Decentralized Bilevel Optimization: A Perspective from Transient Iteration Complexity

Boao Kong[*]   KONGBOAO@STU.PKU.EDU.CN
Center for Data Science, Peking University
Beijing, China

Shuchen Zhu[*]   SHUCHENZHU@STU.PKU.EDU.CN
Center for Data Science, Peking University
Beijing, China

Songtao Lu   STLU@CSE.CUHK.EDU.HK
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Hong Kong SAR

Xinmeng Huang   XINMENGH@SAS.UPENN.EDU
Graduate Group in Applied Mathematics and Computational Science
University of Pennsylvania
Philadelphia, USA

Kun Yuan[†]   KUNYUAN@PKU.EDU.CN
Center for Machine Learning Research, Peking University
AI for Science Institute
Beijing, China

Editor: Shiqian Ma

Abstract

Stochastic bilevel optimization (SBO) is becoming increasingly essential in machine learning due to its versatility in handling nested structures. To address large-scale SBO, decentralized approaches have emerged as effective paradigms in which nodes communicate with immediate neighbors without a central server, thereby improving communication efficiency and enhancing algorithmic robustness. However, most decentralized SBO algorithms focus solely on asymptotic convergence rates, overlooking transient iteration complexity-the number of iterations required before asymptotic rates dominate, which results in limited understanding of the influence of network topology, data heterogeneity, and the nested bilevel algorithmic structures. To address this issue, this paper introduces **D-SOBA**, a **D**ecentralized **S**tochastic **O**ne-loop **B**ilevel **A**lgorithm framework. D-SOBA comprises two variants: D-SOBA-SO, which incorporates second-order Hessian and Jacobian matrices, and D-SOBA-FO, which relies entirely on first-order gradients. We provide a comprehensive non-asymptotic convergence analysis and establish the transient iteration complexity of D-SOBA. This provides the first theoretical understanding of how network topology, data heterogeneity, and nested bilevel structures influence decentralized SBO. Extensive experimental results demonstrate the efficiency and theoretical advantages of D-SOBA.

---

[*]. Equal contribution.

[†]. Corresponding author.

Keywords: bilevel optimization, decentralized optimization, convergence rate, transient iteration complexity, non-asymptotic analysis.

1. Introduction

Stochastic bilevel optimization, which tackles problems with nested optimization structures, has gained growing interest. This two-level structure provides a flexible and potent framework for addressing a broad range of tasks, spanning from hyperparameter optimization (Bertinetto et al., 2018; Franceschi et al., 2018; Snell et al., 2017) to reinforcement learning (Hong et al., 2023), adversarial learning (Zhang et al., 2022), continue learning (Borsos et al., 2020), and imitation learning (Arora et al., 2020). State-of-the-art performance in these tasks is typically achieved with extremely large datasets, which necessitates efficient distributed algorithms for stochastic bilevel optimization across multiple computing nodes.

This paper considers N nodes connected through a given graph topology. Each node i privately owns an upper-level loss function $f_i : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ and a lower-level loss function $g_i : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$. The goal of all nodes is to collaboratively find a solution to the following distributed stochastic bilevel optimization problem:

$$\min_{x \in \mathbb{R}^d} \quad \Phi(x) := f(x, y^\star(x)) := \frac{1}{N} \sum_{i=1}^{N} f_i\left(x, y^\star(x)\right) \tag{1a}$$

$$\text{s.t.} \quad y^\star(x) := \arg\min_{y \in \mathbb{R}^p} \left\{ g(x, y) := \frac{1}{N} \sum_{i=1}^{N} g_i(x, y) \right\} \tag{1b}$$

where f_i and g_i are defined as the expectation of $F(x, y; \xi_i)$ and $G(x, y; \zeta_i)$:

$$f_i(x, y) := \mathbb{E}_{\xi_i \sim \mathcal{D}_{f_i}}[F(x, y; \xi_i)], \quad g_i(x, y) := \mathbb{E}_{\zeta_i \sim \mathcal{D}_{g_i}}[G(x, y; \zeta_i)]. \tag{2}$$

The random variables ξ_i and ζ_i represent data samples available at node i, drawn from the local upper-level distribution $\mathcal{D}_{f_i}$ and the lower-level distribution $\mathcal{D}_{g_i}$, respectively. Throughout this paper, we assume that these local data distributions vary across nodes, potentially causing data heterogeneity issues during training.

1.1 Decentralized Bilevel Algorithms

Conventional distributed approaches for solving problem (1) typically follow a centralized paradigm (Huang et al., 2023; Tarzanagh et al., 2022; Yang et al., 2023b), where a central server synchronizes updates across the network by computing a globally averaged upper- or lower-level stochastic gradient to update the model parameters. This global averaging step is commonly implemented using mechanisms such as Parameter Server (Andersen and Park, 2014; Smola and Narayanamurthy, 2010) or Ring-Allreduce (Gibiansky, 2017). However, these techniques incur substantial bandwidth costs or high latency (Ying et al., 2021b, Table I), which severely limits the scalability of centralized bilevel optimization.

Decentralized bilevel optimization is a new paradigm to approach problem (1) eliminating the global averaging step, wherein each node maintains a local model updated by communicating with its immediate neighbors. This distinctive feature results in a noteworthy reduction in communication costs compared to centralized approaches. Beyond the communication

efficiency, decentralized bilevel algorithms also exhibit enhanced robustness to node and link failures, persisting in converging to the desired solution as long as the graph remains connected. In contrast, centralized bilevel algorithms inevitably break down when the central server crashes. For these reasons, considerable research efforts, *e.g.*, (Chen et al., 2022, 2023a; Dong et al., 2023; Gao et al., 2023; Liu et al., 2022b; Lu et al., 2022b; Niu et al., 2023; Yang et al., 2022; Zhang et al., 2023a), have been dedicated to developing decentralized bilevel algorithms with theoretical guarantees and empirical effectiveness.

1.2 Limitations and Open Questions

Limitations in existing literature. Significant progress has been made in developing decentralized stochastic bilevel algorithms, with most studies focusing on their *asymptotic* convergence rates, which are achieved after a sufficiently large number of iterations. References (Lu et al., 2022b; Yang et al., 2022) demonstrate that decentralized bilevel algorithms attain the same asymptotic convergence rates as their centralized bilevel counterparts or decentralized single-level algorithms, highlighting their strong theoretical guarantees. However, an exclusive focus on asymptotic convergence rates presents several critical limitations:

L1. In practical scenarios, computational constraints—such as limited time and resource budgets—often restrict the number of iterations that can be performed, preventing algorithms from reaching their asymptotic stage. It remains unknown the number of transient iterations required before the algorithm attains its asymptotic rate.

L2. Asymptotic convergence rates offer limited insight into the impact of nested algorithmic structures, network topologies, and data heterogeneity on decentralized bilevel algorithms, as these factors are typically associated with the algorithms' transient convergence stage rather than its asymptotic steady-state rate.

L3. Asymptotic convergence rates are inadequate as a metric for distinguishing between various decentralized bilevel algorithms, as most existing methods share the same asymptotic convergence rate (Lu et al., 2022b; Yang et al., 2022). This makes it challenging to determine which algorithm delivers superior theoretical performance.

Open questions. To address these limitations and advance our understanding of decentralized bilevel optimization, several fundamental open questions need to be answered:

Q1. Can we quantify the transient iteration complexity of decentralized stochastic bilevel algorithms—that is, the number of iterations needed for the asymptotic convergence rate to dominate? Furthermore, can transient iteration complexity serve as a criterion to differentiate between various decentralized bilevel algorithms?

Q2. Can we elucidate how network topologies and data heterogeneity influence transient iteration complexity? What benefits can we achieve by enhancing network connectivity and reducing data heterogeneity?

Q3. Can we clarify how the nested bilevel algorithmic structure affects transient iteration complexity? Does the lower-level optimization introduce substantial challenges to the non-asymptotic phase?

1.3 Main Results

This paper proposes a new decentralized stochastic bilevel optimization method, focusing on its non-asymptotic convergence analysis and transient iteration complexity. This case study sheds light on the aforementioned open questions.

Transient iteration complexity. Before presenting the main results, we will formally introduce the concept of transient iteration complexity. Although stochastic decentralized bilevel algorithms can asymptotically achieve the same linear speedup rate of $\mathcal{O}(1/\sqrt{NT})$ as their centralized counterparts (Gao et al., 2023; Lu et al., 2022b; Yang et al., 2022), they typically require more iterations to reach this stage due to inexact averaging in decentralized communication settings. *Transient iteration complexity* refers to the number of iterations a decentralized algorithm undergoes before the asymptotic convergence rate becomes dominant, as illustrated in Fig. 1. A lower transient complexity indicates that the decentralized
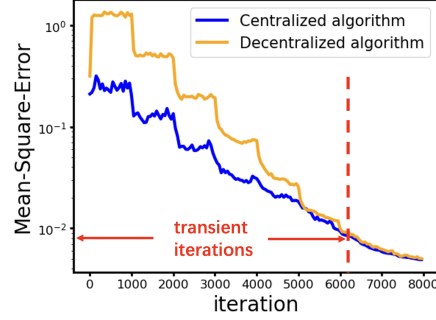

Figure 1: Decentralized algorithm (with stage-wise decaying learning rate) has to experience sufficiently massive transient iterations to achieve the same asymptotic rate as centralized approach.

approach can more rapidly align with the performance of its centralized counterpart.

Contributions. This paper addresses the above open questions through the results below:

C1. We propose **D-SOBA**, a **D**ecentralized **S**tochastic **O**ne-loop **B**ilevel **A**lgorithm framework. A key feature of D-SOBA is its elimination of the inner loop typically required for estimating the lower-level solution $y^\star(x)$ or the Hessian inverse of $g_i(x, y)$. The D-SOBA framework is versatile and can be implemented in two variants: **D-SOBA-SO** and **D-SOBA-FO**, which differ in their approaches to Hessian/Jacobian-vector product evaluations. D-SOBA-SO maintains **S**econd-**O**rder stochastic Hessian/Jacobian-vector products, while D-SOBA-FO performs hypergradient evaluation using fully **F**irst-**O**rder gradients through finite-difference techniques (Allen-Zhu and Li, 2018; Carmon et al., 2018). By avoiding direct computation and storage of Hessian/Jacobian matrices or their vector products, D-SOBA-FO offers greater efficiency and practicality in high-dimensional applications such as meta-learning (Franceschi et al., 2018; Rajeswaran et al., 2019; Kayaalp et al., 2022; Sow et al., 2022).

C2. We provide a comprehensive non-asymptotic convergence analysis and derive the transient iteration complexity for the D-SOBA framework. On one hand, we demonstrate that D-SOBA achieves an asymptotic gradient complexity of $\mathcal{O}(1/(N\epsilon^2))$ where ϵ is the desired convergence accuracy, outperforming existing results by at least a factor of $\log(1/\epsilon)$. On the other hand, the transient iteration complexity allows us to clarify the theoretical superiority of D-SOBA over other bilevel algorithms in the transient iteration regime. This addresses open question Q1. Notably, all convergence results for the D-SOBA framework extend naturally to both D-SOBA-SO and D-SOBA-FO.

C3. Our transient iteration complexity is the *first* result in decentralized bilevel optimizaiton to quantify how network topology and data heterogeneity jointly affect the non-

Table 1: Comparison between different decentralized stochastic bilevel algorithms. Notation T indicates the number of (outer) iterations, $1 - \rho \in (0,1]$ measures the connectivity of the underlying graph, N is the number of nodes, G^2 denotes gradient upper bound, and b^2 denotes the magnitude of gradient dissimilarity. We also list the result of single-level DSGD in the bottom line for reference.

Algorithm	S.L.$^\blacklozenge$	L.S.$^\diamondsuit$	F.O.$^\heartsuit$	Asymptotic complexity†	Transient complexity‡	Assumption$^\triangleleft$
(Chen et al., 2022)	✗	✗	✗	$\frac{1}{\varepsilon^3}$	N. A.	LC f_i
(Chen et al., 2023a)	✗	✗	✗	$\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon})$	N. A.	LC f_i
(Lu et al., 2022b)	✗	✔	✗	$\frac{1}{N\varepsilon^2}\log(\frac{1}{\varepsilon})$	N. A.	LC f_i
(Yang et al., 2022)	✗	✔	✗	$\frac{1}{N\varepsilon^2}\log(\frac{1}{\varepsilon})$	$\frac{N^3 G^4}{(1-\rho)^4}$ $^\triangleright$	BG ∇f_i
(Gao et al., 2023, Thm 1)*	✗	✗	✗	$\frac{1}{(1-\rho)^2\varepsilon^2}\log(\frac{1}{\varepsilon})$	N. A.	BG ∇f_i
(Gao et al., 2023, Thm 2)*	✗	✔	✗	$\frac{1}{N\varepsilon^2}\log(\frac{1}{\varepsilon})$	$\frac{N^3}{(1-\rho)^8}$	BG $\nabla f_i, \nabla g_i$ BG $\nabla^2 g_i$
(Wang et al., 2024)	✗	✔	✔	$\frac{1}{(1-\rho^2)^{7/2}N\varepsilon^7}$	N. A.	LC f_i $^\clubsuit$
D-SOBA-SO(ours)	✔	✔	✗	$\frac{1}{N\varepsilon^2}$	$\max\left\{\frac{N^3}{(1-\rho)^2}, \frac{N^3 b^2}{(1-\rho)^4}\right\}$	**BGD** $\nabla f_i, \nabla g_i$ **BGD** $\nabla^2 g_i$
D-SOBA-FO(ours)	✔	✔	✔	$\frac{1}{N\varepsilon^2}$	$\max\left\{\frac{N^3}{(1-\rho)^2}, \frac{N^3 b^2}{(1-\rho)^4}\right\}$	**BGD** $\nabla f_i, \nabla g_i$ **BGD** $\nabla^2 g_i$
Single-level DSGD (Chen et al., 2021b)	✔	✔	N. A.	$\frac{1}{N\varepsilon^2}$	$\max\left\{\frac{N^3}{(1-\rho)^2}, \frac{N^3 b^2}{(1-\rho)^4}\right\}$	BGD ∇f_i

$^\blacklozenge$ The algorithm has the single-loop structure.
$^\diamondsuit$ The algorithm achieves linear speedup, i.e., it achieves the asymptotic rate of $1/\sqrt{NT}$.
$^\heartsuit$ The algorithm only uses first-order gradients and does not need computation of the Hessian/Jacobian matrix or the Hessian/Jacobian-vector product.
† #gradient/Hessian evaluations to achieve an ε-stationary solution when $\epsilon \to 0$ (smaller is better).
‡ #transient iterations an algorithm experiences before the asymptotic rate dominates (smaller is better).
$^\triangleleft$ Additional assumptions beyond convexity, smoothness, and stochastic gradient. Lipschitz continuous (LC) function and bounded gradients (BG) are more restrictive than bounded gradient dissimilarity (BGD).
$^\triangleright$ G is the uniform upper bound of gradients assumed in (Yang et al., 2022). It typically holds that $b \ll G$ where b gauges the magnitude of the gradient dissimilarity, i.e., $\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i - \nabla f\|^2 \le b^2$.
$^\clubsuit$ The algorithm relies on the Lipschitz continuity assumption about f_i with respect to y, and the condition $\|\frac{1}{N}\sum_{i=1}^{N}\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) + \alpha(\nabla_1 g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 g_i(x_i^{(t)}, v_i^{(t)}))\|^2 = \mathcal{O}(\alpha^2)$, where $v_i^{(t)}$ is an intermediate variable. In general cases, this condition essentially requires that each $\nabla_1 f_i$ is uniformly bounded.
* Asymptotic rate/transient complexity are not given in (Gao et al., 2023). We derive them in Appendix C.

asymptotic convergence stage. It highlights that severe data heterogeneity exacerbates the impact of network topology, while poorly connected networks amplify the adverse effects of data heterogeneity. Neglecting either factor result in an incomplete transient-stage analysis. Additionally, we quantify the theoretical benefits achievable through enhanced network connectivity and reduced data heterogeneity, thereby addressing question Q2. All these results apply to both D-SOBA-SO and D-SOBA-FO.

C4. Our comprehensive convergence analysis elucidates the impact of the nested bilevel algorithmic structure on transient iteration complexity. On one hand, we demonstrate that decentralized single-level and bilevel algorithms share the same asymptotic con-

vergence rate and transient iteration complexity with respect to network topology and data heterogeneity. On the other hand, we clarify that the nested algorithmic structure significantly exacerbates the influence of the lower-level optimization's condition number. This highlights the challenges inherent to decentralized bilevel optimization, thereby addressing open question Q3.

Comparison with existing algorithms. All our established results, along with those of existing decentralized stochastic bilevel optimization algorithms, are summarized in Table 1. Notably, both D-SOBA-SO and D-SOBA-FO achieve an asymptotic gradient complexity of $\mathcal{O}(1/(N\epsilon^2))$, surpassing existing methods by at least a factor of $\log(1/\epsilon)$. In particular, the fully first-order decentralized bilevel algorithm proposed in (Wang et al., 2024) attains a complexity of $\mathcal{O}(1/(N\epsilon^7))$, which is less efficient than our proposed D-SOBA-FO. Furthermore, the transient iteration complexity of the proposed algorithms is considerably better than that of (Yang et al., 2022) and (Gao et al., 2023) when the data heterogeneity b^2 is trivial and the network topology is sparse (i.e., $1 - \rho \to 0$). In addition, it is observed that our algorithms attain the same asymptotic and transient-iteration complexities as single-level DSGD (Chen et al., 2021b) when the influence of the condition number can be neglected. Finally, Table 1 shows our results hold under more relaxed assumptions than those of existing methods.

Note. This paper focuses on bilevel algorithms based on vanilla decentralized SGD (Sayed, 2014; Lian et al., 2017). Since decentralized SGD is inherently influenced by data heterogeneity, studying decentralized bilevel algorithms within this framework enables a quantitative analysis of how data heterogeneity impacts decentralized bilevel optimization. This establishes a foundation for understanding the quantitative benefits of reducing data heterogeneity in decentralized bilevel optimization. By contrast, skipping the analysis of decentralized bilevel SGD and directly delving into decentralized bilevel gradient tracking, as in references (Zhang et al., 2023a; Dong et al., 2023; Wang et al., 2024), has resulted in an inability to quantify the improvement in convergence rates achieved by removing data heterogeneity.

1.4 Related Works

Bilevel optimization. Bilevel optimization (Bracken and McGill, 1973) has shown extensive applications in operations research, signal processing, and machine learning (Colson et al., 2007; Vicente and Calamai, 1994; Zhang et al., 2023b). A central challenge in bilevel optimization is the estimation of the hypergradient $\nabla\Phi(x)$. To address this, various algorithms have emerged, leveraging approaches such as approximate implicit differentiation (Domke, 2012; Ghadimi and Wang, 2018; Grazzi et al., 2020; Ji et al., 2021), iterative differentiation (Domke, 2012; Franceschi et al., 2018; Grazzi et al., 2020; Ji et al., 2021; Maclaurin et al., 2015), and Neumann series (Chen et al., 2021a; Hong et al., 2023; Xiao et al., 2022; Li et al., 2022). However, these approaches incur additional inner-loops to estimate the lower-level solution $y^\star(x)$ and the Hessian inverse of $g(x, y)$, resulting in extra computational overhead and worsening the iteration complexity. A recent work (Dagréou et al., 2022) develops a novel single-loop framework for stochastic bilevel optimization, in which the Hessian inverse is removed from the algorithmic structure and the lower and upper variables are updated simultaneously. Moreover, first-order algorithms for bilevel optimization have gained an increasing focus to solve large-scale optimization tasks. An effective approach is to view the lower-level loss as a penalty and add it to the upper-level loss (Liu et al., 2022a; Kwon

et al., 2023; Lu and Mei, 2024). Moreover, Yang et al. (2023a) also uses finite-difference of stochastic gradients to evaluate Hessian/Jacobian, reducing computation and storage cost especially in large-scaled scenarios.

Decentralized optimization. Decentralized optimization is useful in situations where the centralized control of all nodes by a single server is either practically infeasible or prohibitively expensive. Early well-known algorithms include decentralized gradient descent (Nedic and Ozdaglar, 2009; Yuan et al., 2016), diffusion strategies (Chen and Sayed, 2012), dual averaging (Duchi et al., 2011), EXTRA (Shi et al., 2015), Exact-Diffusion (Li et al., 2019; Yuan et al., 2018), gradient tracking (Di Lorenzo and Scutari, 2016; Nedic et al., 2017; Xu et al., 2015), and decentralized ADMM methods (Chang et al., 2014; Shi et al., 2014). In the stochastic context, decentralized SGD is established in (Lian et al., 2017) to achieve the same asymptotic linear speedup as centralized SGD. Since then, many efforts have extended decentralized SGD to directed topologies (Nedić and Olshevsky, 2014), time-varying topologies (Koloskova et al., 2020; Nedić and Olshevsky, 2014), and data-heterogeneous scenarios (Lin et al., 2021; Lu et al., 2019; Tang et al., 2018; Xin et al., 2020; Yuan et al., 2023, 2021). Lower bounds and optimal complexities are also recently established for stochastic decentralized optimization (Lu and De Sa, 2021; Yuan et al., 2022).

Decentralized stochastic bilevel optimization (SBO). Decentralized SBO algorithms are studied in (Chen et al., 2022, 2023a; Gao et al., 2023; Lu et al., 2022b; Yang et al., 2022; Wang et al., 2024) with solid theoretical guarantees and strong empirical performance. Nevertheless, these algorithms entail computationally expensive inner-loop updates and often exhibit a limited focus on the non-asymptotic convergence stage in their analyses, as discussed in Sec. 1.2. Table 1 presents a comparison of our results with existing works in terms of both asymptotic and transient iteration complexities. Specifically, Wang et al. (2024) achieve the hyper-gradient evaluation only by first-order gradients. Meanwhile, (Lu et al., 2022a; Niu et al., 2023) are also proposed to solve the personalized bilevel problem:

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{N} \sum_{i=1}^{N} f_i\left(x, y_i^\star(x)\right), \qquad \text{s.t.} \quad y_i^\star(x) := \arg\min_{y \in \mathbb{R}^p} g_i(x, y), \quad \forall\, 1 \leq i \leq N. \qquad (3)$$

where each node has a personalized lower-level cost function. This differs from problem (1), where both the upper- and lower-level cost functions are globally averaged.

Concurrent works. Recently, a single-loop decentralized stochastic bilevel optimization method, SPARKLE, was introduced in (Zhu et al., 2024). Building on our findings on the impact of data heterogeneity[†], SPARKLE quantifies the performance gains achieved by mitigating data heterogeneity through gradient tracking. Another recent work (Wen et al., 2024) proposed a fully first-order algorithm for decentralized *deterministic* bilevel optimization, successfully avoiding Hessian/Jacobian-vector product computations. However, this algorithm relies on a stronger assumption of the Lipschitz continuity of the Hessian matrices $\nabla^2 f_i$ and $\nabla^2 g_i$ and does not clarify performance in stochastic settings.

---

†. The main results of this paper were first presented in arXiv reports (Kong et al., 2024).

2. Preliminaries

2.1 Notations

For a second-order differentiable function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$, we denote its partial gradients as $\nabla_1 f(x, y) \in \mathbb{R}^d$ and $\nabla_2 f(x, y) \in \mathbb{R}^p$ with respect to x and y, respectively. The corresponding partial Hessians are $\nabla_{12}^2 f(x, y) \in \mathbb{R}^{d \times p}$ and $\nabla_{22}^2 f(x, y) \in \mathbb{R}^{p \times p}$. Additionally, we use $\nabla_x f(x, y^\star(x))$ to denote the gradient of f with respect to x, treating y as a function of x. We let $\| \cdot \|$ denote the ℓ_2 norm for both vectors and matrices, and $\| \cdot \|_F$ the Frobenius norm for matrices. The vector $\mathbb{1}_N \in \mathbb{R}^N$ represents an all-ones vector. For local variables $\{x_i^{(t)}\}_{i=1}^N$, the subscript i denotes the node index, while the superscript t indicates the iteration. Their average, $\sum_{i=1}^N x_i^{(t)}/N$, is denoted as $\bar{x}^{(t)}$. Finally, we write $a \lesssim b$ if there exists a constant $C > 0$ such that $a \leq Cb$.

2.2 Assumptions

With the notations introduced above, we next state the assumptions used in the paper.

Assumption 1 (SMOOTHNESS). *There exist positive constants $L_{\nabla f}$, $L_{\nabla g}$, $L_{\nabla^2 g}$, L_f and μ_g such that for any $1 \leq i \leq N$,*
   1. *$\nabla f_i, \nabla g_i, \nabla^2 g_i$ are $L_{\nabla f}$, $L_{\nabla g}$, $L_{\nabla^2 g}$ Lipschitz continuous, respectively;*
   2. *$\|\nabla_2 f_i(x, y^\star(x))\| \leq L_f < \infty$ for all $x \in \mathbb{R}^d$ in which $y^\star(x)$ is defined in problem (1b);*
   3. *$g_i(x, \cdot)$ is μ_g-strongly convex for any given $x \in \mathbb{R}^d$.*

It is noteworthy that the second condition of Assumption 1 relaxes the restrictive assumptions of Lipschitz continuity of f or, equivalently, the boundedness of $\nabla_2 f$ used in (Gao et al., 2023; Lu et al., 2022b).

Due to the heterogeneity of local data distributions, the local functions $\{(f_i, g_i)\}_{i=1}^N$ are non-identical across different nodes. As this paper focuses on bilevel optimization algorithms built upon vanilla decentralized SGD, we bound gradient dissimilarity using the following standard assumption (Koloskova et al., 2020; Lian et al., 2017):

Assumption 2 (GRADIENT DISSIMILARITY). *There exist constants $b_1, b_2 \geq 0$ such that for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^p$, it holds that:*

$$\frac{1}{N} \sum_{i=1}^N \|\nabla_1 f_i(x, y) - \nabla_1 f(x, y)\|^2 \leq b_1^2, \quad \frac{1}{N} \sum_{i=1}^N \|\nabla_2 f_i(x, y) - \nabla_2 f(x, y)\|^2 \leq b_1^2;$$

$$\frac{1}{N} \sum_{i=1}^N \|\nabla_2 g_i(x, y) - \nabla_2 g(x, y)\|^2 \leq b_2^2, \quad \frac{1}{N} \sum_{i=1}^N \left\|\nabla_{12}^2 g_i(x, y) - \nabla_{12}^2 g(x, y)\right\|^2 \leq b_2^2,$$

$$\frac{1}{N} \sum_{i=1}^N \left\|\nabla_{22}^2 g_i(x, y) - \nabla_{22}^2 g(x, y)\right\|^2 \leq b_2^2.$$

We also make the following standard assumption for stochastic gradients and Hessians.

