = \arg\min_{\tau\in\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^{n} g_i(\lambda, \tau). \end{aligned} \tag{10}$$

Here f_i and g_i denote the validation loss and training loss on node i, respectively. The goal is to find the best hyperparameter λ under the constraint that $\tau^*(\lambda)$ is the optimal model parameter of the lower level problem. Due to the space limit, the details of the setup of the experiments are given in the "Appendix".

4.1 Synthetic data

We first conduct logistic regression with l^2 regularization on synthetic heterogeneous data (e.g., [9, 24]). We plot the logarithm of the norm of the gradient in Fig. 1a. From this figure we see that all three algorithms: DBO (Algorithm 3), DBOGT (Algorithm 4), and DSBO (Algorithm 5) can reduce the gradient to an acceptable level. Moreover, DBO and DBOGT have similar performance, and they are both slightly better than DSBO. We also include the test accuracy in Fig. 1b, which indicates similarly good performance in terms of accuracy.

4.2 Real-world data

We now conduct the DSBO algorithm on a logistic regression problem on 20 Newsgroup dataset[2] [24]. In Fig. 1c we plot the test accuracy of every iteration. From this figure we see that the DSBO algorithm is able to get good test accuracy under different settings of stepsizes.

Finally we apply deterministic DBO and DBOGT algorithms on a data hypercleaning problem [13, 53] for MNIST dataset [54]. The purpose is to demonstrate the advantage of the gradient tracking technique. The Fig. 2a shows that the perofromance of DBO and DBOGT are similar when the stepsizes are small. However, the Fig. 2b

[2] http://qwone.com/~jason/20Newsgroups/.

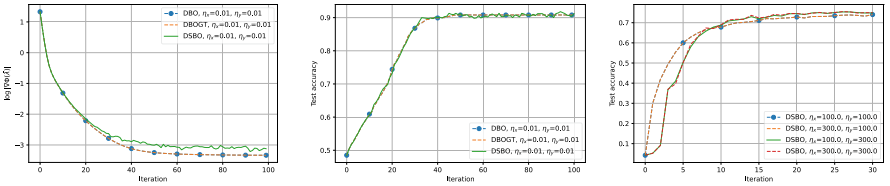


Fig. 1 **a**, **b** Logistic regression on synthetic data. **c** Logistic regression on 20 Newsgroup

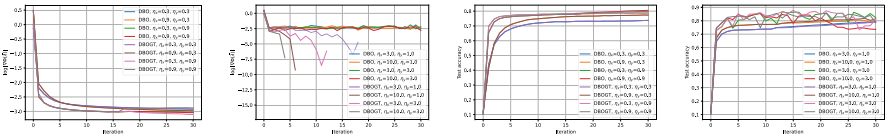


Fig. 2 Data hyper-cleaning on MNIST

shows that DBOGT converges much faster than DBO when the stepsizes are relatively large. This supports the conclusions in Theorem 3.2. We also include the test accuracy results in Fig. 2c, d, from which we can find that our test performance are comparable with [13].

5 Conclusion

In this paper we propose both deterministic and stochastic algorithms for solving decentralized bilevel optimization problems. We obtain sublinear convergence rates when the lower level function is generated by homogeneous data. Moreover, at the price of computing Jacobian matrices, we propose decentralized algorithms with sublinear convergence rates when the lower level function is generated by heterogeneous data. Numerical experiments demonstrate that the proposed algorithms are efficient. It is still an open question whether one can design decentralized optimization algorithms without assuming data homogeneity and Jacobian computation. We leave this as a future work.

Appendix 1: Details about experiments and other results

In this section we provide details about our experiments as well as results about training and test loss. For each experiment, we set our network topology as a special ring network, where $W = (w_{i,j})$ and the only nonzero elements are given by:

$$w_{i,i} = a,\ w_{i,i+1} = w_{i,i-1} = \frac{1-a}{2},\quad \text{for some } a \in (0,1).$$

Here we overload the notation and set $w_{n,n+1} = w_{n,1}, w_{1,0} = w_{1,n}$. Note that a is the unique parameter that determines the weight matrix and will be specified in each experiment.

Synthetic data

Logistic regression on synthetic data

In this experiment, on node i we have:

$$f_i(\lambda, \tau^*(\lambda)) = \sum_{(x_e, y_e) \in \mathcal{D}_i'} \psi(y_e x_e^\top \tau^*(\lambda)),$$
$$g_i(\lambda, \tau) = \sum_{(x_e, y_e) \in \mathcal{D}_i} \psi(y_e x_e^\top \tau) + \frac{1}{2} \tau^\top \mathrm{diag}(e^\lambda) \tau,$$

where e^λ is element-wise, $\mathrm{diag}(v)$ denotes the diagonal matrix generated by vector v, and $\psi(x) = \log(1 + e^{-x})$. $\mathcal{D}_i'$ and $\mathcal{D}_i$ represent validation set and training set on node i. Following the setup in [24], we first randomly generate $\tau^* \in \mathbb{R}^p$ and the noise vector $\epsilon \in \mathbb{R}^p$. For the data point (x_e, y_e) on node i, each element of x_e is sampled from the normal distribution with mean 0, variance i^2. y_e is then set by $y_e = \mathrm{sign}(x_e^\top \tau^* + m\epsilon)$, where sign denotes the sign function and $m = 0.1$ denotes the noise rate. In the experiment we choose $p = q = 50$, and the number of inner-loop and outer-loop iterations as 10 and 100 respectively. N, the number of iterations of the JHIP oracle 1 is 20. The stepsizes are $\eta_x = \eta_y = \gamma = 0.01$. The number of agents n is chosen as 20, and the weight parameter $a = 0.4$ (Fig. 3).

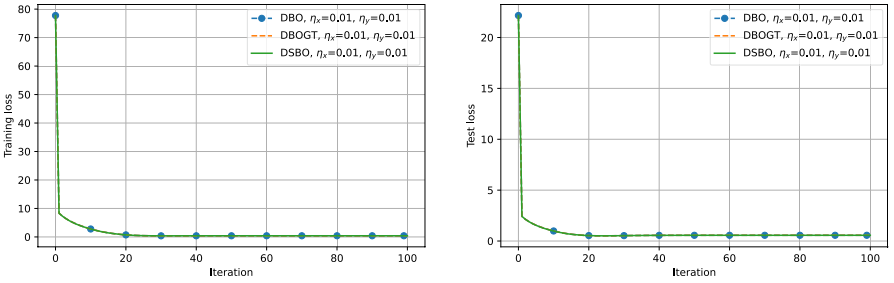
Real-world data

Logistic regression on 20 Newsgroup dataset

In this experiment, on node i we have:

$$f_i(\lambda, \tau^*(\lambda)) = \frac{1}{|\mathcal{D}_{val}^{(i)}|} \sum_{(x_e, y_e) \in \mathcal{D}_{val}^{(i)}} L(x_e^\top \tau^*, y_e),$$
$$g_i(\lambda, \tau) = \frac{1}{|\mathcal{D}_{tr}^{(i)}|} \sum_{(x_e, y_e) \in \mathcal{D}_{tr}^{(i)}} L(x_e^\top \tau, y_e) + \frac{1}{cp} \sum_{i=1}^{c} \sum_{j=1}^{p} e^{\lambda_j} \tau_{ij}^2,$$

where $c = 20$ denotes the number of topics, $p = 101631$ is the feature dimension, L is the cross entropy loss, $\mathcal{D}_{val}$ and $\mathcal{D}_{tr}$ are the validation and training data sets, respectively. Our codes can be seen as decentralized versions of the one provided in [13].



(a) (b)

Fig. 3 Logistic regression on synthetic data

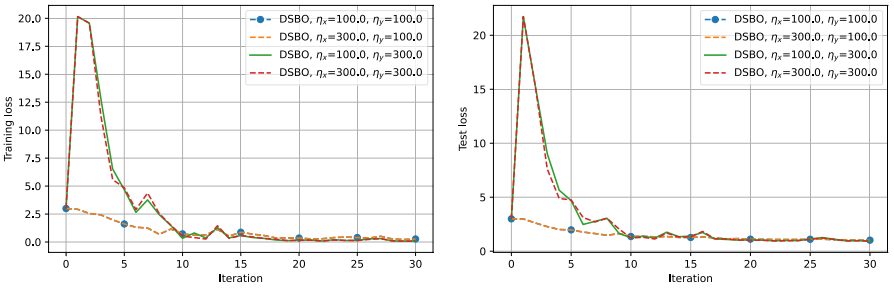
We first set inner and outer stepsizes $\eta_x = \eta_y = 100$ (the same as the ones used in [13]), and then compare its performance with different stepsizes. We set the number of inner-loop iterations $T = 10$, the number of outer-loop iterations $K = 30$, the number of agents $n = 20$, and the weight parameter $a = 0.33$. At the end of jth outer-loop iteration we use the average $\overline{\tau}_j = \frac{1}{n}\sum_{i=1}^{n} \tau_{i,j}$ as the model parameter and then do the classification on the test set to get the test accuracy (Fig. 4).

Data hyper-cleaning on MNIST

In this experiment, on node i we have:

$$f_i(\lambda, \tau) = \frac{1}{|\mathcal{D}_{val}^{(i)}|} \sum_{(x_e, y_e) \in \mathcal{D}_{val}^{(i)}} L(x_e^{\mathsf{T}} \tau, y_e),$$

$$g_i(\lambda, \tau) = \frac{1}{|\mathcal{D}_{tr}^{(i)}|} \sum_{(x_e, y_e) \in \mathcal{D}_{tr}^{(i)}} \sigma(\lambda_e) L(x_e^{\mathsf{T}} \tau, y_e) + C_r \|\tau\|^2,$$



(a) (b)

Fig. 4 Logistic regression on 20 Newsgroup dataset

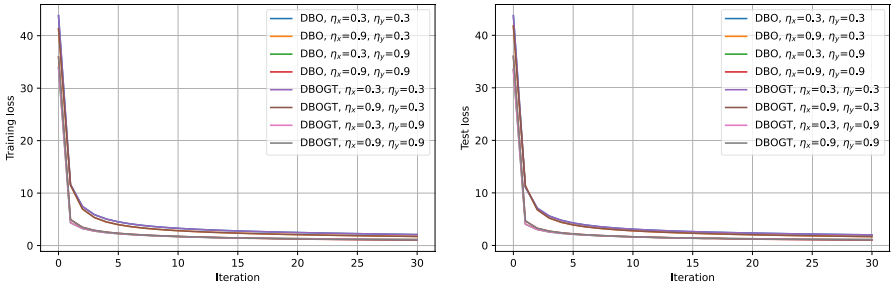


Fig. 5 Data hyper-cleaning on MNIST

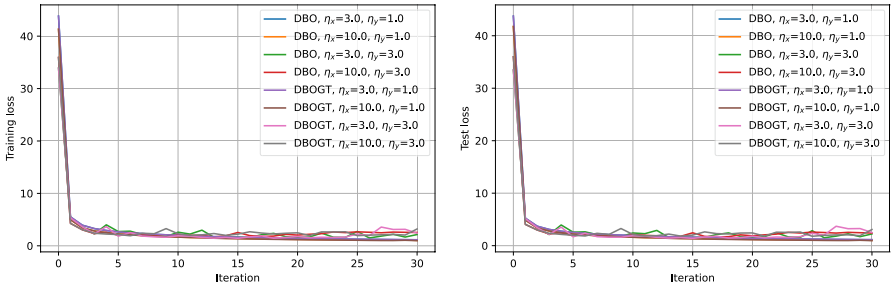


Fig. 6 Data hyper-cleaning on MNIST

where L is the cross-entropy loss and $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The number of inner-loop iterations T and outer-loop iterations K are set as 10 and 30, respectively. The number of agents $n = 20$ and the weight parameter $a = 0.5$. Following [13, 53] the regularization parameter C_r is set as 0.001. We first choose stepsizes similar to those in [13] and then set larger stepsizes. In each iteration we evaluate the norm of the hypergradient at the average of the hyperparameters $\bar{\lambda}$, and plot the logarithm (base 10) of the norm of the hypergradient versus iteration number in Fig. 2 (Figs. 5, 6).

Appendix 2: Convergence analysis

In this section we provide the proofs of convergence results. For convenience, we first list the notation below.

$$W := (w_{ij}) \text{ is symmetric doubly stochastic, and } \rho := \max\left(|\lambda_2|, |\lambda_n|\right) < 1$$

$$X_k := \left(x_{1,k}, x_{2,k}, \ldots, x_{n,k}\right),\ \bar{x}_k := \frac{1}{n}\sum_{i=1}^{n} x_{i,k},$$

$$\partial\Phi(X_k) := \left(\hat{\nabla} f_1(x_{1,k}, y_{1,k}^{(T)}), \ldots, \hat{\nabla} f_n(x_{n,k}, y_{n,k}^{(T)})\right),$$

$$\partial\Phi(X_k;\phi) := \left(\hat{\nabla} f_1(x_{1,k}, y_{1,k}^{(T)};\phi_{1,k}), \ldots, \hat{\nabla} f_n(x_{n,k}, y_{n,k}^{(T)};\phi_{n,k})\right)$$

$$\overline{\partial\Phi(X_k)} := \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}),\ \overline{\partial\Phi(X_k;\phi)} := \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)};\phi_{i,k}),$$

$$q_{i,k} := x_{i,k} - \bar{x}_k,\ r_{i,k} := u_{i,k} - \bar{u}_k,$$

$$Q_k := \left(q_{1,k}, q_{2,k}, \ldots, q_{n,k}\right),\ R_k := \left(r_{1,k}, r_{2,k}, \ldots, r_{n,k}\right) \in \mathbb{R}^{p\times n},$$

$$S_K := \sum_{k=1}^{K} \|Q_k\|^2,\ T_K := \sum_{j=0}^{K} \|\nabla\Phi(\bar{x}_j)\|^2,\ E_K := \sum_{j=1}^{K}\sum_{i=1}^{n} \|x_{i,j} - x_{i,j-1}\|^2,$$

$$A_K := \sum_{j=0}^{K}\sum_{i=1}^{n} \|y_{i,j}^{(T)} - y_i^*(x_{i,j})\|^2, B_K := \sum_{j=0}^{K}\sum_{i=1}^{n} \|v_{i,j}^* - v_{i,j}^{(0)}\|^2,$$

$$v_{i,j}^* = \left(\nabla_y^2 g_i(x_{i,j}, y_i^*(x_{i,j}))\right)^{-1} \nabla_y f_i(x_{i,j}, y_i^*(x_{i,j})),$$

$$\delta_y := (1-\eta_y\mu)^2,\ \delta_\kappa := \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2.$$

We first introduce a few lemmas that are useful in the proofs.

Lemma 1 *For any $p, q, r \in \mathbb{N}_+$ and matrix $A \in \mathbb{R}^{p\times q}, B \in \mathbb{R}^{q\times r}$, we have*:

$$\|AB\| \le \min\left(\|A\|_2 \cdot \|B\|, \|A\| \cdot \|B^{\mathrm{T}}\|_2\right).$$

Lemma 2 *For any matrix $A = (a_1, a_2, \ldots, a_q) \in \mathbb{R}^{p\times q}$, we have*:

$$\|a_j\|^2 \le \|A\|_2^2 \le \|A\|^2 = \sum_{i=1}^{q} \|a_i\|^2,\ \forall j \in \{1, 2, \ldots, q\}.$$

For one-step gradient descent, we have the following result (see, e.g., Lemma 10 in [34] and Lemma 3 in [46]).

Lemma 3 *Suppose $f(x)$ is μ-strongly convex and $L-$smooth. For any x and $\eta < \frac{2}{\mu+L}$, define $x^+ = x - \eta\nabla f(x),\ x^* = \arg\min f(x)$. Then we have*

$$\|x^+ - x^*\| \le (1-\eta\mu)\|x - x^*\|.$$

The following lemma is a common result in decentralized optimization (e.g., [15, Lemma 4]).

Lemma 4 *Suppose Assumption 2.2 holds. We have for any integer $k \geq 0$,*

$$\left\| W^k - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right\|_2 \leq \rho^k.$$

Proof Assume $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1$ are eigenvalues of W. Since $W^k \mathbf{1}_n \mathbf{1}_n^\top = \mathbf{1}_n \mathbf{1}_n^\top W^k$, we know W^k and $\mathbf{1}_n \mathbf{1}_n^\top$ are simultaneously diagonalizable. Hence there exists an orthogonal matrix P such that

$$W^k = P\text{diag}(\lambda_i^k)P^{-1}, \quad \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} = P\text{diag}(1, 0, 0, \ldots, 0)P^{-1},$$

and thus:

$$\left\| W^k - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right\|_2 = \left\| P(\text{diag}(\lambda_i^k) - \text{diag}(1, 0, 0, \ldots, 0))P^{-1} \right\|_2 \leq \max\left(|\lambda_2|^k, |\lambda_n|^k \right).$$

By definition of *rho*, the proof is complete. □

The following three lemmas are adopted from Lemma 2.2 in [12]:

Lemma 5 (Hypergradient) *Define* $\Phi_i(x) := f_i(x, y^*(x))$, *where* $y^*(x) = \arg\min_{y \in \mathbb{R}^q} g(x, y)$. *Under Assumption 2.1 we have*:

$$\nabla \Phi_i(x) = \nabla_x f_i(x, y^*(x)) - \nabla_{xy} g(x, y^*(x)) \left(\nabla_y^2 g(x, y^*(x)) \right)^{-1} \nabla_y f_i(x, y^*(x)).$$

Moreover, $\nabla \Phi_i$ *is Lipschitz continuous*:

$$\|\nabla \Phi_i(x_1) - \nabla \Phi_i(x_2)\| \leq L_\Phi \|x_1 - x_2\|,$$

with the Lipschitz constant given by:

$$L_\Phi = L + \frac{2L^2 + L_{g,2} L_{f,0}^2}{\mu} + \frac{L L_{f,0} L_{g,2} + L^3 + L_{g,2} L_{f,0} L}{\mu^2} + \frac{L_{g,2} L^2 L_{f,0}}{\mu^3} = \Theta(\kappa^3).$$

Remark if Assumption 2.3 does not hold, then this hypergradient is completely different from the local hypergradient:

$$\nabla f_i(x, y_i^*(x)) = \nabla_x f_i(x, y_i^*(x)) - \nabla_{xy} g_i(x, y_i^*(x)) \left(\nabla_y^2 g_i(x, y_i^*(x)) \right)^{-1} \nabla_y f_i(x, y_i^*(x)), \tag{11}$$

where $y_i^*(x) = \arg\min_{y \in \mathbb{R}^q} g_i(x, y)$.

Lemma 6 *Define*:

$$\bar{\nabla} f_i(x, y) = \nabla_x f_i(x, y) - \nabla_{xy} g(x, y) \left(\nabla_y^2 g(x, y) \right)^{-1} \nabla_y f_i(x, y).$$

Under the Assumption 2.1 we have:

$$\|\bar{\nabla} f_i(x,y) - \bar{\nabla} f_i(\tilde{x},\tilde{y})\| \le L_f \|(x,y) - (\tilde{x},\tilde{y})\|,$$

where the Lipschitz constant is given by:

$$L_f = L + \frac{L^2}{\mu} + L_{f,0}\left(\frac{L_{g,2}}{\mu} + \frac{L_{g,2}L}{\mu^2}\right) = \Theta(\kappa).$$

Lemma 7 *Suppose Assumption 2.1 holds. We have*:

$$\|y_i^*(x_1) - y_i^*(x_2)\| \le \kappa \|x_1 - x_2\|, \quad \forall i \in \{1,2,\dots,n\}.$$

These lemmas reveal some nice properties of functions in bilevel optimization under Assumption 2.1. We will make use of these lemmas in our theoretical analysis.