Assumption 3 (STOCHASTICITY). *There exists a constant $\sigma \geq 0$ such that for any given $(x, y) \in \mathbb{R}^d \times \mathbb{R}^p$ and $1 \leq i \leq N$:*

1. the gradient oracles satisfy:

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_{f_i}}[\nabla_1 F(x, y; \xi_i)] = \nabla_1 f_i(x, y), \quad \mathbb{E}_{\xi_i \sim \mathcal{D}_{f_i}}\left[\|\nabla_1 F(x, y; \xi_i) - \nabla_1 f_i(x, y)\|^2\right] \le \sigma^2,$$

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_{f_i}}[\nabla_2 F(x, y; \xi_i)] = \nabla_2 f_i(x, y), \quad \mathbb{E}_{\xi_i \sim \mathcal{D}_{f_i}}\left[\|\nabla_2 F(x, y; \xi_i) - \nabla_2 f_i(x, y)\|^2\right] \le \sigma^2,$$

$$\mathbb{E}_{\zeta_i \sim \mathcal{D}_{g_i}}[\nabla_2 G(x, y; \zeta_i)] = \nabla_2 g_i(x, y), \quad \mathbb{E}_{\zeta_i \sim \mathcal{D}_{g_i}}\left[\|\nabla_2 G(x, y; \zeta_i) - \nabla_2 g_i(x, y)\|^2\right] \le \sigma^2;$$

2. the Hessian/Jacobian oracles satisfy:

$$\mathbb{E}_{\zeta_i \sim \mathcal{D}_{g_i}}\left[\nabla_{12}^2 G(x, y; \zeta_i)\right] = \nabla_{12}^2 g_i(x, y), \quad \mathbb{E}_{\zeta_i \sim \mathcal{D}_{g_i}}\left[\left\|\nabla_{12}^2 G(x, y; \zeta_i) - \nabla_{12}^2 g_i(x, y)\right\|^2\right] \le \sigma^2,$$

$$\mathbb{E}_{\zeta_i \sim \mathcal{D}_{g_i}}\left[\nabla_{22}^2 G(x, y; \zeta_i)\right] = \nabla_{22}^2 g_i(x, y), \quad \mathbb{E}_{\zeta_i \sim \mathcal{D}_{g_i}}\left[\left\|\nabla_{22}^2 G(x, y; \zeta_i) - \nabla_{22}^2 g_i(x, y)\right\|^2\right] \le \sigma^2.$$

This work studies decentralized algorithms over networks of N nodes interconnected by a graph $\mathcal{G}$ with a set of edges $\mathcal{E}$. Node i is connected to node j if $(i, j) \in \mathcal{E}$. To facilitate decentralized communication, we introduce the mixing matrix $W = [w_{i,j}]_{i,j=1}^N \in \mathbb{R}^{N \times N}$ in which each weight $w_{i,j}$ gauges information flowing from node j to node i. Furthermore, we set $w_{i,j} = 0$ for $(j, i) \notin \mathcal{E}$. The following assumption on the mixing matrix is widely used for decentralized algorithms (Koloskova et al., 2020; Lian et al., 2017; Yuan et al., 2016).

Assumption 4 (MIXING MATRIX). *The mixing matrix W is doubly stochastic, i.e.,*

$$\mathbb{1}_N^\top W = \mathbb{1}_N^\top, \quad W \mathbb{1}_N = \mathbb{1}_N.$$

Moreover, we assume $\rho := \left\|W - \mathbb{1}_N \mathbb{1}_N^\top / N\right\|_2 \in [0, 1)$.

Remark 1 (SPECTRAL GAP). *In decentralized algorithms, the quantity $1 - \rho$ is commonly known as the spectral gap (Lu and De Sa, 2021; Yuan et al., 2023) of W, which serves as a metric for measuring the connectivity of the network topology. Notably, as $1 - \rho \to 1$, it indicates that the topology is well-connected (e.g., for a fully connected graph, the mixing matrix is $W = \mathbb{1}_N \mathbb{1}_N^\top / N$ with $\rho = 0$). Conversely, as $1 - \rho \to 0$, it suggests that the topology is potentially sparse (Lian et al., 2017; Ying et al., 2021a).*

Remark 2 (NON-SYMMETRIC MIXING MATRIX). *While Assumption 4 requires the mixing matrix W to be doubly stochastic, it does not need to be symmetric. This contrasts with the recent work (Zhu et al., 2024), where the SPARKLE algorithm is proposed and analyzed under the assumption of a symmetric and doubly-stochastic mixing matrix.*

3. Decentralized Stochastic One-loop Bilevel Algorithm (D-SOBA)

Major challenges of decentralized SBO. In this section, we present our D-SOBA algorithm for decentralized SBO. The primary challenge in SBO stems from the estimation of the hypergradient $\nabla \Phi(x)$, *i.e.*, $\nabla_x f(x, y^\star(x))$, due to the implicit dependence of $y^\star(x)$ on x. Under Assumption 1 and the principles of implicit function theory (Griewank and Walther, 2008), $\nabla \Phi(x)$ can be expressed as:

$$\nabla \Phi(x) = \nabla_1 f(x, y^\star(x)) - \left(\nabla_{12}^2 g(x, y^\star(x)) \cdot \left[\nabla_{22}^2 g(x, y^\star(x))\right]^{-1} \cdot \nabla_2 f(x, y^\star(x))\right)$$

which is computationally expensive due to the need for inverting the partial Hessian. Moreover, the Hessian-inversion $\left[\nabla_{22}^2 g(x, y^\star(x))\right]^{-1} = \left[\frac{1}{N}\sum_{i=1}^N \nabla_{22}^2 g_i(x, y^\star(x))\right]^{-1}$ cannot be easily accessed through decentralized communication even in the absence of stochastic noise affecting the estimate of $\nabla_{22}^2 g_i(x, y^\star(x))$. This challenge can be partially mitigated by incorporating auxiliary inner loops to approximate the lower-level solution $y^\star(x)$ and evaluate the Hessian inversion using the Neumann series, as demonstrated in works such as (Chen et al., 2022, 2023a; Lu et al., 2022b; Yang et al., 2022). However, introducing these auxiliary inner loops results in sub-optimal convergence, as shown in Table 1. Moreover, it complicates algorithmic implementation and may impose a substantial burden on communication and computation.

Centralized SOBA. The challenge of Hessian inversion can be effectively addressed by a novel framework known as SOBA. SOBA, initially proposed in (Dagréou et al., 2022), is a single-node algorithm. We now extend it to solve the distributed problem (1) in the centralized setup. To begin with, SOBA introduces

$$z^\star(x) = \left[\nabla_{22}^2 g(x, y^\star(x))\right]^{-1} \nabla_2 f(x, y^\star(x)) \tag{5}$$

which can be regarded as the solution to minimizing the following optimization problem

$$\frac{1}{2}z^\top \nabla_{22}^2 g(x, y^\star) z - z^\top \nabla_2 f(x, y^\star) \overset{(1)}{=} \frac{1}{N}\sum_{i=1}^N \left\{ \frac{1}{2}z^\top \nabla_{22}^2 g_i(x, y^\star) z - z^\top \nabla_2 f_i(x, y^\star) \right\}, \tag{6}$$

where $y^\star$ denotes $y^\star(x)$. It is worth noting that while $z^\star(x)$ in (5) cannot be written as a finite sum across nodes, problem (6) involves only simple sums.

To save computations, we can approximately solve (6) using one-step (stochastic) gradient descent. This, combined with one-step (stochastic) gradient descent to update the upper- and lower-level variables (x, y), forms the centralized single-loop framework for solving (1):

$$\underbrace{x^{(t+1)} = x^{(t)} - \frac{\alpha_t}{N}\sum_{i=1}^N D_{x,i}^{(t)}}_{\text{minimize (1a)}}, \quad \underbrace{y^{(t+1)} = y^{(t)} - \frac{\beta_t}{N}\sum_{i=1}^N D_{y,i}^{(t)}}_{\text{minimize (1b)}}, \quad \underbrace{z^{(t+1)} = z^{(t)} - \frac{\gamma_t}{N}\sum_{i=1}^N D_{z,i}^{(t)}}_{\text{minimize (6)}}, \tag{7}$$

where $D_{x,i}^{(t)}, D_{y,i}^{(t)}$ and $D_{z,i}^{(t)}$ are defined as follows:

$$D_{x,i}^{(t)}(x^{(t)}, y^{(t)}, z^{(t)}) = \nabla_1 f_i(x^{(t)}, y^{(t)}) - \nabla_{12}^2 g_i(x^{(t)}, y^{(t)})z^{(t)}, \tag{8a}$$

$$D_{y,i}^{(t)}(x^{(t)}, y^{(t)}, z^{(t)}) = \nabla_2 g_i(x^{(t)}, y^{(t)}), \tag{8b}$$

$$D_{z,i}^{(t)}(x^{(t)}, y^{(t)}, z^{(t)}) = \nabla_{22}^2 g_i(x^{(t)}, y^{(t)})z^{(t)} - \nabla_2 f_i(x^{(t)}, y^{(t)}), \tag{8c}$$

and α_t, β_t and γ_t are learning rates. We denote recursion (7) as centralized SOBA since a central server is required to collect $D_{x,i}^{(t)}, D_{y,i}^{(t)}$, and $D_{z,i}^{(t)}$ across the entire network, as well as update variables $x^{(t)}, y^{(t)}$ and $z^{(t)}$. It is noteworthy that (7) does not require any inner loop to approximate the lower-level solution $y^\star(x)$ or evaluate the Hessian inversion of $g(x, y)$.

Decentralized SOBA (D-SOBA) Framework. Inspired by (Chen and Sayed, 2012; Nedic and Ozdaglar, 2009), we extend centralized SOBA (7) to decentralized setup:

$$x_i^{(t+1)} = \sum_{j \in \mathcal{N}_i} w_{ij}(x_j^{(t)} - \alpha_t D_{x,j}^{(t)}), \tag{9a}$$

$$y_i^{(t+1)} = \sum_{j \in \mathcal{N}_i} w_{ij}(y_j^{(t)} - \beta_t D_{y,j}^{(t)}), \tag{9b}$$

$$z_i^{(t+1)} = \sum_{j \in \mathcal{N}_i} w_{ij}(z_j^{(t)} - \gamma_t D_{z,j}^{(t)}), \tag{9c}$$

where $x_i^{(t)}$, $y_i^{(t)}$, and $z_i^{(t)}$ are local variables maintained by node i at iteration t, and $D_{x,i}^{(t)}$, $D_{y,i}^{(t)}$, and $D_{z,i}^{(t)}$ are simplified notations for $D_{x,i}^{(t)}(x^{(t)}, y^{(t)}, z^{(t)})$, $D_{y,i}^{(t)}(x^{(t)}, y^{(t)}, z^{(t)})$, and $D_{z,i}^{(t)}(x^{(t)}, y^{(t)}, z^{(t)})$, respectively, as defined in (8). The set $\mathcal{N}_i$ includes node i and all its immediate neighbors. We refer to recursions (9) as the D-SOBA framework, which forms the foundation of the D-SOBA-SO and D-SOBA-FO algorithms.

D-SOBA-SO Algorithm. The quantities $D_{x,i}^{(t)}$, $D_{y,i}^{(t)}$, and $D_{z,i}^{(t)}$ in the decentralized SOBA framework (9) involve second-order matrices such as $\nabla_{12}^2 g_i(x^{(t)}, y^{(t)})$ and $\nabla_{22}^2 g_i(x^{(t)}, y^{(t)})$. When these second-order matrices are explicitly maintained in the algorithm, the resulting method is referred to as D-SOBA-SO. In stochastic settings, D-SOBA-SO samples stochastic estimates of the gradients, Jacobians, and Hessians in (8). In particular, suppose D-SOBA-SO independently samples a minibatch of data, $\xi_i^{(t)} \sim \mathcal{D}_{f_i}$ and $\zeta_i^{(t)} \sim \mathcal{D}_{g_i}$, it approximates

$$D_{x,i}^{(t)}(x_i^{(t)}, y_i^{(t)}, z_i^{(t)}) \approx \nabla_1 F(x_i^{(t)}, y_i^{(t)}; \xi_i^{(t)}) - \nabla_{12}^2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}) z_i^{(t)}, \tag{10a}$$

$$D_{y,i}^{(t)}(x_i^{(t)}, y_i^{(t)}, z_i^{(t)}) \approx \nabla_2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}), \tag{10b}$$

$$D_{z,i}^{(t)}(x_i^{(t)}, y_i^{(t)}, z_i^{(t)}) \approx \nabla_{22}^2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}) z_i^{(t)} - \nabla_2 F(x_i^{(t)}, y_i^{(t)}; \xi_i^{(t)}). \tag{10c}$$

The detailed implementation of D-SOBA-SO is listed in Algorithm 1. Furthermore, we impose a moving average on the update of x_i in Algorithm 1. As shown in (Chen et al., 2023b), the moving average step enables a more stable and finer-grained direction for estimating hypergradients. This strategy plays a crucial role in reducing the order of bias from sample noise in the convergence analysis, as well as relaxing the technical assumptions. For notational simplicity in the proof, we introduce a scaling parameter τ, such that the upper-level step size is given by $\tau\alpha$.

D-SOBA-FO Algorithm. In high-dimensional bilevel optimization, storing and computing the second-order matrices $\nabla_{12}^2 g_i(x^{(t)}, y^{(t)})$ and $\nabla_{22}^2 g_i(x^{(t)}, y^{(t)})$ is highly expensive, particularly in large-scale machine learning applications, such as data reweighting in large language models (Pan et al., 2024). Finite difference approximation is an efficient approach for estimating higher-order gradients using lower-order information (Allen-Zhu and Li, 2018; Carmon et al., 2018; Yang et al., 2023a). Here, we introduce finite difference Hessian/Jacobian-vector product approximations to D-SOBA. Given the Lipschitz continuity of $\nabla_{12}^2 g_i$ under

11

Algorithm 1 D-SOBA-SO and D-SOBA-FO

Require: Initialize $x_i^{(0)} \in \mathbb{R}^d$, $y_i^{(0)} \in \mathbb{R}^p$, $z_i^{(0)} \in \mathbb{R}^p$, and $h_i^{(0)} \in \mathbb{R}^d$ for any $i = 1, 2, \cdots, N$,
the step size $\{\alpha_t\}_{t=0}^{T-1}, \{\beta_t\}_{t=0}^{T-1}, \{\gamma_t\}_{t=0}^{T-1}, \{\theta_t\}_{t=0}^{T-1}, \tau$, the mixing matrix W, $\{\delta_t\}_{t=0}^{T-1}$.
  for $t = 0, 1, \cdots, T-1$ **do**
    for each node $i = 1, 2, \cdots, N$ in parallel **do**
      Get stochastic sample $\xi_i^{(t)} \sim \mathcal{D}_{f_i}$ and $\zeta_i^{(t)} \sim \mathcal{D}_{g_i}$.

$$p_{H,i}^{(t)} := \nabla_{22}^2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}) z_i^{(t)};$$

$$p_{J,i}^{(t)} := \nabla_{12}^2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}) z_i^{(t)}.$$

$$p_{H,i}^{(t)} := \frac{1}{2\delta_t} \left(\nabla_2 G(x_i^{(t)}, y_i^{(t)} + \delta_t z_i^{(t)}, \zeta_i^{(t)}) - \nabla_2 G(x_i^{(t)}, y_i^{(t)} - \delta_t z_i^{(t)}, \zeta_i^{(t)}) \right);$$

$$p_{J,i}^{(t)} := \frac{1}{2\delta_t} \left(\nabla_1 G(x_i^{(t)}, y_i^{(t)} + \delta_t z_i^{(t)}, \zeta_i^{(t)}) - \nabla_1 G(x_i^{(t)}, y_i^{(t)} - \delta_t z_i^{(t)}, \zeta_i^{(t)}) \right).$$

$$x_i^{(t+1)} := \sum_{j \in \mathcal{N}_i} w_{ij}(x_j^{(t)} - \tau \alpha_t h_j^{(t)});$$
$$y_i^{(t+1)} := \sum_{j \in \mathcal{N}_i} w_{ij} \left(y_j^{(t)} - \beta_t \nabla_2 G(x_j^{(t)}, y_j^{(t)}; \zeta_j^{(t)}) \right);$$
$$z_i^{(t+1)} := \sum_{j \in \mathcal{N}_i} w_{ij} \left(z_j^{(t)} - \gamma_t \left(p_{H,j}^{(t)} - \nabla_2 F(x_j^{(t)}, y_j^{(t)}; \xi_j^{(t)}) \right) \right);$$
$$\omega_i^{(t+1)} := \nabla_1 F(x_i^{(t)}, y_i^{(t)}; \xi_i^{(t)}) - p_{J,i}^{(t)};$$
$$h_i^{(t+1)} := (1 - \theta_t) h_i^{(t)} + \theta_t \omega_i^{(t+1)}.$$

    end for
  end for

Assumption 1, we obtain the following approximation for some perturbation $\delta_t > 0$:

$$\nabla_1 g_i(x_i^{(t)}, y_i^{(t)} + \delta_t z_i^{(t)}) = \nabla_1 g_i(x_i^{(t)}, y_i^{(t)}) + \delta_t \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} + \mathcal{O}\left(\delta_t^2 \|z_i^{(t)}\|^2\right), \quad (11a)$$

$$\nabla_1 g_i(x_i^{(t)}, y_i^{(t)} - \delta_t z_i^{(t)}) = \nabla_1 g_i(x_i^{(t)}, y_i^{(t)}) - \delta_t \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} + \mathcal{O}\left(\delta_t^2 \|z_i^{(t)}\|^2\right). \quad (11b)$$

Thus, the term $(2\delta_t)^{-1}(\nabla_1 g_i(x_i^{(t)}, y_i^{(t)} + \delta_t z_i^{(t)}) - \nabla_1 g_i(x_i^{(t)}, y_i^{(t)} - \delta_t z_i^{(t)}))$ provides an approximation of the Jacobian-vector product $\nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)}$, particularly when δ_t is small:

$$\frac{1}{2\delta_t}(\nabla_1 g_i(x_i^{(t)}, y_i^{(t)} + \delta_t z_i^{(t)}) - \nabla_1 g_i(x_i^{(t)}, y_i^{(t)} - \delta_t z_i^{(t)}))$$
$$= \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} + \mathcal{O}\left(\delta_t \|z_i^{(t)}\|^2\right) \approx \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)}.$$

Similarly, $\nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)}$ can also be evaluated by first-order gradients as follows:

$$\frac{1}{2\delta_t}(\nabla_2 g_i(x_i^{(t)}, y_i^{(t)} + \delta_t z_i^{(t)}) - \nabla_2 g_i(x_i^{(t)}, y_i^{(t)} - \delta_t z_i^{(t)}))$$
$$= \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} + \mathcal{O}\left(\delta_t \|z_i^{(t)}\|^2\right) \approx \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)}.$$

By replacing the true gradients in the above expressions with stochastic approximations, we introduce D-SOBA-FO , which enables Hessian/Jacobian evaluation while requiring only first-order gradients. See Algorithm 1 for details.

4. Theoretical Analysis

Here we establishes the convergence guarantees for D-SOBA-SO and D-SOBA-FO. We define the notation $\mathcal{F}^{(t)} = \sigma \left[\bigcup_{\tau=0}^{t} \left(\bigcup_{i=1}^{N} \{x_i^{(\tau)}, y_i^{(\tau)}, z_i^{(\tau)}, h_i^{(\tau)}\} \cup \{\alpha_\tau, \beta_\tau, \gamma_\tau, \theta_\tau, \delta_\tau\} \right) \right]$ to represent the filtration of the t-th iteration, i.e., the σ-field generated by all elements with superscripts within the first t iterations. Additionally, we denote the conditional expectation with respect to $\mathcal{F}^{(t)}$ as $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | \mathcal{F}^{(t)}]$. The following lemma quantifies the error introduced by the finite-difference approximation in D-SOBA-FO:

Proposition 1. *Suppose Assumptions 1 and 3 hold. Then, the terms $p_{H,i}^{(t)}$ and $p_{J,i}^{(t)}$ obtained from D-SOBA-FO satisfy (see proof in Lemma 3):*

$$\mathbb{E}_t \left\| p_{H,i}^{(t)} - \mathbb{E}_t[p_{H,i}^{(t)}] \right\|^2 \leq \sigma^2 \left\| z_i^{(t)} \right\|^2, \quad \left\| \mathbb{E}_t[p_{H,i}^{(t)}] - \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} \right\|^2 \leq \iota^2; \quad (12a)$$

$$\mathbb{E}_t \left\| p_{J,i}^{(t)} - \mathbb{E}_t[p_{J,i}^{(t)}] \right\|^2 \leq \sigma^2 \left\| z_i^{(t)} \right\|^2, \quad \left\| \mathbb{E}_t[p_{J,i}^{(t)}] - \nabla_{12} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} \right\|^2 \leq \iota^2, \quad (12b)$$

where $\iota^2 = \frac{1}{3} L_{\nabla^2 g}^2 \delta_t^2 \left\| z_i^{(t)} \right\|^4$.

Proposition 1 demonstrates that the error in the finite-difference Hessian/Jacobian-vector approximation is primarily dominated by sample noise when δ_t is sufficiently small.

We then present the unified convergence results for both D-SOBA-SO and D-SOBA-FO, stated in the following theorem:

Theorem 1. *Suppose Assumptions 1, 2, 3, and 4 hold. If $\alpha = \Theta(\sqrt{N/T})$ and $\tau = \Theta(\kappa^{-4})$, then there exist constants $c_1, c_2, c_3 > 0$ such that, for all $0 \leq t < T$, setting $\alpha_t \equiv \alpha$, $\beta_t \equiv c_1 \alpha_t$, $\gamma_t \equiv c_2 \alpha_t$, and $\theta_t \equiv c_3 \alpha_t$, the iterates $\{\bar{x}^{(t)}\}_{t \geq 0}$ generated by Algorithm 1 (including both D-SOBA-SO and D-SOBA-FO) satisfy (see proof in Appendix A):*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \nabla \Phi(\bar{x}^{(t)}) \right\|^2 \right]$$

$$\lesssim \underbrace{\frac{\kappa^5}{\sqrt{NT}}}_{\text{A.rate}} + \underbrace{\frac{\rho^{\frac{2}{3}} \kappa^6}{(1-\rho)^{\frac{1}{3}} T^{\frac{2}{3}}} + \frac{\rho^{\frac{2}{3}} (b_1^{\frac{2}{3}} \kappa^{\frac{16}{3}} + b_2^{\frac{2}{3}} \kappa^6)}{(1-\rho)^{\frac{2}{3}} T^{\frac{2}{3}}} + \frac{\rho \kappa^6}{(1-\rho)T} + \frac{\kappa^4}{(1-\rho)NT} + \frac{\kappa^{13}}{T} + \frac{\kappa^7}{NT}}_{\text{extra overhead}},$$

$$(13)$$

where $\kappa := \max\{L_f, L_{\nabla f}, L_{\nabla g}, L_{\nabla^2 g}\}/\mu_g$ denotes the condition number.

D-SOBA-SO and D-SOBA-FO achieve the same convergence rate. According to Theorem 1, D-SOBA-SO and D-SOBA-FO achieve the same convergence rate. Moreover,

Proposition 1 shows that the error introduced by the finite difference approximation is negligible in the overall error, implying that first-order hyper-gradient evaluation does not substantially impact convergence. To the best of our knowledge, this is the first result demonstrating that fully first-order decentralized bilevel algorithms can achieve a convergence rate of the same order as their second-order counterparts.

Two-Stage Convergence of D-SOBA. The convergence rate established in Theorem 1 reveals a two-stage behavior for both D-SOBA-SO and D-SOBA-FO, which can be summarized as follows:

- **Stage 1:** When T is small, the $1/\sqrt{NT}$ term does not dominate the convergence rate; instead, higher-order terms prevail, leading to slower convergence compared to centralized algorithms. This stage corresponds to the transient iteration period of the decentralized optimization process.

- **Stage 2:** When T is sufficiently large, the $1/\sqrt{NT}$ term becomes dominant. The decentralized algorithm achieves linear speedup and convergence performance comparable to that of centralized approaches.

Asymptotic linear speedup. An algorithm is said to achieve *linear speedup* if the term $1/\sqrt{NT}$ dominates the convergence rate (Koloskova et al., 2020; Lian et al., 2017) as T becomes sufficiently large. In this regime, the number of iterations required to reach an ε-stationary solution is $1/(N\varepsilon^2)$, decreasing linearly with the number of computing nodes. As shown in (13), D-SOBA eventually attains linear speedup, whereas the algorithms in (Chen et al., 2022, 2023a) achieve only a significantly slower asymptotic rate of $1/\sqrt{T}$. Furthermore, the algorithm that rely solely on first-order gradients proposed in (Wang et al., 2024) achieve a complexity of $1/(N\epsilon^7)$, which is worse than that of the proposed D-SOBA-FO.

Transient iteration complexity. While D-SOBA attains the asymptotic rate of $1/(NT)$ as $T \to \infty$, it undergoes additional transient iterations, quantified by the extra overhead terms in (13), before reaching this rate. Transient iteration complexity (Pu et al., 2021) refers to the number of iterations an algorithm has to experience before reaching its asymptotic linear-speedup stage, that is, iterations $1, \cdots, T$ where T is relatively small so that non-NT terms still dominate the rate. Here, the smoothness coefficients as well as the conditional number κ are viewed as constants in the transient analysis since we mainly consider the influence of N, ρ, b_1, b_2 in transient complexity analysis. Transient iteration complexity measures the non-asymptotic stage in decentralized stochastic algorithms, see Fig. 1. With the explicit asymptotic rate and extra overhead characterized in (13), we establish the transient iteration complexity of D-SOBA algorithms.

Corollary 1 (TRANSIENT ITERATION COMPLEXITY)**.** *Under the same assumptions as in Theorem 1, the transient iteration complexity of D-SOBA is (Proof in Appendix A.8)*

$$\mathcal{O}\left(\max\left\{\frac{N^3}{(1-\rho)^2}, \frac{N^3 b^2}{(1-\rho)^4}\right\}\right), \tag{14}$$

where $b := \max\{b_1, b_2\}$.

Many existing works, such as (Chen et al., 2022, 2023a; Lu et al., 2022a; Wang et al., 2024), fail to establish the transient iteration complexity as their analysis ignores non-dominant convergence terms. Transient iteration complexity is crucial for differentiating between various decentralized bilevel algorithms. While (Yang et al., 2022) and (Gao et al., 2023) achieve nearly the same asymptotic complexity as D-SOBA (differing by a factor of $\log(1/\epsilon)$), their transient complexity is worse than that of D-SOBA, especially when the data heterogeneity b^2 is small and the network topology is sparse (i.e., $1 - \rho \to 0$). This addresses **Open Question Q1** listed in Section 1.2.