Lemma 8 *Suppose Assumption 2.1 holds. If the iterates satisfy*:

$$\bar{x}_{k+1} = \bar{x}_k - \eta_x \overline{\partial \Phi(X_k)}, \quad \text{where } 0 < \eta_x \le \frac{1}{L_\Phi},$$

then we have the following inequality holds:

$$\begin{aligned}\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla\Phi(\bar{x}_k)\|^2 \le{}& \frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) \\ &+ \frac{1}{K+1}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2.\end{aligned} \tag{12}$$

Proof Since $\Phi(x)$ is L_Φ-smooth, we have:

$$\begin{aligned}&\Phi(\bar{x}_{k+1}) - \Phi(\bar{x}_k) \\ &\le \nabla\Phi(\bar{x}_k)^{\mathsf{T}}(-\eta_x\overline{\partial\Phi(X_k)}) + \frac{L_\Phi\eta_x^2}{2}\|\overline{\partial\Phi(X_k)}\|^2 \\ &= \frac{L_\Phi\eta_x^2}{2}\|\overline{\partial\Phi(X_k)}\|^2 - \eta_x\nabla\Phi(\bar{x}_k)^{\mathsf{T}}\overline{\partial\Phi(X_k)} \\ &= \frac{L_\Phi\eta_x^2}{2}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 + \left(\frac{L_\Phi\eta_x^2}{2} - \eta_x\right)\|\nabla\Phi(\bar{x}_k)\|^2 \\ &\quad + (L_\Phi\eta_x^2 - \eta_x)\nabla\Phi(\bar{x}_k)^{\mathsf{T}}(\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)) \\ &\le \frac{L_\Phi\eta_x^2}{2}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 + \left(\frac{L_\Phi\eta_x^2}{2} - \eta_x\right)\|\nabla\Phi(\bar{x}_k)\|^2 \\ &\quad + (\eta_x - L_\Phi\eta_x^2)\left(\frac{1}{2}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 + \frac{1}{2}\|\nabla\Phi(\bar{x}_k)\|^2\right) \\ &= \frac{\eta_x}{2}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 - \frac{\eta_x}{2}\|\nabla\Phi(\bar{x}_k)\|^2,\end{aligned}$$

where the second inequality is due to Young's inequality and $\eta_x \leq \frac{1}{L_\Phi}$. Therefore, we have:

$$\|\nabla\Phi(\bar{x}_k)\|^2 \leq \frac{2}{\eta_x}(\Phi(\bar{x}_k) - \Phi(\bar{x}_{k+1})) + \|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2. \tag{13}$$

Summing (13) over $k = 0, \ldots, K$, yields:

$$\sum_{k=0}^{K} \|\nabla\Phi(\bar{x}_k)\|^2 \leq \frac{2}{\eta_x}(\Phi(\bar{x}_0) - \Phi(\bar{x}_{k+1})) + \sum_{k=0}^{K} \|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2,$$

which completes the proof. □

We have the following lemma which provides an upper bound for E_K:

Lemma 9 *In each iteration, if we have* $\bar{x}_{k+1} = \bar{x}_k - \eta_x\overline{\partial\Phi(X_k)}$, *then the following inequality holds*:

$$E_K \leq 8S_K + 4n\eta_x^2 \sum_{j=0}^{K-1} \|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 T_{K-1}.$$

Proof By the definition of E_K, we have:

$$\begin{aligned}
E_K &= \sum_{j=1}^{K}\sum_{i=1}^{n} \|x_{i,j} - x_{i,j-1}\|^2 = \sum_{j=1}^{K}\sum_{i=1}^{n} \|x_{i,j} - \bar{x}_j + \bar{x}_j - \bar{x}_{j-1} + \bar{x}_{j-1} - x_{i,j-1}\|^2 \\
&= \sum_{j=1}^{K}\sum_{i=1}^{n} \|q_{i,j} - \eta_x(\overline{\partial\Phi(X_{j-1})} - \nabla\Phi(\bar{x}_{j-1})) - \eta_x\nabla\Phi(\bar{x}_{j-1}) - q_{i,j-1}\|^2 \\
&\leq 4\sum_{j=1}^{K}\sum_{i=1}^{n}(\|q_{i,j}\|^2 + \eta_x^2\|\overline{\partial\Phi(X_{j-1})} - \nabla\Phi(\bar{x}_{j-1}))\|^2 \\
&\quad + \eta_x^2\|\nabla\Phi(\bar{x}_{j-1})\|^2 + \|q_{i,j-1}\|^2) \\
&\leq 4\sum_{j=1}^{K}(\|Q_j\|^2 + \|Q_{j-1}\|^2 + n\eta_x^2\|\overline{\partial\Phi(X_{j-1})} - \nabla\Phi(\bar{x}_{j-1})\|^2 + n\eta_x^2\|\nabla\Phi(\bar{x}_{j-1})\|^2) \\
&\leq 8S_K + 4n\eta_x^2\sum_{j=0}^{K-1}(\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + \|\nabla\Phi(\bar{x}_j)\|^2) \\
&= 8S_K + 4n\eta_x^2\sum_{j=0}^{K-1} \|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 T_{K-1},
\end{aligned}$$

where the second inequality is by the definition of Q_j, the third inequality is by the definition of S_K and $Q_0 = 0$, the last equality is by the definition of T_{K-1}. □

Next we give bounds for A_K and B_K.

Lemma 10 *Suppose Assumptions 2.1 and 2.3 hold. If η_y, T and N in Algorithm 3 and 4 satisfy*:

$$0 < \eta_y < \frac{2}{\mu + L}, \quad \delta_y^T < \frac{1}{3}, \quad \delta_\kappa^N < \frac{1}{8\kappa}, \tag{14}$$

then the following inequalities hold:

$$A_K \le 3\delta_y^T(c_1 + 2\kappa^2 E_K), \quad B_K \le 2c_2 + 2d_1 A_{K-1} + 2d_2 E_K,$$

where the constants are defined as follows:

$$\begin{aligned} c_1 &= \sum_{i=1}^n \|y_{i,0}^{(0)} - y_i^*(x_{i,0})\|^2, \ c_2 = \sum_{i=1}^n \|v_{i,0}^* - v_{i,0}^{(0)}\|^2, \\ d_1 &= 4(1+\sqrt{\kappa})^2\left(\kappa + \frac{L_{g,2}L_{f,0}}{\mu^2}\right)^2 = \Theta(\kappa^3), \\ d_2 &= 2\left(\kappa^2 + \frac{2L_{f,0}\kappa}{\mu} + \frac{2L_{f,0}\kappa^2}{\mu}\right)^2 = \Theta(\kappa^4). \end{aligned} \tag{15}$$

Proof For each term in A_K we have

$$\begin{aligned} \|y_{i,j}^{(T)} - y_i^*(x_{i,j})\|^2 &= \|y_{i,j}^{(T-1)} - \eta_y \nabla_y g(x_{i,j}, y_{i,j}^{(T-1)}) - y_i^*(x_{i,j})\|^2 \\ &\le (1 - \eta_y\mu)^2 \|y_{i,j}^{(T-1)} - y_i^*(x_{i,j})\|^2 \le \delta_y^T \|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2, \end{aligned} \tag{16}$$

where the first inequality uses Lemma 3. We further have:

$$\begin{aligned} \|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2 &= \|y_{i,j-1}^{(T)} - y_i^*(x_{i,j-1}) + y_i^*(x_{i,j-1}) - y_i^*(x_{i,j})\|^2 \\ &\le 2(\|y_{i,j-1}^{(T)} - y_i^*(x_{i,j-1})\|^2 + \|y_i^*(x_{i,j-1}) - y_i^*(x_{i,j})\|^2) \\ &\le 2\delta_y^T \|y_{i,j-1}^{(0)} - y_i^*(x_{i,j-1})\|^2 + 2\kappa^2 \|x_{i,j-1} - x_{i,j}\|^2 \\ &< \frac{2}{3}\|y_{i,j-1}^{(0)} - y_i^*(x_{i,j-1})\|^2 + 2\kappa^2 \|x_{i,j-1} - x_{i,j}\|^2, \end{aligned}$$

where the second inequality is by (16) and Lemma 7, and the last inequality is by the condition (14). Taking summation on both sides, we get

$$
\begin{aligned}
&\sum_{j=1}^{K}\sum_{i=1}^{n}\|y_{i,j}^{(0)}-y_i^*(x_{i,j})\|^2\\
&\quad\le\frac{2}{3}\sum_{j=1}^{K}\sum_{i=1}^{n}\|y_{i,j-1}^{(0)}-y_i^*(x_{i,j-1})\|^2+2\kappa^2\sum_{j=1}^{K}\sum_{i=1}^{n}\|x_{i,j}-x_{i,j-1}\|^2\\
&\quad\le\frac{2}{3}\sum_{j=0}^{K}\sum_{i=1}^{n}\|y_{i,j}^{(0)}-y_i^*(x_{i,j})\|^2+2\kappa^2E_K\\
&\quad\le\frac{2}{3}c_1+\frac{2}{3}\sum_{j=1}^{K}\sum_{i=1}^{n}\|y_{i,j}^{(0)}-y_i^*(x_{i,j})\|^2+2\kappa^2E_K,
\end{aligned}
$$

which directly implies:

$$
\sum_{j=1}^{K}\sum_{i=1}^{n}\|y_{i,j}^{(0)}-y_i^*(x_{i,j})\|^2\le 2c_1+6\kappa^2E_K. \tag{17}
$$

Combining (16) and (17) leads to:

$$
\begin{aligned}
A_K=\sum_{j=0}^{K}\sum_{i=1}^{n}\|y_{i,j}^{T}-y_i^*(x_{i,j})\|^2&\le\delta_y^T\sum_{j=0}^{K}\sum_{i=1}^{n}\|y_{i,j}^{(0)}-y_i^*(x_{i,j})\|^2\\
&\le\delta_y^T(c_1+2c_1+6\kappa^2E_K)=3\delta_y^T(c_1+2\kappa^2E_K).
\end{aligned}
$$

We then consider the bound for B_K. Recall that:

$$
v_{i,k}^*=\left(\nabla_y^2g_i(x_{i,k},y_i^*(x_{i,k}))\right)^{-1}\nabla_yf_i(x_{i,k},y_i^*(x_{i,k})),
$$

which is the solution of the linear system $\nabla_y^2g_i(x_{i,k},y_i^*(x_{i,k}))v=\nabla_yf_i(x_{i,k},y_i^*(x_{i,k}))$ in the AID-based approach in Algorithm 2. Note that $v_{i,k}^*$ is a function of $x_{i,k}$, and it is $(\kappa^2+\frac{2L_{f,0}L}{\mu^2}+\frac{2L_{f,0}L\kappa}{\mu^2})$-Lipschitz continuous with respect to $x_{i,k}$ [13]. For each term in B_K, we have:

$$
\begin{aligned}
&\|v_{i,j}^*-v_{i,j}^{(0)}\|^2\\
&\quad\le 2(\|v_{i,j-1}^*-v_{i,j-1}^{(N)}\|^2+\|v_{i,j}^*-v_{i,j-1}^*\|^2)\\
&\quad\le 4(1+\sqrt{\kappa})^2\left(\kappa+\frac{L_{g,2}L_{f,0}}{\mu^2}\right)^2\|y_{i,j-1}^{(T)}-y_i^*(x_{i,j-1})\|^2\\
&\qquad+4\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|v_{i,j-1}^*-v_{i,j-1}^{(0)}\|^2+2\left(\kappa^2+\frac{2L_{f,0}L(1+\kappa)}{\mu^2}\right)^2\|x_{i,j}-x_{i,j-1}\|^2,
\end{aligned}
$$

where the second inequality follows [13, Lemma 4]. Taking summation over i,j, we get

$$\sum_{j=1}^{K}\sum_{i=1}^{n}\|v_{i,j}^{*}-v_{i,j}^{(0)}\|^2 \le d_1A_{K-1}+4\kappa\delta_{\kappa}^{N}B_{K-1}+d_2E_K \le d_1A_{K-1}+\frac{1}{2}B_K+d_2E_K, \tag{18}$$

where the last inequality holds since we pick N such that $4\kappa\delta_{\kappa}^{N}<\frac{1}{2}$. Therefore, we can get:

$$B_K \le c_2+d_1A_{K-1}+\frac{1}{2}B_K+d_2E_K \quad \Rightarrow \quad B_K \le 2c_2+2d_1A_{K-1}+2d_2E_K,$$

which completes the proof. □

The following lemmas give bounds on $\sum\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2$ in (13).We first consider the case when the Assumption 2.3 holds. In this case, the outer loop computes the hypergradient via AID based approach. Therefore, we borrow [13, Lemma 3] and restate it as follows.

Lemma 11 [13, *Lemma* 3] *Suppose Assumptions* 2.1 *and* 2.3 *hold*, *then we have*:

$$\begin{aligned}&\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)})-\nabla f_i(x_{i,j},y_i^{*}(x_{i,j}))\|^2\\ &\qquad\le \Gamma\|y_i^{*}(x_{i,j})-y_{i,j}^{(T)}\|^2+6L^2\kappa\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2N}\|v_{i,j}^{*}-v_{i,j}^{(0)}\|^2\end{aligned}$$

where the constant Γ *is*

$$\Gamma=3L^2+\frac{3L_{g,2}^2L_{f,0}}{\mu^2}+6L^2(1+\sqrt{\kappa})^2\left(\kappa+\frac{L_{g,2}L_{f,0}}{\mu^2}\right)^2=\Theta(\kappa^3).$$

Next, we bound $\sum\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2$ under Assumption 2.3.

Lemma 12 *Suppose Assumptions* 2.1 *and* 2.3 *hold*. *We have*:

$$\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2 \le \frac{2L_{\Phi}^2}{n}S_K+\frac{2\Gamma}{n}A_K+\frac{12L^2\kappa}{n}\delta_{\kappa}^{N}B_K. \tag{19}$$

Proof Under Assumption 2.3 we know $g_i=g$, and thus from (5) and (6) we have

$$\nabla\Phi_i(\bar{x}_k)=\nabla f_i(\bar{x}_k,y^{*}(\bar{x}_k)).$$

Therefore, we have

$$
\begin{aligned}
\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 &= \frac{1}{n^2}\left\|\sum_{i=1}^{n}\left(\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla f_i(\bar{x}_k, y^*(\bar{x}_k))\right)\right\|^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\|\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla f_i(\bar{x}_k, y^*(\bar{x}_k))\|^2 \\
&\le \frac{2}{n}\sum_{i=1}^{n}(\|\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla f_i(x_{i,k}, y_i^*(x_{i,k}))\|^2 \\
&\quad + \|\nabla f_i(x_{i,k}, y_i^*(x_{i,k})) - \nabla f_i(\bar{x}_k, y^*(\bar{x}_k))\|^2) \\
&\le \frac{2}{n}\sum_{i=1}^{n}(\Gamma\|y_i^*(x_{i,k}) - y_{i,k}^{(T)}\|^2 + 6L^2\kappa\delta_\kappa^N\|v_{i,k}^* - v_{i,k}^{(0)}\|^2 + L_\Phi^2\|x_{i,k} - \bar{x}_k\|^2) \\
&\le \frac{2\Gamma}{n}\sum_{i=1}^{n}\|y_i^*(x_{i,k}) - y_{i,k}^{(T)}\|^2 + \frac{12L^2\kappa}{n}\delta_\kappa^N\sum_{i=1}^{n}\|v_{i,k}^* - v_{i,k}^{(0)}\|^2 + \frac{2L_\Phi^2}{n}\|Q_k\|^2,
\end{aligned}
$$

where the first inequality follows from the convexity of $\|\cdot\|^2$, the third inequality follows from Lemma 11 and Assumption 2.3, the last inequality is by Lemma 5:

$$
\|\nabla f_i(x_{i,k}, y^*(x_{i,k})) - \nabla f_i(\bar{x}_k, y^*(\bar{x}_k))\|^2 = \|\nabla\Phi_i(x_{i,k}) - \nabla\Phi_i(\bar{x}_k)\|^2 \le L_\Phi^2\|q_{i,k}\|^2.
$$

Taking summation on both sides, we get:

$$
\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 \le \frac{2L_\Phi^2}{n}S_K + \frac{2\Gamma}{n}A_K + \frac{12L^2\kappa}{n}\delta_\kappa^N B_K.
$$

□

We now consider the case when Assumption 2.3 does not hold. In this case, our target in the lower level problem is

$$
y^*(\bar{x}_k) = \arg\min_y \frac{1}{n}\sum_{i=1}^{n} g_i(\bar{x}_k, y). \tag{20}
$$

However, the update in our decentralized algorithm (e.g. line 8 of Algorithm 3) aims at solving

$$
\tilde{y}_k^* := \arg\min_y \frac{1}{n}\sum_{i=1}^{n} g_i(x_{i,k}, y), \tag{21}
$$

which is completely different from our target (20). To resolve this problem, we introduce the following lemma to characterize the difference:

Lemma 13 *The following inequality holds*:

$$
\|\tilde{y}_k^* - y^*(\bar{x}_k)\| \le \frac{\kappa}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\| \le \frac{\kappa}{\sqrt{n}}\|Q_k\|.
$$

Proof By optimality conditions of (20) and (21), we have:

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(x_{i,k},\tilde{y}_k^*)=0,\quad \frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(\bar{x}_k,y^*(\bar{x}_k))=0.$$

Combining with the strongly convexity and the smoothness of g_i yields:

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(\bar{x}_k,\tilde{y}_k^*)\right\|\\&\quad=\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(\bar{x}_k,\tilde{y}_k^*)-\frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(\bar{x}_k,y^*(\bar{x}_k))\right\|\geq\mu\|\tilde{y}_k^*-y^*(\bar{x}_k)\|,\\&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(\bar{x}_k,\tilde{y}_k^*)\right\|\\&\quad=\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(\bar{x}_k,\tilde{y}_k^*)-\frac{1}{n}\sum_{i=1}^{n}\nabla_y g_i(x_{i,k},\tilde{y}_k^*)\right\|\leq\frac{L}{n}\sum_{i=1}^{n}\|x_{i,k}-\bar{x}_k\|.\end{aligned}$$

Therefore, we obtain the following inequality:

$$\|\tilde{y}_k^*-y^*(\bar{x}_k)\|\leq\frac{\kappa}{n}\sum_{i=1}^{n}\|x_{i,k}-\bar{x}_k\|=\frac{\kappa}{n}\sum_{i=1}^{n}\|q_{i,k}\|\leq\frac{\kappa}{\sqrt{n}}\|Q_k\|,$$

where the last inequality is by Cauchy–Schwarz inequality. □

Notice that in the inner loop of Algorithms 3, 4 and 5, i.e., Lines 4–11 of Algorithms 3 and 4, and Lines 4–10 of Algorithm 5, $y_{i,k}^{(T)}$ converges to $\tilde{y}_k^*$ and the rates are characterized by [34, 36, 46, 55] (e.g., Corollary 4.7. in [55], Theorem 10 in [35] and Theorem 1 in [46]). We include all the convergence rates here.

Lemma 14 *Suppose Assumption 2.3 does not hold. We have*:

- In Algorithm 3 and 4 there exists a constant η_y such that

$$\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2\leq C_1\alpha_1^T.$$

- In Algorithm 5 there exists $\eta_y^{(t)}=\mathcal{O}(\frac{1}{t})$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2\right]\leq\frac{C_2}{T}.$$

Here C_1, C_2 are positive constants and $\alpha_1\in(0,1)$.

Besides, the JHIP oracle (Algorithm 1) also performs standard decentralized optimization with gradient tracking in deterministic case (Algorithms 3, 4) and stochastic case (Algorithm 5). We have:

Lemma 15 *In Algorithm 1, we have*:

- For deterministic case, there exists a constant γ such that if $\gamma_t \equiv \gamma$ then

$$\|Z_i^{(t)} - Z^*\|^2 \le C_3 \alpha_2^t. \quad \text{(See [34])}.$$

- For stochastic case and there exists a diminishing stepsize sequence $\gamma_t = \mathcal{O}(\frac{1}{t})$, such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i^{(t)} - Z^*\|^2\right] \le \frac{C_4}{t}. \quad \text{(See [36])}.$$

Here C_3, C_4 are positive constant, and $\alpha_2 \in (0,1)$. Here the optimal solution is denoted by $(Z^*)^\top = \left(\sum_{i=1}^n J_i\right)\left(\sum_{i=1}^n H_i\right)^{-1}$.

For simplicity we define:

$$C = \max\left(C_1, C_2, C_3, C_4\right), \quad \alpha = \max\left(\alpha_1, \alpha_2\right).$$

Since the objective functions mentioned in Lemma 14 (the lower level function g) and 15 (the objective in (9)) are strongly convex, we know C and α only depend on L, μ, ρ and the stepsize (only when it is a constant). For example α_2 in Lemma 15 only depends on the spectral radius of H_i, smallest eigenvalue of H_i, ρ and γ.

For heterogeneous data (i.e., no Assumption 2.3) on g we have a different error estimation. We first notice that for each JHIP oracle, the following lemma holds:

Lemma 16 *Suppose Assumptions 2.1 holds. In Algorithm 3 and 4 we have*:

$$\begin{aligned}\left\|\left(Z_k^*\right)^\top - \nabla_{xy} g(\bar{x}_k, \tilde{y}_k^*)\left(\nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)\right)^{-1}\right\|_2^2 \\ \le \frac{2L_{g,2}^2(1+\kappa^2)}{\mu^2}\left(\frac{1}{n}\|Q_k\|^2 + \frac{1}{n}\sum_{j=1}^{n}\|y_{j,k}^{(T)} - \tilde{y}_k^*\|^2\right),\end{aligned}$$

where Z_k^* denotes the optimal solution of Algorithm 1 *in iteration k*:

$$\left(Z_k^*\right)^\top = \left(\frac{1}{n}\sum_{j=1}^{n}\nabla_{xy} g_j(x_{j,k}, y_{j,k}^{(T)})\right)\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_y^2 g_j(x_{j,k}, y_{j,k}^{(T)})\right)^{-1}.$$

Proof Notice that we have

$$
\begin{aligned}
&\|\left(Z_k^*\right)^{\top} - \nabla_{xy} g(\bar{x}_k, \tilde{y}_k^*)\left[\nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)\right]^{-1}\|_2^2 \\
&\quad \le 2\left\|\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_{xy} g_j(x_{j,k}, y_{j,k}^{(T)}) - \nabla_{xy} g(\bar{x}_k, \tilde{y}_k^*)\right)\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_y^2 g_j(x_{j,k}, y_{j,k}^{(T)})\right)^{-1}\right\|_2^2 \\
&\qquad + 2\left\|\nabla_{xy} g(\bar{x}_k, \tilde{y}_k^*)\left[\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_y^2 g_j(x_{j,k}, y_{j,k}^{(T)})\right)^{-1} - \left(\nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)\right)^{-1}\right]\right\|_2^2 \\
&\quad \le \frac{2L_{g,2}^2}{n\mu^2}\sum_{j=1}^{n}(\|x_{j,k} - \bar{x}_k\|^2 + \|y_{j,k}^{(T)} - \tilde{y}_k^*\|^2) \\
&\qquad + \frac{2L_{g,1}^2 L_{g,2}^2}{n\mu^4}\sum_{j=1}^{n}(\|x_{j,k} - \bar{x}_k\|^2 + \|y_{j,k}^{(T)} - \tilde{y}_k^*\|^2) \\
&\quad \le \frac{2L_{g,2}^2(1+\kappa^2)}{\mu^2}\left(\frac{1}{n}\|Q_k\|^2 + \frac{1}{n}\sum_{j=1}^{n}\|y_{j,k}^{(T)} - \tilde{y}_k^*\|^2\right)
\end{aligned}
$$

where the second inequality holds due to Assumption 2.1 and the following inequality:

$$
\begin{aligned}
&\left\|\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_y^2 g_j(x_{j,k}, y_{j,k}^{(T)})\right)^{-1} - \left(\nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)\right)^{-1}\right\|_2^2 \\
&\quad = \left\|\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_y^2 g_j(x_{j,k}, y_{j,k}^{(T)})\right)^{-1} \cdot \right. \\
&\qquad \left.\left(\nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*) - \frac{1}{n}\sum_{j=1}^{n}\nabla_y^2 g_j(x_{j,k}, y_{j,k}^{(T)})\right)\left(\nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)\right)^{-1}\right\|_2^2 \\
&\quad \le \frac{L_{g,2}^2}{n\mu^4}\sum_{j=1}^{n}(\|x_{j,k} - \bar{x}_k\|^2 + \|y_{j,k}^{(T)} - \tilde{y}_k^*\|^2).
\end{aligned}
$$

□

Lemma 17 *Suppose Assumption 2.1 holds. In Algorithm 3 and 4 we have*:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \bar{\nabla} f_i(x_{i,k}, \tilde{y}_k^*)\|^2 \\
&\quad \le \frac{18L_{f,0}^2 L_{g,2}^2(1+\kappa^2)}{n\mu^2}\|Q_k\|^2 + \frac{6L_{f,0}^2}{n}\sum_{i=1}^{n}\|Z_{i,k}^{(N)} - Z_k^*\|^2 \\
&\qquad + \left(6 + 6L^2\kappa^2 + \frac{12L_{f,0}^2 L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)} - \tilde{y}_k^*\|^2\right).
\end{aligned} \tag{22}
$$

Proof Note that

$$\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) = \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \left(Z_{i,k}^{(N)}\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}),$$
$$\bar{\nabla} f_i(x_{i,k}, \tilde{y}_k^*) = \nabla_x f_i(x_{i,k}, \tilde{y}_k^*) - \nabla_{xy} g(x_{i,k}, \tilde{y}_k^*) \nabla_y^2 g(x_{i,k}, \tilde{y}_k^*)^{-1} \nabla_y f_i(x_{i,k}, \tilde{y}_k^*).$$

Then we know

$$\begin{aligned}
&\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \bar{\nabla} f_i(x_{i,k}, \tilde{y}_k^*) \\
&\quad = \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla_x f_i(x_{i,k}, \tilde{y}_k^*) \\
&\qquad - \left(Z_{i,k}^{(N)}\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}) + \left(Z_k^*\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}) \\
&\qquad - \left(Z_k^*\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}) + \left(Z_k^*\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, \tilde{y}_k^*) \\
&\qquad - \left(Z_k^*\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, \tilde{y}_k^*) + \nabla_{xy} g(\bar{x}_k, \tilde{y}_k^*) \nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)^{-1} \nabla_y f_i(x_{i,k}, \tilde{y}_k^*) \\
&\qquad - \nabla_{xy} g(\bar{x}_k, \tilde{y}_k^*) \nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)^{-1} \nabla_y f_i(x_{i,k}, \tilde{y}_k^*) \\
&\qquad + \nabla_{xy} g(x_{i,k}, \tilde{y}_k^*) \nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)^{-1} \nabla_y f_i(x_{i,k}, \tilde{y}_k^*) \\
&\qquad - \nabla_{xy} g(x_{i,k}, \tilde{y}_k^*) \nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)^{-1} \nabla_y f_i(x_{i,k}, \tilde{y}_k^*) \\
&\qquad + \nabla_{xy} g(x_{i,k}, \tilde{y}_k^*) \nabla_y^2 g(x_{i,k}, \tilde{y}_k^*)^{-1} \nabla_y f(x_{i,k}, \tilde{y}_k^*),
\end{aligned}$$

which gives

$$\begin{aligned}
&\|\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \bar{\nabla} f_i(x_{i,k}, \tilde{y}_k^*)\|^2 \\
&\quad \le 6\Big(\|y_{i,k}^{(T)} - \tilde{y}_k^*\|^2 + L_{f,0}^2 \|Z_{i,k}^{(N)} - Z_k^*\|^2 + L^2 \|\left(Z_k^*\right)^{\mathsf{T}}\|_2^2 \|y_{i,k}^{(T)} - \tilde{y}_k^*\|^2 \\
&\qquad + L_{f,0}^2 \|\left(Z_k^*\right)^{\mathsf{T}} - \nabla_{xy} g(\bar{x}_k, \tilde{y}_k^*) \nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)^{-1}\|_2^2 + \frac{L_{g,2}^2 L_{f,0}^2}{\mu^2} \|x_{i,k} - \bar{x}_k\|^2 \\
&\qquad + L^2 L_{f,0}^2 \|\nabla_y^2 g(\bar{x}_k, \tilde{y}_k^*)^{-1} - \nabla_y^2 g(x_{i,k}, \tilde{y}_k^*)^{-1}\|_2^2 \Big) \\
&\quad \le 6\Bigg(\|y_{i,k}^{(T)} - \tilde{y}_k^*\|^2 + L_{f,0}^2 \|Z_{i,k}^{(N)} - Z_k^*\|^2 + \frac{L^4}{\mu^2} \|y_{i,k}^{(T)} - \tilde{y}_k^*\|^2 \\
&\qquad + \frac{2L_{f,0}^2 L_{g,2}^2 (1+\kappa^2)}{\mu^2} \left(\frac{1}{n} \|Q_k\|^2 + \frac{1}{n} \sum_{j=1}^n \|y_{j,k}^{(T)} - \tilde{y}_k^*\|^2 \right) \\
&\qquad + \frac{L_{g,2}^2 L_{f,0}^2}{\mu^2} \|x_{i,k} - \bar{x}_k\|^2 + \frac{L^2 L_{f,0}^2 L_{g,2}^2}{\mu^4} \|x_{i,k} - \bar{x}_k\|^2 \Bigg).
\end{aligned}$$

The second inequality uses Lemma 16, Assumption 2.1. Taking summation on both sides and using Lemma 15, we know

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\hat{\nabla}f_i(x_{i,k},y_{i,k}^{(T)})-\bar{\nabla}f_i(x_{i,k},\tilde{y}_k^*)\|^2\\
&\quad\le\frac{18L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{n\mu^2}\|Q_k\|^2+\frac{6L_{f,0}^2}{n}\sum_{i=1}^{n}\|Z_{i,k}^{(N)}-Z_k^*\|^2\\
&\qquad+\left(6+6L^2\kappa^2+\frac{12L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2\right).
\end{aligned}$$

□

Lemma 18 *Suppose Assumption 2.3 does not hold, then in Algorithms 3 and 4 we have*:

$$\begin{aligned}
\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2\le\frac{(1+\kappa^2)}{n}\cdot\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2}+2L_f^2\right)\|Q_k\|^2\\
+12C\left[\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T+L_{f,0}^2\alpha^N\right].
\end{aligned}\tag{23}$$

Proof We have

$$\begin{aligned}
&\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2=\frac{1}{n^2}\left\|\sum_{i=1}^{n}\left(\hat{\nabla}f_i(x_{i,k},y_{i,k}^{(T)})-\nabla f_i(\bar{x}_k,y^*(\bar{x}_k))\right)\right\|^2\\
&\le\frac{1}{n}\sum_{i=1}^{n}\|\hat{\nabla}f_i(x_{i,k},y_{i,k}^{(T)})-\nabla f_i(\bar{x}_k,y^*(\bar{x}_k))\|^2\\
&\le\frac{2}{n}\sum_{i=1}^{n}(\|\hat{\nabla}f_i(x_{i,k},y_{i,k}^{(T)})-\bar{\nabla}f_i(x_{i,k},\tilde{y}_k^*))\|^2+\|\bar{\nabla}f_i(x_{i,k},\tilde{y}_k^*)-\nabla f_i(\bar{x}_k,y^*(\bar{x}_k))\|^2)\\
&\le\frac{36L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{n\mu^2}\|Q_k\|^2\\
&\quad+12\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2\\
&\quad+\frac{12L_{f,0}^2}{n}\sum_{i=1}^{n}\|Z_{i,k}^{(N)}-Z_k^*\|^2+\frac{2}{n}\sum_{i=1}^{n}(L_f^2\|x_{i,k}-\bar{x}_k\|^2+L_f^2\|\tilde{y}_k^*-y^*(\bar{x}_k)\|^2)\\
&\le12\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2\\
&\quad+\frac{12L_{f,0}^2}{n}\sum_{i=1}^{n}\|Z_{i,k}^{(N)}-Z_k^*\|^2+\frac{(1+\kappa^2)}{n}\cdot\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2}+2L_f^2\right)\|Q_k\|^2,
\end{aligned}\tag{24}$$

where the third inequality is due to Lemma 17 and Lemma 6, and the fourth inequality is by Lemma 13. Notice that $\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2$ in the first term denotes the error of the inner loop iterates. In both DBO (Algorithm 3) and DBOGT (Algorithm 4), the inner loop performs a decentralized gradient descent with gradient tracking. By Lemmas 14 and 15, we have the error bounds $\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2 \le C\alpha^T$ and $\frac{1}{n}\sum_{i=1}^{n}\|Z_{i,k}^{(N)}-Z_k^*\|^2 \le C\alpha^N$, which complete the proof. □

Proof of the DBO convergence

In this section we will prove the following convergence result of the DBO algorithm:

Theorem 19 *In Algorithm 3, suppose Assumptions 2.1 and 2.2 hold. If Assumption 2.3 holds, then by setting* $0<\eta_x\le\frac{1-\rho}{130L_\Phi}$, $0<\eta_y<\frac{2}{\mu+L}$, $T=\Theta(\kappa\log\kappa)$, $N=\Theta(\sqrt{\kappa}\log\kappa)$, *we have*:

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2 \le \frac{4}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x\Phi(x))+\eta_x^2\cdot\frac{1272L_\Phi^2L_{f,0}^2(1+\kappa)^2}{(1-\rho)^2}+\frac{C_1}{K+1}.$$

If Assumption 2.3 does not hold, then by setting $0<\eta_x\le\frac{1}{L_\Phi}$, $\eta_y^{(t)}=\mathcal{O}(\frac{1}{t})$, *we have*:

$$\begin{aligned}\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2 \le{}& \frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x\Phi(x))\\ &+\eta_x^2\left(\frac{18L_{f,0}^2L_{g,2}^2}{\mu^2}+L_f^2\right)\frac{4(1+\kappa^2)((1+\kappa)^2+C\alpha^N)L_{f,0}^2}{(1-\rho)^2}+\tilde{C}_1,\end{aligned}$$

where $C_1=\Theta(1)$, $C=\Theta(1)$ *and* $\tilde{C}_1=\mathcal{O}(\alpha^T+\alpha^N)$.

We first consider bounding the consensus error estimation for DBO:

Lemma 20 *In Algorithm 3, we have*

$$S_K:=\sum_{k=1}^{K}\|Q_k\|^2<\frac{\eta_x^2}{(1-\rho)^2}\sum_{j=0}^{K-1}\sum_{i=1}^{n}\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)})\|^2.$$

Proof Note that the x update can be written as

$$X_k=X_{k-1}W-\eta_x\partial\Phi(X_{k-1}),$$

which indicates

$$\bar{x}_k=\bar{x}_{k-1}-\eta_x\overline{\partial\Phi(X_{k-1})}.$$

By definition of $q_{i,k}$, we have

$$
\begin{aligned}
q_{i,k+1} &= \sum_{j=1}^{n} w_{ij} x_{j,k} - \eta_x \hat{\nabla} f(x_{i,k}, y_{i,k}^{(T)}) - (\bar{x}_k - \eta_x \overline{\partial \Phi(X_k)}) \\
&= \sum_{j=1}^{n} w_{ij}(x_{j,k} - \bar{x}_k) - \eta_x (\hat{\nabla} f(x_{i,k}, y_{i,k}^{(T)}) - \overline{\partial \Phi(X_k)}) \\
&= Q_k W e_i - \eta_x \partial \Phi(X_k) \left(e_i - \frac{\mathbf{1}_n}{n} \right),
\end{aligned}
$$

where the last equality uses the fact that W is symmetric. Therefore, for Q_{k+1} we have

$$
\begin{aligned}
Q_{k+1} &= Q_k W - \eta_x \partial \Phi(X_k) \left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \\
&= \left(Q_{k-1} W - \eta_x \partial \Phi(X_{k-1}) \left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \right) W - \eta_x \partial \Phi(X_k) \left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \\
&= Q_0 W^{k+1} - \eta_x \sum_{i=0}^{k} \left(\partial \Phi(X_i) \left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) W^{k-i} \right) \\
&= -\eta_x \sum_{i=0}^{k} \partial \Phi(X_i) \left(W^{k-i} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right),
\end{aligned}
$$

where the last equality is obtained by $Q_0 = 0$ and $\mathbf{1}_n \mathbf{1}_n^\top W = \mathbf{1}_n \mathbf{1}_n^\top$. By Cauchy–Schwarz inequality, we have the following estimate

$$
\begin{aligned}
\|Q_{k+1}\|^2 &= \eta_x^2 \| \sum_{i=0}^{k} \partial \Phi(X_i) \left(W^{k-i} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \|^2 \\
&\le \eta_x^2 \left(\sum_{i=0}^{k} \| \partial \Phi(X_i) \left(W^{k-i} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \| \right)^2 \\
&\le \eta_x^2 \left(\sum_{i=0}^{k} \| \partial \Phi(X_i) \| \| \left(W^{k-i} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \|_2 \right)^2 \\
&\le \eta_x^2 \left(\sum_{i=0}^{k} \rho^{k-i} \| \partial \Phi(X_i) \|^2 \right) \left(\sum_{i=0}^{k} \frac{1}{\rho^{k-i}} \left\| W^{k-i} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right\|_2^2 \right) \\
&\le \eta_x^2 \left(\sum_{i=0s}^{k} \rho^{k-i} \| \partial \Phi(X_i) \|^2 \right) \left(\sum_{i=0}^{k} \rho^{k-i} \right) < \frac{\eta_x^2}{1-\rho} \left(\sum_{i=0}^{k} \rho^{k-i} \| \partial \Phi(X_i) \|^2 \right) \\
&= \frac{\eta_x^2}{1-\rho} \left(\sum_{j=0}^{k} \rho^{k-j} \sum_{i=1}^{n} \| \hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)}) \|^2 \right) = \frac{\eta_x^2}{1-\rho} \sum_{i=1}^{n} \sum_{j=0}^{k} \rho^{k-j} \| \hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)}) \|^2.
\end{aligned}
$$

where the fourth inequality is obtained by Lemma 4. Summing the above inequality yields

$$\begin{aligned} S_K &= \sum_{k=0}^{K-1} \|Q_{k+1}\|^2 < \frac{\eta_x^2}{1-\rho} \sum_{k=0}^{K-1} \sum_{i=1}^{n} \sum_{j=0}^{k} \rho^{k-j} \|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 \\ &= \frac{\eta_x^2}{1-\rho} \sum_{j=0}^{K-1} \sum_{i=1}^{n} \sum_{k=j}^{K-1} \rho^{k-j} \|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 \\ &< \frac{\eta_x^2}{(1-\rho)^2} \sum_{j=0}^{K-1} \sum_{i=1}^{n} \|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2, \end{aligned} \tag{25}$$

where the second equality holds since we can change the order of summation. □

Case 1: Assumption 2.3 holds

We first consider the case when Assumption 2.3 holds.