Joint influence of graph and heterogeneity. To our knowledge, Corollary 1 presents the first result that quantifies how network topology and data heterogeneity jointly affect the non-asymptotic convergence in decentralized SBO. First, expression (14) reveals that a sparse topology with $1 - \rho \to 0$ can significantly amplify the influence of data heterogeneity b^2. Second, a large data heterogeneity b^2 also exacerbates the adverse impact of sparse topologies from $\mathcal{O}((1-\rho)^{-2})$ to $\mathcal{O}((1-\rho)^{-4})$. Prior to our work, no results have clarified the interplay between network topology and data heterogeneity. Furthermore, expression (14) quantifies the benefits that can be achieved by enhancing network connectivity and reducing data heterogeneity. For instance, if the influence of data heterogeneity b^2 can be eliminated, the transient complexity can be improved from $\mathcal{O}((1-\rho)^{-4})$ to $\mathcal{O}((1-\rho)^{-2})$. This addresses **Open Question Q2** and motivates recent efforts (Zhu et al., 2024) to develop more effective decentralized bilevel algorithms that mitigate the influence of data heterogeneity by leveraging advanced methods such as EXTRA (Shi et al., 2015), Exact-Diffusion (Yuan et al., 2018), and Gradient-Tracking (Xu et al., 2015; Nedic et al., 2017).

Influence of the data heterogeneity across different levels. Theorem 1 demonstrates that the heterogeneity at the lower level exerts a more pronounced influence on the convergence rate, particularly when the lower-level problem possesses a large condition number. This effect arises because the corresponding κ parameter associated with b_2 exhibits a higher order of magnitude compared to that of b_1, as derived from Equation (13). Specifically, the increased sensitivity to lower-level heterogeneity can be attributed to the amplification of errors in optimization processes under ill-conditioned scenarios. However, when κ is treated as a constant and the analysis focuses on transient iterations, both b_1 and b_2 contribute equally to the transient complexity. This uniformity occurs because, under such assumptions, the dominant terms in the complexity bound become symmetric with respect to the heterogeneity parameters, leading to comparable impacts during the transient regime.

Comparison with single-level DSGD. Our comprehensive convergence analysis enables the first detailed comparison between bilevel and single-level optimization. We find that both the asymptotic rate and transient iteration complexity of D-SOBA are identical to those of single-level decentralized SGD (Chen et al., 2021b) with respect to network size N, data heterogeneity b^2, and spectral gap $1 - \rho$. This suggests that the nested lower- and upper-level structure does not introduce substantial challenges to decentralized stochastic optimization in terms of both asymptotic rate and transient iteration complexity, as illustrated in the comparison in Table 1. However, upon examining the convergence rate in (13), we observe that the presence of the lower-level optimization exacerbates the influence of the condition number. While single-level DSGD achieves an asymptotic rate of $1/\sqrt{NT}$, D-SOBA results in $\kappa^5/\sqrt{NT}$. Prior to our work, all previous works listed in Table 1 fail to explicitly capture

the influence of κ on both asymptotic and non-asymptotic terms. With all these results, **Open Question Q3** has been addressed.

Non-asymptotic concensus error. With the theoretical analysis of the convergence rate, we can also present an upper-bound for the concensus error of D-SOBA as follows (see proof in Appendix B):

Corollary 2 (CONCENSUS ERROR). *Suppose Assumptions 1, 2, 3, and 4 hold. If we take the hyperparameters as Theorem 2, then the concensus error of D-SOBA satisfies that:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2}{n} + \frac{\left\| \mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|_F^2}{n} + \frac{\left\| \mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|_F^2}{n} \right]$$
$$\lesssim \frac{N}{T} \left(\frac{b^2}{(1-\rho)^2} + \frac{1}{1-\rho} \right) + \frac{N^{\frac{1}{2}}}{T^{\frac{3}{2}}(1-\rho)^2} + \frac{1}{(1-\rho)NT}. \tag{15}$$

Convergence in deterministic scenario. When gradient or Hessian information can be accessed noiselessly, *i.e.*, σ_{f1}, σ_{g1}, and σ_{g2} are all zero, D-SOBA has enhanced convergence. Following the argument in Theorem 1, we can readily derive the following result:

Corollary 3 (DETERMINISTIC CONVERGENCE). *Suppose Assumptions 1, 2, and 4 hold, and variances $\sigma_{f1} = \sigma_{g1} = \sigma_{g2} = 0$. D-SOBA-SO and D-SOBA-FO achieve the following conergence rate (see proof in Appendix A.9):*

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla \Phi(\bar{x}^{(t)}) \right\|^2 \lesssim \frac{\rho^{\frac{2}{3}}(b_1^{\frac{2}{3}} \kappa^{\frac{16}{3}} + b_2^{\frac{2}{3}} \kappa^6)}{T^{\frac{2}{3}}(1-\rho)^{\frac{2}{3}}} + \frac{\rho \kappa^6}{(1-\rho)T} + \frac{\kappa^4}{(1-\rho)NT} + \frac{\kappa^{13}}{T}. \tag{16}$$

Corollary 3 implies that D-SOBA achieves an asymptotic rate of $\mathcal{O}(T^{-2/3}(1-\rho)^{-2/3})$. In the same deterministic setting, the concurrent single-loop decentralized bilevel algorithm (Dong et al., 2023) achieves a rate of at best $\mathcal{O}(T^{-1}(1-\rho)^{-3/2})$, which is inferior to ours when the communication topology is sufficiently sparse (i.e., as $(1-\rho) \to 0$). Additionally, the single-loop algorithm employs a gradient tracking technique, which introduces extra communication overhead and does not address performance in stochastic settings.

5. Experiments

In this section, we present experiments to validate the influence of convergence caused by communication topologies, data heterogeneity, moving average parameter θ, and finite-difference parameter δ. We also compare D-SOBA to other decentralized SBO algorithms and show the benefit of finite-difference Hessian/Jacobian-vector approximation of D-SOBA-FO in high-dimensional settings.

5.1 Linear regression with a synthetic dataset.

Here, we consider problem (1) with the upper- and lower-level loss functions defined as

$$f_i(x, y) = \mathbb{E}_{\xi_i, \zeta_i} \left[\left(\xi_i^\top y - \zeta_i \right)^2 \right],$$

$$g_i(x, y) = \mathbb{E}_{\xi_i, \zeta_i} \left[\left(\xi_i^\top y - \zeta_i \right)^2 + |x| \|y\|^2 \right],$$

where $y \in \mathbb{R}^p$ denotes the regression parameter, $|x| \geq 0$ is a ridge regularization parameter. This setting exemplifies tuning the regularization parameter with the upper-/lower-level objectives associated with validation and training datasets, respectively. Each node i observes data samples (ξ_i, ζ_i) in a streaming manner in which ξ_i is a p-dimensional vector with all elements drawn independently and randomly form the uniform distribution $U\left(-2 \times 1.5^{1/3}, 2 \times 1.5^{1/3}\right)$. Then ζ_i is generated by $\zeta_i = \xi_i^\top \omega_i + \delta_i$, where $\delta_i \sim \mathcal{N}(0, 1)$. Moreover, ω_i is generated by $\omega_i = \omega + \varepsilon_i$, where ω is a given vector whose elements are generated independently and randomly form the uniform distribution $U(0, 10)$. Here we adjust $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, 2I_p)$ to represent severe heterogeneity across nodes while $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 0.5I_p)$ yields mild heterogeneity.

We set $p = 10$ and the number of nodes was set to 9 and 20. The communication topologies are set to Ring, 2D-Torus (Ying et al., 2021a) topologies (3×3 when $N = 9$ and 4×5 when $N = 20$) as well as a fully connected network (i.e. centralized SOBA). The step-sizes $\alpha_t, \beta_t, \gamma_t$ are initialized as 0.1 and multiplied by 0.8 per $1,000$ iterations while θ_t is set to 0.2. We repeat all the cases 50 times and plot the average of all trials.

Fig.2 illustrates the performance of D-SOBA-SO. For Fig. 2, we can observe that D-SOBA with a decentralized communication topology achieves linear speedup if the corresponding convergence curve matches with the convergence curve with a fully-connected graph. Thus, Fig.2 illustrates the two-stage convergence of the D-SOBA algorithm.

It can be also observed that for both $N = 9$ and $N = 20$, the transient stage of D-SOBA-SO over decentralized networks becomes longer as data heterogeneity increases, highlighting the crucial impact of data heterogeneity. Additionally, D-SOBA-SO over the 2D-torus graph exhibits a shorter transient stage than that with the ring graph, indicating that networks with a smaller spectral gap (i.e., worse connectivity) can significantly slow down convergence. Moreover, the transient iterations of both the 2D-torus and ring networks increase as the network size N grows. All these results are consistent with our transient iteration complexity derived in Corollary 1.

5.2 Data hyper-cleaning on FashionMNIST dataset.

Here we consider a data hyper-cleaning problem (Shaban et al., 2019) for a Fashion MNIST dataset (Xiao et al., 2017), which consists of 60000 images for training and 10000 images for testing. Then we split the 60000 training images into the training set and validation set, which contains 50000 images and 10000 images separately.

In the data hyper-cleaning problem, we aim to train a classifier in a corrupted setting in which the label of each training data is replaced by a random class number with a probability p (i.e. the corruption rate) on a decentralized network with $N = 10$ clients. It can be viewed as a bilevel optimization problem (1) whose loss functions on the upper and lower level of
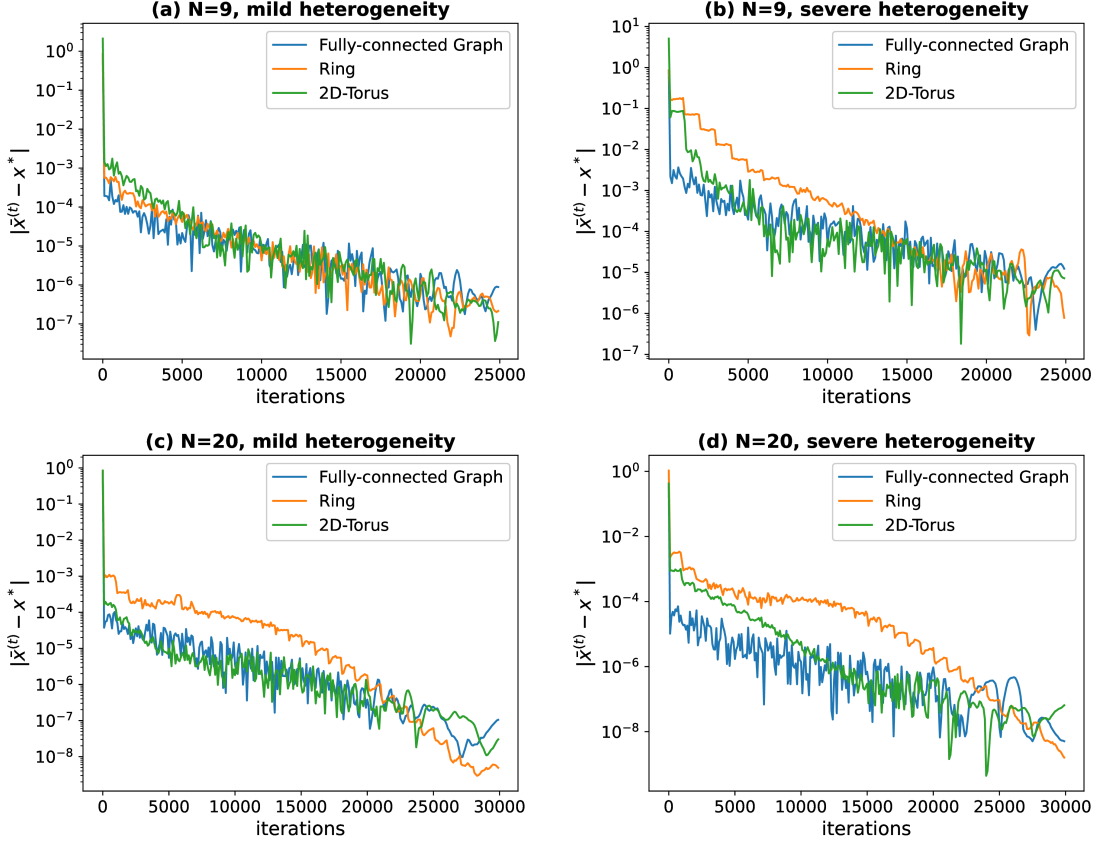
Figure 2: Convergence performance of D-SOBA-SO over various networks under different data heterogeneity.

the i-th node denote:

$$f_i(x,y) = \frac{1}{\left|\mathcal{D}_{val}^{(i)}\right|} \sum_{(\xi_e,\zeta_e)\in D_{val}^{(i)}} L(\phi(\xi_e;y),\zeta_e), \tag{18a}$$

$$g_i(x,y) = \frac{1}{\left|\mathcal{D}_{tr}^{(i)}\right|} \sum_{(\xi_e,\zeta_e)\in \mathcal{D}_{tr}^{(i)}} \sigma(x_e)L(\phi(\xi_e;y),\zeta_e) + C \left\|y\right\|^2, \tag{18b}$$

where ϕ denotes the parameters of a two-layer MLP network with a 300-dim hidden layer and ReLU activation while y denotes its parameters. L denotes cross-entropy loss, $\sigma(u) = (1 + e^{-u})^{-1}$ denotes the sigmoid function. $\mathcal{D}_{val}^{(i)}$ and $\mathcal{D}_{tr}^{(i)}$ denote the validation and training sets of the i-th client, which are sampled randomly by Dirichlet distribution with parameters $\alpha = 0.1$ in non-i.i.d. cases (Lin et al., 2021). Following with the setting in (Shaban et al., 2019), we set $C = 0.001$. The step-sizes in all the gradient updates are all set to 0.1. And all the experiments are repeated 10 times.

Comparison between D-SOBA and other existing algorithms. We firstly compare D-SOBA-SO and D-SOBA-FO with MA-DSBO (Chen et al., 2023a) and Gossip DSBO (Yang et al., 2022) by solving (18) over an exponential graph (Ying et al., 2021a) with

18
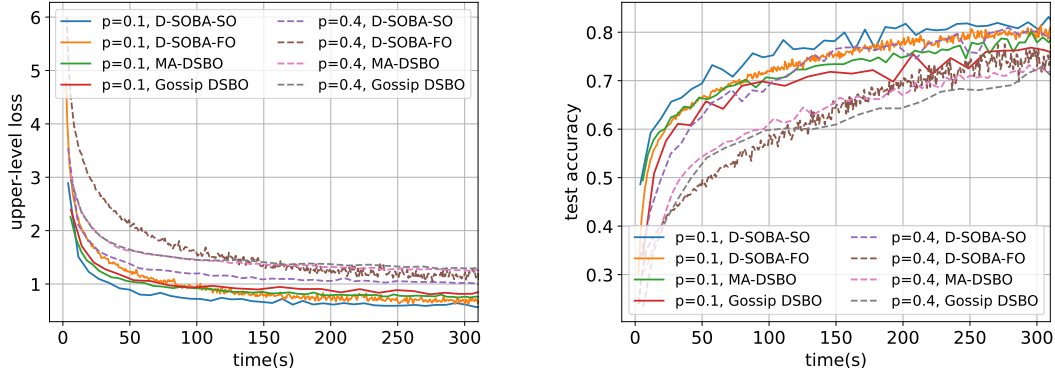
Figure 3: The upper-level loss (left) and test accuracy (right) of different decentralized stochastic bilevel optimization algorithms.

$p = 0.1, 0.4$, respectively. For all algorithms, the step-sizes are set to 0.1 and the batch size are set to 200. Meanwhile, the moving average parameters θ_t of D-SOBA-SO, D-SOBA-FO and MA-DSBO are set to 0.8. For MA-DSBO, we set the number of inner-loop iterations $T = 5$ and the number of outer-loop iterations $K = 5$. For Gossip DSBO, we set the number of Hessian-inverse estimation iterations $T = 5$. At the end of the update of outer parameters, we use the average of y among all clients to do the classification of the test set.

Fig. 3 presents the upper-level loss and test accuracy of D-SOBA-SO, D-SOBA-FO, MA-DSBO, and Gossip DSBO when the corruption rate p equals 0.1 and 0.4 over an exponential graph (Ying et al., 2021a) with $N = 10$ nodes, respectively. We observe that both D-SOBA-SO and D-SOBA-FO achieves higher test accuracy compared to the other two algorithms.

Moreover, although the first-order variant in D-SOBA-FO reduces computational overhead, it may still lead to suboptimal performance than D-SOBA-SO in Figure 3 as the finite difference of first-order gradients in D-SOBA-FO exactly approximates a Hessian/Jacobian–vector product. Then D-SOBA-FO may introduce additional error into the hyper-gradient evaluation. In contrast, D-SOBA-SO computes exact Hessian-/Jacobian–vector products directly from stochastic samples, resulting in fewer iterations to reach a given accuracy compared to first-order methods. On the other hand, the Hessian/Jacobian–vector product is computed via PyTorch's forward-mode automatic differentiation in D-SOBA-SO, which is significantly faster than first computing the full Hessian/Jacobian matrix and then multiplying it by the vector. This computational advantage prevents its per-iteration time from becoming prohibitively slow compared to D-SOBA-FO.

Convergence of D-SOBA with different communication topologies. Then we conduct the D-SOBA-SO with different topologies including Adjust Ring, 2D-Torus, Exponential graph as well as the centralized case in the non-i.i.d setting, while the weight matrix
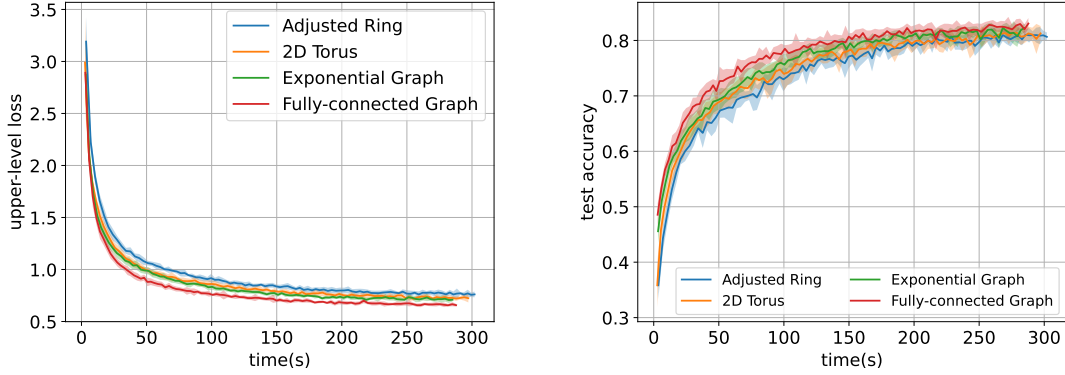
Figure 4: The upper-level loss (left) and test accuracy (right) of D-SOBA with different communication topologies.
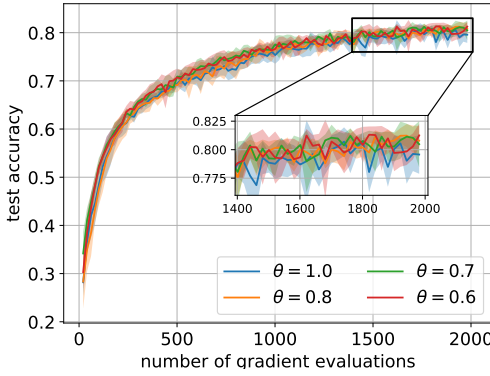

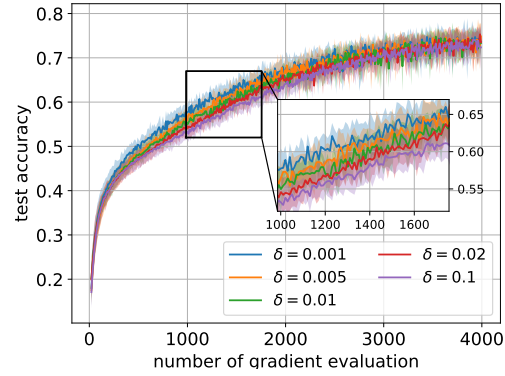Figure 5: The test accuracy of D-SOBA-SO with different moving-average parameter θ_t.

Figure 6: The test accuracy of D-SOBA-FO with different moving-average parameter δ_t.

of Adjust Ring $W = [w_{ij}]_{N \times N}$ satisfies:

$$
w_{ij} = \begin{cases} 0.4, & \text{if } (j-i)\%N \in \{\pm 1\}, \\ 0.2, & \text{if } j = i, \\ 0, & \text{else.} \end{cases}
$$

We set $p = 0.2$, batch size equal to 200, and repeat all the cases 10 times and illustrate the mean of all trials. Figure 4 illustrates the upper-level training loss and test accuracy of different cases, from which we can observe how the two-stage convergence of decentralized communication affect the practical optimization tasks. Specifically, D-SOBA with sparser topologies exhibits slower convergence in the upper-level loss and then comes with suboptimal test accuracy.

Influence of moving-average iterations on convergence. Here we aim to find the influence of moving-average parameter θ on the convergence. Let the batch size equal 100
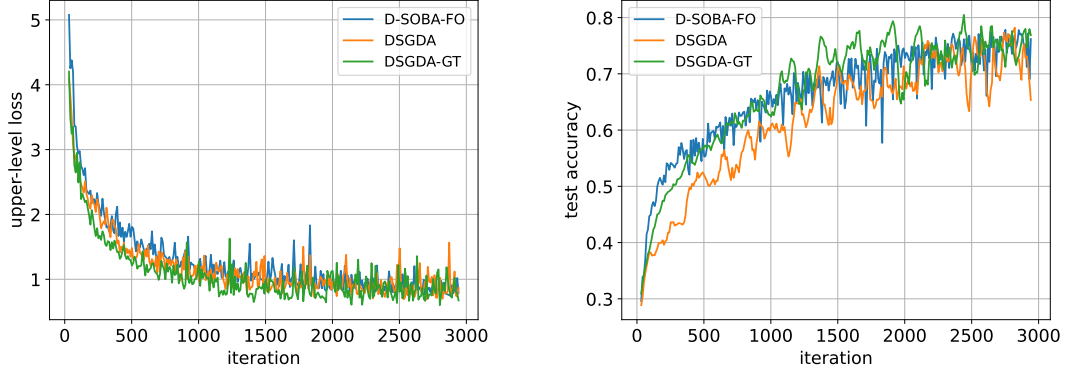
20

Figure 7: The upper-level loss (left) and test accuracy (right) of different first-order decentralized stochastic bilevel optimization algorithms.
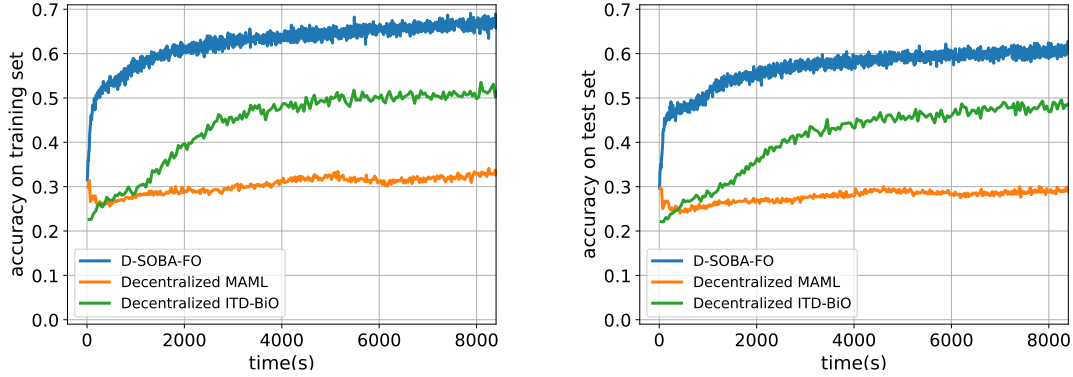

Figure 8: The accuracy on training set (left) and testing set (right) of different algorithms for the meta-learning problem.

and $p = 0.2$, we run D-SOBA-SO on an Adjusted Ring with $\theta \in \{1.0, 0.8, 0.7, 0.6\}$. We repeat each case 10 times and obtain the test accuracy. As the Figure 5 demonstrates, the cases when $\theta = 0.7, 0.8$ achieve a higher average test accuracy than when $\theta = 0.6, 1.0$, which implies the trade-off of the selection θ in the convergence.

Influence of finite-difference parameter δ in D-SOBA-FO. To find how the hyperparameter δ in D-SOBA-FO affect the convergence, we run D-SOBA-FO on an Adjusted Ring with the batch size equal 100 and $p = 0.3$. And the term $\theta = 0.8$. Moreover, the term $\delta = 0.001, 0.005, 0.01, 0.02, 0.1$, respectively. As Fig. 6 shows, a smaller δ_t leads to a faster convergence at the beginning of the iteration. However, the influence of δ on the convergence is insignificant as all the cases finally achieve similar test accuracy. This result illustrates how δ_t affect the convergence, which has been discussed in Section 4.

Comparison between D-SOBA-FO and first-order baselines. Additionally, we compare D-SOBA-FO with DSGDA† as well as DSGDA-GT (Wang et al., 2024) on an Adjusted Ring topology, with batch size 50 and $p = 0.2$. Key hyperparameters: $\delta = 0.0001$ (D-SOBA-FO), inner iterations $T = 1$ per gradient step (DSGDA), $\theta_t = 0.8$ (D-SOBA-FO moving average), and $\alpha = 1$ (DSGDA). All methods use step size 0.03. Figure 7 demonstrates comparable performance between D-SOBA-FO and DSGDA in both upper-level loss minimization and test accuracy.

5.3 Decentralized meta-learning

Consider a meta-learning problem (Finn et al., 2017). We consider R tasks $\{\mathcal{T}_r, r = 1, \cdots, R\}$. Each task $\mathcal{T}_r$ has its own loss function $l_r(x, y_r) = \mathbb{E}_{\xi \in \mathcal{D}_r} L(x, y_r, \xi)$, where y_r denotes the task-specific parameters and x denotes global parameters shared by all the tasks. Meta-learning problem aims to find good global parameter $x^\star$ to bulid the embedded features. Then each task can adapt its task-specific parameters y_i.

Here, we train a meta-learning framework in a decentralized scenario (Kayaalp et al., 2022) over N nodes. For $i = 1, 2, \cdots, N$, let $\mathcal{D}_{r,i}$ denote the data set received by the i-th node of the r-th task $\mathcal{T}_r$. And we split $\mathcal{D}_{r,i}$ to $\mathcal{D}_{r,i}^{\text{train}}$ and $\mathcal{D}_{r,i}^{\text{val}}$ for training and validation, respectively. Then we find the optimal parameter by solving the decentralized SBO problem (1) with the upper- and lower-level loss functions defined as:

$$
\begin{aligned}
f_i(x, y) &= \frac{1}{R} \sum_{r=1}^{R} \left[\frac{1}{|\mathcal{D}_{r,i}^{\text{val}}|} \sum_{\xi \in \mathcal{D}_{r,i}^{\text{val}}} L(x, y_r, \xi) \right], \\
g_i(x, y) &= \frac{1}{R} \sum_{r=1}^{R} \left[\frac{1}{|\mathcal{D}_{r,i}^{\text{train}}|} \sum_{\xi \in \mathcal{D}_{r,i}^{\text{train}}} L(x, y_r, \xi) + \mathcal{R}(y_r) \right],
\end{aligned}
\tag{19}
$$

where $\mathcal{R}(y_r)$ denotes a strongly-convex regularization term of y_r.

In this experiment, we compared D-SOBA-FO with ITD-BiO (Ji et al., 2021) and MAML (Finn et al., 2017) with decentralized communication over a 5-way 10-shot task over a network of $N = 8$ nodes connected by Ring graph. The dataset is miniImageNet (Vinyals et al., 2016), which is generated from ImageNet (Russakovsky et al., 2015) and consists of 100 classes with each class containing 600 images of size 84×84. Following the setting of, we split these classes into 64 for training, 16 for validation, and 20 for testing. Each training and validation class splits its data by Dirichlet distribution with parameters $\alpha = 0.1$ (Lin et al., 2021). We use a four-layer CNN with four convolution blocks in which each block sequentially consists of a 3×3 convolution with 32 filters, batch normalizationm, ReLU activation, and 2×2 max pooling. L denotes cross-entropy loss and $\mathcal{R}(y) = 0.001||y||^2$. The number of tasks $R = 1000$. The batch size is set to 32. The parameters of the last linear layer are set to be task-specific and the other parameters are global. We choose the stepsize 0.002 for upper-level iterations and 0.01 for lower-level iterations for all algorithms. The stepsize of auxiliary variables in D-SOBA-FO is 0.01 and the term δ_t in D-SOBA-FO is 0.001.

$\dagger$. DSGDA is the non-tracking version of DSGDA-GT that proposed in (Wang et al., 2024).

Figure 8 the average accuracy on training set for all nodes and the test accuracy of the three algorithms. We can observe that D-SOBA-FO outperforms than the other algorithms in this large-scaled problems. Thus, this experiment indicates the advantage of D-SOBA-FO in large-scaled scenarios.

6. Conclusion and Limitations

This paper introduces D-SOBA-SO and D-SOBA-FO, two single-loop decentralized SBO algorithms that achieve state-of-the-art asymptotic convergence rate and gradient/Hessian complexity under more relaxed assumptions compared to existing methods. Especially, D-SOBA-FO evaluates the Hessian/Jacobian-vector only by first-order gradients. We also clarify, for the first time, the joint influence of network topology and data heterogeneity on bilevel algorithms through an analysis of transient iteration complexity. Nevertheless, D-SOBA relies on the assumption of bounded gradient dissimilarity, which might be violated when local distribution $\mathcal{D}_{f_i}$ or $\mathcal{D}_{g_i}$ differs drastically across nodes.

Acknowledgements

Appendix A. Proof of the Convergence Rate for Algorithm 1.

Here we present the proof of Theorem 1 as well as two corollaries including the transient time and the convergence rate in the deterministic case of Algorithm 1, which provides the theoretical analysis on the convergence of Algorithm 1.

A.1 Notations

We first introduce some notations which will be used in the proof as follows. Firstly, for $0 \leq t \leq T$, let $\bar{y}_\star^{(t)} = \arg\min_y g(\bar{x}^{(t)}, y)$ denote the minimizer of (1b) with respect to $\bar{x}^{(t)}$. Then, let $\bar{z}_\star^{(t)} = (\nabla_{22}^2 g(\bar{x}^{(t)}, \bar{y}_\star^{(t)}))^{-1} \nabla_2 f(\bar{x}^{(t)}, \bar{y}_\star^{(t)})$ denote the minimizer of (6) with respect to $\bar{x}^{(t)}, \bar{y}_\star^{(t)}$. Denote $q_i^{(t)} = \nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)}$, and $\bar{q}^{(t)} = \frac{1}{N} \sum_{i=1}^n q_i^{(t)}$.

For iteration t, we use the following notations to represent the stochastic gradients in Algorithm 1:

$$u_{1,i}^{(t)} := \nabla_1 F(x_i^{(t)}, y_i^{(t)}; \xi_i^{(t)}), \quad u_{2,i}^{(t)} := \nabla_2 F(x_i^{(t)}, y_i^{(t)}; \xi_i^{(t)}), \quad v_{2,i}^{(t)} := \nabla_2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}).$$

Remark. From the definition of $\mathcal{F}^{(t)}$, we know that $h_i^{(t+1)}$ and $x_i^{(t+1)}$ $(1 \leq i \leq N)$ are both $\mathcal{F}^{(t)}$-measurable. Then from the update of $x_i^{(t+1)}$ in Algorithm 1, we know that $x_i^{(t+1)}$ is actually $\mathcal{F}^{(t)}$-measurable.

Then we present the matrix form of iterators on different clients as follows:

$$\mathbf{x}^{(t)} = \left(x_1^{(t)}, x_2^{(t)}, \cdots, x_N^{(t)}\right)_{N \times d}^\top, \quad \bar{\mathbf{x}}^{(t)} = \left(\bar{x}^{(t)}, \bar{x}^{(t)}, \cdots, \bar{x}^{(t)}\right)_{N \times d}^\top,$$

$$\mathbf{y}^{(t)} = \left(y_1^{(t)}, y_2^{(t)}, \cdots, y_N^{(t)}\right)_{N \times p}^\top, \quad \bar{\mathbf{y}}^{(t)} = \left(\bar{y}^{(t)}, \bar{y}^{(t)}, \cdots, \bar{y}^{(t)}\right)_{N \times p}^\top,$$

$$\mathbf{z}^{(t)} = \left(z_1^{(t)}, z_2^{(t)}, \cdots, z_N^{(t)}\right)_{N \times p}^\top, \quad \bar{\mathbf{z}}^{(t)} = \left(\bar{z}^{(t)}, \bar{z}^{(t)}, \cdots, \bar{z}^{(t)}\right)_{N \times p}^\top,$$

$$\mathbf{h}^{(t)} = \left(h_1^{(t)}, h_2^{(t)}, \cdots, h_N^{(t)}\right)_{N \times d}^\top,$$

$$\mathbb{E}_{t-1}[\omega^{(t)}] = \left(\mathbb{E}_{t-1}[\omega_1^{(t)}], \mathbb{E}_{t-1}[\omega_2^{(t)}], \cdots, \mathbb{E}_{t-1}[\omega_N^{(t)}]\right)_{N \times d}^\top,$$

$$\mathbb{E}_{t-1}[\overline{\omega^{(t)}}] = \left(\mathbb{E}_{t-1}[\bar{\omega}^{(t)}], \mathbb{E}_{t-1}[\bar{\omega}^{(t)}], \cdots, \mathbb{E}_{t-1}[\bar{\omega}^{(t)}]\right)_{N \times d}^\top,$$

$$\mathbf{v}^{(t)} = \left(v_1^{(t)}, v_2^{(t)}, \cdots, v_N^{(t)}\right)_{N \times p}^\top,$$

$$\mathbf{r}^{(t)} = \left(p_{H,1}^{(t)} - u_{2,1}^{(t)}, p_{H,2}^{(t)} - u_{2,2}^{(t)}, \cdots, p_{H,N}^{(t)} - u_{2,N}^{(t)}\right)_{N \times d}^\top.$$

The gradient of lower-level loss functions of different nodes can also be presented as the following matrix forms:

$$\nabla_2 \mathbf{g}^{(t+1)} = \left(\nabla_2 g_1(x_1^{(t)}, y_1^{(t)}), \nabla_2 g_2(x_2^{(t)}, y_2^{(t)}), \cdots, \nabla_2 g_N(x_N^{(t)}, y_N^{(t)})\right)_{N \times p}^\top.$$

Besides, we denote Δ_t^2 as the consensus error of the t-th round of iteration, $i.e.$,

$$\Delta_t^2 = \kappa^2 \left[\left\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_F^2 + \left\|\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right\|_F^2\right] + \left\|\mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)}\right\|_F^2.$$

We also assume that the step-sizes are fixed over different rounds, which means that there exist constants $c_1, c_2, c_3 > 0$ that

$$\alpha_1 = \alpha_2 = \cdots = \alpha_T = \alpha, \qquad \beta_1 = \beta_2 = \cdots = \beta_T = \beta = c_1\alpha,$$
$$\gamma_1 = \gamma_2 = \cdots = \gamma_T = \gamma = c_2\alpha, \quad \theta_1 = \theta_2 = \cdots = \theta_T = \theta = c_3\alpha.$$

A.2 Technical Lemmas

Lemma 1. *Suppose $f(x)$ is μ-strongly convex and L- smooth. For any x and $\eta < \frac{2}{\mu+L}$, define $x^+ = x - \eta\nabla f(x), x^\star = \arg\min f(x)$. Then we have*

$$\left\| x^+ - x^\star \right\| \leq (1 - \eta\mu) \left\| x - x^\star \right\|.$$

Proof. See Lemma 10 in Qu and Li (2017). □

Lemma 2. *Suppose Assumption 1 holds, then $\nabla\Phi(x)$ is $L_{\nabla\Phi}$-Lipschitz continuous, where*

$$L_{\nabla\Phi} = L_{\nabla f} + \frac{2L_{\nabla f}L_{\nabla g} + L_f^2 L_{\nabla^2 g}}{\mu_g} + \frac{2L_f L_{\nabla g}L_{\nabla^2 g} + L_{\nabla f}L_{\nabla g}^2}{\mu_g^2} + \frac{L_f L_{\nabla^2 g}L_{\nabla g}^2}{\mu_g^3} = \mathcal{O}(\kappa^3).$$

Also, if we define $y^\star(x) = \arg\min_y g(x, y)$, $z^\star(x) = (\nabla_{22}^2 g(x, y^\star(x)))^{-1}\nabla_2 f(x, y^\star(x))$, then $y^\star(x)$ are $L_{y^\star}$-Lipschitz continuous, and $z^\star(x)$ are $L_{z^\star}$-Lipschitz continuous, where

$$L_{y^\star} = \frac{L_{\nabla g}}{\mu_g} = \mathcal{O}(\kappa), \quad L_{z^\star} = \sqrt{1 + L_{y^\star}^2}\left(\frac{L_{\nabla f}}{\mu_g} + \frac{L_f L_{\nabla^2 g}}{\mu_g^2}\right) = \mathcal{O}(\kappa^3).$$

And we also have:

$$\|z^\star(x)\| \leq \frac{L_f}{\mu_g}, \quad \forall x \in \mathbb{R}^d. \tag{21}$$

Proof. Due to Assumption 1 and its remark, the proof of this lemma is the same as the single agent case which is shown in (Chen et al., 2023b, Lemma B.2.). □

Remark 3. *If we define $L := \max\left\{L_f, L_{\nabla f}, L_{\nabla g}, L_{\nabla^2 g}\right\}$, then the order of $L_{\nabla\Phi}$ is actually $\mathcal{O}(L\kappa^3)$. We follow the existing works Zhu et al. (2024); Chen et al. (2023a) to view $\kappa = L/\mu$ and the smooth term $L := \max\{L_f, L_{\nabla f}, L_{\nabla g}, L_{\nabla^2 g}\}$ as a constant.*

A.3 Error Analysis of the Hessian/Jacobian vector product evaluation

Here, we analysis the error of the Hessian/Jacobian vector product evaluation. The following lemma presents the error of the Hessian/Jacobian vector product evaluation in both D-SOBA-SO and D-SOBA-FO.

Lemma 3. *Suppose Assumption 1 and 3 hold, then the terms $p_{H,i}^{(t)}$, $p_{J,i}^{(t)}$ obtained from D-SOBA-SO satisfy:*

$$\mathbb{E}_t\left\|p_{H,i}^{(t)} - \mathbb{E}_t[p_{H,i}^{(t)}]\right\|^2 \leq \sigma^2\left\|z_i^{(t)}\right\|^2, \quad \left\|\mathbb{E}_t[p_{H,i}^{(t)}] - \nabla_{22}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}\right\|^2 \leq \iota^2; \tag{22a}$$

$$\mathbb{E}_t\left\|p_{J,i}^{(t)} - \mathbb{E}_t[p_{J,i}^{(t)}]\right\|^2 \leq \sigma^2\left\|z_i^{(t)}\right\|^2, \quad \left\|\mathbb{E}_t[p_{J,i}^{(t)}] - \nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}\right\|^2 \leq \iota^2, \tag{22b}$$

for all $\iota > 0$. If we choose $\delta_t \leq \sqrt{3}\iota L_{\nabla^2 g}^{-1}\|z_i^{(t)}\|^{-2}$ in D-SOBA-FO, (22a) and (22b) also hold.

Proof. Firstly, we consider D-SOBA-FO. From assumption 3, we know that the Hessian of g_i with respect to a stochastic sample $\zeta_i^{(t)} \in \mathcal{D}_{g_i}$ exists for all $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^p$. Then, for the term $p_{H,i}^{(t)}$ obtained from D-SOBA-FO, we have:

$$
\begin{aligned}
&\mathbb{E}_t \left\| p_{H,i}^{(t)} - \mathbb{E}_t[p_{H,i}^{(t)}] \right\|^2 \\
=& \mathbb{E}_t \left\| \frac{1}{2\delta_t} \int_{-1}^{1} \left(\nabla_{22}^2 G(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)}); \zeta_i^{(t)}) - \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) \right) (\delta_t z_i^{(t)}) dt \right\|^2 \\
\leq& \frac{1}{4\delta_t^2} \int_{-1}^{1} 2\mathbb{E}_t \left\| \nabla_{22}^2 G(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)}); \zeta_i^{(t)}) - \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) \right\|^2 \left\| \delta_t z_i^{(t)} \right\|^2 dt \\
\leq& \frac{1}{\delta_t^2} \cdot \sigma^2 \left\| \delta_t z_i^{(t)} \right\|^2 \leq \sigma^2 \left\| z_i^{(t)} \right\|^2.
\end{aligned}
\tag{23}
$$

Then, we consider the error between $\mathbb{E}_t[p_{H,i}^{(t)}]$ and $\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}$. We can obtain:

$$
\begin{aligned}
&\left\| \mathbb{E}_t[p_{H,i}^{(t)}] - \nabla_{22}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)} \right\|^2 \\
=& \left\| \frac{1}{2\delta_t} \int_{-1}^{1} \left(\nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) - \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)}) \right) (\delta_t z_i^{(t)}) dt \right\|^2 \\
\leq& \frac{1}{4\delta_t^2} \int_{-1}^{1} 2 \left\| \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) - \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)}) \right\|^2 \left\| \delta_t z_i^{(t)} \right\|^2 dt \leq \frac{1}{3} L_{\nabla^2 g}^2 \delta_t^2 \left\| z_i^{(t)} \right\|^4.
\end{aligned}
\tag{24}
$$

Similarly, for $p_{J,i}^{(t)}$, we have:

$$
\begin{aligned}
&\mathbb{E}_t \left\| p_{J,i}^{(t)} - \mathbb{E}_t[p_{J,i}^{(t)}] \right\|^2 \\
=& \mathbb{E}_t \left\| \frac{1}{2\delta_t} \int_{-1}^{1} \left(\nabla_{12}^2 G(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)}); \zeta_i^{(t)}) - \nabla_{12}^2 g(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) \right) (\delta_t z_i^{(t)}) dt \right\|^2 \\
\leq& \frac{1}{4\delta_t^2} \int_{-1}^{1} 2\mathbb{E}_t \left\| \nabla_{12}^2 G(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)}); \zeta_i^{(t)}) - \nabla_{12}^2 g(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) \right\|^2 \left\| \delta_t z_i^{(t)} \right\|^2 dt \\
\leq& \frac{1}{\delta_t^2} \cdot \sigma^2 \left\| \delta_t z_i^{(t)} \right\|^2 \leq \sigma^2 \left\| z_i^{(t)} \right\|^2.
\end{aligned}
\tag{25}
$$

And the error between $\mathbb{E}_t[p_{J,i}^{(t)}]$ and $\nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}$ can be bounded by:

$$
\begin{aligned}
&\left\| \mathbb{E}_t[p_{J,i}^{(t)}] - \nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)} \right\|^2 \\
=& \left\| \frac{1}{2\delta_t} \int_{-1}^{1} \left(\nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) - \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) \right) (\delta_t z_i^{(t)}) dt \right\|^2 \\
\leq& \frac{1}{4\delta_t^2} \int_{-1}^{1} \left\| 2\nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)} + t(\delta_t z_i^{(t)})) - \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) \right\|^2 \left\| \delta_t z_i^{(t)} \right\|^2 dt \leq \frac{1}{3} L_{\nabla^2 g}^2 \delta_t^2 \left\| z_i^{(t)} \right\|^4.
\end{aligned}
\tag{26}
$$

Let $\delta_t \leq \sqrt{3}\iota L_{\nabla^2 g}^{-1} \|z_i^{(t)}\|^{-2}$, then (22a) and (22b) hold in D-SOBA-FO.

Then we consider the scenario of D-SOBA-SO. From Assumption 3, we have:

$$\mathbb{E}_t \left\| p_{H,i}^{(t)} - \mathbb{E}_t[p_{H,i}^{(t)}] \right\|^2 \leq \mathbb{E}_t \left[\left\| \nabla_{22}^2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}) - \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)}) \right\|^2 \right] \left\| z_i^{(t)} \right\|^2 \leq \sigma^2 \left\| z_i^{(t)} \right\|^2 ;$$
$$(27a)$$

$$\mathbb{E}_t \left\| p_{J,i}^{(t)} - \mathbb{E}_t[p_{J,i}^{(t)}] \right\|^2 \leq \mathbb{E}_t \left[\left\| \nabla_{12}^2 G(x_i^{(t)}, y_i^{(t)}; \zeta_i^{(t)}) - \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)}) \right\|^2 \right] \left\| z_i^{(t)} \right\|^2 \leq \sigma^2 \left\| z_i^{(t)} \right\|^2 .$$
$$(27b)$$

We also have $\left\| \mathbb{E}_t[p_{H,i}^{(t)}] - \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} \right\|^2 = \left\| \mathbb{E}_t[p_{J,i}^{(t)}] - \nabla_{12} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} \right\|^2 = 0$ from Assumption 3 for D-SOBA-SO. So (22a) and (22b) also hold in D-SOBA-SO. □

With Lemma 3, we can directly obtain that Proposition 1 is also satisfied.

A.4 Descent Lemma

In this subsection, we firstly present the descent lemma.

Lemma 4 (Descent Lemma). *Suppose Assumption 1 and 3 hold, and the step-sizes α, γ satisfy:*

$$\alpha \leq \min \left\{ \frac{c_3}{2\tau(c_3 L_{\nabla\Phi} + L_{\nabla\eta})}, \frac{\tau^2}{20c_3} \right\}, \quad 0 < \tau < \sqrt{20c_3} \leq 1, \quad \gamma\mu_g \leq 1. \quad (28)$$

Then in Algorithm 1, we have:

$$\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \bar{h}^{(t)} \right\|^2 \right] + \alpha \sum_{t=0}^{T} \mathbb{E}\left[\left\| \bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)}) \right\|^2 \right]$$

$$\leq \frac{2}{\tau}\mathbb{E}[E_0] + \frac{1}{c_3}\mathbb{E}\left[\left\| \bar{h}^{(0)} - \nabla\Phi(\bar{x}^{(0)}) \right\|^2 \right] + \frac{4L_{\nabla\Phi}^2 \tau^2}{c_3^2} \alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \bar{h}^{(t)} \right\|^2 \right]$$

$$+ \frac{1}{2} \cdot \alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)}) \right\|^2 \right] + \left(60L^2 + \frac{6(1+c_3)\alpha\sigma^2}{N} \right) \frac{1}{N} \cdot \alpha \sum_{t=0}^{T-1} \mathbb{E}[\Delta_t^2]$$

$$+ \frac{6(1+c_3)T\alpha^2}{N} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2 + 10T\alpha\iota^2 + 60 \left(\frac{L_f^2 L_{\nabla^2 g}^2}{\mu_g^2} + L_{\nabla f}^2 \right) \sum_{t=0}^{T-1} \alpha\mathbb{E}\left[\left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right]$$

$$+ \left(60L_{\nabla^2 g}^2 + \frac{6(1+c_3)\alpha\sigma^2}{N} \right) \sum_{t=0}^{T-1} \alpha\mathbb{E}\left[\left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 \right].$$
$$(29)$$

Proof. Define

$$\eta(x, h, \tau) = \min_{y \in \mathbb{R}^p} \left[\langle h, y - x \rangle + \frac{1}{2\tau} \|y - x\|^2 \right]. \quad (30)$$

27

From (Ghadimi et al., 2020, Lemma 3.2), the minimum problem in Eq. (30) has the solution $y^\star = x - \tau h$, and $\eta(x, h, \tau)$ is $L_{\nabla \eta}$-smooth, where

$$L_{\nabla \eta} = 2\sqrt{\left(1 + \frac{1}{\tau}\right)^2 + \left(1 + \frac{\tau}{2}\right)^2}. \tag{31}$$

As $\Phi(x)$ is $L_{\nabla \Phi}$-smooth, we have:

$$\Phi(\bar{x}^{(t+1)}) - \Phi(\bar{x}^{(t)}) \le \tau \alpha \langle \nabla \Phi(\bar{x}^{(t)}), -\bar{h}^{(t)} \rangle + \frac{L_{\nabla \Phi}}{2} \left\| \bar{x}^{(t+1)} - \bar{x}^{(t)} \right\|^2. \tag{32}$$

Also, according to (Chen et al., 2023b, Lemma B.8.), we have:

$$\begin{aligned}
& \eta(\bar{x}^{(t)}, \bar{h}^{(t)}, \tau) - \eta(\bar{x}^{(t+1)}, \bar{h}^{(t+1)}, \tau) \\
& \le -\theta \tau \left\| \bar{h}^{(t)} \right\|^2 + \theta \tau \langle \bar{\omega}^{(t+1)}, \bar{h}^{(t)} \rangle + \frac{L_{\nabla \eta}}{2} \left(\left\| \bar{x}^{(t+1)} - \bar{x}^{(t)} \right\|^2 + \left\| \bar{h}^{(t+1)} - \bar{h}^{(t)} \right\|^2 \right).
\end{aligned} \tag{33}$$

Define $E_t = \Phi(\bar{x}^{(t)}) - \inf_x \Phi(x) - c_3^{-1} \eta(\bar{x}^{(t)}, \bar{h}^{(t)}, \tau)$. Substituting (32) into (33) and taking the expectation with respect to $\mathcal{F}^{(t)}$, we have:

$$\begin{aligned}
& \alpha \left\| \bar{h}^{(t)} \right\|^2 \\
& \le \frac{1}{c_3 \tau} \left(\mathbb{E}_t \left[\eta(\bar{x}^{(t+1)}, \bar{h}^{(t+1)}, \tau) \right] - \eta(\bar{x}^{(t)}, \bar{h}^{(t)}, \tau) \right) + \frac{1}{\tau} \left(\Phi(\bar{x}^{(t)}) - \mathbb{E}_t \left[\Phi(\bar{x}^{(t+1)}) \right] \right) \\
& \quad - \alpha \mathbb{E}_t \left[\left\langle \nabla \Phi(\bar{x}^{(t)}) - \bar{\omega}^{(t+1)}, \bar{h}^{(t)} \right\rangle \right] + \frac{c_3 L_{\nabla \Phi} + L_{\nabla \eta}}{2 c_3 \tau} \left\| \bar{x}^{(t+1)} - \bar{x}^{(t)} \right\|^2 \\
& \quad + \frac{L_{\nabla \eta}}{2 c_3 \tau} \mathbb{E}_t \left[\left\| \bar{h}^{(t+1)} - \bar{h}^{(t)} \right\|^2 \right] \\
& \le \frac{1}{\tau} (E_t - \mathbb{E}_t [E_{t+1}]) + \alpha \left(\left\| \nabla \Phi(\bar{x}^{(t)}) - \mathbb{E}_t[\bar{\omega}^{(t+1)}] \right\|^2 + \frac{1}{4} \left\| \bar{h}^{(t)} \right\|^2 \right) + \frac{\alpha}{4} \left\| \bar{h}^{(t)} \right\|^2 \\
& \quad + \frac{\alpha^2}{4} \mathbb{E}_t \left[\left\| \nabla \Phi(\bar{x}^{(t)}) - \bar{\omega}^{(t+1)} \right\|^2 \right] + \frac{\alpha^2}{4} \mathbb{E}_t \left[\left\| \bar{h}^{(t)} - \nabla \Phi(\bar{x}^{(t)}) \right\|^2 \right] \\
& \le \frac{1}{\tau} (E_t - \mathbb{E}_t [E_{t+1}]) + \alpha \left(\left\| \nabla \Phi(\bar{x}^{(t)}) - \mathbb{E}_t[\bar{\omega}^{(t+1)}] \right\|^2 + \frac{1}{4} \left\| \bar{h}^{(t)} \right\|^2 \right) + \frac{\alpha}{4} \left\| \bar{h}^{(t)} \right\|^2 \\
& \quad + \frac{\alpha}{4} \mathbb{E}_t \left[\left\| \nabla \Phi(\bar{x}^{(t)}) - \bar{\omega}^{(t+1)} \right\|^2 \right] + \frac{\alpha^2}{4} \left\| \bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}] \right\|^2 \\
& \quad + \frac{\alpha}{4} \mathbb{E}_t \left[\left\| \bar{h}^{(t)} - \nabla \Phi(\bar{x}^{(t)}) \right\|^2 \right],
\end{aligned} \tag{34}$$

where the first inequality uses the fact that $\bar{x}^{(t)}$ and $\bar{h}^{(t)}$ are measurable with respect to $\mathcal{F}_t$. The second inequality uses the assumption $\alpha \le \frac{c_3}{2\tau(c_3 L_{\nabla \Phi} + L_{\nabla \eta})}$, $\frac{c_3}{\tau^2} \le \frac{1}{20}$, and the fact that $L_{\nabla \eta} \le \frac{5}{\tau}$. The third inequality uses the assumption that $\alpha \le 1$.