Lemma 21 *Suppose Assumptions 2.1 and 2.3 hold, then we have*:

$$\|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 \le 2(L^2 \kappa \delta_\kappa^N \|v_{i,j}^{(0)} - v_{i,j}^*\|^2 + (1+\kappa)^2 L_{f,0}^2).$$

Proof Notice that we have:

$$\begin{aligned} \|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 &\le 2\|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)}) - \bar{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 + 2\|\bar{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 \\ &\le 2\|\nabla_{xy} g_i(x_{i,j}, y_{i,j}^{(T)})(v_{i,j}^{(N)} - v_{i,j}^*)\|^2 \\ &\quad + 2\|\nabla_x f_i(x_{i,j}, y_{i,j}^{(T)}) - \nabla_{xy} g_i(x_{i,j}, y_{i,j}^{(T)}) \Big(\nabla_y^2 g_i(x_{i,j}, y_{i,j}^{(T)}) \Big)^{-1} \nabla_y f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 \\ &\le 2(L^2 \|v_{i,j}^{(N)} - v_{i,j}^*\|^2 + (L_{f,0} + \frac{L}{\mu} L_{f,0})^2) \le 2(L^2 \kappa \delta_\kappa^N \|v_{i,j}^{(0)} - v_{i,j}^*\|^2 + (1+\kappa)^2 L_{f,0}^2), \end{aligned}$$

where the second inequality is via the Assumption 2.1, and the last inequality is based on the convergence result of CG for the quadratic programming, e.g., eq. (17) in [24]. □

Next we obtain the upper bound for S_K.

Lemma 22 *Suppose Assumptions 2.1 and 2.3 hold, then we have*:

$$S_K < \frac{2\eta_x^2}{(1-\rho)^2} (L^2 \kappa \delta_\kappa^N B_{K-1} + nK(1+\kappa)^2 L_{f,0}^2).$$

Proof By Lemmas 20 and 21, we have:

$$\begin{aligned} S_K <& \frac{\eta_x^2}{(1-\rho)^2} \sum_{j=0}^{K-1} \sum_{i=1}^{n} \|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 \\ \leq& \frac{\eta_x^2}{(1-\rho)^2} \sum_{j=0}^{K-1} \sum_{i=1}^{n} 2(L^2 \kappa \delta_\kappa^N \|v_{i,j}^{(0)} - v_{i,j}^*\|^2 + (1+\kappa)^2 L_{f,0}^2) \\ =& \frac{2\eta_x^2}{(1-\rho)^2} (L^2 \kappa \delta_\kappa^N B_{K-1} + nK(1+\kappa)^2 L_{f,0}^2), \end{aligned}$$

which completes the proof. □

We are ready to prove the main results in Theorem 19. We first summarize main results in Lemmas 22, 10 and 9:

$$\begin{aligned} & S_K < \frac{2\eta_x^2}{(1-\rho)^2} (L^2 \kappa \delta_\kappa^N B_{K-1} + nK(1+\kappa)^2 L_{f,0}^2), \\ & A_K \leq 3\delta_y^T (c_1 + 2\kappa^2 E_K),\ B_K \leq 2c_2 + 2d_1 A_{K-1} + 2d_2 E_K, \\ & E_K \leq 8S_K + 4n\eta_x^2 \sum_{j=0}^{K-1} \|\overline{\partial \Phi(X_j)} - \nabla \Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 T_{K-1}. \end{aligned} \tag{26}$$

The next lemma proves the first part of Theorem 19.

Lemma 23 *Suppose the assumptions of Lemma 10 hold. Furthermore, if we set* $N = \Theta(\sqrt{\kappa} \log \kappa), T = \Theta(\kappa \log \kappa), \eta_x = \mathcal{O}(\kappa^{-3})$ *such that*:

$$\begin{aligned} \delta_\kappa^N &< \min\left(\frac{L_\Phi^2}{L^2 \kappa (4d_1 \kappa^2 + 2d_2)}, \kappa^{-6} \right) = \Theta(\kappa^{-6}), \\ \delta_y^T &< \min\left(\frac{L_\Phi^2}{12\Gamma \kappa^2}, \kappa^{-5}, \frac{1}{3} \right) = \Theta(\kappa^{-5}),\ \eta_x < \frac{1-\rho}{130 L_\Phi}, \end{aligned}$$

we have:

$$\frac{1}{K+1} \sum_{j=0}^{K} \|\nabla \Phi(\bar{x}_j)\|^2 \leq \frac{4}{\eta_x (K+1)} (\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \eta_x^2 \cdot \frac{1272 L_\Phi^2 L_{f,0}^2 (1+\kappa)^2}{(1-\rho)^2} + \frac{C_1}{K+1},$$

where the constant is given by:

$$\begin{aligned} C_1 &= 106 L_\Phi^2 \cdot \frac{6\eta_x^2}{(1-\rho)^2} L^2 \kappa \delta_\kappa^N (2c_2 + 2d_1 c_1) + \frac{18 L^2 \kappa \delta_\kappa^N (2c_2 + 2d_1 c_1) + 9\Gamma c_1 \delta_y^T}{n} \\ &= \Theta(\eta_x^2 \delta_\kappa^N \kappa^{12} + \kappa^5 \delta_y^T) = \Theta(1). \end{aligned}$$

Proof For B_K we know:

$$\begin{aligned} B_K &\le 2c_2 + 2d_1 A_K + 2d_2 E_K \le 2c_2 + \frac{2}{3} d_1 (3c_1 + 6\kappa^2 E_K) + 2d_2 E_K \\ &= 2c_2 + 2d_1 c_1 + (4d_1\kappa^2 + 2d_2) E_K. \end{aligned} \tag{27}$$

We first eliminate B_K in the upper bound of S_K. Pick N, T such that:

$$\delta_\kappa^N \cdot (4d_1\kappa^2 + 2d_2) \cdot L^2\kappa < L_\Phi^2 \quad \Rightarrow \quad \delta_\kappa^N < \frac{L_\Phi^2}{L^2\kappa(4d_1\kappa^2 + 2d_2)}. \tag{28}$$

Therefore, we have

$$\begin{aligned} S_K &\le \frac{2\eta_x^2}{(1-\rho)^2}(L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + L^2\kappa\delta_\kappa^N(4d_1\kappa^2 + 2d_2)E_K + nK(1+\kappa)^2 L_{f,0}^2) \\ &\le \frac{2\eta_x^2}{(1-\rho)^2}(L_\Phi^2 E_K + L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + nK(1+\kappa)^2 L_{f,0}^2), \end{aligned}$$

where in the first inequality we use (27) to eliminate B_K. Next we eliminate E_K in this bound. By the definition of η_x, we know:

$$\eta_x < \frac{(1-\rho)}{4\sqrt{2}L_\Phi} \quad \Rightarrow \quad \frac{16\eta_x^2 L_\Phi^2}{(1-\rho)^2} < \frac{1}{2},$$

which, together with (26), yields

$$\begin{aligned} S_K &\le \frac{2\eta_x^2}{(1-\rho)^2}(L_\Phi^2(8S_K + 4n\eta_x^2 \sum_{j=0}^{K-1} \|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 T_{K-1}) \\ &\quad + L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + nK(1+\kappa)^2 L_{f,0}^2) \\ &< \frac{1}{2}S_K + \frac{2\eta_x^2}{(1-\rho)^2}(4n\eta_x^2 L_\Phi^2 \sum_{j=0}^{K-1} \|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 L_\Phi^2 T_{K-1} \\ &\quad + L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + nK(1+\kappa)^2 L_{f,0}^2). \end{aligned} \tag{29}$$

The above inequality indicates

$$\begin{aligned} S_K \le\ & \frac{4\eta_x^2}{(1-\rho)^2}\left(4n\eta_x^2 L_\Phi^2 \sum_{j=0}^{K-1} \|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 L_\Phi^2 T_{K-1}\right) \\ &+ \frac{4\eta_x^2}{(1-\rho)^2}\left(L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + nK(1+\kappa)^2 L_{f,0}^2\right). \end{aligned} \tag{30}$$

Note that we have

$$\delta_y^T < \frac{L_\Phi^2}{12\Gamma\kappa^2} \quad \Rightarrow \quad \delta_y^T \cdot 6\kappa^2 \cdot 2\Gamma < L_\Phi^2. \tag{31}$$

Define

$$\Lambda = \frac{12L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)+6\Gamma c_1\delta_y^T}{n}.$$

By Lemma 12,

$$\begin{aligned}
\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2 &\le \frac{2L_\Phi^2}{n}S_K+\frac{2\Gamma}{n}A_K+\frac{12L^2\kappa}{n}\delta_\kappa^N B_K\\
&\le \frac{2L_\Phi^2}{n}S_K+\left(\frac{2\Gamma}{n}\cdot 6\kappa^2\delta_y^T+\frac{12L^2\kappa}{n}\cdot\delta_\kappa^N\cdot(4d_1\kappa^2+2d_2)\right)E_K\\
&\quad+\frac{12L^2\kappa\delta_\kappa^N(2c_2+6d_1c_1\delta_y^T)+6\Gamma c_1\delta_y^T}{n}\\
&\le \frac{2L_\Phi^2}{n}S_K+\left(\frac{L_\Phi^2}{n}+\frac{12L_\Phi^2}{n}\right)E_K+\frac{12L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)+6\Gamma c_1\delta_y^T}{n}\\
&\le \frac{2L_\Phi^2}{n}S_K+\frac{13L_\Phi^2}{n}\left(8S_K+4n\eta_x^2\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)}-\nabla\Phi(\bar{x}_j)\|^2+4n\eta_x^2T_{K-1}\right)+\Lambda\\
&< \frac{106L_\Phi^2}{n}S_K+52\eta_x^2L_\Phi^2\left(\sum_{j=0}^{K}\|\overline{\partial\Phi(X_j)}-\nabla\Phi(\bar{x}_j)\|^2+T_K\right)+\Lambda\\
&\le \left(\frac{106L_\Phi^2}{n}\cdot\frac{16nL_\Phi^2\eta_x^4}{(1-\rho)^2}+52\eta_x^2L_\Phi^2\right)\left(\sum_{j=0}^{K}\|\overline{\partial\Phi(X_j)}-\nabla\Phi(\bar{x}_j)\|^2+T_K\right)\\
&\quad+\frac{106L_\Phi^2}{n}\cdot\frac{4\eta_x^2}{(1-\rho)^2}\left(L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)+nK(1+\kappa)^2L_{f,0}^2\right)+\Lambda,
\end{aligned}$$

where the second inequality is by (26) and (27), the third inequality is by (28) and (31), the fourth inequality is obtained by (26) and the last inequality is by (30). Note that the definition of η_x also indicates:

$$106L_\Phi^2\cdot\frac{16L_\Phi^2\eta_x^4}{(1-\rho)^2}+52\eta_x^2L_\Phi^2<\frac{1}{3}.$$

Therefore,

$$\begin{aligned}
\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2 &< \frac{1}{3}\left(\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2+T_K\right)\\
&\quad+\frac{106L_\Phi^2}{n}\cdot\frac{4\eta_x^2}{(1-\rho)^2}(L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)+nK(1+\kappa)^2L_{f,0}^2)+\Lambda,
\end{aligned}$$

which leads to

$$\sum_{k=0}^{K} \|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2$$
$$\leq \frac{1}{2}T_K + 106L_\Phi^2 \cdot \frac{6\eta_x^2}{(1-\rho)^2}\left(\frac{L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)}{n} + K(1+\kappa)^2L_{f,0}^2\right)$$
$$+ \frac{18L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1) + 9\Gamma c_1\delta_y^T}{n}.$$

Combining this bound with (12), we can obtain

$$T_K \leq \frac{2}{\eta_x}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \sum_{k=0}^{K} \|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2$$
$$\leq \frac{2}{\eta_x}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \eta_x^2 \cdot \frac{636L_\Phi^2L_{f,0}^2(1+\kappa)^2}{(1-\rho)^2}K + \frac{1}{2}T_K + \frac{1}{2}C_1,$$

which implies

$$\frac{1}{K+1}\sum_{j=0}^{K} \|\nabla\Phi(\bar{x}_j)\|^2$$
$$\leq \frac{4}{\eta_x(K+1)}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \eta_x^2 \cdot \frac{1272L_\Phi^2L_{f,0}^2(1+\kappa)^2}{(1-\rho)^2} + \frac{C_1}{K+1}.$$

The constant C_1 satisfies

$$\frac{1}{2}C_1 = 106L_\Phi^2 \cdot \frac{6\eta_x^2L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)}{n(1-\rho)^2} + \frac{18L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1) + 9\Gamma c_1\delta_y^T}{n}$$
$$= \mathcal{O}(\eta_x^2\delta_\kappa^N\kappa^{12} + \kappa^5\delta_y^T) = \mathcal{O}(1).$$

Moreover, we notice that by setting

$$N = \Theta(\sqrt{\kappa}\log\kappa),\ T = \Theta(\kappa\log\kappa),\ \eta_x = \Theta(K^{-\frac{1}{3}}\kappa^{-\frac{8}{3}}),\ \eta_y = \frac{1}{\mu+L},$$

for sufficiently large K the conditions on algorithm parameters in Lemma 23 hold and

$$\frac{1}{K+1}\sum_{j=0}^{K} \|\nabla\Phi(\bar{x}_j)\|^2 = \mathcal{O}\left(\frac{\kappa^{\frac{8}{3}}}{K^{\frac{2}{3}}}\right),$$

which proves the first case of Theorems 3.1 and 19. □

Case 2: Assumption 2.3 does not hold

Now we consider the case when Assumption 2.3 does not hold.

$$S_K < \frac{\eta_x^2}{(1-\rho)^2}\sum_{j=0}^{K-1}\sum_{i=1}^{n}\|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 < \frac{\eta_x^2 L_{f,0}^2}{(1-\rho)^2} nK\big(2(1+\kappa)^2 + 2C\alpha^N\big).$$

Lemma 24

Proof The first inequality follows from Lemma 20. For the second one observe that:

$$\begin{aligned}
\|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\| &= \left\|\nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \left(Z_{i,k}^{(N)}\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\right\| \\
&\le \|\nabla_x f_i(x_{i,k}, y_{i,k}^{(T)})\| + \|(Z_{i,k}^{(N)} - Z_k^*)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\| + \|\left(Z_k^*\right)^{\mathsf{T}} \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\| \\
&\le \left(1 + \left\|\left(Z_{i,k}^{(N)}\right)^{\mathsf{T}} - \left(Z_k^*\right)^{\mathsf{T}}\right\|_2 + \kappa\right) L_{f,0},
\end{aligned}$$

where we use $\left(Z_{i,k}^{(N)}\right)^{\mathsf{T}}$ to denote the output of Algorithm 1 in outer loop iteration k of agent i, and $\left(Z_k^*\right)^{\mathsf{T}}$ denotes the optimal solution. By Cauchy–Schwarz inequality we know:

$$\begin{aligned}
\|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})\|^2 &\le (1 + \kappa + \|\left(Z_{i,k}^{(N)}\right)^{\mathsf{T}} - \left(Z_k^*\right)^{\mathsf{T}}\|_2)^2 L_{f,0}^2 \\
&\le (2(1+\kappa)^2 + 2\|\left(Z_{i,k}^{(N)}\right)^{\mathsf{T}} - \left(Z_k^*\right)^{\mathsf{T}}\|_2^2) L_{f,0}^2 \\
&\le (2(1+\kappa)^2 + 2C\alpha^N) L_{f,0}^2,
\end{aligned}$$

which completes the proof. □

Taking summation on both sides of (23) and applying Lemma 24 we know:

$$\begin{aligned}
&\sum_{k=0}^{K} \|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 \\
&\le \frac{(1+\kappa^2)}{n} \cdot \left(\frac{36L_{f,0}^2 L_{g,2}^2}{\mu^2} + 2L_f^2\right) S_K \\
&\quad + 12(K+1)C\left[\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2 L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right] \\
&\le \left(\frac{36L_{f,0}^2 L_{g,2}^2}{\mu^2} + 2L_f^2\right) \frac{(1+\kappa^2)\eta_x^2 L_{f,0}^2}{(1-\rho)^2} K(2(1+\kappa)^2 + 2C\alpha^N) + (K+1)\tilde{C}_1,
\end{aligned}$$

where we define:

$$\tilde{C}_1 = 12C\left[\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2 L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right] = \mathcal{O}(\alpha^T + \alpha^N).$$

The above inequality together with (12) gives

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla\Phi(\bar{x}_k)\|^2 \\
&\quad\le \frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \frac{1}{K+1}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 \\
&\quad\le \frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) \\
&\qquad + \frac{4\eta_x^2(1+\kappa^2)L_{f,0}^2}{(1-\rho)^2}((1+\kappa)^2 + C\alpha^N)\left(\frac{18L_{f,0}^2L_{g,2}^2}{\mu^2} + L_f^2\right) + \tilde{C}_1.
\end{aligned}$$

Moreover, if we choose

$$N = \Theta(\log K),\ T = \Theta(\log K),\ \eta_x = \Theta(K^{-\frac{1}{3}}\kappa^{-\frac{8}{3}}),\ \eta_y^{(t)} = \Theta(1)$$

then we can get:

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2 = \mathcal{O}\left(\frac{\kappa^{\frac{8}{3}}}{K^{\frac{2}{3}}}\right),$$

which proves the second case of Theorems 3.1 and 19.

Proof of the convergence of DBOGT

In this section we will prove the following convergence result of Algorithm 4

Theorem 25 *In Algorithm 4, suppose Assumptions 2.1 and 2.2 hold. If Assumption 2.3 holds, then by setting* $0 < \eta_x < \frac{(1-\rho)^2}{8L_\Phi}$, $0 < \eta_y < \frac{2}{\mu+L}$, $T = \Theta(\kappa\log\kappa)$, $N = \Theta(\sqrt{\kappa}\log\kappa)$, *we have*:

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2 \le \frac{4}{\eta_x(K+1)}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \frac{C_2}{K+1}.$$

If Assumption 2.3 does not hold, then by setting

$$0 < \eta_x < \min\left(\frac{(1-\rho)^2}{14\kappa L_f}, \frac{\mu(1-\rho)^2}{21 L_{f,0} L_{g,2} \kappa}\right),\ \eta_y = \Theta(1),$$

we have:

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2 \le \frac{4}{\eta_x(K+1)}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \tilde{C}_2.$$

Here $C_2 = \Theta(1)$ *and* $\tilde{C}_2 = \Theta(\alpha^T + \alpha^N + \frac{1}{K+1})$.

We first bound the consensus estimation error in the following lemma.