28

Then, taking the expectation and summation on both sides of (34), we have:

$$
\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right]
$$

$$
\leq \frac{2}{\tau}\mathbb{E}[E_0] + \alpha^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2\right] + 3\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right]
$$

$$
+ \frac{1}{2}\cdot\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] \tag{35}
$$

$$
\leq \frac{2}{\tau}\mathbb{E}[E_0] + \alpha^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2\right] + 6\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \bar{q}^{(t)}\right\|^2\right]
$$

$$
+ 6\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + \frac{1}{2}\cdot\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right].
$$

Then we consider the term $\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2$. Note that:

$$
\begin{aligned}
\bar{h}^{(t+1)} - \nabla\Phi(\bar{x}^{(t+1)}) =& (1-\theta)\left(\bar{h}^{(t+1)} - \nabla\Phi(\bar{x}^{(t)})\right) + \theta\left(\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right) \\
&+ \nabla\Phi(\bar{x}^{(t)}) - \nabla\Phi(\bar{x}^{(t+1)}) + \theta\left(\bar{\omega}^{(t+1)} - \bar{q}^{(t)}\right).
\end{aligned} \tag{36}
$$

Taking the conditional expectation with respect to $\mathcal{F}^{(t)}$ and noting that $\bar{x}^{(t+1)}$ is measurable with respect to $\mathcal{F}_t$, we have:

$$
\mathbb{E}_t\left[\left\|\bar{h}^{(t+1)} - \nabla\Phi(\bar{x}^{(t+1)})\right\|^2\right]
$$

$$
\leq (1-\theta)\mathbb{E}_t\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + 2\theta\mathbb{E}_t\left[\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \bar{q}^{(t)}\right\|^2\right]
$$

$$
+ 2\theta\mathbb{E}_t\left[\left\|\left(\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right) + \frac{1}{\theta}\left(\nabla\Phi(\bar{x}^{(t)}) - \nabla\Phi(\bar{x}^{(t+1)})\right)\right\|^2\right] + \theta^2\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2
$$

$$
\leq (1-\theta)\mathbb{E}_t\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + 2\theta\mathbb{E}_t\left[\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \bar{q}^{(t)}\right\|^2\right] + 4\theta\mathbb{E}_t\left[\left\|\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right]
$$

$$
+ \frac{4L_{\nabla\Phi}^2\tau^2\alpha^2}{\theta}\mathbb{E}_t\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \theta^2\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2
$$

$$
\tag{37}
$$

where the first inequality is due to the convex of 2-norm $\|\cdot\|^2$, and the last equality is from the Lipchitz continuity of $\nabla\Phi$.

Taking expectation on both sides and then taking summation from $t = 0$ to $T - 1$ and, we have:

$$\alpha \sum_{t=0}^{T} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right]$$

$$\leq \frac{1}{c_3}\mathbb{E}\left[\left\|\bar{h}^{(0)} - \nabla\Phi(\bar{x}^{(0)})\right\|^2\right] + 2\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \bar{q}^{(t)}\right\|^2\right]$$

$$+ 4\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + \frac{4L_{\nabla\Phi}^2\tau^2}{c_3^2}\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] \tag{38}$$

$$+ c_3\alpha^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2\right].$$

Then, we consider the term $\left\|\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2$. Note that:

$$\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)}) = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f_i(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right) - \nabla_{22}g_i(x_i^{(t)}, y_i^{(t)})(z_i^{(t)} - \bar{z}_\star^{(t)})\right]$$

$$- \frac{1}{N}\sum_{i=1}^{N}\left(\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22}g_n(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right)\bar{z}_\star^{(t)}. \tag{39}$$

Then, taking the norm on both sides and using Assumption 1, we have:

$$\left\|\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2$$

$$\leq 3 \cdot \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f_i(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right\|^2 + 3 \cdot \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)})\right\|^2 \left\|z_i^{(t)} - \bar{z}_\star^{(t)}\right\|^2$$

$$+ 3 \cdot \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22}g_i(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right\|^2 \left\|\bar{z}_\star^{(t)}\right\|^2$$

$$\leq 6\left(\frac{L_f^2 L_{\nabla^2 g}^2}{\mu_g^2} + L_{\nabla f}^2\right)\frac{1}{N}\left(\left\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_F^2 + \left\|\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right\|_F^2 + N\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right)$$

$$+ 6L_{\nabla^2 g}^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\left(\left\|z_i^{(t)} - \bar{z}^{(t)}\right\|^2 + \left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right)$$

$$\leq 6L^2\frac{1}{N}\Delta_t^2 + 6\left(\frac{L_f^2 L_{\nabla^2 g}^2}{\mu_g^2} + L_{\nabla f}^2\right)\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2 + 6L_{\nabla^2 g}^2\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2. \tag{40}$$

Next, considering the term $\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \bar{q}^{(t)}\right\|^2$, we have:

$$\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \bar{q}^{(t)}\right\|^2 \leq \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbb{E}_t[p_{J,i}^{(t)}] - \nabla_{12}^2 g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}\right\|^2 \leq \iota^2. \tag{41}$$

Finally, we consider the term $\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2$. From Assumption 3 and Eq. (22b), we have:

$$
\begin{aligned}
\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2 \leq & \frac{2}{N^2} \sum_{i=1}^{N} \left[\left\|p_{J,i}^{(t)} - \mathbb{E}_t[p_{J,i}^{(t)}]\right\|^2 + \left\|u_{1,i}^t - \mathbb{E}_t[u_{1,i}^t]\right\|^2 \right] \\
\leq & \frac{2}{N^2} \sum_{i=1}^{N} \left(1 + \left\|z_i^{(t)}\right\|^2 \right) \sigma^2 \\
\leq & \frac{6}{N} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2 + \frac{6\sigma^2}{N} \left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2 + \frac{6\sigma^2}{N} \frac{1}{N} \sum_{i=1}^{N} \left\|z_i^{(t)} - \bar{z}^{(t)}\right\|^2 .
\end{aligned}
\tag{42}
$$

Adding (35) and (38) together, and then plugging (40) into it, we have:

$$
\begin{aligned}
& \alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \alpha \sum_{t=0}^{T} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] \\
\leq & \frac{2}{\tau} \mathbb{E}[E_0] + \frac{1}{c_3} \mathbb{E}\left[\left\|\bar{h}^{(0)} - \nabla\Phi(\bar{x}^{(0)})\right\|^2\right] + \frac{4L_{\nabla\Phi}^2 \tau^2}{c_3^2} \alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] \\
& + 10\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\mathbb{E}_t[\bar{\omega}^{(t+1)}] - \bar{q}^{(t)}\right\|^2\right] + \frac{1}{2} \cdot \alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] \\
& + 10\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{q}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + (1+c_3)\alpha^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{\omega}^{(t+1)} - \mathbb{E}_t[\bar{\omega}^{(t+1)}]\right\|^2\right] \\
\leq & \frac{2}{\tau} \mathbb{E}[E_0] + \frac{1}{c_3} \mathbb{E}\left[\left\|\bar{h}^{(0)} - \nabla\Phi(\bar{x}^{(0)})\right\|^2\right] + \frac{4L_{\nabla\Phi}^2 \tau^2}{c_3^2} \alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] \\
& + \frac{1}{2} \cdot \alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + \left(60L^2 + \frac{6(1+c_3)\alpha\sigma^2}{N} \right) \frac{1}{N} \cdot \alpha \sum_{t=0}^{T-1} \mathbb{E}[\Delta_t^2] \\
& + \frac{6(1+c_3)T\alpha^2}{N} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2 + 10T\alpha\iota^2 + 60 \left(\frac{L_f^2 L_{\nabla^2 g}^2}{\mu_g^2} + L_{\nabla f}^2 \right) \sum_{t=0}^{T-1} \alpha\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] \\
& + \left(60L_{\nabla^2 g}^2 + \frac{6(1+c_3)\alpha\sigma^2}{N} \right) \sum_{t=0}^{T-1} \alpha\mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right] .
\end{aligned}
\tag{43}
$$

Then we complete the proof of this lemma. $\square$

A.5 Estimation bias of y, z

In this section, we first estimate the convergence error induced by y and z, and then estimate the error between $\nabla\Phi(\bar{x}^{(t+1)})$ and $\bar{h}^{(t+1)}$.

Lemma 5. *Suppose Assumption 1 and 3 hold, and the step-size β satisfies:*

$$
\beta(\mu_g + L_{\nabla g}) \leq 1.
\tag{44}
$$

Then in Algorithm 1, we have:

$$\sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right]$$

$$\leq C_{yx} \sum_{t=0}^{T-1} \tau^2 \alpha \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + C_{yD} \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\frac{1}{N}\Delta_t^2\right] + C_{y1}\frac{T\alpha^2}{N} + C_{y2}, \tag{45}$$

where the constants are defined as:

$$C_{yx} = \frac{3L_{y^\star}^2}{c_1^2 \mu_g^2} = \mathcal{O}\left(\frac{\kappa^4}{c_1^2}\right), \quad C_{yD} = \frac{6L_{\nabla g}^2}{\mu_g^2 \kappa^2} = \mathcal{O}\left(1\right),$$

$$C_{y1} = \frac{2c_1\sigma_{g1}^2}{\mu_g} = \mathcal{O}\left(c_1 \kappa\right), \quad C_{y2} = \frac{1}{c_1\mu_g}\mathbb{E}\left[\left\|\bar{y}^{(0)} - \bar{y}_\star^{(0)}\right\|^2\right] = \mathcal{O}\left(\frac{\kappa}{c_1}\right).$$

Proof. Firstly, from Jensen's Inequality and the Lipchitz continuity of $y^\star(x)$, we have:

$$\left\|\bar{y}^{(t+1)} - \bar{y}_\star^{(t+1)}\right\|^2 \leq \left(1 + \frac{\beta\mu_g}{2}\right)\left\|\bar{y}^{(t+1)} - \bar{y}_\star^{(t)}\right\|^2 + \left(1 + \frac{2}{\beta\mu_g}\right)\left\|\bar{y}_\star^{(t)} - \bar{y}_\star^{(t+1)}\right\|^2$$

$$\leq \left(1 + \frac{\beta\mu_g}{2}\right)\left\|\bar{y}^{(t+1)} - \bar{y}_\star^{(t)}\right\|^2 + \left(\alpha^2 + \frac{2\alpha^2}{\beta\mu_g}\right)L_{y^\star}^2 \tau^2 \left\|\bar{h}^{(t)}\right\|^2. \tag{46}$$

Then, as g is μ_g-strongly convex with respect to y, we have:

$$\mathbb{E}_t\left[\left\|\bar{y}^{(t+1)} - \bar{y}_\star^{(t)}\right\|^2\right]$$

$$= \left\|\bar{y}^{(t)} - \beta \cdot \frac{1}{N}\sum_{i=1}^{N}\nabla_2 g_i(x_i^{(t)}, y_i^{(t)}) - \bar{y}_\star^{(t)}\right\|^2 + \beta^2 \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_t\left[\left\|\nabla_2 g_i(x_i^{(t)}, y_i^{(t)}) - v_i^{(t)}\right\|^2\right]$$

$$\leq \left(1 + \frac{\beta\mu_g}{2}\right)\left\|\bar{y}^{(t)} - \beta\nabla_2 g(\bar{x}^{(t)}, \bar{y}^{(t)}) - \bar{y}_\star^{(t)}\right\|^2 \tag{47}$$

$$+ \left(\beta^2 + \frac{2\beta}{\mu_g}\right)\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla_2 g_n(x_i^{(t)}, y_i^{(t)}) - \nabla_2 g_n(\bar{x}^{(t)}, \bar{y}^{(t)})\right\|^2 + \frac{\beta^2\sigma_{g1}^2}{N}$$

$$\leq \left(1 + \frac{\beta\mu_g}{2}\right)(1 - \beta\mu_g)^2\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2 + \frac{3\beta L_{\nabla g}^2}{N\mu_g\kappa^2}\Delta_t^2 + \frac{\beta^2\sigma^2}{N},$$

where the first equation is due to Assumption 3, the first inequality is from Jensen's inequality and Assumption 3, and the last inequality is from $\beta\mu_g < 1$ and Lemma 1.

Taking the conditional expectation of (46) and using (47), we have:

$$
\begin{aligned}
&\mathbb{E}_t \left[\left\| \bar{y}^{(t+1)} - \bar{y}_\star^{(t+1)} \right\|^2 \right] \\
&\leq \left(1 + \frac{\beta\mu_g}{2} \right)^2 (1 - \beta\mu_g)^2 \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 + \left(1 + \frac{\beta\mu_g}{2} \right) \left(\frac{3\beta L_{\nabla g}^2}{N\mu_g\kappa^2} \Delta_t^2 + \frac{\beta^2\sigma^2}{N} \right) \\
&\quad + \left(\alpha^2 + \frac{2\alpha^2}{\beta\mu_g} \right) L_{y^\star}^2 \tau^2 \left\| \bar{h}^{(t)} \right\|^2 \\
&\leq (1 - \beta\mu_g) \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 + \frac{6\beta L_{\nabla g}^2}{N\mu_g\kappa^2} \Delta_t^2 + \frac{2\beta^2\sigma^2}{N} + \frac{3\tau^2\alpha^2 L_{y^\star}^2}{\beta\mu_g} \left\| \bar{h}^{(t)} \right\|^2,
\end{aligned}
$$

(48)

where the second inequality holds because $-\frac{3}{4}(\beta\mu_g)^2 + \frac{1}{2}(\beta\mu_g)^3 + \frac{1}{4}(\beta\mu_g)^4 \leq 0$ when $0 \leq \beta\mu_g \leq 1$. Taking summation and expectation on both sides, we have:

$$
\begin{aligned}
&\beta\mu_g \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right] \\
&\leq \frac{3\tau^2\alpha^2 L_{y^\star}^2}{\beta\mu_g} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \bar{h}^{(t)} \right\|^2 \right] + \frac{6\beta L_{\nabla g}^2}{\mu_g\kappa^2} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{N}\Delta_t^2 \right] + \frac{2T\beta^2\sigma^2}{N} + \mathbb{E} \left[\left\| \bar{y}^{(0)} - \bar{y}_\star^{(0)} \right\|^2 \right].
\end{aligned}
$$

(49)

Finally, multiplying $\frac{1}{c_1\mu_g}$ on both sides of (49), we have:

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \alpha\mathbb{E} \left[\left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right] \\
&\leq \frac{3\tau^2 L_{y^\star}^2}{c_1^2\mu_g^2} \sum_{t=0}^{T-1} \alpha\mathbb{E} \left[\left\| \bar{h}^{(t)} \right\|^2 \right] + \frac{6L_{\nabla g}^2}{\mu_g^2} \sum_{t=0}^{T-1} \alpha\mathbb{E} \left[\frac{1}{N}\Delta_t^2 \right] + \frac{2c_1\sigma^2}{\mu_g}\frac{T\alpha^2}{N} + \frac{1}{c_1\mu_g}\mathbb{E} \left[\left\| \bar{y}^{(0)} - \bar{y}_\star^{(0)} \right\|^2 \right] \\
&= C_{yx} \sum_{t=0}^{T-1} \tau^2\alpha\mathbb{E} \left[\left\| \bar{h}^{(t)} \right\|^2 \right] + C_{yD} \sum_{t=0}^{T-1} \alpha\mathbb{E} \left[\frac{1}{N}\Delta_t^2 \right] + C_{y1}\frac{T\alpha^2}{N} + C_{y2},
\end{aligned}
$$

(50)

where $C_{yx} = \frac{3L_{y^\star}^2}{c_1^2\mu_g^2}$, $C_{yD} = \frac{6L_{\nabla g}^2}{\mu_g^2\kappa^2}$, $C_{y1} = \frac{2c_1\sigma^2}{\mu_g}$, $C_{y2} = \frac{1}{c_1\mu_g}\mathbb{E} \left[\left\| \bar{y}^{(0)} - \bar{y}_\star^{(0)} \right\|^2 \right]$. $\qquad\square$

Lemma 6. *Suppose Assumption 1 and 3 hold. If the step-size γ satisfies:*

$$
-\frac{1}{6}\gamma\mu_g + \frac{6}{N}\gamma^2\sigma^2 + \frac{1}{2}\gamma^3\mu_g^3 \leq 0, \quad \gamma\mu_g \leq \min\left\{ 1, \frac{NL^2}{\mu_g\sigma^2} \right\},
$$

(51)

in Algorithm 1, we have:

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \alpha\mathbb{E} \left[\left\| \bar{z}^{(t+1)} - \bar{z}_\star^{(t+1)} \right\|^2 \right] \\
&\leq C_{zx} \sum_{t=0}^{T-1} \tau^2\alpha\mathbb{E} \left[\left\| \bar{h}^{(t)} \right\|^2 \right] + C_{zD} \sum_{t=0}^{T-1} \alpha\mathbb{E} \left[\frac{1}{N}\Delta_t^2 \right] + C_{z1}\frac{T\alpha^2}{N} + C_{z2} + C_{z3}T\alpha\iota^2,
\end{aligned}
$$

(52)

where the constants are defined as:

$$C_{zx} = \frac{32}{\mu_g^2} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{yx} + \frac{8 L_{z^\star}^2}{c_2^2 \mu_g^2} = \mathcal{O}\left(\frac{\kappa^8}{c_1^2} + \frac{\kappa^8}{c_2^2} \right),$$

$$C_{zD} = \frac{32}{\mu_g^2} \left[\left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{yD} + 2L^2 \right] = \mathcal{O}(\kappa^4),$$

$$C_{z1} = \frac{32}{\mu_g^2} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{y1} + \frac{8 c_2}{\mu_g} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2 = \mathcal{O}(c_1 \kappa^5 + c_2 \kappa^3),$$

$$C_{z2} = \frac{32}{\mu_g^2} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{y2} + \frac{1}{c_2 \mu_g} \mathbb{E}\left[\left\| \bar{z}^{(0)} - \bar{z}_\star^{(0)} \right\|^2 \right] = \mathcal{O}\left(\frac{\kappa^5}{c_1} + \frac{\kappa}{c_2} \right),$$

$$C_{z3} = \frac{16}{\mu_g^2} = \mathcal{O}(\kappa^2).$$

Proof. For the term $\bar{z}^{(t+1)} - \bar{z}_\star^{(t+1)}$, we have:

$$
\begin{aligned}
\left\| \bar{z}^{(t+1)} - \bar{z}_\star^{(t+1)} \right\|^2 &\leq \left(1 + \frac{\gamma \mu_g}{3} \right) \left\| \bar{z}^{(t+1)} - \bar{z}_\star^{(t)} \right\|^2 + \left(1 + \frac{3}{\gamma \mu_g} \right) \left\| \bar{z}_\star^{(t)} - \bar{z}_\star^{(t+1)} \right\|^2 \\
&\leq \left(1 + \frac{\gamma \mu_g}{3} \right) \left\| \bar{z}^{(t+1)} - \bar{z}_\star^{(t)} \right\|^2 + \left(\alpha^2 + \frac{3\alpha^2}{\gamma \mu_g} \right) \tau^2 L_{z^\star}^2 \left\| \bar{h}^{(t)} \right\|^2,
\end{aligned}
\tag{53}
$$

where the first inequality is due to Jensen's Inequality and the second is due to the Lipchitz continuity of $z^\star(x)$. Then, from the definition of $\bar{z}_\star^{(t)}$, we have:

$$
\begin{aligned}
&\bar{z}^{(t+1)} - \bar{z}_\star^{(t)} \\
=& \frac{1}{N} \sum_{i=1}^N (I - \gamma \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}))(\bar{z}^{(t)} - \bar{z}_\star^{(t)}) - \frac{\gamma}{N} \sum_{i=1}^N \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)})(z_i^{(t)} - \bar{z}^{(t)}) \\
&- \frac{\gamma}{N} \sum_{i=1}^N \left[\left(\nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22} g(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) \right) \bar{z}_\star^{(t)} - \left(\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) \right) \right] \\
&- \frac{\gamma}{N} \sum_{i=1}^N \left[(p_{H,i} - \mathbb{E}_t[p_{H,i}]) - (u_{1,i}^{(t+1)} - \nabla_1 f_i(x_i^{(t)}, y_i^{(t)})) - (\mathbb{E}_t[p_{H,i}] - \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)}) \right],
\end{aligned}
\tag{54}
$$

Taking the conditional expectation with respect to $\mathcal{F}^{(t)}$ on both sides, with Jensen's Inequality, and Assumption 3, we have:

$$
\begin{aligned}
&\mathbb{E}_t\left[\left\|\bar{z}^{(t+1)} - \bar{z}_\star^{(t)}\right\|^2\right] \\
&\leq \left(1 + \frac{\gamma\mu_g}{2}\right)\frac{1}{N}\sum_{i=1}^N\left\|(I - \gamma\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}))(\bar{z}^{(t)} - \bar{z}_\star^{(t)})\right\|^2 \\
&\quad + 4\gamma^2 L_{\nabla g}^2\left(1 + \frac{2}{\gamma\mu_g}\right)\cdot\frac{1}{N}\left\|\mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)}\right\|_F^2 \\
&\quad + 4\gamma^2\left(1 + \frac{2}{\gamma\mu_g}\right)\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22}g(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right)\right\|^2\left\|\bar{z}_\star^{(t)}\right\|^2 \\
&\quad + 4\gamma^2\left(1 + \frac{2}{\gamma\mu_g}\right)\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right)\right\|^2 \\
&\quad + 4\gamma^2\left(1 + \frac{2}{\gamma\mu_g}\right)\left\|\frac{1}{N}\sum_{i=1}^N\left(\mathbb{E}_t[p_{H,i}] - \nabla_{22}^2 g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}\right)\right\|^2 \\
&\quad + \frac{2\gamma^2}{N^2}\sum_{i=1}^N\mathbb{E}_t\left[\|p_{H,i} - \mathbb{E}_t[p_{H,i}]\|^2\right] + \frac{2\gamma^2}{N^2}\sum_{i=1}^N\mathbb{E}_t\left[\left\|u_{1,i}^{(t+1)} - \nabla_1 f_i(x_i^{(t)}, y_i^{(t)})\right\|^2\right].
\end{aligned}
\tag{55}
$$

From the Lipchitz continuity of $\nabla_{22}g_i$, we have:

$$
\begin{aligned}
&\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22}g_i(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right)\right\|^2\left\|\bar{z}_\star^{(t)}\right\|^2 \\
&\leq \frac{2L_f^2}{\mu_g^2}\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22}g_i(\bar{x}^{(t)}, \bar{y}^{(t)})\right)\right\|^2 \\
&\quad + \frac{2L_f^2}{\mu_g^2}\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla_{22}g_i(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_{22}g_i(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right)\right\|^2 \\
&\leq \frac{2L_f^2}{N\mu_g^2}\sum_{i=1}^N\left(\left\|\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22}g_i(\bar{x}^{(t)}, \bar{y}^{(t)})\right\|^2 + \left\|\nabla_{22}g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22}g_i(\bar{x}^{(t)}, \bar{y}^{(t)})\right\|^2\right) \\
&\leq \frac{2L_{\nabla^2 g}^2 L_f^2}{N\mu_g^2}\left(\left\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_F^2 + \left\|\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right\|_F^2 + N\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right).
\end{aligned}
\tag{56}
$$

Similarly, from the Lipchitz continuity of $\nabla_1 f_i$, we have:

$$
\begin{aligned}
&\left\| \frac{1}{N} \sum_{i=1}^{N} \left(\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f_i(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) \right) \right\|^2 \\
&\leq 2 \left\| \frac{1}{N} \sum_{i=1}^{N} \left(\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f_i(\bar{x}^{(t)}, \bar{y}^{(t)}) \right) \right\|^2 \\
&\quad + 2 \left\| \frac{1}{N} \sum_{i=1}^{N} \left(\nabla_1 f_i(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_1 f_i(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) \right) \right\|^2 \\
&\leq \frac{2 L_{\nabla f}^2}{N} \left(\left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 + \left\| \mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|_F^2 + N \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right).
\end{aligned}
\tag{57}
$$

We also note that g_i is μ_g-strongly convex with respect to y and $\gamma \mu_g \leq 1$, then:

$$
\left\| (I - \gamma \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}))(\bar{z}^{(t)} - \bar{z}_\star^{(t)}) \right\|^2 \leq (1 - \gamma \mu_g)^2 \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2
\tag{58}
$$

Substituting (22a), (56), (57), and (58) into (55), and then using Assumption 3, we get:

$$
\begin{aligned}
&\mathbb{E}_t \left[\left\| \bar{z}^{(t+1)} - \bar{z}_\star^{(t)} \right\|^2 \right] \\
&\leq \left(1 + \frac{\gamma \mu_g}{2} \right) (1 - \gamma \mu_g)^2 \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 + \frac{12 \gamma L_{\nabla g}^2}{\mu_g} \cdot \frac{1}{N} \left\| \mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|_F^2 \\
&\quad + \frac{24 \gamma}{N \mu_g} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) \left(\left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 + \left\| \mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|_F^2 + N \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right) \\
&\quad + \frac{12 \gamma}{\mu_g} \iota^2 + \frac{2 \gamma^2}{N} \sigma^2 + \frac{6 \gamma^2}{N} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| z_i^{(t)} - \bar{z}^{(t)} \right\|^2 + \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 + \left\| \bar{z}_\star^{(t)} \right\|^2 \right) \sigma^2 \\
&\leq \left(\left(1 + \frac{\gamma \mu_g}{2} \right) (1 - \gamma \mu_g)^2 + \frac{6 \gamma^2 \sigma^2}{N} \right) \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 + \frac{24 \gamma}{\mu_g} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \\
&\quad + \frac{24 \gamma}{\mu_g} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) \cdot \frac{1}{N} \left(\left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 + \left\| \mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|_F^2 \right) \\
&\quad + \left(\frac{12 \gamma L_{\nabla g}^2}{\mu_g} + \frac{6 \gamma^2 \sigma^2}{N} \right) \cdot \frac{1}{N} \left\| \mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|_F^2 + \frac{6 \gamma^2}{N} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2 + \frac{12 \gamma}{\mu_g} \iota^2 \\
&\leq \left(1 - \frac{4 \gamma \mu_g}{3} \right) \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 + \frac{24 \gamma}{\mu_g} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \\
&\quad + \frac{48 \gamma L^2}{\mu_g} \cdot \frac{1}{N} \Delta_t^2 + \frac{6 \gamma^2}{N} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2 + \frac{12 \gamma}{\mu_g} \iota^2,
\end{aligned}
\tag{59}
$$

where the last equation uses the assumption $-\frac{1}{6}\gamma\mu_g + \frac{6}{N}\gamma^2\sigma^2 + \frac{1}{2}\gamma^3\mu_g^3 \le 0$, $\gamma\mu_g \le \frac{NL^2}{\mu_g\sigma_{g2}^2}$,

and the fact that $\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \le (1 + \kappa^2)L^2 \le 2\kappa^2 L^2$.