Lemma 26 *In Algorithm* 4, *we have the following inequality holds*:

$$S_K \le \frac{\eta_x^2}{(1-\rho)^4}\left(\sum_{j=1}^{K-1}\sum_{i=1}^{n}\|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)}) - \hat{\nabla} f_i(x_{i,j-1}, y_{i,j-1}^{(T)})\|^2 + \|\partial\Phi(X_0)\|^2\right).$$

Proof From the updates of x and u, we have:

$$\bar{u}_k = \bar{u}_{k-1} + \overline{\partial\Phi(X_k)} - \overline{\partial\Phi(X_{k-1})}, \quad \bar{u}_0 = \overline{\partial\Phi(X_0)}, \quad \bar{x}_{k+1} = \bar{x}_k - \eta_x\bar{u}_k,$$

which implies:

$$\bar{u}_k = \overline{\partial\Phi(X_k)}, \quad \bar{x}_{k+1} = \bar{x}_k - \eta_x\overline{\partial\Phi(X_k)}.$$

Hence by definition of $q_{i,k+1}$:

$$\begin{aligned} q_{i,k+1} &= x_{i,k+1} - \bar{x}_{k+1} = \sum_{j=1}^{n} w_{ij}x_{j,k} - \eta_x u_{i,k} - \bar{x}_k + \eta_x\bar{u}_k \\ &= \sum_{j=1}^{n} w_{ij}(x_{j,k} - \bar{x}_k) - \eta_x(u_{i,k} - \bar{u}_k) \\ &= \sum_{j=1}^{n} w_{ij}q_{j,k} - \eta_x r_{i,k} = Q_k W e_i - \eta_x R_k e_i. \end{aligned}$$

Therefore, we can write the update of the matrix Q_{k+1} as

$$Q_{k+1} = Q_k W - \eta_x R_k, \quad Q_1 = -\eta_x R_0.$$

Note that Q_{k+1} takes the form of

$$Q_{k+1} = (Q_{k-1}W - \eta_x R_{k-1})W - \eta_x R_k = -\eta_x\sum_{i=0}^{k} R_i W^{k-i}. \tag{32}$$

We then compute $r_{i,k}$ as following

$$\begin{aligned} r_{i,k+1} &= u_{i,k+1} - \bar{u}_{k+1} \\ &= \sum_{j=1}^{n} w_{ij} u_{j,k} + \hat{\nabla} f_i(x_{i,k+1}, y_{i,k+1}^{(T)}) - \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \bar{u}_k - (\overline{\partial\Phi(X_{k+1})} - \overline{\partial\Phi(X_k)}) \\ &= \sum_{j=1}^{n} w_{ij}(u_{j,k} - \bar{u}_k) + (\partial\Phi(X_{k+1}) - \partial\Phi(X_k))\left(e_i - \frac{\mathbf{1}_n}{n}\right) \\ &= R_k W e_i + (\partial\Phi(X_{k+1}) - \partial\Phi(X_k))\left(e_i - \frac{\mathbf{1}_n}{n}\right). \end{aligned}$$

The matrix R_{k+1} can be written as

$$\begin{aligned} R_{k+1} &= R_k W + (\partial\Phi(X_{k+1}) - \partial\Phi(X_k))(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}) \\ &= R_0 W^{k+1} + \sum_{j=0}^{k} (\partial\Phi(X_{j+1}) - \partial\Phi(X_j))\left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}\right) W^{k-j} \\ &= \partial\Phi(X_0)\left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}\right) W^{k+1} + \sum_{j=0}^{k} (\partial\Phi(X_{j+1}) - \partial\Phi(X_j))\left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}\right) W^{k-j} \\ &= \sum_{j=0}^{k+1} (\partial\Phi(X_j) - \partial\Phi(X_{j-1}))\left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}\right) W^{k+1-j}, \end{aligned} \tag{33}$$

where the third equality holds because of the initialization $u_{i,0} = \hat{\nabla} f_i(x_{i,0}, y_{i,0}^{(T)})$ and we denote $\partial\Phi(X_{-1}) = 0$. Plugging (33) into (32) yields

$$\begin{aligned} Q_{k+1} &= -\eta_x \sum_{i=0}^{k} \sum_{j=0}^{i} (\partial\Phi(X_j) - \partial\Phi(X_{j-1}))\left(I - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}\right) W^{k-j} \\ &= -\eta_x \sum_{j=0}^{k} \sum_{i=j}^{k} (\partial\Phi(X_j) - \partial\Phi(X_{j-1}))\left(W^{k-j} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}\right) \\ &= -\eta_x \sum_{j=0}^{k} (k+1-j)(\partial\Phi(X_j) - \partial\Phi(X_{j-1}))\left(W^{k-j} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n}\right), \end{aligned}$$

where the second equality is obtained by $\mathbf{1}_n \mathbf{1}_n^\top W = \mathbf{1}_n \mathbf{1}_n^\top$ and switching the order of the summations. Therefore, we have

$$
\begin{aligned}
\|Q_{k+1}\|^2 &= \eta_x^2 \left\| \sum_{j=0}^{k} (k+1-j)(\partial\Phi(X_j) - \partial\Phi(X_{j-1})) \left(W^{k-j} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \right\|^2 \\
&\leq \eta_x^2 \left(\sum_{j=0}^{k} \left\| (k+1-j)(\partial\Phi(X_j) - \partial\Phi(X_{j-1})) \left(W^{k-j} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \right\| \right)^2 \\
&\leq \eta_x^2 \left(\sum_{j=0}^{k} \left\| (k+1-j)(\partial\Phi(X_j) - \partial\Phi(X_{j-1})) \right\| \left\| W^{k-j} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right\|_2 \right)^2 \\
&\leq \eta_x^2 \left(\sum_{j=0}^{k} \rho^{k-j}(k+1-j) \|\partial\Phi(X_j) - \partial\Phi(X_{j-1})\|^2 \right) \cdot \\
&\quad \left(\sum_{j=0}^{k} \frac{(k+1-j)}{\rho^{k-j}} \left\| W^{k-j} - \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right\|_2^2 \right) \\
&\leq \eta_x^2 \left(\sum_{j=0}^{k} \rho^{k-j}(k+1-j) \|\partial\Phi(X_j) - \partial\Phi(X_{j-1})\|^2 \right) \left(\sum_{j=0}^{k} (k+1-j)\rho^{k-j} \right) \\
&< \frac{\eta_x^2}{(1-\rho)^2} \left(\sum_{j=0}^{k} \rho^{k-j}(k+1-j) \|\partial\Phi(X_j) - \partial\Phi(X_{j-1})\|^2 \right),
\end{aligned}
\tag{34}
$$

where the second inequality is by Lemma 1, the third inequality is by Lemma 4, and the last inequality uses the fact that:

$$
\sum_{j=0}^{k} (k+1-j)\rho^{k-j} = \sum_{m=0}^{k} (m+1)\rho^m = \frac{1-(k+2)\rho^{k+1}+(k+1)\rho^{k+2}}{(1-\rho)^2} < \frac{1}{(1-\rho)^2}.
\tag{35}
$$

Summing (34) over $k = 0, \ldots, K-1$, we get:

$$
\begin{aligned}
S_K &= \sum_{k=0}^{K-1} \|Q_{k+1}\|^2 \\
&\leq \frac{\eta_x^2}{(1-\rho)^2} \left(\sum_{k=0}^{K-1} \sum_{j=0}^{k} \rho^{k-j}(k+1-j) \|\partial\Phi(X_j) - \partial\Phi(X_{j-1})\|^2 \right) \\
&= \frac{\eta_x^2}{(1-\rho)^2} \left(\sum_{j=0}^{K-1} \sum_{k=j}^{K-1} \rho^{k-j}(k+1-j) \|\partial\Phi(X_j) - \partial\Phi(X_{j-1})\|^2 \right) \\
&< \frac{\eta_x^2}{(1-\rho)^4} \sum_{j=0}^{K-1} \|\partial\Phi(X_j) - \partial\Phi(X_{j-1})\|^2 \\
&= \frac{\eta_x^2}{(1-\rho)^4} \left(\sum_{j=1}^{K-1} \sum_{i=1}^{n} \|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)}) - \hat{\nabla} f_i(x_{i,j-1}, y_{i,j-1}^{(T)})\|^2 + \|\partial\Phi(X_0)\|^2 \right),
\end{aligned}
$$

which completes the proof. □

Case 1: Assumption 2.3 holds

When Assumption 2.3 holds, we have the following lemmas.

Lemma 27 *Under Assumption 2.3, the following inequality holds for Algorithm 4:*

$$\sum_{j=1}^{K-1}\sum_{i=1}^{n}\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)})-\hat{\nabla}f_i(x_{i,j-1},y_{i,j-1}^{(T)})\|^2 \le 6\Gamma A_{K-1}+36L^2\kappa\delta_\kappa^N B_{K-1}+3L_\Phi^2 E_{K-1}.$$

Moreover, we have:

$$S_K \le \frac{\eta_x^2}{(1-\rho)^4}(6\Gamma A_{K-1}+36L^2\kappa\delta_\kappa^N B_{K-1}+3L_\Phi^2 E_{K-1}+\|\partial\Phi(X_0)\|^2).$$

Proof For each term, we know that for $j \ge 1$:

$$\begin{aligned}
&\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)})-\hat{\nabla}f_i(x_{i,j-1},y_{i,j-1}^{(T)})\|^2\\
&\le 3(\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)})-\nabla\Phi_i(x_{i,j})\|^2+\|\nabla\Phi_i(x_{i,j})-\nabla\Phi_i(x_{i,j-1})\|^2\\
&\quad+\|\nabla\Phi_i(x_{i,j-1})-\hat{\nabla}f_i(x_{i,j-1},y_{i,j-1}^{(T)})\|^2)\\
&\le 3(\Gamma(\|y_i^*(x_{i,j})-y_{i,j}^{(T)}\|^2+\|y_i^*(x_{i,j-1})-y_{i,j-1}^{(T)}\|^2)\\
&\quad+6L^2\kappa\delta_\kappa^N(\|v_{i,j}^*-v_{i,j}^{(0)}\|^2+\|v_{i,j-1}^*-v_{i,j-1}^{(0)}\|^2)+L_\Phi^2\|x_{i,j}-x_{i,j-1}\|^2),
\end{aligned}$$

where the last inequality uses Lemmas 11 and 5. Taking summation ($j=1,2,\ldots,K-1$ and $i=1,2,\ldots,n$) on both sides, we have:

$$\sum_{j=1}^{K-1}\sum_{i=1}^{n}\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)})-\hat{\nabla}f_i(x_{i,j-1},y_{i,j-1}^{(T)})\|^2 \le 6\Gamma A_{K-1}+36L^2\kappa\delta_\kappa^N B_{K-1}+3L_\Phi^2 E_{K-1}.$$

Together with Lemma 26, we can prove the second inequality for S_K. □

The above lemma together with Lemma 10 and 9 gives

$$\begin{aligned}
&S_K \le \frac{\eta_x^2}{(1-\rho)^4}(6\Gamma A_{K-1}+36L^2\kappa\delta_\kappa^N B_{K-1}+3L_\Phi^2 E_{K-1}+\|\partial\Phi(X_0)\|^2)\\
&A_K \le \delta_y^T(3c_1+6\kappa^2 E_K)\quad B_K \le 2c_2+2d_1A_{K-1}+2d_2E_K\\
&E_K \le 8S_K+4n\eta_x^2\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)}-\nabla\Phi(\bar{x}_j)\|^2+4n\eta_x^2T_{K-1}.
\end{aligned} \tag{36}$$

Now we can obtain the following result.

Lemma 28 *Suppose Assumptions 2.1, 2.2 and 2.3 hold. Set*:

$$\delta_y^T < \min\left(\frac{L_\Phi^2}{72\kappa^2\Gamma}, \kappa^{-5}\right) = \Theta(\kappa^{-5}),$$
$$\delta_\kappa^N < \min\left(\frac{L_\Phi^2}{72L^2\kappa(4d_1\kappa^2+2d_2)}, \kappa^{-4}\right) = \Theta(\kappa^{-4}),\ \eta_x < \frac{(1-\rho)^2}{8L_\Phi}.$$

For Algorithm 4, we have:

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla\Phi(\bar{x}_k)\|^2 \le \frac{4}{\eta_x(K+1)}(\Phi(\bar{x}_0) - \inf_x \Phi(x)) + \frac{C_2}{K+1},$$

where the constant is defined as:

$$\begin{aligned}\frac{1}{2}C_2 &= \frac{15\eta_x^2L_\Phi^2}{n(1-\rho)^4}(\|\partial\Phi(X_0)\|^2 + 18\Gamma c_1\delta_y^T + 36L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)) \\ &\quad + \frac{18L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1) + 9\Gamma c_1\delta_y^T}{n} \\ &= \Theta(\eta_x^2\kappa^6 + (\eta_x^2\kappa^6+1)(\kappa^5\delta_y^T + \kappa^4\delta_\kappa^N)) = \Theta(1).\end{aligned}$$

Proof We first bound B_K as

$$\begin{aligned}B_K &\le 2c_2 + 2d_1A_K + 2d_2E_K \le 2c_2 + \frac{2}{3}d_1(3c_1 + 6\kappa^2E_K) + 2d_2E_K \\ &= 2c_2 + 2d_1c_1 + (4d_1\kappa^2+2d_2)E_K.\end{aligned} \tag{37}$$

Next we eliminate A_K and B_K in the upper bound of S_K. Choose N, T such that

$$\delta_y^T\cdot 6\kappa^2\cdot 6\Gamma < \frac{L_\Phi^2}{2},\quad \delta_\kappa^N\cdot(4d_1\kappa^2+2d_2)\cdot 36L^2\kappa < \frac{L_\Phi^2}{2},$$

which implies

$$\delta_y^T < \frac{L_\Phi^2}{72\kappa^2\Gamma},\quad \delta_\kappa^N < \frac{L_\Phi^2}{72L^2\kappa(4d_1\kappa^2+2d_2)}. \tag{38}$$

By (36) and the, we have

$$S_K \le \frac{\eta_x^2}{(1-\rho)^4}(4L_\Phi^2E_{K-1} + \|\partial\Phi(X_0)\|^2 + 18\Gamma c_1\delta_y^T + 36L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)).$$

Next we eliminate E_{K-1} in this bound. The definition of η_x gives $\eta_x < \frac{(1-\rho)^2}{8L_\Phi}$, which implies $\frac{32L_\Phi^2\eta_x^2}{(1-\rho)^4} < \frac{1}{2}$. Together with (36) and $E_{K-1} \le E_K$, we have:

$$
\begin{aligned}
S_K &\le \frac{\eta_x^2}{(1-\rho)^4}\left(4L_\Phi^2(8S_K + 4n\eta_x^2\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 T_{K-1}\right)\\
&\quad + \|\partial\Phi(X_0)\|^2 + 18\Gamma c_1\delta_y^T + 36L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1))\\
&\le \frac{1}{2}S_K + \frac{\eta_x^2}{(1-\rho)^4}\left(4L_\Phi^2(4n\eta_x^2\sum_{j=0}^{K}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 T_K\right)\\
&\quad + \|\partial\Phi(X_0)\|^2 + 18\Gamma c_1\delta_y^T + 36L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1)),
\end{aligned}
$$

which immediately implies

$$
\begin{aligned}
S_K &< \frac{2\eta_x^2}{(1-\rho)^4}(16n\eta_x^2L_\Phi^2\left(\sum_{j=0}^{K}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + T_K\right)\\
&\quad + \|\partial\Phi(X_0)\|^2 + 18\Gamma c_1\delta_y^T + 36L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1)).
\end{aligned}
\tag{39}
$$

Moreover, by (19) we have

$$
\begin{aligned}
\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 &\le \frac{2L_\Phi^2}{n}S_K + \frac{2\Gamma}{n}A_K + \frac{12L^2\kappa}{n}\delta_\kappa^N B_K\\
&\le \frac{2L_\Phi^2}{n}S_K + \left(\frac{L_\Phi^2}{6n} + \frac{L_\Phi^2}{6n}\right)E_K + \frac{12L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + 6\Gamma c_1\delta_y^T}{n}\\
&\le \frac{2L_\Phi^2}{n}S_K + \frac{L_\Phi^2}{3n}\left(8S_K + 4n\eta_x^2\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + 4n\eta_x^2 T_{K-1}\right)\\
&\quad + \frac{12L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + 6\Gamma c_1\delta_y^T}{n}\\
&< \frac{5L_\Phi^2}{n}S_K + \frac{4\eta_x^2L_\Phi^2}{3}\left(\sum_{j=0}^{K}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + T_K\right)\\
&\quad + \frac{12L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + 6\Gamma c_1\delta_y^T}{n}\\
&\le \left(\frac{5L_\Phi^2}{n}\cdot\frac{32nL_\Phi^2\eta_x^4}{(1-\rho)^4} + \frac{4\eta_x^2L_\Phi^2}{3}\right)\left(\sum_{j=0}^{K}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + T_K\right)\\
&\quad + \frac{5L_\Phi^2}{n}\cdot\frac{2\eta_x^2}{(1-\rho)^4}\left(\|\partial\Phi(X_0)\|^2 + 18\Gamma c_1\delta_y^T + 36L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1)\right)\\
&\quad + \frac{12L^2\kappa\delta_\kappa^N(2c_2 + 2d_1c_1) + 6\Gamma c_1\delta_y^T}{n},
\end{aligned}
$$

where the second inequality is by (36), (37) and (38), and the third inequality uses (36). Note that η_x satisfies:

$$
\eta_x < \frac{(1-\rho)^2}{8L_\Phi} \quad \Rightarrow \quad \frac{160L_\Phi^4\eta_x^4}{(1-\rho)^4} + \frac{8\eta_x^2L_\Phi^2}{3} < \frac{1}{3}.
$$

Therefore, we have:

$$\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2\le\frac{1}{3}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2+\frac{1}{3}T_K+\frac{10\eta_x^2L_\Phi^2}{n(1-\rho)^4}(\|\partial\Phi(X_0)\|^2+18\Gamma c_1\delta_y^T+36L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1))+\frac{12L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)+6\Gamma c_1\delta_y^T}{n},$$

which leads to

$$\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2\le\frac{1}{2}T_K+\frac{15\eta_x^2L_\Phi^2}{n(1-\rho)^4}\Big(\|\partial\Phi(X_0)\|^2+18\Gamma c_1\delta_y^T+36L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)\Big)+\frac{18L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)+9\Gamma c_1\delta_y^T}{n}.$$

Recall (12), we have

$$\frac{1}{K+1}T_K\le\frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x\Phi(x))+\frac{1}{K+1}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2\le\frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x\Phi(x))+\frac{1}{2(K+1)}T_K+\frac{1}{2(K+1)}C_2.$$

Therefore, we get

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2\le\frac{4}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x\Phi(x))+\frac{C_2}{K+1},$$

where the constant is defined as following

$$\begin{aligned}\frac{1}{2}C_2&=\frac{15\eta_x^2L_\Phi^2}{n(1-\rho)^4}(\|\partial\Phi(X_0)\|^2+18\Gamma c_1\delta_y^T+36L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1))+\frac{18L^2\kappa\delta_\kappa^N(2c_2+2d_1c_1)+9\Gamma c_1\delta_y^T}{n}\\&=\Theta(\eta_x^2\kappa^6+(\eta_x^2\kappa^6+1)(\kappa^5\delta_y^T+\kappa^4\delta_\kappa^N))=\Theta(1).\end{aligned}$$

Then if we choose

$$T=\Theta(\kappa\log\kappa),N=\Theta(\sqrt{\kappa}\log\kappa),\eta_x=\Theta(\kappa^{-3}),\eta_y=\frac{1}{\mu+L},$$

then the restrictions on algorithm parameters in Lemma 28 hold and we have

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2=\mathcal{O}\Big(\frac{1}{K}\Big),$$

which proves the first case of Theorems 3.2 and 25. □

Case 2: Assumption 2.3 does not hold

We first give a bound for $\|\tilde{y}_j^*-\tilde{y}_{j-1}^*\|$ in the following lemma.

Lemma 29 *Recall that* $\tilde{y}_j^*=\arg\min\frac{1}{n}\sum_{i=1}^n g_i(x_{i,j},y)$. *We have*:

$$\|\tilde{y}_j^*-\tilde{y}_{j-1}^*\|^2\le\frac{\kappa^2}{n}\sum_{i=1}^n\|x_{i,j}-x_{i,j-1}\|^2.$$

Proof The proof technique is similar to Lemma 13. Consider:

$$\begin{aligned}
&\|\frac{1}{n}\sum_{i=1}^n\nabla_y g_i(x_{i,j-1},\tilde{y}_j^*)\|\\
&\quad=\|\frac{1}{n}\sum_{i=1}^n\nabla_y g_i(x_{i,j-1},\tilde{y}_j^*)-\frac{1}{n}\sum_{i=1}^n\nabla_y g_i(x_{i,j-1},\tilde{y}_{j-1}^*)\|\ge\mu\|\tilde{y}_j^*-\tilde{y}_{j-1}^*\|,\\
&\|\frac{1}{n}\sum_{i=1}^n\nabla_y g_i(x_{i,j-1},\tilde{y}_j^*)\|\\
&\quad=\|\frac{1}{n}\sum_{i=1}^n\nabla_y g_i(x_{i,j-1},\tilde{y}_j^*)-\frac{1}{n}\sum_{i=1}^n\nabla_y g_i(x_{i,j},\tilde{y}_j^*)\|\le\frac{L}{n}\sum_{i=1}^n\|x_{i,j}-x_{i,j-1}\|,
\end{aligned}$$

which implies:

$$\|\tilde{y}_j^*-\tilde{y}_{j-1}^*\|^2\le\frac{\kappa^2}{n^2}\left(\sum_{i=1}^n\|x_{i,j}-x_{i,j-1}\|\right)^2\le\frac{\kappa^2}{n}\sum_{i=1}^n\|x_{i,j}-x_{i,j-1}\|^2.$$

□

Lemma 30 *Suppose* η_x *satisfies*

$$\eta_x\le\frac{\mu(1-\rho)^2}{21L_{f,0}L_{g,2}\kappa}.\tag{40}$$

When the Assumption 2.3 does not hold, we have for Algorithm 4:

$$S_K \le \frac{2\eta_x^2}{(1-\rho)^4}\left[3L_f^2(1+\kappa^2)\sum_{j=1}^{K-1}\sum_{i=1}^{n}E_{K-1} + \|\partial\Phi(X_0)\|^2\right] + \frac{72nKC\eta_x^2}{(1-\rho)^4}\left(\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right)$$

Proof We first notice that

$$\begin{aligned}&\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)}) - \hat{\nabla}f_i(x_{i,j-1},y_{i,j-1}^{(T)})\|^2\\ &\le 3\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)}) - \bar{\nabla}f_i(x_{i,j},\tilde{y}_j^*)\|^2 + 3\|\bar{\nabla}f_i(x_{i,j},\tilde{y}_j^*) - \bar{\nabla}f_i(x_{i,j-1},\tilde{y}_{j-1}^*)\|^2\\ &\quad + 3\|\bar{\nabla}f_i(x_{i,j-1},\tilde{y}_{j-1}^*) - \hat{\nabla}f_i(x_{i,j-1},y_{i,j-1}^{(T)})\|^2.\end{aligned}$$

Taking summation on both sides and using Lemma 17, we have

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{K-1}\sum_{i=1}^{n}\|\hat{\nabla}f_i(x_{i,j},y_{i,j}^{(T)}) - \hat{\nabla}f_i(x_{i,j-1},y_{i,j-1}^{(T)})\|^2\\ &\le \frac{108L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{n\mu^2}S_{K-1}\\ &\quad + 36(K-1)\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2\\ &\quad + 36(K-1)CL_{f,0}^2\alpha^N + \frac{3L_f^2}{n}\sum_{j=1}^{K-1}\sum_{i=1}^{n}(\|x_{i,j}-x_{i,j-1}\|^2 + \|\tilde{y}_j^*-\tilde{y}_{j-1}^*\|^2)\\ &\le \frac{(1-\rho)^4}{2n\eta_x^2}S_{K-1} + 36KC\left(\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right)\\ &\quad + \frac{3L_f^2(1+\kappa^2)}{n}E_{K-1},\end{aligned}$$

where the second inequality uses Lemma 14, 29 and (40). This completes the proof together with Lemma 26. □

Lemma 31 *When the Assumption 2.3 does not hold, we further have for Algorithm 4:*

$$\begin{aligned}&\frac{1}{K+1}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2\\ &\le \frac{(1+\kappa^2)}{n(K+1)}\cdot\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2}+2L_f^2\right)S_K\\ &\quad + 12C\left[\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right].\end{aligned}$$

Proof Note that the above inequality is a direct result of Lemma 18. □

Now we are ready to provide the convergence rate. Recall that from Lemma 30, 9 and inequality (12), we have:

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla\Phi(\bar{x}_k)\|^2 \\
&\quad\le \frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x \Phi(x))+\frac{1}{K+1}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2, \\
S_K &\le \frac{2\eta_x^2}{(1-\rho)^4}\left(3L_f^2(1+\kappa^2)E_{K-1}+\|\partial\Phi(X_0)\|^2\right) \\
&\quad+\frac{72nKC\eta_x^2}{(1-\rho)^4}\left(\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T+L_{f,0}^2\alpha^N\right), \\
E_K &\le 8S_K+4n\eta_x^2\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)}-\nabla\Phi(\bar{x}_j)\|^2+4n\eta_x^2T_{K-1}.
\end{aligned} \tag{41}$$

The following lemma proves the convergence results in Theorem 25.

Lemma 32 *Suppose the Assumption 2.3 does not hold. We set η_x as*

$$\eta_x < \min\left(\frac{(1-\rho)^2}{14\kappa L_f},\ \frac{\mu(1-\rho)^2}{21L_{f,0}L_{g,2}\kappa}\right). \tag{42}$$

Then we have:

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla\Phi(\bar{x}_k)\|^2 \le \frac{6}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x\Phi(x))+\frac{\|\partial\Phi(X_0)\|^2}{K+1}+\tilde{C}_2,$$

where the constant is given by:

$$\begin{aligned}
\frac{\tilde{C}_2}{6} &= 6L^2(1+\kappa^2)C\alpha^T+6L_{f,0}^2C\alpha^N \\
&\quad+2L_f^2(1+\kappa^2)\cdot\frac{2\eta_x^2}{(1-\rho)^4}\left[6L^2(1+\kappa^2)nC\alpha^T+6nL_{f,0}^2C\alpha^N+\frac{\|\partial\Phi(X_0)\|^2}{K+1}\right] \\
&= \Theta\left(\alpha^T+\alpha^N+\frac{1}{K+1}\right).
\end{aligned}$$

Proof We first eliminate E_{K-1} in the upper bound of S_K. Note that (42) implies

$$\frac{2\eta_x^2}{(1-\rho)^4}\cdot 3L_f^2(1+\kappa^2)\cdot 8 < \frac{1}{2},$$

which together with $E_{K-1}\le E_K$ and the upper bounds of S_K and E_K in (41) gives

$$S_K \le \frac{1}{2}\left(S_K + \frac{\eta_x^2}{2}\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + \frac{\eta_x^2}{2}T_{K-1}\right) + \frac{2\eta_x^2}{(1-\rho)^4}\|\partial\Phi(X_0)\|^2 + \frac{2\eta_x^2}{(1-\rho)^4}\left(36nKC\left(\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right)\right).$$

Hence we know

$$S_K \le \frac{\eta_x^2}{2}\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + \frac{\eta_x^2}{2}T_{K-1} + \frac{4\eta_x^2}{(1-\rho)^4}\|\partial\Phi(X_0)\|^2 + \frac{4\eta_x^2}{(1-\rho)^4}\left(36nKC\left(\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right)\right).$$

By Lemma 31, we have

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 \\
&\le \frac{(1+\kappa^2)}{n(K+1)}\cdot\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2} + 2L_f^2\right)S_K \\
&\quad + 12C\left[\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right] \\
&\le \frac{1}{3(K+1)}\left(\sum_{j=0}^{K-1}\|\overline{\partial\Phi(X_j)} - \nabla\Phi(\bar{x}_j)\|^2 + T_{K-1}\right) + \frac{\tilde{C}_2}{3},
\end{aligned} \tag{43}$$

where the second inequality holds since we have (42), which implies

$$\eta_x^2(1+\kappa^2)\cdot\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2} + 2L_f^2\right) \le \frac{1}{4}.$$

The constant is defined as:

$$\begin{aligned}
\frac{\tilde{C}_2}{3} &= 12C\left[\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right] + \frac{1}{(1-\rho)^4}\frac{\|\partial\Phi(X_0)\|^2}{n(K+1)} \\
&\quad + \frac{1}{(1-\rho)^4}\left(36C\left(\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\alpha^T + L_{f,0}^2\alpha^N\right)\right) \\
&= \Theta\left(\alpha^T + \alpha^N + \frac{1}{K+1}\right).
\end{aligned}$$

From (43) we know

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)} - \nabla\Phi(\bar{x}_k)\|^2 < \frac{\tilde{C}_2}{2} + \frac{1}{2(K+1)}T_{K-1}.$$

Combining the above inequality, Lemma 8, and $T_{K-1} \leq T_K$, we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla\Phi(\bar{x}_k)\|^2 < \frac{2}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x \Phi(x)) + \frac{\tilde{C}_2}{2} + \frac{1}{2(K+1)}T_K.$$

Hence

$$\frac{1}{K+1}T_K < \frac{4}{\eta_x(K+1)}(\Phi(\bar{x}_0)-\inf_x \Phi(x)) + \tilde{C}_2.$$

Furthermore, by setting

$$N = \Theta(\log K),\ T = \Theta(\log K),\ \eta_x = \Theta(\kappa^{-3}),\ \eta_y = \Theta(1)$$

we have

$$\frac{1}{K+1}\sum_{j=0}^{K}\|\nabla\Phi(\bar{x}_j)\|^2 = \mathcal{O}\Big(\frac{1}{K}\Big),$$

which proves the second case of Theorems 3.2 and 25. □

Proof of the convergence of DSBO

In this section we will prove the convergence result of the DSBO algorithm.

Theorem 33 *In Algorithm 5, suppose Assumptions 2.1 and 2.2 hold. If Assumption 2.3 holds, then by setting* $M = \Theta(\log K)$, $T = \Omega(\kappa \log \kappa)$, $\beta \leq \min\left(\frac{\mu}{\mu^2+\sigma_{g,2}^2}, \frac{1}{L}\right)$, $\eta_x \leq \frac{1}{L_\Phi}$, $\eta_y < \frac{2}{\mu+L}$, *we have*:

$$\begin{aligned}&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right]\\ &\leq \frac{2}{\eta_x(K+1)}(\mathbb{E}\left[\Phi(\bar{x}_0)\right]-\inf_x\Phi(x)) + \frac{3\eta_y L_f^2\sigma_{g,1}^2}{\mu} + \frac{3\eta_x^2 L_\Phi^2}{(1-\rho)^2}\tilde{C}_f^2 + L\eta_x\tilde{\sigma}_f^2 + C_3.\end{aligned}$$

If Assumption 2.3 does not hold, then by setting $\eta_x \leq \frac{1}{L_\Phi}$, $\eta_y^{(t)} = \mathcal{O}(\frac{1}{t})$, *we have*:

$$\begin{aligned}&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right]\\ &\leq \frac{2}{\eta_x(K+1)}(\mathbb{E}\left[\Phi(\overline{x_0})\right]-\inf_x\Phi(x)) + \left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2}+2L_f^2\right)\frac{\eta_x^2(1+\kappa^2)}{(1-\rho)^2}\tilde{C}_f^2\\ &\quad + L\eta_x\left(4\sigma_f^2(1+\kappa^2)+(8L_{f,0}^2+4\sigma_f^2)\frac{C}{N}\right)+\tilde{C}_3.\end{aligned}$$

Here $C = \Theta(1),\ C_3 = \Theta(\eta_x^2 + \frac{1}{K+1})$ *and* $\tilde{C}_3 = \mathcal{O}\Big(\frac{1}{T} + \alpha^N\Big)$.

We first define the following filtration:

$$\mathcal{F}_k = \sigma\Bigg(\bigcup_{i=1}^{n}\{x_{i,0}, x_{i,1}, \dots, x_{i,k}\}\Bigg),$$
$$\mathcal{G}_{i,j}^{(t)} = \sigma\Big(\{y_{i,l}^{(s)} : 0 \le l \le j, 0 \le s \le t\} \bigcup \{x_{i,l} : 0 \le l \le j\}\Big).$$

Then in both cases we have the following lemma.

Lemma 34 *If* $\eta_x \le \frac{1}{L_\Phi}$, *then we have*:

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] \\
&\le \frac{2}{\eta_x}(\mathbb{E}\left[\Phi(\bar{x}_k)\right] - \mathbb{E}\left[\Phi(\bar{x}_{k+1})\right]) + \mathbb{E}\left[\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right] \\
&\quad + L\eta_x\mathbb{E}\|\left[\overline{\partial\Phi(X_k;\phi)}\right] - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2.
\end{aligned}$$

Proof In each iteration of Algorithm 5, we have:

$$\bar{x}_{k+1} = \bar{x}_k - \eta_x\overline{\partial\Phi(X_k;\phi)}. \tag{44}$$

The L_Φ-smoothness of Φ indicates that

$$\Phi(\bar{x}_{k+1}) - \Phi(\bar{x}_k) \le \nabla\Phi(\bar{x}_k)^\top(-\eta_x\overline{\partial\Phi(X_k;\phi)}) + \frac{L_\Phi\eta_x^2}{2}\|\overline{\partial\Phi(X_k;\phi)}\|^2.$$

Taking conditional expectation with respect to $\mathcal{F}_k$ on both sides, we have the following

$$\begin{aligned}
&\mathbb{E}\left[\Phi(\bar{x}_{k+1})|\mathcal{F}_k\right] - \Phi(\bar{x}_k) \\
&\le \nabla\Phi(\bar{x}_k)^\top(-\eta_x\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]) + \frac{L_\Phi\eta_x^2}{2}\mathbb{E}\left[\|\overline{\partial\Phi(X_k;\phi)}\|^2|\mathcal{F}_k\right] \\
&= -\frac{\eta_x}{2}(\|\nabla\Phi(\bar{x}_k)\|^2 + \|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2 - \|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2) \\
&\quad + \frac{L_\Phi\eta_x^2}{2}(\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2 + \mathbb{E}\left[\|\overline{\partial\Phi(X_k;\phi)} - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2|\mathcal{F}_k\right]) \\
&= \left(\frac{L_\Phi\eta_x^2}{2} - \frac{\eta_x}{2}\right)\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2 \\
&\quad + \frac{L_\Phi\eta_x^2}{2}\mathbb{E}\left[\|\overline{\partial\Phi(X_k;\phi)} - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2|\mathcal{F}_k\right] \\
&\quad - \frac{\eta_x}{2}(\|\nabla\Phi(\bar{x}_k)\|^2 - \|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2) \\
&\le \frac{L_\Phi\eta_x^2}{2}\mathbb{E}\left[\|\overline{\partial\Phi(X_k;\phi)} - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2|\mathcal{F}_k\right] \\
&\quad - \frac{\eta_x}{2}(\|\nabla\Phi(\bar{x}_k)\|^2 - \|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2),
\end{aligned}$$

where the second inequality holds since we pick $\eta_x \leq \frac{1}{L}$. Thus we can take expectation again and use tower property to obtain:

$$\begin{aligned}
&\frac{\eta_x}{2}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] \\
&\quad\leq \mathbb{E}\left[\Phi(\bar{x}_k)\right] - \mathbb{E}\left[\Phi(\bar{x}_{k+1})\right] + \frac{\eta_x}{2}\mathbb{E}\left[\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right] \\
&\qquad+ \frac{L_\Phi\eta_x^2}{2}\mathbb{E}\|\left[\overline{\partial\Phi(X_k;\phi)}\right] - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2.
\end{aligned} \tag{45}$$

which completes the proof. □

Case 1: Assumption 2.3 holds

Lemma 35 *Suppose $\beta \leq \frac{1}{L}$ and Assumption 2.3 holds, we have*:

$$\left\|\mathbb{E}\left[\hat{\nabla}f_i(x_{i,k}, y_{i,k}^{(T)};\phi_{i,k})|\mathcal{F}_k\right] - \bar{\nabla}f_i(x_{i,k}, y_{i,k}^{(T)})\right\|_2 \leq L_{f,0}(1-\beta\mu)^M\kappa.$$

Proof We first consider the expectation

$$\begin{aligned}
&\mathbb{E}\left[\hat{\nabla}f_i(x_{i,k}, y_{i,k}^{(T)};\phi_{i,k})|\mathcal{F}_k\right] \\
&\quad= \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) \\
&\qquad- \beta\nabla_{xy}g(x_{i,k}, y_{i,k}^{(T)})\sum_{j=0}^{M-1}\left(I - \beta\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)})\right)^j\nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}).
\end{aligned} \tag{46}$$

Notice that for the finite sum we have:

$$\begin{aligned}
\beta\sum_{j=0}^{M-1}\left(I - \beta\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)})\right)^j &= \beta\left(\beta\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)})\right)^{-1}\left(I - (I - \beta\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)}))^M\right) \\
&= \left(\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)})\right)^{-1}\left(I - (I - \beta\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)}))^M\right),
\end{aligned}$$

which implies:

$$\left\|\beta\sum_{j=0}^{M-1}\left(I - \beta\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)})\right)^j - \left(\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)})\right)^{-1}\right\|_2 \leq \frac{(1-\beta\mu)^M}{\mu}. \tag{47}$$

The above inequality and the fact that

$$\bar{\nabla}f_i(x_{i,k}, y_{i,k}^{(T)}) = \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla_{xy}g(x_{i,k}, y_{i,k}^{(T)})\left(\nabla_y^2 g(x_{i,k}, y_{i,k}^{(T)})\right)^{-1}\nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})$$

imply

$$\left\|\mathbb{E}\left[\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k})|\mathcal{F}_k\right] - \bar{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)})\right\|_2 \leq L_{f,0}(1-\beta\mu)^M \kappa,$$

which completes the proof. □

Lemma 36 *Under Assumption 2.3, we have*:

$$\begin{aligned}&\sum_{k=0}^{K} \|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2 \\ &\quad\leq 3\left((K+1)L_{f,0}^2(1-\beta\mu)^{2M}\kappa^2 + \frac{L_f^2}{n}A_K + \frac{L_\Phi^2}{n}S_K\right).\end{aligned} \tag{48}$$

Proof We first bound each component of the gradient error as

$$\begin{aligned}&\|\mathbb{E}\left[\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k})|\mathcal{F}_k\right] - \nabla\Phi_i(\bar{x}_k)\|^2 \\ &\quad\leq 3(\|\mathbb{E}\left[\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k})|\mathcal{F}_k\right] - \bar{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)})\|^2 \\ &\qquad + \|\bar{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) - \nabla f_i(x_{i,k}, y_i^*(x_{i,k}))\|^2 + \|\nabla f_i(x_{i,k}, y_i^*(x_{i,k})) - \nabla\Phi_i(\bar{x}_k)\|^2) \\ &\quad\leq 3(L_{f,0}^2(1-\beta\mu)^{2M}\kappa^2 + L_f^2\|y_{i,k}^{(T)} - y_i^*(x_{i,k})\|^2 + L_\Phi^2\|x_{i,k} - \bar{x}_k\|^2),\end{aligned}$$

where the second inequality is obtained by Lemmas 35 and 5. Taking summation on both sides over $i = 1, \ldots, n$, we have:

$$\begin{aligned}&\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2 \\ &\quad\leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbb{E}\left[\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k})|\mathcal{F}_k\right] - \nabla\Phi_i(\bar{x}_k)\|^2 \\ &\quad\leq 3\left(L_{f,0}^2(1-\beta\mu)^{2M}\kappa^2 + \frac{L_f^2}{n}\sum_{i=1}^{n}\|y_{i,k}^{(T)} - y_i^*(x_{i,k})\|^2 + \frac{L_\Phi^2}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2\right).\end{aligned}$$

Taking summation on both sides over $k = 0, \ldots, K$, we know

$$\begin{aligned}&\sum_{k=0}^{K} \|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2 \\ &\quad\leq 3\left((K+1)L_{f,0}^2(1-\beta\mu)^{2M}\kappa^2 + \frac{L_f^2}{n}A_K + \frac{L_\Phi^2}{n}S_K\right),\end{aligned}$$

which completes the proof. □

The following lemma characterizes the variance of the hypergradient estimation.