Plugging (59) into (53) and noting that $1 + \frac{\gamma\mu_g}{3} \le \frac{4}{3}$, we have:

$$
\mathbb{E}_t\left[\left\|\bar{z}^{(t+1)} - \bar{z}_\star^{(t+1)}\right\|^2\right]
$$
$$
\le \left(1 + \frac{\gamma\mu_g}{3}\right)\left(1 - \frac{4\gamma\mu_g}{3}\right)\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2 + \left(\alpha^2 + \frac{3\alpha^2}{\gamma\mu_g}\right)\tau^2 L_{z^\star}^2\left\|\bar{h}^{(t)}\right\|^2 + \frac{64\gamma}{\mu_g}L^2\frac{1}{N}\Delta_t^2
$$
$$
+ \frac{32\gamma}{\mu_g}\left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2\right)\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2 + \frac{8\gamma^2}{N}\left(1 + \frac{L_f^2}{\mu_g^2}\right)\sigma^2 + \frac{16\gamma}{\mu_g}\iota^2 \qquad (60)
$$
$$
\le (1 - \gamma\mu_g)\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2 + \frac{8L_{z^\star}^2\tau^2\alpha^2}{\gamma\mu_g}\left\|\bar{h}^{(t)}\right\|^2 + \frac{64\gamma}{\mu_g}L^2\frac{1}{N}\Delta_t^2
$$
$$
+ \frac{32\gamma}{\mu_g}\left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2\right)\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2 + \frac{8\gamma^2}{N}\left(1 + \frac{L_f^2}{\mu_g^2}\right)\sigma^2 + \frac{16\gamma}{\mu_g}\iota^2.
$$

Taking summation and expectation on both sides, we obtain:

$$
\gamma\mu_g\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{z}^{(t+1)} - \bar{z}_\star^{(t+1)}\right\|^2\right]
$$
$$
\le \mathbb{E}\left[\left\|\bar{z}^{(0)} - \bar{z}_\star^{(0)}\right\|^2\right] + \frac{8L_{z^\star}^2\tau^2\alpha^2}{\gamma\mu_g}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \frac{64\gamma}{\mu_g}L^2\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right] \qquad (61)
$$
$$
+ \frac{32\gamma}{\mu_g}\left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \frac{8T\gamma^2}{N}\left(1 + \frac{L_f^2}{\mu_g^2}\right)\sigma^2 + \frac{16T\gamma}{\mu_g}\iota^2.
$$

Finally, from (45), we have:

$$
\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{z}^{(t+1)} - \bar{z}_\star^{(t+1)}\right\|^2\right]
$$
$$
\le \frac{1}{c_2\mu_g}\mathbb{E}\left[\left\|\bar{z}^{(0)} - \bar{z}_\star^{(0)}\right\|^2\right] + \frac{8L_{z^\star}^2\tau^2}{c_2^2\mu_g^2}\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \frac{64}{\mu_g^2}L^2\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right]
$$
$$
+ \frac{32}{\mu_g^2}\left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2\right)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \frac{T\alpha^2}{N}\frac{8c_2}{\mu_g}\left(1 + \frac{L_f^2}{\mu_g^2}\right)\sigma^2 + T\alpha\iota^2\frac{16}{\mu_g^2}
$$
$$
\le C_{zx}\sum_{t=0}^{T-1}\tau^2\alpha\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + C_{zD}\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right] + C_{z1}\frac{T\alpha^2}{N} + C_{z2} + C_{z3}T\alpha\iota^2,
$$
$$
\tag{62}
$$

where

$$C_{zx} = \frac{32}{\mu_g^2} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{yx} + \frac{8 L_{z^\star}^2}{c_2^2 \mu_g^2},$$

$$C_{zD} = \frac{32}{\mu_g^2} \left[\left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{yD} + 2L^2 \right],$$

$$C_{z1} = \frac{32}{\mu_g^2} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{y1} + \frac{8 c_2}{\mu_g} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2,$$

$$C_{z2} = \frac{32}{\mu_g^2} \left(\frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} + L_{\nabla f}^2 \right) C_{y2} + \frac{1}{c_2 \mu_g} \mathbb{E} \left[\left\| \bar{z}^{(0)} - \bar{z}_\star^{(0)} \right\|^2 \right],$$

$$C_{z3} = \frac{16}{\mu_g^2}.$$

$\square$

Now we can obtain the estimation errors for y, z in the following lemma:

Lemma 7. *Suppose Assumption 1, 2, 3, and 4 hold. If the step-sizes β, γ satisfy (44), (51), in Algorithm 1 we have:*

$$(1 + \kappa^2) \sum_{t=0}^{T-1} \alpha \mathbb{E} \left[\left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right] + \sum_{t=0}^{T-1} \alpha \mathbb{E} \left[\left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 \right]$$

$$\leq ((1 + \kappa^2) C_{yx} + C_{zx}) \sum_{t=0}^{T-1} \tau^2 \alpha \mathbb{E} \left[\left\| \bar{h}^{(t)} \right\|^2 \right] + ((1 + \kappa^2) C_{y1} + C_{z1}) \frac{T\alpha^2}{N} + ((1 + \kappa^2) C_{y2} + C_{z2})$$

$$+ C_{z3} T \alpha^2 \iota^2 + ((1 + \kappa^2) C_{yD} + C_{zD}) \sum_{t=0}^{T-1} \alpha \mathbb{E} \left[\frac{1}{N} \Delta_t^2 \right].$$

(63)

A.6 Consensus Lemmas

Then we analyze the consensus error Δ_t^2. In the beginning, we present some notations. The iteration of x in Algorithm 1 can be written as follows:

$$\mathbf{x}^{(t+1)} = W \left(\mathbf{x}^{(t)} - \tau \alpha \mathbf{h}^{(t)} \right). \tag{64}$$

By left-multiplying $\frac{1}{N} \mathbf{1} \mathbf{1}^\top$ on both sides of (64) and using Assumption 4, we have:

$$\bar{\mathbf{x}}^{(t+1)} = \frac{1}{N} \mathbf{1} \mathbf{1}^\top \left(\mathbf{x}^{(t)} - \tau \alpha \mathbf{h}^{(t)} \right). \tag{65}$$

Then, from (64) and (65), we have

$$\begin{aligned} \mathbf{x}^{(t+1)} - \bar{\mathbf{x}}^{(t+1)} &= \left(W - \frac{1}{N} \mathbf{1} \mathbf{1}^\top \right) \left(\mathbf{x}^{(t)} - \tau \alpha \mathbf{h}^{(t)} \right) \\ &= \left(W - \frac{1}{N} \mathbf{1} \mathbf{1}^\top \right) \left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} - \tau \alpha \mathbf{h}^{(t)} \right). \end{aligned} \tag{66}$$

Similarly, the consensus term in the iteration of y can be written as:

$$\mathbf{y}^{(t+1)} - \bar{\mathbf{y}}^{(t+1)} = \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right) \left(\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} - \beta\mathbf{v}^{(t+1)} \right), \tag{67}$$

and the consensus term in the iteration of z can be written as:

$$\mathbf{z}^{(t+1)} - \bar{\mathbf{z}}^{(t+1)} = \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right) \left(\mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)} - \gamma\mathbf{r}^{(t+1)} \right). \tag{68}$$

The following lemma 8 gives the upper bound of the consensus error of x:

Lemma 8. *Suppose Assumption 1, 2, 3, and 4 hold. Then in Algorithm 1, we have:*

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1} \left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 \right] \\
&\leq \left(\frac{208\tau^2\alpha^2\rho^2 L^2}{(1-\rho)^2} + \frac{6\tau^2\alpha^2\rho^2\sigma^2}{1-\rho} \right) \mathbb{E}\left[\sum_{t=0}^{T-1} \Delta_t^2 \right] + \frac{192\tau^2\alpha^2\rho^2}{(1-\rho)^2} NT \left(b_1^2 + \frac{L_f^2}{\mu_g^2} b_2^2 \right) \\
&\quad + \left(\frac{192N\tau^2\alpha^2\rho^2}{(1-\rho)^2} L_{\nabla g}^2 + \frac{6N\tau^2\alpha^2\rho^2}{1-\rho}\sigma^2 \right) \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 \right] + \frac{12\tau^2\alpha^2\rho^2}{(1-\rho)^2} NT\iota^2 \\
&\quad + \frac{6\tau^2\alpha^2\rho^2}{1-\rho} NT \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2 + \frac{1}{1-\rho} \left\| \mathbf{x}^{(0)} - \bar{\mathbf{x}}^{(0)} \right\|_F^2.
\end{aligned}
\tag{69}
$$

Proof. From the definition of $\mathcal{F}^{(t)}$, we know that for all $i = 1, 2, \cdots, N$, the term $x_i^{(t)}$ is measureable with respect to $\mathcal{F}^{(t-1)}$. We have:

$$
\begin{aligned}
&\mathbb{E}_{t-1}\left[\left\| \left(\mathbf{x}^{(t+1)} - \bar{\mathbf{x}}^{(t+1)} \right) - (1-\theta)\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right) \right\|_F^2 \right] \\
=&\mathbb{E}_{t-1}\left[\left\| \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right)\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} - \tau\alpha\mathbf{h}^{(t)} \right) - (1-\theta)\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right) \right\|_F^2 \right] \\
=& \left\| \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right)\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} - \tau\alpha(1-\theta)\mathbf{h}^{(t-1)} - \tau\alpha\theta\mathbb{E}_{t-1}[\omega^{(t)}] \right) - (1-\theta)(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}) \right\|_F^2 \\
&+ \tau^2\alpha^2\theta^2\rho^2 \sum_{i=1}^{N} \mathbb{E}_{t-1}\left[\left\| \omega_i^{(t)} - \mathbb{E}_{t-1}[\omega_i^{(t)}] \right\|^2 \right].
\end{aligned}
\tag{70}
$$

where the last equation is due to Assumption 3. Note that

$$-\tau\alpha \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right)\mathbf{h}^{(t-1)} = \left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right) - \left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \right)\left(\mathbf{x}^{(t-1)} - \bar{\mathbf{x}}^{(t-1)} \right). \tag{71}$$

It follows that:

$$
\begin{aligned}
& \mathbb{E}_{t-1}\left[\left\|\left(\mathbf{x}^{(t+1)} - \bar{\mathbf{x}}^{(t+1)}\right) - (1-\theta)\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right)\right\|_F^2\right] \\
& \leq \frac{1}{\rho}\left\|\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left[\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} - (1-\theta)\left(\mathbf{x}^{(t-1)} - \bar{\mathbf{x}}^{(t-1)}\right)\right]\right\|_F^2 \\
& \quad + \frac{\tau^2\alpha^2\theta^2}{1-\rho}\left\|\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbb{E}_{t-1}[\omega^{(t)}] - \overline{\mathbb{E}_{t-1}[\omega^{(t)}]}\right)\right\|_F^2 \\
& \quad + \tau^2\alpha^2\theta^2\rho^2\sum_{i=1}^N \mathbb{E}_{t-1}\left[\left\|\omega_i^{(t)} - \mathbb{E}_{t-1}[\omega_i^{(t)}]\right\|_F^2\right] \\
& \leq \rho\left\|\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right) - (1-\theta)\left(\mathbf{x}^{(t-1)} - \bar{\mathbf{x}}^{(t-1)}\right)\right\|_F^2 + \frac{\tau^2\alpha^2\theta^2\rho^2}{1-\rho}\sum_{i=1}^N \left\|\mathbb{E}_{t-1}[\omega_i^{(t)}] - \mathbb{E}_{t-1}[\bar{\omega}^{(t)}]\right\|^2 \\
& \quad + \tau^2\alpha^2\theta^2\rho^2\sum_{i=1}^N \mathbb{E}_{t-1}\left[\left\|\omega_i^{(t)} - \mathbb{E}_{t-1}[\omega_i^{(t)}]\right\|_F^2\right],
\end{aligned}
\tag{72}
$$

where the first inequality is due to Jensen's inequality and Assumption 4. Then taking expectation and summation on both sides of (72) from 0 to $T-1$ and using the fact that $x_i^{(0)} = x_i^{(1)} = 0$, we have:

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right) - (1-\theta)\left(\mathbf{x}^{(t-1)} - \bar{\mathbf{x}}^{(t-1)}\right)\right\|_F^2\right] \\
& \leq \frac{\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2}\sum_{t=0}^{T-1}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathbb{E}_{t-1}[\omega_i^{(t)}] - \mathbb{E}_{t-1}[\bar{\omega}^{(t)}]\right\|^2\right] \\
& \quad + \frac{\tau^2\alpha^2\theta^2\rho^2}{1-\rho}\sum_{t=0}^{T-1}\sum_{i=1}^N \mathbb{E}\left[\left\|\omega_i^{(t)} - \mathbb{E}_{t-1}[\omega_i^{(t)}]\right\|_F^2\right] + \frac{1}{1-\rho}\mathbb{E}\left[\left\|\mathbf{x}^{(0)} - \bar{\mathbf{x}}^{(0)}\right\|_F^2\right] \\
& \leq \frac{2\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2}\sum_{t=0}^{T-1}\sum_{i=1}^N \mathbb{E}\left[\left\|\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \frac{1}{N}\sum_{m=1}^N \nabla_1 f_m(x_m^{(t)}, y_m^{(t)})\right\|^2\right] \\
& \quad + \frac{2\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2}\sum_{t=0}^{T-1}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathbb{E}_{t-1}[p_{J,i}^{(t)}] - \mathbb{E}_{t-1}[\bar{p}_{J,i}^{(t)}]\right\|^2\right] \\
& \quad + \frac{\tau^2\alpha^2\theta^2\rho^2}{1-\rho}\sum_{t=0}^{T-1}\sum_{i=1}^N \mathbb{E}\left[\left\|\omega_i^{(t)} - \mathbb{E}_{t-1}[\omega_i^{(t)}]\right\|_F^2\right] + \frac{1}{1-\rho}\mathbb{E}\left[\left\|\mathbf{x}^{(0)} - \bar{\mathbf{x}}^{(0)}\right\|_F^2\right],
\end{aligned}
\tag{73}
$$

where we denote $\mathbf{x}^{(-1)} = \bar{\mathbf{x}}^{(-1)} = \mathbf{0}_{N\times d}$

Note that:

$$
\begin{aligned}
&\sum_{i=1}^{N} \mathbb{E}\left[\left\|\mathbb{E}_{t-1}[p_{J,i}^{(t)}] - \mathbb{E}_{t-1}[\bar{p}_{J,i}^{(t)}]\right\|^2\right] \\
&\leq 3\sum_{i=1}^{N} \mathbb{E}\left[\left\|\mathbb{E}_{t-1}[p_{J,i}^{(t)}] - \nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}\right\|^2\right] \\
&\quad + 3\sum_{i=1}^{N} \mathbb{E}\left[\left\|\nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)} - \frac{1}{N}\sum_{m=1}^{N} \nabla_{12}g_m(x_m^{(t)}, y_m^{(t)})z_m^{(t)}\right\|^2\right] \\
&\quad + 3\sum_{i=1}^{N} \mathbb{E}\left[\left\|\frac{1}{N}\sum_{m=1}^{N} \mathbb{E}_{t-1}[p_{J,m}^{(t)}] - \frac{1}{N}\sum_{m=1}^{N} \nabla_{12}g_m(x_m^{(t)}, y_m^{(t)})z_m^{(t)}\right\|^2\right] \\
&\leq 6\sum_{i=1}^{N} \mathbb{E}\left[\left\|\mathbb{E}_{t-1}[p_{J,i}^{(t)}] - \nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}\right\|^2\right] \\
&\quad + 6\sum_{i=1}^{N} \mathbb{E}\left[\left\|\nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)} - \frac{1}{N}\sum_{m=1}^{N} \nabla_{12}g_m(x_m^{(t)}, y_m^{(t)})z_m^{(t)}\right\|^2\right].
\end{aligned}
\tag{74}
$$

Then we can obtain:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T-1} \left\|\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right) - (1-\theta)\left(\mathbf{x}^{(t-1)} - \bar{\mathbf{x}}^{(t-1)}\right)\right\|_F^2\right] \\
&\leq \frac{2\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2} \sum_{t=1}^{T-1}\sum_{i=1}^{N} \mathbb{E}\left[\left\|\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \frac{1}{N}\sum_{m=1}^{N} \nabla_1 f_m(x_m^{(t)}, y_m^{(t)})\right\|^2\right] \\
&\quad + \frac{12\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2} \sum_{t=1}^{T-1}\sum_{i=1}^{N} \mathbb{E}\left[\left\|\mathbb{E}_{t-1}[p_{J,i}^{(t)}] - \nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)}\right\|^2\right] \\
&\quad + \frac{12\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2} \sum_{t=1}^{T-1}\sum_{i=1}^{N} \mathbb{E}\left[\left\|\nabla_{12}g_i(x_i^{(t)}, y_i^{(t)})z_i^{(t)} - \frac{1}{N}\sum_{m=1}^{N} \nabla_{12}g_m(x_m^{(t)}, y_m^{(t)})z_m^{(t)}\right\|^2\right] \\
&\quad + \frac{2\tau^2\alpha^2\theta^2\rho^2}{1-\rho} \sum_{t=1}^{T-1}\sum_{i=1}^{N} \left(\mathbb{E}\left[\left\|u_{i,x}^{(t)} - \mathbb{E}_t[u_{i,x}^{(t)}]\right\|^2\right] + \mathbb{E}\left[\left\|p_{J,i}^{(t)} - \mathbb{E}_{t-1}[\bar{p}_{J,i}^{(t)}]\right\|^2\right]\right).
\end{aligned}
\tag{75}
$$

For the first term on the right-hand side of (75), we have:

$$
\sum_{i=1}^{N} \left\| \nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \frac{1}{N} \sum_{m=1}^{N} \nabla_1 f_m(x_m^{(t)}, y_m^{(t)}) \right\|^2
$$

$$
= \sum_{i=1}^{N} \left\| \left[\nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f(\bar{x}^{(t)}, \bar{y}^{(t)}) \right] - \frac{1}{N} \sum_{m=1}^{N} \left[\nabla_1 f_m(x_m^{(t)}, y_m^{(t)}) - \nabla_1 f(\bar{x}^{(t)}, \bar{y}^{(t)}) \right] \right\|^2
$$

$$
\leq 4 \sum_{i=1}^{N} \left\| \nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f(\bar{x}^{(t)}, \bar{y}^{(t)}) \right\|^2
$$

$$
\leq 8 \sum_{i=1}^{N} \left\| \nabla_1 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_1 f_i(\bar{x}^{(t)}, \bar{y}^{(t)}) \right\|^2 + 8 \sum_{i=1}^{N} \left\| \nabla_1 f_i(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_1 f(\bar{x}^{(t)}, \bar{y}^{(t)}) \right\|^2
$$

$$
\leq 8 L_{\nabla f}^2 \Delta_t^2 + 8 N b_1^2.
\tag{76}
$$

For the third term of (75), we have:

$$
\sum_{i=1}^{N} \left\| \nabla_{12} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} - \frac{1}{N} \sum_{m=1}^{N} \nabla_{12} g_m(x_m^{(t)}, y_m^{(t)}) z_m^{(t)} \right\|^2
$$

$$
\leq 4 \sum_{i=1}^{N} \left\| \nabla_{12} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} - \nabla_{12} g(\bar{x}^{(t)}, \bar{y}^{(t)}) \bar{z}_\star^{(t)} \right\|^2
$$

$$
\leq 16 \sum_{i=1}^{N} \left\| \nabla_{12} g_i(x_i^{(t)}, y_i^{(t)}) \left(z_i^{(t)} - \bar{z}^{(t)} \right) \right\|^2 + 16 \sum_{i=1}^{N} \left\| \nabla_{12} g_i(x_i^{(t)}, y_i^{(t)}) \left(\bar{z}^{(t)} - \bar{z}_\star^{(t)} \right) \right\|^2
$$

$$
+ 16 \sum_{i=1}^{N} \left\| \left(\nabla_{12} g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{12} g_i(\bar{x}^{(t)}, \bar{y}^{(t)}) \right) \bar{z}_\star^{(t)} \right\|^2
$$

$$
+ 16 \sum_{i=1}^{N} \left\| \left(\nabla_{12} g_i(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_{12} g(\bar{x}^{(t)}, \bar{y}^{(t)}) \right) \bar{z}_\star^{(t)} \right\|^2
\tag{77}
$$

$$
\leq 16 L_{\nabla g}^2 \left\| \mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|_F^2 + 16 L_{\nabla g}^2 \sum_{i=1}^{N} \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2
$$

$$
+ 16 \kappa^2 L_{\nabla g}^2 \left(\left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 + \left\| \mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|_F^2 \right) + \frac{16 L_f^2}{\mu_g^2} N b_2^2
$$

$$
\leq 16 L_{\nabla g}^2 \sum_{i=1}^{N} \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 + 16 L_{\nabla g}^2 \Delta_t^2 + \frac{16 L_f^2}{\mu_g^2} N b_2^2.
$$

Finally, we consider the last term of (75). From Assumption 3 and Eq. (22b), we obtain:

$$
\begin{aligned}
&\sum_{i=1}^{N}\left(\mathbb{E}\left[\left\|u_{i,x}^{(t)}-\mathbb{E}_t[u_{i,x}^{(t)}]\right\|_F^2\right]+\mathbb{E}\left[\left\|p_{J,i}^{(t)}-\mathbb{E}_{t-1}[\bar{p}_{J,i}^{(t)}]\right\|_F^2\right]\right)\\
&\leq\sum_{i=1}^{N}\left(1+\left\|z_i^{(t)}\right\|^2\right)\sigma^2\\
&\leq\sum_{i=1}^{N}\left(1+3\left\|z_i^{(t)}-\bar{z}^{(t)}\right\|^2+3\left\|\bar{z}^{(t)}-\bar{z}_\star^{(t)}\right\|^2+3\left\|\bar{z}_\star^{(t)}\right\|^2\right)\sigma^2\\
&\leq 3N\left(1+\frac{L_f^2}{\mu_g^2}\right)\sigma^2+3N\sigma^2\left\|\bar{z}^{(t)}-\bar{z}_\star^{(t)}\right\|^2+3\sigma^2\Delta_t^2.
\end{aligned}
\tag{78}
$$

Substituting (76), (77), and (78) into (75), then using Assumption 3 and Eq. (22b), we have:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\left(\mathbf{x}^{(t)}-\bar{\mathbf{x}}^{(t)}\right)-(1-\theta)\left(\mathbf{x}^{(t-1)}-\bar{\mathbf{x}}^{(t-1)}\right)\right\|_F^2\right]\\
&\leq\left(\frac{208\tau^2\alpha^2\theta^2\rho^2 L^2}{(1-\rho)^2}+\frac{6\tau^2\alpha^2\theta^2\rho^2\sigma^2}{1-\rho}\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\Delta_t^2\right]+\frac{192\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2}NT\left(b_1^2+\frac{L_f^2}{\mu_g^2}b_2^2\right)\\
&\quad+\left(\frac{192N\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2}L_{\nabla g}^2+\frac{6N\tau^2\alpha^2\theta^2\rho^2}{1-\rho}\sigma^2\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{z}^{(t)}-\bar{z}_\star^{(t)}\right\|^2\right]+\frac{12\tau^2\alpha^2\theta^2\rho^2}{(1-\rho)^2}NT\iota^2\\
&\quad+\frac{6\tau^2\alpha^2\theta^2\rho^2}{1-\rho}NT\left(1+\frac{L_f^2}{\mu_g^2}\right)\sigma^2+\frac{1}{1-\rho}\mathbb{E}\left[\left\|\mathbf{x}^{(0)}-\bar{\mathbf{x}}^{(0)}\right\|_F^2\right].
\end{aligned}
\tag{79}
$$

Next, for $t=1,2,\cdots,T-1$, we have the following inequality thanks to Jensen's inequality:

$$
\left\|\mathbf{x}^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|_F^2\leq\frac{1}{\theta}\left\|(\mathbf{x}^{(t)}-\bar{\mathbf{x}}^{(t)})-(1-\theta)(\mathbf{x}^{(t-1)}-\bar{\mathbf{x}}^{(t-1)})\right\|_F^2+(1-\theta)\left\|\mathbf{x}^{(t-1)}-\bar{\mathbf{x}}^{(t-1)}\right\|_F^2.
\tag{80}
$$

Taking summation over $i=1,2,\cdots,N$ and $t=1,2,\cdots,T-1$ on both sides of (80), then using the fact that $x_i^{(0)}=x_i^{(1)}=0$ and (79), we have:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\mathbf{x}^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|_F^2\right]\\
&\leq\left(\frac{208\tau^2\alpha^2\rho^2 L^2}{(1-\rho)^2}+\frac{6\tau^2\alpha^2\rho^2\sigma^2}{1-\rho}\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\Delta_t^2\right]+\frac{192\tau^2\alpha^2\rho^2}{(1-\rho)^2}NT\left(b_1^2+\frac{L_f^2}{\mu_g^2}b_2^2\right)\\
&\quad+\left(\frac{192N\tau^2\alpha^2\rho^2}{(1-\rho)^2}L_{\nabla g}^2+\frac{6N\tau^2\alpha^2\rho^2}{1-\rho}\sigma^2\right)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{z}^{(t)}-\bar{z}_\star^{(t)}\right\|^2\right]+\frac{12\tau^2\alpha^2\rho^2}{(1-\rho)^2}NT\iota^2\\
&\quad+\frac{6\tau^2\alpha^2\rho^2}{1-\rho}NT\left(1+\frac{L_f^2}{\mu_g^2}\right)\sigma^2+\frac{1}{1-\rho}\left\|\mathbf{x}^{(0)}-\bar{\mathbf{x}}^{(0)}\right\|_F^2.
\end{aligned}
\tag{81}
$$

43

□

Then, the following lemma 9 gives the upper bound of the consensus error of y.