Lemma 37 *Suppose β in Algorithm 2 satisfies*

$$\beta \leq \min\left(\frac{\mu}{\mu^2+\sigma_{g,2}^2},\ \frac{1}{L}\right) \tag{49}$$

Under Assumptions 2.1–2.4, we have:

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{E}\left[\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k})|\mathcal{F}_k\right] - \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k})\|^2\right] &\leq \tilde{\sigma}_f^2,\\ \mathbb{E}\left[\|\left[\overline{\partial\Phi(X_k;\phi)}\right] - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2\right] &\leq \frac{\tilde{\sigma}_f^2}{n},\end{aligned} \tag{50}$$

where the constants are defined as

$$\tilde{\sigma}_f^2 = \sigma_{f,1}^2 + \frac{2(\sigma_{g,2}^2+L^2)(\sigma_{f,1}^2+L_{f,0}^2)}{\mu^2} = \mathcal{O}(\kappa^2).$$

Proof We first notice that in the stochastic case of Algorithm 2 under Assumption 2.3, for each agent i we have

$$H_M \cdot \nabla_y f_i(x,y;\phi^{(0)}) = \beta \sum_{s=0}^{M-1}\prod_{n=1}^{s}(I - \beta\nabla_y^2 g_i(x,y;\phi^{(M+1-n)}))\nabla_y f_i(x,y;\phi^{(0)}). \tag{51}$$

For $m = 1, 2, \ldots, M-1$ we define

$$\begin{aligned}&A = \nabla_y^2 g_i(x,y),\ A_m = \nabla_y^2 g_i(x,y;\phi^{(m+1)}),\ b_0 = \nabla_y f_i(x,y;\phi^{(0)}),\\ &x_m = \beta\sum_{s=0}^{m-1}\prod_{n=1}^{s}(I-\beta A_{m-n})b_0,\ x_0 = 0,\end{aligned}$$

which gives

$$x_{m+1} = (I - \beta A_m)x_m + \beta b_0. \tag{52}$$

For simplicity in the proof of this lemma we denote by $\mathbb{E}_0$ the conditional expectation given $\phi^{(0)}$. In other words we have $\mathbb{E}_0[x] = \mathbb{E}\left[x|\phi^{(0)}\right]$ for any random vector (or matrix) x. From (52) we know

$$\|\mathbb{E}_0\left[x_m\right]\| = \beta\left\|\sum_{n=1}^{M-1}(I-\beta A)^n b_0\right\| = \left\|A^{-1}\left(I-(I-\beta A)^M\right)b_0\right\| \leq \frac{\|b_0\|}{\mu}. \tag{53}$$

Combining (52) and (53), we know

$$
\begin{aligned}
&\mathbb{E}_0\left[\|x_{m+1}-\mathbb{E}_0\left[x_{m+1}\right]\|^2\right]\\
&\quad=\mathbb{E}_0\left[\|(I-\beta A)(x_m-\mathbb{E}_0\left[x_m\right])+\beta(A-A_m)x_m\|^2\right]\\
&\quad=\mathbb{E}_0\left[\|(I-\beta A)(x_m-\mathbb{E}_0\left[x_m\right])\|^2\right]+\beta^2\mathbb{E}_0\left[\|(A-A_m)x_m\|^2\right]\\
&\quad\le(1-\beta\mu)^2\mathbb{E}_0\left[\|x_m-\mathbb{E}_0\left[x_m\right]\|^2\right]+\beta^2\sigma_{g,2}^2(\mathbb{E}_0\left[\|x_m-\mathbb{E}_0\left[x_m\right]\|^2\right]+\|\mathbb{E}_0\left[x_m\right]\|^2)\\
&\quad\le(1-\beta\mu)\mathbb{E}_0\left[\|x_m-\mathbb{E}_0\left[x_m\right]\|^2\right]+\frac{\beta^2\sigma_{g,2}^2\|b_0\|^2}{\mu^2}\\
&\quad\le(1-\beta\mu)^{m+1}\mathbb{E}\left[\|x_0-\mathbb{E}_0\left[x_0\right]\|^2\right]+\frac{\beta^2\sigma_{g,2}^2\|b_0\|^2}{\mu^2}\left(\sum_{i=0}^{m}(1-\beta\mu)^i\right)\le\frac{\beta\sigma_{g,2}^2\|b_0\|^2}{\mu^3}.
\end{aligned}
$$

The second equality uses the independence, the second inequality uses (49), and the third inequality repeats the second inequality for m times. From the above inequality we know that the variance of x_m, namely, (51), has bounded variance since

$$
\mathbb{E}\left[\|x_M-\mathbb{E}_0\left[x_M\right]\|^2\right]\le\frac{\beta\sigma_{g,2}^2\mathbb{E}\left[\|b_0\|^2\right]}{\mu^3}\le\frac{\beta\sigma_{g,2}^2(\sigma_{f,1}^2+L_{f,0}^2)}{\mu^3}\le\frac{\sigma_{f,1}^2+L_{f,0}^2}{\mu^2},
$$

where the second inequality uses Assumption 2.1 and the third inequality uses (49). We further know from the above conclusion and (53) that

$$
\begin{aligned}
&\mathbb{E}\left[\|x_M-\mathbb{E}\left[x_M\right]\|^2\right]\\
&\quad\le\mathbb{E}\left[\|x_M\|^2\right]=\mathbb{E}\left[\|x_M-\mathbb{E}_0\left[x_M\right]\|^2\right]+\mathbb{E}\left[\|\mathbb{E}_0\left[x_M\right]\|^2\right]\le\frac{2(\sigma_{f,1}^2+L_{f,0}^2)}{\mu^2}.
\end{aligned}
\tag{54}
$$

Hence in Algorithm 2 (stochastic case under Assumption 2.3) we have the following decomposition:

$$
\begin{aligned}
\hat{\nabla}f_i-\mathbb{E}\left[\hat{\nabla}f_i\right]&=\nabla_x f_i(x,y;\phi^{(0)})-\nabla_x f_i(x,y)+\nabla_{xy}g_i(x,y)\mathbb{E}\left[x_M\right]-\nabla_{xy}g_i(x,y;\phi^{(1)})x_M\\
&=\nabla_x f_i(x,y;\phi^{(0)})-\nabla_x f_i(x,y)+(\nabla_{xy}g_i(x,y)-\nabla_{xy}g_i(x,y;\phi^{(1)}))x_M\\
&\quad+\nabla_{xy}g_i(x,y)(\mathbb{E}\left[x_M\right]-x_M),
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\nabla}f_i-\mathbb{E}\left[\hat{\nabla}f_i\right]\right\|^2|x,y\right]\\
&\quad=\mathbb{E}\left[\|\nabla_x f_i(x,y;\phi^{(0)})-\nabla_x f_i(x,y)\|^2|x,y\right]\\
&\qquad+\mathbb{E}\left[\|(\nabla_{xy}g_i(x,y)-\nabla_{xy}g_i(x,y;\phi^{(1)}))x_M\|^2|x,y\right]\\
&\qquad+\mathbb{E}\left[\|\nabla_{xy}g_i(x,y)(\mathbb{E}\left[x_M\right]-x_M)\|^2|x,y\right]\\
&\quad\le\sigma_{f,1}^2+(\sigma_{g,2}^2+L^2)\mathbb{E}\left[\|x_M\|^2\right]\le\sigma_{f,1}^2+\frac{2(\sigma_{g,2}^2+L^2)(\sigma_{f,1}^2+L_{f,0}^2)}{\mu^2}=\tilde{\sigma}_f^2,
\end{aligned}
$$

where the first inequality uses the independence between different samples, the first inequality uses Assumptions 2.1 and 2.4 and the second inequality uses (54). Hence

we know the first inequality of (50) holds. Furthermore the second inequality of (50) is true since for any n independent random vectors $v_1, \ldots, v_n$ with variance bounded by σ_v^2 if we define $\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i$ we have

$$\mathbb{E}\left[\|\bar{v} - \mathbb{E}[\bar{v}]\|^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\|v_i - \mathbb{E}\left[v_i\right]\|^2\right] \le \frac{\sigma_v^2}{n}.$$

□

The following lemmas give the estimation bound of A_K and S_K in the stochastic case.

Lemma 38 *In Algorithm 5, we have*

$$\mathbb{E}\left[S_K\right] < \frac{\eta_x^2}{(1-\rho)^2}\sum_{j=0}^{K-1}\sum_{i=1}^n \mathbb{E}\left[\|\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)}; \phi_{i,j})\|^2\right] \le \frac{\eta_x^2 nK}{(1-\rho)^2}\tilde{C}_f^2,$$

where the constant is defined as

$$\tilde{C}_f^2 = \left(L_{f,0} + \frac{LL_{f,1}}{\mu} + \frac{LL_{f,1}}{\mu}\right)^2 + \tilde{\sigma}_f^2 = \mathcal{O}(\kappa^2).$$

Proof Observe that in this stochastic case, we can replace $\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)})$ with $\hat{\nabla} f_i(x_{i,j}, y_{i,j}^{(T)}; \phi_{i,j})$ in Lemma 20 to get the first inequality. For the second inequality, we adopt the bound in Lemma 2 of [14]. □

Lemma 39 *Set parameters in Algorithm 5 as*

$$\eta_y < \frac{2}{\mu + L}, \quad \delta_y^T \le \frac{1}{3}. \tag{55}$$

Then we have the following inequalities

$$\mathbb{E}\left[A_K\right] \le \delta_y^T(2\mathbb{E}\left[c_1\right] + 6\kappa^2\mathbb{E}\left[E_K\right]) + \frac{\eta_y nK\sigma_{g,1}^2}{\mu}, \ \mathbb{E}\left[E_K\right] \le \frac{9n\eta_x^2 K\tilde{C}_f^2}{(1-\rho)^2}.$$

Proof The proof is based on Lemma 10. Taking conditional expectation with respect to the filtration $\mathcal{G}_{i,j}^{(t-1)}$, we get

$$\begin{aligned}
&\mathbb{E}\left[\|y_{i,j}^{(t)} - y_i^*(x_{i,j})\|^2 | \mathcal{G}_{i,j}^{(t-1)}\right] \\
&= \mathbb{E}\left[\|y_{i,j}^{(t-1)} - \eta_y \nabla_y g(x_{i,j}, y_{i,j}^{(t-1)}; \xi_{i,j}^{(t-1)}) - y_i^*(x_{i,j})\|^2 | \mathcal{G}_{i,j}^{(t-1)}\right] \\
&= \|y_{i,j}^{(t-1)} - \eta_y \nabla_y g(x_{i,j}, y_{i,j}^{(t-1)}) - y_i^*(x_{i,j})\|^2 \\
&\quad + \eta_y^2 \mathbb{E}\left[\|\nabla_y g(x_{i,j}, y_{i,j}^{(t-1)}) - \nabla_y g(x_{i,j}, y_{i,j}^{(t-1)}; \xi_{i,j}^{(t-1)})\|^2 | \mathcal{G}_{i,j}^{(t-1)}\right] \\
&\le (1 - \eta_y\mu)^2 \|y_{i,j}^{(t-1)} - y_i^*(x_{i,j})\|^2 + \eta_y^2\sigma_{g,1}^2,
\end{aligned}$$

where the inequality uses Lemma 3. Taking expectation on both sides and using the tower property, we have

$$\begin{aligned}
&\mathbb{E}\left[\|y_{i,j}^{(T)} - y_i^*(x_{i,j})\|^2\right] \\
&\le (1-\eta_y\mu)^2\mathbb{E}\left[\|y_{i,j}^{(T-1)} - y_i^*(x_{i,j})\|^2\right] + \eta_y^2\sigma_{g,1}^2 \\
&\le (1-\eta_y\mu)^{2T}\mathbb{E}\left[\|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2\right] + \eta_y^2\sigma_{g,1}^2\sum_{s=0}^{T-1}(1-\eta_y\mu)^{2s} \\
&\le \delta_y^T\mathbb{E}\left[\|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2\right] + \frac{\eta_y\sigma_{g,1}^2}{\mu}.
\end{aligned} \tag{56}$$

Moreover, by the warm-start strategy, we have $y_{i,j}^{(0)} = y_{i,j-1}^{(T)}$ and thus

$$\begin{aligned}
&\mathbb{E}\left[\|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2\right] \\
&= \mathbb{E}\left[\|y_{i,j-1}^{(T)} - y_i^*(x_{i,j-1}) + y_i^*(x_{i,j-1}) - y_i^*(x_{i,j})\|^2\right] \\
&\le 2\mathbb{E}\left[\|y_{i,j-1}^{(T)} - y_i^*(x_{i,j-1})\|^2\right] + 2\mathbb{E}\left[\|y_i^*(x_{i,j-1}) - y_i^*(x_{i,j})\|^2\right] \\
&\le 2\delta_y^T\mathbb{E}\left[\|y_{i,j-1}^{(0)} - y_i^*(x_{i,j-1})\|^2\right] + 2\kappa^2\mathbb{E}\left[\|x_{i,j-1} - x_{i,j}\|^2\right] \\
&\le \frac{2}{3}\mathbb{E}\left[\|y_{i,j-1}^{(0)} - y_i^*(x_{i,j-1})\|^2\right] + 2\kappa^2\mathbb{E}\left[\|x_{i,j-1} - x_{i,j}\|^2\right],
\end{aligned} \tag{57}$$

where the second inequality is by Lemma 7 and (57) and the last inequality is by (55). Taking summation over i, j, we have:

$$\begin{aligned}
&\sum_{j=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2\right] \\
&\le \frac{2}{3}\sum_{j=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\|y_{i,j-1}^{(0)} - y_i^*(x_{i,j-1})\|^2\right] + 2\kappa^2\mathbb{E}\left[E_K\right] \\
&\le \frac{2}{3}\mathbb{E}\left[c_1\right] + \frac{2}{3}\sum_{j=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2\right] + 2\kappa^2\mathbb{E}\left[E_K\right],
\end{aligned}$$

which leads to

$$\sum_{j=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2\right] \le 2\mathbb{E}\left[c_1\right] + 6\kappa^2\mathbb{E}\left[E_K\right]. \tag{58}$$

Combining (58) with (56) and taking summation over i, j, we have

$$\mathbb{E}\left[A_K\right] \leq \delta_y^T \sum_{j=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\|y_{i,j}^{(0)} - y_i^*(x_{i,j})\|^2\right] + \frac{\eta_y nK\sigma_{g,1}^2}{\mu}$$
$$\leq \delta_y^T(2\mathbb{E}\left[c_1\right] + 6\kappa^2\mathbb{E}\left[E_K\right]) + \frac{\eta_y nK\sigma_{g,1}^2}{\mu}.$$

Recall that for E_K we have:

$$\begin{aligned}
E_K &= \sum_{j=1}^{K}\sum_{i=1}^{n}\|x_{i,j} - x_{i,j-1}\|^2 \\
&= \sum_{j=1}^{K}\sum_{i=1}^{n}\|x_{i,j} - \bar{x}_j + \bar{x}_j - \bar{x}_{j-1} + \bar{x}_{j-1} - x_{i,j-1}\|^2 \\
&= \sum_{j=1}^{K}\sum_{i=1}^{n}\|q_{i,j} - \eta_x\overline{\partial\Phi(X_{j-1};\phi)} - q_{i,j-1}\|^2 \\
&\leq 3\sum_{j=1}^{K}\sum_{i=1}^{n}(\|q_{i,j}\|^2 + \eta_x^2\|\overline{\partial\Phi(X_{j-1};\phi)}\|^2 + \|q_{i,j-1}\|^2) \\
&\leq 3\sum_{j=1}^{K}(\|Q_j\|^2 + \|Q_{j-1}\|^2 + \eta_x^2\|\overline{\partial\Phi(X_{j-1};\phi)}\|^2 \\
&\leq 6S_K + \frac{3\eta_x^2}{n}\sum_{j=0}^{K-1}\sum_{i=1}^{n}\|\hat{\nabla}f_i(x_{i,j}, y_{i,j}^{(T)};\phi_{i,j})\|^2.
\end{aligned}$$

Taking expectation on both sides yields

$$\mathbb{E}\left[E_K\right] \leq 6\mathbb{E}\left[S_K\right] + \frac{3\eta_x^2}{n}\sum_{j=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left[\|\hat{\nabla}f_i(x_{i,j}, y_{i,j}^{(T)};\phi_{i,j})\|^2\right]$$
$$\leq \frac{6\eta_x^2 nK}{(1-\rho)^2}\tilde{C}_f^2 + 3\eta_x^2 K\tilde{C}_f^2 = \frac{9n\eta_x^2 K\tilde{C}_f^2}{(1-\rho)^2},$$

which completes the proof. □

Next, we prove the main convergence results in Theorem 33. Taking expectation on both sides in (48), we have:

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right] \\
&\leq 3\left(L_{f,0}^2(1-\beta\mu)^{2M}\kappa^2 + \frac{L_f^2}{n(K+1)}\mathbb{E}\left[A_K\right] + \frac{L_\Phi^2}{n(K+1)}\mathbb{E}\left[S_K\right]\right) \qquad (59) \\
&\leq C_3 + \frac{3\eta_y L_f^2\sigma_{g,1}^2}{\mu} + \frac{3\eta_x^2 L_\Phi^2}{(1-\rho)^2}\tilde{C}_f^2,
\end{aligned}$$

where the constant is defined as:

$$\begin{aligned} C_3 &= 3L_{f,0}^2(1-\beta\mu)^{2M}\kappa^2 + \frac{3L_f^2}{n(K+1)}\delta_y^T(2\mathbb{E}\left[c_1\right] + 6\kappa^2\mathbb{E}\left[E_K\right]) \\ &\le 3L_{f,0}^2(1-\beta\mu)^{2M}\kappa^2 + \frac{3L_f^2}{n(K+1)}\delta_y^T\left(2\mathbb{E}\left[c_1\right] + \frac{54\kappa^2 n\eta_x^2 K\tilde{C}_f^2}{(1-\rho)^2}\right) \\ &= \Theta(\delta_\beta^M\kappa^2 + \eta_x^2\delta_y^T\kappa^8). \end{aligned}$$

Here we denote $\delta_\beta = (1-\beta\mu)^2$ for simplicity. Therefore, we set $M = \Theta(\log K)$ and $T = \Theta(\log\kappa)$ such that $C_3 = \Theta(\eta_x^2 + \frac{1}{K+1})$. Recall that (45) yields:

$$\begin{aligned} &\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] \\ &\le \frac{2}{\eta_x}\left(\mathbb{E}\left[\Phi(\bar{x}_k)\right] - \mathbb{E}\left[\Phi(\bar{x}_{k+1})\right]\right) + \mathbb{E}\left[\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2\right] \\ &\quad + L\eta_x\mathbb{E}\|\left[\overline{\partial\Phi(X_k;\phi)}\right] - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2. \end{aligned}$$

Taking summation on both sides and using (59) and Lemma 37, we have

$$\begin{aligned} \frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] \le \frac{2}{\eta_x(K+1)}(\mathbb{E}\left[\Phi(\bar{x}_0)\right] - \inf_x\Phi(x)) + \frac{3\eta_y L_f^2\sigma_{g,1}^2}{\mu} \\ + \frac{3\eta_x^2 L_\Phi^2}{(1-\rho)^2}\tilde{C}_f^2 + \frac{L\eta_x\tilde{\sigma}_f^2}{n} + C_3. \end{aligned}$$

By setting

$$M = \Theta(\log K),\ T = \Theta(K^{\frac{1}{2}}),\ \eta_x = \Theta(K^{-\frac{1}{2}}),\ \eta_y = \Theta(K^{-\frac{1}{2}})$$

we know that the restrictions on algorithm parameters in Lemmas 35, 37, and 39 hold and we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right),$$

which proves the first case of Theorems 3.3 and 33.