Lemma 9. *Suppose Assumption 1, 2, 3, and 4 hold. Then in Algorithm 1, we have:*

$$
\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right\|_F^2\right] \leq \frac{3\rho^2\beta^2}{\kappa^2(1-\rho)^2}L_{\nabla g}^2\mathbb{E}\left[\sum_{t=0}^{T-1}\Delta_t^2\right] + \frac{3N\rho^2\beta^2}{(1-\rho)^2}b_2^2
$$
$$
+\frac{3N\rho^2\beta^2}{(1-\rho)^2}L_{\nabla g}^2\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \frac{NT\rho^2\beta^2}{1-\rho}\sigma^2 + \frac{1}{1-\rho}\left\|\mathbf{y}^{(0)} - \bar{\mathbf{y}}^{(0)}\right\|_F^2.
\tag{82}
$$

Proof. We consider the conditional expectation of $\left\|\mathbf{y}^{(t+1)} - \bar{\mathbf{y}}^{(t+1)}\right\|_F^2$ with respect to $\mathcal{F}^{(t)}$:

$$
\mathbb{E}_t\left[\left\|\mathbf{y}^{(t+1)} - \bar{\mathbf{y}}^{(t+1)}\right\|_F^2\right]
$$
$$
=\left\|\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left[\left(\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right) - \beta\nabla_2\mathbf{g}^{(t+1)}\right]\right\|_F^2 + \rho^2\beta^2\mathbb{E}_t\left[\sum_{i=1}^N\left\|v_{2,i}^{(t+1)} - \nabla_2 g_i(x_i^{(t)}, y_i^{(t)})\right\|^2\right]
$$
$$
\leq\frac{1}{\rho}\left\|\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right)\right\|_F^2 + \frac{\beta^2}{1-\rho}\left\|\left(W - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\nabla_2\mathbf{g}^{(t+1)}\right\|_F^2 + N\rho^2\beta^2\sigma^2
$$
$$
\leq\rho\left\|\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right\|_F^2 + \frac{\rho^2\beta^2}{1-\rho}\sum_{i=1}^N\left\|\nabla_2 g_i(x_i^{(t)}, y_i^{(t)})\right\|^2 + N\rho^2\beta^2\sigma^2,
\tag{83}
$$

where the first inequality is from Assumption 3. For the second term on the right-hand side of (83), we have:

$$
\sum_{i=1}^N\left\|\nabla_2 g_i(x_i^{(t)}, y_i^{(t)})\right\|^2
$$
$$
\leq 3\sum_{i=1}^N\left\|\nabla_2 g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_2 g_i(\bar{x}^{(t)}, \bar{y}^{(t)})\right\|^2 + 3N\left\|\nabla_2 g(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_2 g(\bar{x}^{(t)}, \bar{y}_\star^{(t)})\right\|^2
$$
$$
+ 3\sum_{i=1}^N\left\|\nabla_2 g_i(\bar{x}^{(t)}, \bar{y}^{(t)}) - \frac{1}{N}\sum_{m=1}^N\nabla_2 g_m(\bar{x}^{(t)}, \bar{y}^{(t)})\right\|^2
\tag{84}
$$
$$
\leq 3L_{\nabla g}^2\sum_{i=1}^N\left[\left\|x_i^{(t)} - \bar{x}^{(t)}\right\|^2 + \left\|y_i^{(t)} - \bar{y}^{(t)}\right\|^2\right] + 3Nb_2^2 + 3NL_{\nabla g}^2\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2
$$
$$
\leq\frac{3L_{\nabla g}^2}{\kappa^2}\Delta_t^2 + 3Nb_2^2 + 3NL_{\nabla g}^2\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2,
$$

where the first inequality is due to the fact that $\nabla_2 g(\bar{x}^{(t)}, \bar{y}^{(t)}) = \frac{1}{N}\sum_{i=1}^N\nabla_2 g_i(\bar{x}^{(t)}, \bar{y}^{(t)})$ as well as $\nabla_2 g(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) = 0$.

Finally, taking summation and expectation on both sides of (83), we have:

$$
\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\mathbf{y}^{(t+1)}-\bar{\mathbf{y}}^{(t+1)}\right\|_F^2\right]
$$

$$
\leq \frac{\rho^2\beta^2}{(1-\rho)^2}\mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left\|\nabla_2 g_i(x_i^{(t)},y_i^{(t)})\right\|^2\right] + \frac{N\rho^2\beta^2\sigma_{g1}^2}{1-\rho} + \frac{1}{1-\rho}\left\|\mathbf{y}^{(0)}-\bar{\mathbf{y}}^{(0)}\right\|_F^2
$$

$$
\leq \frac{3\rho^2\beta^2}{\kappa^2(1-\rho)^2}L_{\nabla g}^2\mathbb{E}\left[\sum_{t=0}^{T-1}\Delta_t^2\right] + \frac{3N\rho^2\beta^2}{(1-\rho)^2}b_2^2 + \frac{3NT\rho^2\beta^2}{(1-\rho)^2}L_{\nabla g}^2\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\bar{y}^{(t)}-\bar{y}_\star^{(t)}\right\|^2\right]
$$

$$
+ \frac{NT\rho^2\beta^2}{1-\rho}\sigma^2 + \frac{1}{1-\rho}\left\|\mathbf{y}^{(0)}-\bar{\mathbf{y}}^{(0)}\right\|_F^2. \tag{85}
$$

$\square$

Moreover, the following lemma 10 gives the upper bound of the consensus error of z.

Lemma 10. *Suppose Assumption 1, 2, 3, and 4 hold. Then in Algorithm 1, we have:*

$$
\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\mathbf{z}^{(t)}-\bar{\mathbf{z}}^{(t)}\right\|_F^2\right]
$$

$$
\leq \left(\frac{64\rho^2\gamma^2}{(1-\rho)^2}L^2 + \frac{6\rho^2\gamma^2}{1-\rho}\sigma^2\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\Delta_t^2\right] + \frac{66\rho^2\gamma^2}{(1-\rho)^2}NT\left(b_1^2 + \frac{L_f^2}{\mu_g^2}b_2^2\right)
$$

$$
+ \frac{8N\rho^2\gamma^2}{(1-\rho)^2}\left(L_{\nabla f}^2 + \frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2}\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\bar{y}^{(t)}-\bar{y}_\star^{(t)}\right\|^2\right] + \frac{1}{1-\rho}\left\|\mathbf{z}^{(0)}-\bar{\mathbf{z}}^{(0)}\right\|_F^2
$$

$$
+ \left(\frac{32N\rho^2\gamma^2}{(1-\rho)^2}L^2 + \frac{6N\rho^2\gamma^2}{1-\rho}\sigma^2\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\bar{z}^{(t)}-\bar{z}_\star^{(t)}\right\|^2\right] + \frac{6NT\rho^2\gamma^2}{1-\rho}\left(1+\frac{L_f^2}{\mu_g^2}\right)\sigma^2. \tag{86}
$$

Proof. Considering the conditional expectation of $\left\|\mathbf{z}^{(t+1)}-\bar{\mathbf{z}}^{(t+1)}\right\|_F^2$ with respect to $\mathcal{F}^{(t)}$:

$$
\mathbb{E}_t\left[\left\|\mathbf{z}^{(t+1)}-\bar{\mathbf{z}}^{(t+1)}\right\|_F^2\right]
$$

$$
\leq \frac{1}{\rho}\left\|\left(W-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^{(t)}-\bar{\mathbf{z}}^{(t)}\right)\right\|_F^2
$$

$$
+ \frac{\rho^2\gamma^2}{1-\rho}\sum_{i=1}^{N}\left\|\mathbb{E}_t[p_{H,i}^{(t)}-u_{2,i}^{(t)}]-\left(\nabla_{22}^2 g_i(x_i^{(t)},y_i^{(t)})z_i^{(t)}-\nabla_2 f_i(x_i^{(t)},y_i^{(t)})\right)\right\|^2
$$

$$
+ 2\rho^2\gamma^2\sum_{i=1}^{N}\left\|p_{H,i}^{(t)}-\mathbb{E}_t[p_{H,i}^{(t)}]\right\|^2 + 2\rho^2\gamma^2\sum_{i=1}^{N}\left\|u_{2,i}^{(t)}-\mathbb{E}_t[u_{2,i}^{(t)}]\right\|^2 \tag{87}
$$

$$
\leq \rho\left\|\mathbf{z}^{(t)}-\bar{\mathbf{z}}^{(t)}\right\|_F^2 + \frac{2\rho^2\gamma^2}{1-\rho}\sum_{i=1}^{N}\left\|\mathbb{E}_t[p_{H,i}^{(t)}]-\nabla_{22}^2 g_i(x_i^{(t)},y_i^{(t)})z_i^{(t)}\right\|^2
$$

$$
+ 2\rho^2\gamma^2\sum_{i=1}^{N}\mathbb{E}_t\left[\left\|p_{H,i}^{(t)}-\mathbb{E}_t[p_{H,i}^{(t)}]\right\|^2\right] + 2\rho^2\gamma^2\sum_{i=1}^{N}\mathbb{E}_t\left[\left\|u_{2,i}^{(t)}-\mathbb{E}_t[u_{2,i}^{(t)}]\right\|^2\right],
$$

45

where the second inequality is due to $\mathbb{E}_t[u_{2,i}^{(t)}] = \nabla_2 f_i(x_i^{(t)}, y_i^{(t)})$.

Then using $\nabla_{22}^2 g(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) \bar{z}_\star^{(t)} - \nabla_2 f(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) = 0$, we have:

$$
\sum_{i=1}^{N} \left\| \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} - \nabla_2 f_i(x_i^{(t)}, y_i^{(t)}) \right\|^2
$$

$$
\leq 4 \sum_{i=1}^{N} \left\| \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} - \nabla_{22} g(\bar{x}^{(t)}, \bar{y}^{(t)}) \bar{z}_\star^{(t)} \right\|^2
$$

$$
+ 4 \sum_{i=1}^{N} \left\| \nabla_2 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_2 f(\bar{x}^{(t)}, \bar{y}^{(t)}) \right\|^2
$$

$$
+ 4N \left\| \left(\nabla_{22} g(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_{22} g(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) \right) \bar{z}_\star^{(t)} \right\|^2 + 4N \left\| \nabla_2 f(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_2 f(\bar{x}^{(t)}, \bar{y}_\star^{(t)}) \right\|^2
$$

$$
\leq 4 \sum_{i=1}^{N} \left\| \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} - \nabla_{22} g(\bar{x}^{(t)}, \bar{y}^{(t)}) \bar{z}_\star^{(t)} \right\|^2
$$

$$
+ 4 \sum_{i=1}^{N} \left\| \nabla_2 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_2 f(\bar{x}^{(t)}, \bar{y}^{(t)}) \right\|^2 + 4N \left(L_{\nabla f}^2 + \frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} \right) \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2.
$$

$$(88)$$

For the first term of the right-hand side of (88), we have:

$$
\nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} - \nabla_{22} g(\bar{x}^{(t)}, \bar{y}^{(t)}) \bar{z}_\star^{(t)}
$$
$$
= \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)})(z_i^{(t)} - \bar{z}^{(t)}) + \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)})(\bar{z}^{(t)} - \bar{z}_\star^{(t)}) \qquad (89)
$$
$$
+ (\nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) - \nabla_{22} g_i(\bar{x}^{(t)}, \bar{y}^{(t)})) \bar{z}_\star^{(t)} + (\nabla_{22} g_i(\bar{x}^{(t)}, \bar{y}^{(t)}) - \nabla_{22} g(\bar{x}^{(t)}, \bar{y}^{(t)})) \bar{z}_\star^{(t)}.
$$

Taking the norm on both sides and using Assumption 2, we have:

$$
\sum_{i=1}^{N} \left\| \nabla_{22} g_i(x_i^{(t)}, y_i^{(t)}) z_i^{(t)} - \nabla_{22} g(\bar{x}^{(t)}, \bar{y}^{(t)}) \bar{z}_\star^{(t)} \right\|^2
$$

$$
\leq 4 L_{\nabla g}^2 \sum_{i=1}^{N} \left\| z_i^{(t)} - \bar{z}^{(t)} \right\|^2 + 4N L_{\nabla g}^2 \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 \qquad (90)
$$

$$
+ \frac{4 L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} \left(\left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 + \left\| \mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|_F^2 \right) + \frac{4 L_f^2}{\mu_g^2} N b_2^2.
$$

For the second term on the right-hand side of (88), we have:

$$
\sum_{i=1}^{N} \left\| \nabla_2 f_i(x_i^{(t)}, y_i^{(t)}) - \nabla_2 f(\bar{x}^{(t)}, \bar{y}^{(t)}) \right\|^2 \leq 4 L_{\nabla f}^2 \left(\left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 + \left\| \mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)} \right\|_F^2 \right) + 4N b_1^2.
$$

$$(91)$$

Substituting (88), (90) and (91) into (87), then using (22a), we have:

$$
\mathbb{E}_t \left[\left\| \mathbf{z}^{(t+1)} - \bar{\mathbf{z}}^{(t+1)} \right\|_F^2 \right]
$$

$$
\leq \rho \left\| \mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|_F^2 + \frac{8N\rho^2\gamma^2}{1-\rho} \left(L_{\nabla f}^2 + \frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} \right) \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2
$$

$$
+ \left(\frac{32N\rho^2\gamma^2}{1-\rho} L^2 + 6N\rho^2\gamma^2\sigma^2 \right) \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 + \left(\frac{64\rho^2\gamma^2}{1-\rho} L^2 + 6\rho^2\gamma^2\sigma^2 \right) \Delta_t^2
$$

$$
+ \frac{66\rho^2\gamma^2}{1-\rho} N \left(b_1^2 + \frac{L_f^2}{\mu_g^2} b_2^2 \right) + \frac{2N\rho^2\gamma^2}{1-\rho} \iota^2 + 6N\rho^2\gamma^2 \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2.
$$

(92)

Taking the expectation and summation on both sides of (92), we can get:

$$
\mathbb{E} \left[\sum_{t=0}^{T-1} \left\| \mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)} \right\|_F^2 \right]
$$

$$
\leq \left(\frac{64\rho^2\gamma^2}{(1-\rho)^2} L^2 + \frac{6\rho^2\gamma^2}{1-\rho} \sigma^2 \right) \mathbb{E} \left[\sum_{t=0}^{T-1} \Delta_t^2 \right] + \frac{66\rho^2\gamma^2}{(1-\rho)^2} NT \left(b_1^2 + \frac{L_f^2}{\mu_g^2} b_2^2 \right) + \frac{2NT\rho^2\gamma^2}{1-\rho} \iota^2
$$

$$
+ \frac{8N\rho^2\gamma^2}{(1-\rho)^2} \left(L_{\nabla f}^2 + \frac{L_{\nabla^2 g}^2 L_f^2}{\mu_g^2} \right) \mathbb{E} \left[\sum_{t=0}^{T-1} \left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right] + \frac{1}{1-\rho} \left\| \mathbf{z}^{(0)} - \bar{\mathbf{z}}^{(0)} \right\|_F^2
$$

$$
+ \left(\frac{32N\rho^2\gamma^2}{(1-\rho)^2} L^2 + \frac{6N\rho^2\gamma^2}{1-\rho} \sigma^2 \right) \mathbb{E} \left[\sum_{t=0}^{T-1} \left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 \right] + \frac{6NT\rho^2\gamma^2}{1-\rho} \left(1 + \frac{L_f^2}{\mu_g^2} \right) \sigma^2.
$$

(93)

$\square$

Through Lemma 8, Lemma 9, and Lemma 10, we can present an upper bound of the consensus error during the iterations in Algorithm 1, which is concluded by the following lemma 11:

Lemma 11. *Suppose Assumption 1, 2, 3, and 4 hold, and the step-sizes α, β, γ satisfy:*

$$
208(\kappa^2\tau^2\alpha^2 + \kappa^2\beta^2 + \gamma^2)\rho^2 \left(\frac{L^2}{(1-\rho)^2} + \frac{\sigma^2}{1-\rho} \right) \leq \frac{1}{3}.
$$

(94)

Then in Algorithm 1, we have:

$$
\sum_{t=0}^{T-1} \alpha \mathbb{E} \left[\frac{1}{N} \Delta_t^2 \right] \leq \frac{C_{Dy}\rho^2\alpha^2}{(1-\rho)^2} \sum_{t=0}^{T-1} \alpha \mathbb{E} \left[\left\| \bar{y}^{(t)} - \bar{y}_\star^{(t)} \right\|^2 \right] + \frac{C_{Dz}\rho^2\alpha^2}{(1-\rho)^2} \sum_{t=0}^{T-1} \alpha \mathbb{E} \left[\left\| \bar{z}^{(t)} - \bar{z}_\star^{(t)} \right\|^2 \right]
$$

$$
+ \frac{C_{D1}T\rho^2\alpha^3(b_1^2 + \kappa^2 b_2^2)}{(1-\rho)^2} + \frac{C_{D2}T\rho^2\alpha^3\sigma^2}{1-\rho} + \frac{C_{D3}T\rho^2\alpha^3\iota^2}{(1-\rho)^2} + \frac{\alpha}{1-\rho} \cdot \frac{1}{N} \Delta_0^2.
$$

(95)

where the constants are defined as:

$$C_{Dy} = C_{Dz} = 384(\tau^2\kappa^2 + \kappa^2 c_1^2 + c_2^2)(L^2 + \sigma^2) = \mathcal{O}(\tau^2\kappa^2 + \kappa^2 c_1^2 + c_2^2),$$
$$C_{D1} = 384(\tau^2\kappa^2 + \kappa^2 c_1^2 + c_2^2) = \mathcal{O}(\tau^2\kappa^2 + \kappa^2 c_1^2 + c_2^2),$$
$$C_{D2} = 384(\tau^2\kappa^2 + \kappa^2 c_1^2 + c_2^2)\left(1 + \frac{L_f^2}{\mu_g^2}\right) = \mathcal{O}(\kappa^2(\tau^2\kappa^2 + \kappa^2 c_1^2 + c_2^2)),$$
$$C_{D3} = 24(\tau^2\kappa^2 + c_2^2) = \mathcal{O}(\tau^2\kappa^2 + c_2^2).$$

Proof. From (69), (82), and (86), we have:

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\Delta_t^2\right]$$

$$\leq \kappa^2\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_F^2\right] + \kappa^2\mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\mathbf{y}^{(t)} - \bar{\mathbf{y}}^{(t)}\right\|_F^2\right] + \mathbb{E}\left[\sum_{t=0}^{T-1}\left\|\mathbf{z}^{(t)} - \bar{\mathbf{z}}^{(t)}\right\|_F^2\right]$$

$$\leq 208(\kappa^2\tau^2\alpha^2 + \kappa^2\beta^2 + \gamma^2)\rho^2\left(\frac{L^2}{(1-\rho)^2} + \frac{\sigma^2}{1-\rho}\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\Delta_t^2\right]$$

$$+ 192(\kappa^2\tau^2\alpha^2 + \kappa^2\beta^2 + \gamma^2)\rho^2\left(\frac{L^2}{(1-\rho)^2} + \frac{\sigma^2}{1-\rho}\right)N\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2 + \left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right]$$

$$+ \frac{192(\kappa^2\tau^2\alpha^2 + \kappa^2\beta^2 + \gamma^2)\rho^2}{(1-\rho)^2}NT\left(b_1^2 + \frac{L_f^2}{\mu_g^2}b_2^2\right) + \frac{12(\kappa^2\tau^2\alpha^2 + \gamma^2)\rho^2}{(1-\rho)^2}NT\iota^2$$

$$+ \frac{192(\kappa^2\tau^2\alpha^2 + \kappa^2\beta^2 + \gamma^2)\rho^2}{1-\rho}\left(1 + \frac{L_f^2}{\mu_g^2}\right)NT\sigma^2 + \frac{1}{1-\rho}\Delta_0^2. \tag{96}$$

Then, multiplying $\dfrac{\alpha}{N}$ on both sides of (96) and substituting (94) into it, we finally achieve:

$$\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right] \leq \frac{C_{Dy}\rho^2\alpha^2}{(1-\rho)^2}\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \frac{C_{Dz}\rho^2\alpha^2}{(1-\rho)^2}\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right]$$

$$+ \frac{C_{D1}T\rho^2\alpha^3(b_1^2 + \kappa^2 b_2^2)}{(1-\rho)^2} + \frac{C_{D2}T\rho^2\alpha^3\sigma^2}{1-\rho} + \frac{C_{D3}T\rho^2\alpha^3\iota^2}{(1-\rho)^2} + \frac{\alpha}{1-\rho}\cdot\frac{1}{N}\Delta_0^2, \tag{97}$$

where $C_{Dy}, C_{Dz}, C_{D1}, C_{D2}, C_{D3}$ have already been defined above. $\square$

Finally, we combine the conclusion of Lemma 7 and Lemma 11 together, obtaining the following lemma:

Lemma 12. *Suppose Assumption 1, 2, 3, and 4 hold. If α, β, γ satisfy (44), (51), (94), and*

$$\frac{\rho^2\alpha^2}{(1-\rho)^2}\tilde{C}_D((1+\kappa^2)C_{yD} + C_{zD}+1) \leq \frac{1}{3}, \tag{98}$$

where $\tilde{C}_D = \max\{C_{Dy}, C_{Dz}\}$, in Algorithm 1 we have:

$$
(1 + \kappa^2) \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right]
$$

$$
\leq 2((1 + \kappa^2)C_{yx} + C_{zx}) \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + 2((1 + \kappa^2)C_{y1} + C_{z1})\frac{T\alpha^2}{N}
$$

$$
+ 2((1 + \kappa^2)C_{y2} + C_{z2})\alpha + 2C_{z3}T\alpha^2\iota^2 + \frac{2((1 + \kappa^2)C_{yD} + C_{zD})C_{D1}T\rho^2\alpha^3(b_1^2 + \kappa^2 b_2^2)}{(1 - \rho)^2}
$$

$$
+ \frac{2((1 + \kappa^2)C_{yD} + C_{zD})C_{D2}T\rho^2\alpha^3\sigma^2}{1 - \rho} + \frac{2((1 + \kappa^2)C_{yD} + C_{zD})C_{D3}T\rho^2\alpha^3\iota^2}{(1 - \rho)^2}
$$

$$
+ \frac{2((1 + \kappa^2)C_{yD} + C_{zD})\alpha}{1 - \rho} \cdot \frac{1}{N}\Delta_0^2.
$$

$$(99)$$

Proof. Combining Eq. (63) and (95) together, we have:

$$
(1 + \kappa^2) \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right]
$$

$$
\leq ((1 + \kappa^2)C_{yx} + C_{zx}) \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + ((1 + \kappa^2)C_{y1} + C_{z1})\frac{T\alpha^2}{N}
$$

$$
+ ((1 + \kappa^2)C_{y2} + C_{z2})\alpha + C_{z3}T\alpha^2\iota^2
$$

$$
+ \frac{((1 + \kappa^2)C_{yD} + C_{zD})\tilde{C}_D\rho^2\alpha^2}{(1 - \rho)^2} \left(\sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \sum_{t=0}^{T-1} \alpha \mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right] \right)
$$

$$
+ \frac{((1 + \kappa^2)C_{yD} + C_{zD})C_{D1}T\rho^2\alpha^3(b_1^2 + \kappa^2 b_2^2)}{(1 - \rho)^2} + \frac{((1 + \kappa^2)C_{yD} + C_{zD})C_{D2}T\rho^2\alpha^3\sigma^2}{1 - \rho}
$$

$$
+ \frac{((1 + \kappa^2)C_{yD} + C_{zD})C_{D3}T\rho^2\alpha^3\iota^2}{(1 - \rho)^2} + \frac{((1 + \kappa^2)C_{yD} + C_{zD})\alpha}{1 - \rho} \cdot \frac{1}{N}\Delta_0^2.
$$

$$(100)$$

Then, using (98), we finish the proof of this lemma.

$\square$

A.7 The step-sizes selection and the conclusion

In the end, the following lemma presents the proof of Eq. (13)

Lemma 13. *Suppose Assumption 1, 2, 3, and 4 hold. There exist $c_0, c_1, c_2, c_3 > 0$ such that if $\alpha = c_0\sqrt{N/T}$, and $\alpha_t \equiv \alpha$, $\beta_t \equiv c_1\alpha$, $\gamma_t \equiv c_2\alpha$, $\theta_t \equiv c_3\alpha$ for any $t = 0, 1, \cdots, T - 1$, the*

iterates $\bar{x}^{(t)}$ in Algorithm 1 satisfy:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^{(t)})\right\|^2\right]$$

$$\lesssim \frac{\kappa^5}{\sqrt{NT}} + \frac{\rho^{\frac{2}{3}}\kappa^6}{(1-\rho)^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\rho^{\frac{2}{3}}(b_1^{\frac{2}{3}}\kappa^{\frac{16}{3}}+b_2^{\frac{2}{3}}\kappa^6)}{(1-\rho)^{\frac{2}{3}}T^{\frac{2}{3}}} + \frac{\rho\kappa^6}{(1-\rho)T} + \frac{\kappa^4}{(1-\rho)NT} + \frac{\kappa^{13}}{T} + \frac{\kappa^7}{NT},$$

$$(101)$$

Proof. We initially assume that the conclusions of all preceding lemmas are valid, and provide explicit expressions of α, β, γ at the conclusion of the proof.