Case 2: Assumption 2.3 does not hold

Lemma 40 *Suppose the Assumption 2.3 does not hold in Algorithm 5, we have*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2\right]$$
$$\leq 12\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{C}{T}+12CL_{f,0}^2\alpha^N$$
$$+\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2}+2L_f^2\right)\frac{\eta_x^2(1+\kappa^2)}{(1-\rho)^2}\tilde{C}_f^2.$$

Proof Denote by $\hat{Z}_{i,k}^{(N)}$ the output of each stochastic JHIP oracle 1 in Algorithm 5. Then

$$\mathbb{E}\left[\hat{Z}_{i,k}^{(N)}\right]=Z_{i,k}^{(N)},$$

which implies

$$\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]=\overline{\partial\Phi(X_k)}.$$

Hence we can follow the same process in case 2 of DBO to get (24) and thus

$$\sum_{k=0}^{K}\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]-\nabla\Phi(\bar{x}_k)\|^2=\sum_{k=0}^{K}\|\overline{\partial\Phi(X_k)}-\nabla\Phi(\bar{x}_k)\|^2$$
$$\leq 12\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{1}{n}\sum_{k=0}^{K}\sum_{i=1}^{n}\|y_{i,k}^{(T)}-\tilde{y}_k^*\|^2$$
$$+\frac{12L_{f,0}^2}{n}\sum_{k=0}^{K}\sum_{i=1}^{n}\|Z_{i,k}^{(N)}-Z_k^*\|^2+\frac{(1+\kappa^2)}{n}\cdot\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2}+2L_f^2\right)S_K.$$
$$\leq 12\left(1+L^2\kappa^2+\frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{(K+1)C}{T}+12(K+1)CL_{f,0}^2\alpha^N$$
$$+\frac{(1+\kappa^2)}{n}\cdot\left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2}+2L_f^2\right)S_K.$$

The second inequality uses Lemmaa 14 and 15. Taking expectation, multiplying by $\frac{1}{K+1}$, and using Lemma 38 we complete the proof. □

The next lemma characterizes the variance of the gradient estimation.

Lemma 41 *Suppose the Assumption 2.3 does not hold in Algorithm 5, then there exists* $\gamma_t=\mathcal{O}\left(\frac{1}{t}\right)$ *such that*

$$\mathbb{E}\|\overline{\partial\Phi(X_k;\phi)}-\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2\leq 4\sigma_f^2(1+\kappa^2)+(8L_{f,0}^2+4\sigma_f^2)\frac{C}{N}.$$

Proof Recall that we have:

$$\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi) = \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)}) - \left[\hat{Z}_{i,k}^{(N)}\right]^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)})$$
$$\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) = \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \left(Z_{i,k}^{(N)}\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}).$$

By introducing intermediate terms we have

$$\begin{aligned}
&\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi) - \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}) \\
&= \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)}) - \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}) - \left[\hat{Z}_{i,k}^{(N)}\right]^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)}) \\
&\quad + \left(Z_k^*\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)}) - \left(Z_k^*\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)}) \\
&\quad + \left(Z_k^*\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}) - \left(Z_k^*\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}) + \left(Z_{i,k}^{(N)}\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}).
\end{aligned}$$

Hence we know

$$\begin{aligned}
&\|\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi) - \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)})\|^2 \\
&\le 4\|\nabla_x f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)}) - \nabla_x f_i(x_{i,k}, y_{i,k}^{(T)})\|^2 \\
&\quad + 4\|\left(\hat{Z}_{i,k}^{(N)} - Z_k^*\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)})\|^2 \\
&\quad + 4\|\left(Z_k^*\right)^\top (\nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}; \phi_{i,k}^{(0)}) - \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)}))\|^2 \\
&\quad + 4\|\left(Z_k^* - Z_{i,k}^{(N)}\right)^\top \nabla_y f_i(x_{i,k}, y_{i,k}^{(T)})\|^2.
\end{aligned}$$

For the first term and the third term we use $\mathbb{E}\left[\|\nabla f_i(x, y; \phi) - \nabla f_i(x, y)\|^2\right] \le \sigma_f^2$. For the second term (and the fourth term) we use the fact that stochastic (and deterministic) decentralized algorithm achieves sublinear rate (Lemma 15). Without loss of generality we can set C such that: $\max\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\hat{Z}_{i,k}^{(N)} - Z_k^*\|^2\right], \|Z_{i,k}^{(N)} - Z_k^*\|^2\right) \le \frac{C}{N}$. For partial gradients in the second and fourth terms, we use Assumption 2.1 and the fact that

$$\mathbb{E}\left[\|X\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] + \|\mathbb{E}[X]\|^2$$

for any random vector X. Taking summation and expectation on both sides, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)}; \phi) - \hat{\nabla} f_i(x_{i,k}, y_{i,k}^{(T)})\|^2\right] \\
&\le 4\sigma_f^2 + 4(L_{f,0}^2 + \sigma_f^2)\frac{C}{N} + \frac{4L^2}{\mu^2}\sigma_f^2 + 4L_{f,0}^2\frac{C}{N},
\end{aligned}$$

which, together with

$$\mathbb{E}\|\left[\overline{\partial\Phi(X_k;\phi)}\right] - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\hat{\nabla}f_i(x_{i,k},y_{i,k}^{(T)};\phi) - \hat{\nabla}f_i(x_{i,k},y_{i,k}^{(T)})\|^2\right],$$

proves the lemma. □

Now we are ready to give the final proof. Taking summation on both sides of (45) and putting Lemma 40 and 41 together we know:

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right] \\
&\leq \frac{1}{K+1}\Big(\frac{2}{\eta_x}(\mathbb{E}\left[\Phi(\overline{x_0})\right] - \inf_x \Phi(x)) + \sum_{k=0}^{K}\mathbb{E}[\|\mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right] - \nabla\Phi(\bar{x}_k)\|^2]\Big) \\
&\quad + \frac{L\eta_x}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\left[\overline{\partial\Phi(X_k;\phi)}\right] - \mathbb{E}\left[\overline{\partial\Phi(X_k;\phi)}|\mathcal{F}_k\right]\|^2 \\
&\leq \frac{2}{\eta_x(K+1)}(\mathbb{E}\left[\Phi(\overline{x_0})\right] - \inf_x \Phi(x)) \\
&\quad + 12\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{C}{T} + 12CL_{f,0}^2\alpha^N \\
&\quad + \left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2} + 2L_f^2\right)\frac{\eta_x^2(1+\kappa^2)}{(1-\rho)^2}\tilde{C}_f^2 + L\eta_x\left(4\sigma_f^2(1+\kappa^2) + (8L_{f,0}^2 + 4\sigma_f^2)\frac{C}{N}\right) \\
&= \frac{2}{\eta_x(K+1)}(\mathbb{E}\left[\Phi(\overline{x_0})\right] - \inf_x \Phi(x)) + \left(\frac{36L_{f,0}^2L_{g,2}^2}{\mu^2} + 2L_f^2\right)\frac{\eta_x^2(1+\kappa^2)}{(1-\rho)^2}\tilde{C}_f^2 \\
&\quad + L\eta_x\left(4\sigma_f^2(1+\kappa^2) + (8L_{f,0}^2 + 4\sigma_f^2)\frac{C}{N}\right) + \tilde{C}_3,
\end{aligned}$$

which completes the proof. Here the constant is defined as

$$\tilde{C}_3 = 12\left(1 + L^2\kappa^2 + \frac{2L_{f,0}^2L_{g,2}^2(1+\kappa^2)}{\mu^2}\right)\cdot\frac{C}{T} + 12CL_{f,0}^2\alpha^N = \mathcal{O}\left(\frac{1}{T} + \alpha^N\right).$$

By setting

$$N = \Theta(\log K),\ T = \Theta(K^{\frac{1}{2}}),\ \eta_x = \Theta(K^{-\frac{1}{2}}),\ \eta_y^{(t)} = \mathcal{O}\left(\frac{1}{t}\right),$$

we have:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\|\nabla\Phi(\bar{x}_k)\|^2\right]=\mathcal{O}\left(\frac{1}{\sqrt{K}}\right),$$

which proves the second case of Theorems 3.3 and 33.

Acknowledgements We would like to thank the Guest Editor and two anonymous reviewers whose insightful comments have helped improve the presentation of this paper. The research of Shiqian Ma was supported in part by NSF Grants DMS-2243650, CCF-2308597, CCF-2311275 and ECCS-2326591, UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program, and a startup fund from Rice University.

Data availability All data sets used in this paper are publicly available.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Yang, S., Zhang, X., Wang, M.: Decentralized gossip-based stochastic bilevel optimization over communication networks. arXiv:2206.10870 (2022)
2. Gao, H., Gu, B., Thai, M.T.: Stochastic bilevel distributed optimization over a network. arXiv:2206.15025 (2022)
3. Terashita, N., Hara, S.: Personalized decentralized bilevel optimization over stochastic and directed networks. arXiv:2210.02129 (2022)
4. Chen, X., Huang, M., Ma, S., Balasubramanian, K.: Decentralized stochastic bilevel optimization with improved per-iteration complexity. In: International Conference on Machine Learning, pp. 4641–4671. PMLR (2023)
5. Jiao, Y., Yang, K., Wu, T., Song, D., Jian, C.: Asynchronous distributed bilevel optimization. arXiv:2212.10048 (2022)
6. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Adv. Neural Inf. Process. Syst. **30** (2017)
7. Bertinetto, L., Henriques, J.F., Torr, P.H., Vedaldi, A.: Meta-learning with differentiable closed-form solvers. arXiv:1805.08136 (2018)
8. Rajeswaran, A., Finn, C., Kakade, S.M., Levine, S.: Meta-learning with implicit gradients. Adv. Neural Inf. Process. Syst. **32** (2019)
9. Pedregosa, F.: Hyperparameter optimization with approximate gradient. In: International Conference on Machine Learning, pp. 737–746. PMLR (2016)
10. Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., Pontil, M.: Bilevel programming for hyperparameter optimization and meta-learning. In: International Conference on Machine Learning, pp. 1568–1577. PMLR (2018)
11. Hong, M., Wai, H.-T., Wang, Z., Yang, Z.: A two-timescale framework for bilevel optimization: complexity analysis and application to actor-critic. arXiv:2007.05170 (2020)
12. Ghadimi, S., Wang, M.: Approximation methods for bilevel programming. arXiv:1802.02246 (2018)
13. Ji, K., Yang, J., Liang, Y.: Bilevel optimization: convergence analysis and enhanced design. In: International Conference on Machine Learning, pp. 4882–4892. PMLR (2021)
14. Chen, T., Sun, Y., Yin, W.: Closing the gap: tighter analysis of alternating stochastic gradient methods for bilevel problems. Adv. Neural Inf. Process. Syst. **34**, 25294–25307 (2021)
15. Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., Liu, J.: Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. Adv. Neural Inf. Process. Syst. **30**, 5336–5346 (2017)

16. Tang, H., Lian, X., Yan, M., Zhang, C., Liu, J.: d^2: decentralized training over decentralized data. In: International Conference on Machine Learning, pp. 4848–4856. PMLR (2018)
17. Stackelberg, H.V.: Theory of the Market Economy (1952)
18. Bracken, J., McGill, J.T.: Mathematical programs with optimization problems in the constraints. Oper. Res. **21**(1), 37–44 (1973)
19. Bennett, K., Ji, X., Hu, J., Kunapuli, G., Pang, J.: Model selection via bilevel programming. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN'06) Vancouver, BC Canada, pp. 1922–1929 (2006)
20. Kunapuli, G., Bennett, K., Hu, J., Pang, J.-S.: Bilevel model selection for support vector machines. In: CRM Proceedings and Lecture Notes, vol. 45, pp. 129–158 (2008)
21. Kunapuli, G., Bennett, K.P., Hu, J., Pang, J.-S.: Classification model selection via bilevel programming. Optim. Methods Softw. **23**(4), 475–489 (2008)
22. Domke, J.: Generic methods for optimization-based modeling. In: Artificial Intelligence and Statistics, pp. 318–326. PMLR (2012)
23. Gould, S., Fernando, B., Cherian, A., Anderson, P., Cruz, R.S., Guo, E.: On differentiating parameterized argmin and argmax problems with application to bi-level optimization. arXiv:1607.05447 (2016)
24. Grazzi, R., Franceschi, L., Pontil, M., Salzo, S.: On the iteration complexity of hypergradient computation. In: International Conference on Machine Learning, pp. 3748–3758. PMLR (2020)
25. Maclaurin, D., Duvenaud, D., Adams, R.: Gradient-based hyperparameter optimization through reversible learning. In: International Conference on Machine Learning, pp. 2113–2122. PMLR (2015)
26. Chen, T., Sun, Y., Xiao, Q., Yin, W.: A single-timescale method for stochastic bilevel optimization. In: International Conference on Artificial Intelligence and Statistics, pp. 2466–2488. PMLR (2022)
27. Guo, Z., Hu, Q., Zhang, L., Yang, T.: Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. arXiv:2105.02266 (2021)
28. Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., Yang, Z.: A near-optimal algorithm for stochastic bilevel optimization via double-momentum. Adv. Neural Inf. Process. Syst. **34**, 30271–30283 (2021)
29. Yang, J., Ji, K., Liang, Y.: Provably faster algorithms for bilevel optimization. Adv. Neural Inf. Process. Syst. **34**, 13670–13682 (2021)
30. Gan, S., Lian, X., Wang, R., Chang, J., Liu, C., Shi, H., Zhang, S., Li, X., Sun, T., Jiang, J., et al.: Bagua: scaling up distributed learning with system relaxations. arXiv:2107.01499 (2021)
31. Yuan, B., He, Y., Davis, J., Zhang, T., Dao, T., Chen, B., Liang, P.S., Re, C., Zhang, C.: Decentralized training of foundation models in heterogeneous environments. Adv. Neural Inf. Process. Syst. **35**, 25464–25477 (2022)
32. Xu, J., Zhu, S., Soh, Y.C., Xie, L.: Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In: 2015 54th IEEE Conference on Decision and Control (CDC), pp. 2055–2060. IEEE (2015)
33. Di Lorenzo, P., Scutari, G.: Next: in-network nonconvex optimization. IEEE Trans. Signal Inf. Process. Netw. **2**(2), 120–136 (2016)
34. Qu, G., Li, N.: Harnessing smoothness to accelerate distributed optimization. IEEE Trans. Control Netw. Syst. **5**(3), 1245–1260 (2017)
35. Nedic, A., Olshevsky, A., Shi, W.: Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM J. Optim. **27**(4), 2597–2633 (2017)
36. Xin, R., Kar, S., Khan, U.A.: Decentralized stochastic optimization and machine learning: a unified variance-reduction framework for robust performance and fast convergence. IEEE Signal Process. Mag. **37**(3), 102–113 (2020)
37. Altae-Tran, H., Ramsundar, B., Pappu, A.S., Pande, V.: Low data drug discovery with one-shot learning. ACS Cent. Sci. **3**(4), 283–293 (2017)
38. Zhang, X.S., Tang, F., Dodge, H.H., Zhou, J., Wang, F.: Metapred: meta-learning for clinical risk prediction with limited patient electronic health records. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2487–2495 (2019)
39. Kayaalp, M., Vlaski, S., Sayed, A.H.: Dif-MAML: decentralized multi-agent meta-learning. IEEE Open J. Signal Process. **3**, 71–93 (2022)
40. Tarzanagh, D.A., Li, M., Thrampoulidis, C., Oymak, S.: Fednest: Federated bilevel, minimax, and compositional optimization. arXiv:2205.02215 (2022)

41. Li, J., Huang, F., Huang, H.: Local stochastic bilevel optimization with momentum-based variance reduction. arXiv: 2205.01608 (2022)
42. Xian, W., Huang, F., Zhang, Y., Huang, H.: A faster decentralized algorithm for nonconvex minimax problems. Adv. Neural Inf. Process. Syst. **34**, 25865–25877 (2021)
43. Luo, L., Ye, H.: Decentralized stochastic variance reduced extragradient method. arXiv:2202.00509 (2022)
44. Sharma, P., Panda, R., Joshi, G., Varshney, P.K.: Federated minimax optimization: improved convergence analyses and algorithms. arXiv:2203.04850 (2022)
45. Lu, S., Cui, X., Squillante, M.S., Kingsbury, B., Horesh, L.: Decentralized bilevel optimization for personalized client learning. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5543–5547. IEEE (2022)
46. Pu, S., Nedić, A.: Distributed stochastic gradient tracking methods. Math. Program. **187**(1), 409–457 (2021)
47. Mota, J.F., Xavier, J.M., Aguiar, P.M., Püschel, M.: D-ADMM: a communication-efficient distributed algorithm for separable optimization. IEEE Trans. Signal process. **61**(10), 2718–2723 (2013)
48. Chang, T.-H., Hong, M., Wang, X.: Multi-agent distributed optimization via inexact consensus ADMM. IEEE Trans. Signal Process. **63**(2), 482–497 (2014)
49. Shi, W., Ling, Q., Yuan, K., Wu, G., Yin, W.: On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans. Signal Process. **62**(7), 1750–1761 (2014)
50. Aybat, N.S., Wang, Z., Lin, T., Ma, S.: Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. IEEE Trans. Autom. Control **63**(1), 5–20 (2017)
51. Makhdoumi, A., Ozdaglar, A.: Convergence rate of distributed ADMM over networks. IEEE Trans. Autom. Control **62**(10), 5082–5095 (2017)
52. Koloskova, A., Lin, T., Stich, S.U.: An improved analysis of gradient tracking for decentralized machine learning. Adv. Neural Inf. Process. Syst. **34**, 11422–11435 (2021)
53. Shaban, A., Cheng, C.-A., Hatch, N., Boots, B.: Truncated back-propagation for bilevel optimization. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 1723–1732. PMLR (2019)
54. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
55. Olshevsky, A., Paschalidis, I.C., Pu, S.: A non-asymptotic analysis of network independence for distributed stochastic gradient descent. arXiv:1906.02702 (2019)