Assume that the step-size α satisfies:

$$\frac{(1+c_3)\sigma^2\alpha}{N} \leq L^2, \tag{102}$$

From (29), we have:

$$\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \alpha\sum_{t=0}^{T}\mathbb{E}\left[\left\|\bar{h}^{(t)}-\nabla\Phi(\bar{x}^{(t)})\right\|^2\right]$$

$$\leq\frac{2}{\tau}\mathbb{E}[E_0] + \frac{1}{c_3}\mathbb{E}\left[\left\|\bar{h}^{(0)}-\nabla\Phi(\bar{x}^{(0)})\right\|^2\right] + \frac{1}{2}\cdot\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}-\nabla\Phi(\bar{x}^{(t)})\right\|^2\right]$$

$$+ \frac{4L_{\nabla\Phi}^2\tau^2}{c_3^2}\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + 6L^2\frac{1}{N}\cdot\alpha\sum_{t=0}^{T-1}\mathbb{E}[\Delta_t^2] + \frac{6(1+c_3)T\alpha^2}{N}\left(1+\frac{L_f^2}{\mu_g^2}\right)\sigma^2$$

$$+ 10T\alpha\iota^2 + 66L^2(1+\kappa^2)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{y}^{(t)}-\bar{y}_\star^{(t)}\right\|^2\right] + 66L^2\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{z}^{(t)}-\bar{z}_\star^{(t)}\right\|^2\right].$$

$$(103)$$

Then, plugging (95) and (99) into (103), we get:

$$\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \alpha\sum_{t=0}^{T}\mathbb{E}\left[\left\|\bar{h}^{(t)}-\nabla\Phi(\bar{x}^{(t)})\right\|^2\right]$$

$$\leq\frac{2}{\tau}\mathbb{E}[E_0] + \frac{1}{c_3}\mathbb{E}\left[\left\|\bar{h}^{(0)}-\nabla\Phi(\bar{x}^{(0)})\right\|^2\right] + \frac{4L_{\nabla\Phi}^2\tau^2}{c_3^2}\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right]$$

$$+ \frac{1}{2}\cdot\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}-\nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + \left(\frac{6(1+c_3)T\alpha^2}{N}\left(1+\frac{L_f^2}{\mu_g^2}\right) + \frac{66L^2C_{D2}T\rho^2\alpha^3}{1-\rho}\right)\sigma^2$$

$$+ \left(10T\alpha + \frac{66L^2C_{D3}T\rho^2\alpha^3}{(1-\rho)^2}\right)\iota^2 + \frac{66L^2C_{D1}T\rho^2\alpha^3}{(1-\rho)^2}(b_1^2+\kappa^2b_2^2) + \frac{66L^2\alpha}{1-\rho}\cdot\frac{1}{N}\Delta_0^2$$

$$+ 132L^2\left((1+\kappa^2)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{y}^{(t)}-\bar{y}_\star^{(t)}\right\|^2\right] + \sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{z}^{(t)}-\bar{z}_\star^{(t)}\right\|^2\right]\right).$$

$$(104)$$

50

Then, we have:

$$
\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \alpha \sum_{t=0}^{T} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right]
$$

$$
\leq \frac{2}{\tau}\mathbb{E}[E_0] + \frac{1}{c_3}\mathbb{E}\left[\left\|\bar{h}^{(0)} - \nabla\Phi(\bar{x}^{(0)})\right\|^2\right] + \frac{1}{2}\cdot\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right]
$$

$$
+ \left(\frac{4L_{\nabla\Phi}^2}{c_3^2} + 264L^2((1+\kappa^2)C_{yx} + C_{zx})\right)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + 264L^2((1+\kappa^2)C_{y2} + C_{z2})\tau^2\alpha
$$

$$
+ \left(6(1+c_3)(1+\kappa^2)\sigma^2 + 264L^2((1+\kappa^2)C_{y1} + C_{z1})\right)\frac{T\alpha^2}{N} + \left(10 + 264L^2 C_{z3}\right)T\alpha\iota^2
$$

$$
+ \hat{C}_D\left(\frac{C_{D1}T\rho^2\alpha^3(b_1^2 + \kappa^2 b_2^2)}{(1-\rho)^2} + \frac{C_{D2}T\rho^2\alpha^3\sigma^2}{1-\rho} + \frac{C_{D3}T\rho^2\alpha^3\iota^2}{(1-\rho)^2} + \frac{\alpha\Delta_0^2}{(1-\rho)N}\right),
$$

$$(105)$$

where $\hat{C}_D := 66L^2 + 264L^2((1+\kappa^2)C_{yD} + C_{zD})$, and the last inequality is due to $\frac{\tilde{C}_D\rho^2\alpha^2}{(1-\rho)^2} \leq \frac{1}{3}$.
Then, if we take:

$$
C_x = \frac{4L_{\nabla\Phi}^2\tau^2}{c_3^2} + 264L^2\tau^2((1+\kappa^2)C_{yx} + C_{zx}) = \mathcal{O}\left(\frac{\kappa^8\tau^2}{c_1^2} + \frac{\kappa^8\tau^2}{c_2^2} + \frac{\kappa^6\tau^2}{c_3^2}\right),
$$

$$
C_0 = 264L^2((1+\kappa^2)C_{y2} + C_{z2}) + \frac{2}{\tau}\mathbb{E}[E_0] + \frac{1}{c_3}\mathbb{E}\left[\left\|\bar{h}^{(0)} - \nabla\Phi(\bar{x}^{(0)})\right\|^2\right]
$$

$$
= \mathcal{O}\left(\frac{\kappa^5}{c_1} + \frac{\kappa}{c_2} + \frac{1}{c_3} + \frac{1}{\tau}\right),
$$

$$
C_1 = 6(1+c_3)(1+\kappa^2)\sigma^2 + 264L^2((1+\kappa^2)C_{y1} + C_{z1}) = \mathcal{O}(c_1\kappa^5 + c_2\kappa^3 + c_3\kappa^2 + 1),
$$

$$
C_2 = 10 + 264L^2 C_{z3} = \mathcal{O}(\kappa^2),
$$

$$
\hat{C}_D = 66L^2 + 264L^2((1+\kappa^2)C_{yD} + C_{zD}) = \mathcal{O}(\kappa^4),
$$

$$(106)$$

we can get:

$$
\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \alpha \sum_{t=0}^{T} \mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right]
$$

$$
\leq C_x \sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \frac{1}{2}\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right] + C_1\frac{T\alpha^2}{N} + C_2 T\alpha^2\iota^2 + C_0\alpha \quad (107)
$$

$$
+ \hat{C}_D\left[\frac{C_{D1}T\rho^2\alpha^3(b_1^2 + \kappa^2 b_2^2)}{(1-\rho)^2} + \frac{C_{D2}T\rho^2\alpha^3\sigma^2}{1-\rho} + \frac{C_{D3}T\rho^2\alpha^3\iota^2}{(1-\rho)^2} + \frac{\alpha\Delta_0^2}{(1-\rho)N}\right].
$$

Define

$$\alpha_1 = \left(\frac{C_0 N}{C_1 T}\right)^{\frac{1}{2}}, \quad \alpha_{21} = \left(\frac{C_0(1-\rho)^2}{\hat{C}_D C_{D1} \rho^2 T b_1^2}\right)^{\frac{1}{3}}, \quad \alpha_{22} = \left(\frac{C_0(1-\rho)^2}{\hat{C}_D C_{D1} \rho^2 T \kappa^2 b_2^2}\right)^{\frac{1}{3}},$$

$$\alpha_3 = \left(\frac{C_0(1-\rho)}{\hat{C}_D C_{D2} \rho^2 T}\right)^{\frac{1}{3}}, \alpha_4 = \left(\frac{(1-\rho)^2}{208\rho^2(\kappa^2\tau^2 + \kappa^2 c_1^2 + c_2^2)\left(L^2 + \sigma^2\right)}\right)^{\frac{1}{2}},$$

$$\alpha_5 = \left(\frac{(1-\rho)^2}{3\rho^2\tilde{C}_D((1+\kappa^2)C_{yD} + C_{zD}+1)}\right)^{\frac{1}{2}},$$

$$C_{\alpha 1} = c_1(\mu_g + L_{\nabla g}) + 2c_2\mu_g + \frac{2\tau(c_3 L_{\nabla\Phi} + L_{\nabla\eta})}{c_3} + \frac{20c_3}{\tau^2} + 1,$$

$$C_{\alpha 2} = \frac{144c_2\sigma^2}{N\mu_g} + \frac{c_2\mu_g^2\sigma^2}{NL^2} + \frac{(1+c_3)\sigma_{g2}^2}{NL_{\nabla^2 g}^2}.$$

Then take

$$\alpha = \left(\frac{1}{\alpha_1} + \frac{1}{\alpha_{21}} + \frac{1}{\alpha_{22}} + \frac{1}{\alpha_3} + \frac{1}{\alpha_4} + \frac{1}{\alpha_5} + C_{\alpha 1} + C_{\alpha 2}\right)^{-1}. \tag{108}$$

Review that the step-sizes are restricted by:

- Eq. (28): $\quad \alpha \le \min\left\{\frac{c_3}{2\tau(c_3 L_{\nabla\Phi} + L_{\nabla\eta})}, \frac{\tau^2}{20c_3}, 1\right\}, \quad \gamma\mu_g \le 1;$

- Eq. (44): $\quad \beta(\mu_g + L_{\nabla g}) \le 1;$

- Eq. (51): $\quad -\frac{1}{6}\gamma\mu_g + \frac{6}{N}\gamma^2\sigma^2 + \frac{1}{2}\gamma^3\mu_g^3 \le 0, \quad \gamma\mu_g \le \min\left\{1, \frac{NL^2}{\mu_g\sigma^2}\right\};$

- Eq. (94): $\quad 624(\kappa^2\tau^2\alpha^2 + \kappa^2\beta^2 + \gamma^2)\rho^2\left(\frac{L^2}{(1-\rho)^2} + \frac{\sigma^2}{1-\rho}\right) \le \frac{1}{3};$

- Eq. (98): $\quad \frac{\rho^2\alpha^2}{(1-\rho)^2}\tilde{C}_D((1+\kappa^2)C_{yD} + C_{zD}+1) \le \frac{1}{3};$

- Eq. (102): $\quad \frac{(1+c_3)\sigma_{g2}^2\alpha}{N} \le L^2.$

With (108), all above restrictions hold by the following validation:

- Validation of Eq. (28): $\quad$ As $\max\left\{\frac{2\alpha\tau(c_3 L_{\nabla\Phi} + L_{\nabla\eta})}{c_3}, \frac{20c_3\alpha}{\tau^2}, 2\gamma\mu_g, \alpha\right\} \le \alpha C_{\alpha 1} \le 1,$
  thus $\alpha \le \min\left\{\frac{c_3}{2\tau(c_3 L_{\nabla\Phi} + L_{\nabla\eta})}, \frac{\tau^2}{20c_3}, 1\right\}$ and $\gamma\mu_g \le \frac{1}{2} < 1.$

- Validation of Eq. (44): $\quad \beta(\mu_g + L_{\nabla g}) \le \alpha C_{\alpha 1} \le 1.$

- Validation of Eq. (51): As $\frac{\gamma\mu_g^2\sigma^2}{NL^2} \leq \alpha C_{\alpha 2} \leq 1$, $\gamma\mu_g \leq \min\left\{1, \frac{NL^2}{\mu_g\sigma^2}\right\}$.

  Moreover, it holds that $\gamma\mu_g \leq \frac{1}{2}$. Then, $\frac{1}{6}\gamma\mu_g - \frac{1}{2}\gamma^3\mu_g^3 \geq \frac{1}{6}\gamma\mu_g - \frac{1}{8}\gamma\mu_g = \frac{1}{24}\gamma\mu_g$. And

  from $\frac{144\gamma\sigma^2}{N\mu_g} \leq \alpha C_{\alpha 2} \leq 1$, we can obtain that $\frac{6}{N}\gamma^2\sigma^2 \leq \frac{1}{24}\gamma\mu_g$. Thus Eq. (51) holds

  as $\frac{6}{N}\gamma^2\sigma^2 \leq \frac{1}{6}\gamma\mu_g - \frac{1}{2}\gamma^3\mu_g^3$.

- Validation of Eq. (94): As $\alpha/\alpha_4 \leq 1$, it holds that $\frac{624\rho^2(\kappa^2\tau^2 + \kappa^2 c_1^2 + c_2^2)(L^2 + \sigma^2)}{(1-\rho)^2} \leq$
  1. Thus Eq. (94) holds.

- Validation of Eq. (98): As $\alpha/\alpha_5 \leq 1$, it holds that $\frac{3\rho^2\alpha^2\tilde{C}_D((1+\kappa^2)C_{yD} + C_{zD})}{(1-\rho)^2} \leq$
  1. Thus Eq. (98) holds.

- Validation of Eq. (102): As $\alpha C_{\alpha 2} \leq 1$, it holds that $\frac{(1+c_3)\sigma_{g2}^2\alpha}{NL^2} \leq 1$. Then Eq.
  (102) holds.

Thus, all these restrictions hold. Consequently, all preceding lemmas hold.

Taking $c_1 = \Theta(1)$, $c_2 = \Theta(1)$, $c_3 = \Theta(1)$, $\tau = \Theta(\kappa^{-4})$ such that $C_x < \frac{1}{2}$, we get:

$$C_x = \mathcal{O}(1), \quad C_0 = \mathcal{O}\left(\kappa^5\right), \quad C_1 = \mathcal{O}(\kappa^5), \quad C_2 = \mathcal{O}(\kappa^2),$$
$$C_{D1} = \mathcal{O}(\kappa^2), \quad C_{D2} = \mathcal{O}(\kappa^2), \quad \hat{C}_D = \mathcal{O}\left(\kappa^4\right). \tag{109}$$

Then the order of κ in the following terms can be provided as:

$$\alpha_1 = \mathcal{O}(\kappa^0), \quad \alpha_{21} = \mathcal{O}(\kappa^{-\frac{1}{3}}), \quad \alpha_{22} = \mathcal{O}(\kappa^{-1}), \quad \alpha_3 = \mathcal{O}(\kappa^{-1}), \quad \alpha_4 = \mathcal{O}(\kappa^{-1}),$$
$$\alpha_5 = \mathcal{O}(\kappa^{-8}), \quad C_{\alpha 1} = \mathcal{O}(\kappa^8), \quad C_{\alpha 2} = \mathcal{O}(\kappa^2/N). \tag{110}$$

From (107), we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^{(t)})\right\|^2\right]$$
$$\leq \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + \frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\bar{h}^{(t)} - \nabla\Phi(\bar{x}^{(t)})\right\|^2\right]$$
$$\leq 2C_1\frac{\alpha}{N} + 2C_2\alpha\iota^2 + \frac{2C_0}{\alpha T} + 2\hat{C}_D C_{D1}\frac{\rho^2\alpha^2(b_1^2 + \kappa^2 b_2^2)}{(1-\rho)^2} + 2\hat{C}_D C_{D2}\frac{\rho^2\alpha^2}{1-\rho} + 2\hat{C}_D C_{D3}\frac{\rho^2\alpha^2\iota^2}{(1-\rho)^2}$$
$$+ \frac{2\hat{C}_D\Delta_0^2}{(1-\rho)NT}. \tag{111}$$

Substituting (108) into (111) and taking ι satisfying

$$C_2\alpha\iota^2 + \hat{C}_D C_{D3}\frac{\rho^2\alpha^2\iota^2}{(1-\rho)^2} \lesssim \frac{1}{NT},$$

we obtain:

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^{(t)})\right\|^{2}\right]
$$

$$
\leq\frac{2C_{0}}{T}\left(\frac{1}{\alpha_{1}}+\frac{1}{\alpha_{21}}+\frac{1}{\alpha_{22}}+\frac{1}{\alpha_{3}}+\frac{1}{\alpha_{4}}+\frac{1}{\alpha_{5}}+C_{\alpha1}+C_{\alpha2}\right)+2C_{1}\frac{\alpha_{1}}{N}
$$

$$
+\frac{2\hat{C}_{D}C_{D1}\rho^{2}\alpha_{2}^{2}}{(1-\rho)^{2}}(b_{1}^{2}+\kappa^{2}b_{2}^{2})+2\hat{C}_{D}C_{D2}\frac{\rho^{2}\alpha_{3}^{2}}{1-\rho}+2C_{2}\alpha\iota^{2}+2\hat{C}_{D}C_{D3}\frac{\rho^{2}\alpha^{2}\iota^{2}}{(1-\rho)^{2}}+\frac{2\hat{C}_{D}\Delta_{0}^{2}}{(1-\rho)NT}
$$

$$
\lesssim\frac{\kappa^{5}}{\sqrt{NT}}+\frac{\rho^{\frac{2}{3}}\kappa^{6}}{(1-\rho)^{\frac{1}{3}}T^{\frac{2}{3}}}+\frac{\rho^{\frac{2}{3}}(b_{1}^{\frac{2}{3}}\kappa^{\frac{16}{3}}+b_{2}^{\frac{2}{3}}\kappa^{6})}{(1-\rho)^{\frac{2}{3}}T^{\frac{2}{3}}}+\frac{\rho\kappa^{6}}{(1-\rho)T}+\frac{\kappa^{4}}{(1-\rho)NT}+\frac{\kappa^{13}}{T}+\frac{\kappa^{7}}{NT}.
$$

$$(112)$$

$\square$

A.8 Proof of Corollary 1

Proof. According to (13), the linear speed up stage can be achieved if T is sufficiently large to make $1/\sqrt{NT}$ dominate the convergence rate, *i.e.*,:

$$
\frac{1}{\sqrt{NT}}\lesssim\frac{1}{(1-\rho)^{\frac{1}{3}}T^{\frac{2}{3}}},\quad\frac{1}{\sqrt{NT}}\lesssim\frac{b^{\frac{2}{3}}}{(1-\rho)^{\frac{2}{3}}T^{\frac{2}{3}}},\quad\frac{1}{\sqrt{NT}}\lesssim\frac{1}{(1-\rho)T}. \tag{113}
$$

The above constraints are equivalent to $T\lesssim\max\left\{\dfrac{N^{3}}{(1-\rho)^{2}},\dfrac{N^{3}b^{2}}{(1-\rho)^{4}}\right\}$. $\square$

A.9 Proof of Corollary 3

Proof. From (111), if the stochastic noises are all equal to 0, $C_{1}=C_{D2}=0$ but $C_{0},\hat{C}_{D},C_{D1}\neq 0$, the average of square norm of the hyper-gradients, *i.e.*, $\frac{1}{T}\sum_{t=0}^{T}\left\|\nabla\Phi(\bar{x}^{(t)})\right\|^{2}$, can be bounded as:

$$
\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\Phi(\bar{x}^{(t)})\right\|^{2} \tag{114}
$$

$$
\leq\frac{2C_{0}}{\alpha T}+2\hat{C}_{D}C_{D1}\frac{\rho^{2}\alpha^{2}(b_{1}^{2}+\kappa^{2}b_{2}^{2})}{(1-\rho)^{2}}+2C_{2}\alpha\iota^{2}+2\hat{C}_{D}C_{D3}\frac{\rho^{2}\alpha^{2}\iota^{2}}{(1-\rho)^{2}}+\frac{2\hat{C}_{D}\Delta_{0}^{2}}{(1-\rho)NT}. \tag{115}
$$

Define

$$
\alpha_{21}'=\left(\frac{C_{0}(1-\rho)^{2}}{\hat{C}_{D}C_{D1}\rho^{2}b_{1}^{2}T}\right)^{\frac{1}{3}}=\mathcal{O}(\kappa^{\frac{1}{3}}),\quad \alpha_{22}'=\left(\frac{C_{0}(1-\rho)^{2}}{\hat{C}_{D}C_{D1}\rho^{2}b_{2}^{2}\kappa^{2}T}\right)^{\frac{1}{3}}=\mathcal{O}(\kappa^{-1}),
$$

and set

$$
\alpha=\left(\frac{1}{\alpha_{21}'}+\frac{1}{\alpha_{22}'}+\frac{1}{\alpha_{4}}+\frac{1}{\alpha_{5}}+C_{\alpha1}+C_{\alpha2}\right)^{-1}. \tag{116}
$$

54

Then (28), (44), (51), (94), (98), and (102) hold. Taking ι satisfying

$$2C_2\alpha\iota^2 + 2\hat{C}_D C_{D3}\frac{\rho^2\alpha^2\iota^2}{(1-\rho)^2} \lesssim \frac{1}{T},$$

substituting (116) into (114) and using (109), we have:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\Phi(\bar{x}^{(t)})\right\|^2 \\
&\leq\frac{2C_0}{T}\left(\frac{1}{\alpha'_{21}} + \frac{1}{\alpha'_{22}} + \frac{1}{\alpha_4} + \frac{1}{\alpha_5} + C_{\alpha1} + C_{\alpha2}\right) + 2\hat{C}_D C_{D1}\frac{\rho^2(\alpha'_{21})^2}{(1-\rho)^2}b^2 + 2\hat{C}_D C_{D1}\frac{\rho^2(\alpha'_{22})^2}{(1-\rho)^2}\kappa^2 b_2^2 \\
&\quad + 2C_2\alpha\iota^2 + 2\hat{C}_D C_{D3}\frac{\rho^2\alpha^2\iota^2}{(1-\rho)^2} + \frac{2\hat{C}_D\Delta_0^2}{(1-\rho)NT} \\
&\lesssim\frac{\rho^{\frac{2}{3}}(b_1^{\frac{2}{3}}\kappa^{\frac{16}{3}} + b_2^{\frac{2}{3}}\kappa^6)}{T^{\frac{2}{3}}(1-\rho)^{\frac{2}{3}}} + \frac{\rho\kappa^6}{(1-\rho)T} + \frac{\kappa^4}{(1-\rho)NT} + \frac{\kappa^{13}}{T}.
\end{aligned}
$$

$$(117)$$

$\square$

Appendix B. Proof of the non-asymptotic concensus error

In this section, we present the proof of the non-asymptotic concensus error, which has been present in Corollary 2 .

Lemma 14. *Suppose Assumptions 1, 2, 3, and 4 hold. If we take the hyperparameters as Theorem 2, then the concensus error of D-SOBA satisfies that:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right] \lesssim \frac{N}{T}\left(\frac{b^2}{(1-\rho)^2} + \frac{1}{1-\rho}\right) + \frac{N^{\frac{1}{2}}}{T^{\frac{3}{2}}(1-\rho)^2}. \tag{118}$$

Proof. From Eq. (63), it holds that:

$$
\begin{aligned}
&(1+\kappa^2)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right] \\
&\leq((1+\kappa^2)C_{yx} + C_{zx})\sum_{t=0}^{T-1}\tau^2\alpha\mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] + ((1+\kappa^2)C_{y1} + C_{z1})\frac{T\alpha^2}{N} + ((1+\kappa^2)C_{y2} + C_{z2}) \\
&\quad + C_{z3}T\alpha^2\iota^2 + ((1+\kappa^2)C_{yD} + C_{zD})\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right].
\end{aligned}
$$

$$(119)$$

Moreover, from Eq. (29), the definition of C_x in Eq. (106), $C_x \leq \frac{1}{2}$, and the step-sizes taken in Lemma 13, it holds that:

$$
\begin{aligned}
&\alpha \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\bar{h}^{(t)}\right\|^2\right] \\
&\leq \frac{4}{\tau}\mathbb{E}[E_0] + \frac{2}{c_3}\mathbb{E}\left[\left\|\bar{h}^{(0)} - \nabla\Phi(\bar{x}^{(0)})\right\|^2\right] + 132L^2\frac{1}{N}\cdot\alpha\sum_{t=0}^{T-1}\mathbb{E}[\Delta_t^2] \\
&\quad + \frac{12(1+c_3)T\alpha^2}{N}\left(1 + \frac{L_f^2}{\mu_g^2}\right)\sigma^2 + 20T\alpha\iota^2 \\
&\quad + 132L^2\left((1+\kappa^2)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right]\right).
\end{aligned}
\tag{120}
$$

Then, substituting Eq. (120) into Eq. (119), using the step-sizes taken in Lemma 13 and the fact that

$$
132L^2\tau^2((1+\kappa^2)C_{yx} + C_{zx}) \leq \frac{1}{2},
$$

it holds that:

$$
\begin{aligned}
&(1+\kappa^2)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{y}^{(t)} - \bar{y}_\star^{(t)}\right\|^2\right] + \sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\left\|\bar{z}^{(t)} - \bar{z}_\star^{(t)}\right\|^2\right] \\
&\leq \frac{8C_{con}}{\tau}\mathbb{E}[E_0] + \frac{4C_{con}}{c_3}\mathbb{E}\left[\left\|\nabla\bar{h}^{(0)} - \Phi(\bar{x}^{(0)})\right\|^2\right] + \frac{24C_{con}(1+c_3)T\alpha^2}{N}\left(1 + \frac{L_f^2}{\mu_g^2}\right)\sigma^2 \\
&\quad + 40C_{con}T\alpha\iota^2 + 2((1+\kappa^2)C_{y1} + C_{z1})\frac{T\alpha^2}{N} + 2((1+\kappa^2)C_{y2} + C_{z2}) + 2C_{z3}T\alpha^2\iota^2 \\
&\quad + 2((1+\kappa^2)C_{yD} + C_{zD} + 1)\sum_{t=0}^{T-1}\alpha\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right],
\end{aligned}
\tag{121}
$$

where $C_{con} := ((1+\kappa^2)C_{yx} + C_{zx})\tau^2$.

Then, with Eq. (98), it holds from (121) and (95) that

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right] \lesssim \frac{\rho^2\alpha}{T(1-\rho)^2} + \frac{\rho^2\alpha^3}{N(1-\rho)^2} + \frac{\rho^2\alpha^2 b^2}{(1-\rho)^2} + \frac{\rho^2\alpha^2}{1-\rho} + \frac{\rho^2\alpha^2\iota^2}{(1-\rho)^2} + \frac{1}{(1-\rho)NT}.
\tag{122}
$$

As we have taken ι satisfying $C_2\alpha\iota^2 + \hat{C}_D C_{D3}\frac{\rho^2\alpha^2\iota^2}{(1-\rho)^2} \lesssim \frac{1}{NT}$ in Lemma 13. Thus with the step-sizes taken in Lemma 13, the non-asymptotic concensus error of D-SOBA has an upper bound of:

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{N}\Delta_t^2\right] \lesssim \frac{N}{T}\left(\frac{b^2}{(1-\rho)^2} + \frac{1}{1-\rho}\right) + \frac{N^{\frac{1}{2}}}{T^{\frac{3}{2}}(1-\rho)^2} + \frac{1}{(1-\rho)NT}.
$$

$\square$

Appendix C. Asymptotic Rate and Transient Complexity Analysis of MDBO

In this section, we give the theoretical analysis of the asymptotic convergence rate when $T \to \infty$ and the transient time of MDBO (Gao et al., 2023, Algorithm 1) shown in Table 1, which is not explicitly given in their paper.

Firstly, if we take the constraints of (Gao et al., 2023, Theorem 1), then the convergence rate of MDBO satisfies:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla \Phi(\bar{x}^{(t)})\right\|^2\right] \leq \frac{C_1}{(1-\rho)^2 \eta T} + C_2 \eta, \tag{123}$$

where C_1, C_2, C_3 are constants and η denotes the step-size. Here we remove the $\mathcal{O}(\varepsilon)$ terms as ε can be set sufficiently small. Let $\eta = \left(\frac{C_1}{C_2(1-\rho)^2 T}\right)^{\frac{1}{2}}$, then we have:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla \Phi(\bar{x}^{(t)})\right\|^2\right] \lesssim \frac{1}{(1-\rho)\sqrt{T}}. \tag{124}$$

Similarly, if we take the constraints of (Gao et al., 2023, Theorem 2), then the convergence rate of MDBO satisfies:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla \Phi(\bar{x}^{(t)})\right\|^2\right] \leq \frac{C_1}{\eta T} + \frac{C_2 \eta^2}{(1-\rho)^4} + C_3 \frac{\eta}{N}, \tag{125}$$

where C_1, C_2, C_3, C_4 are constants. Taking

$$\eta_1 = \left(\frac{C_1(1-\rho)^4}{C_2 T}\right)^{\frac{1}{3}}, \quad \eta_2 = \left(\frac{C_1 N}{C_3 T}\right)^{\frac{1}{2}}, \quad \eta = \left(\frac{1}{\eta_1} + \frac{1}{\eta_2}\right)^{-1},$$

we have:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla \Phi(\bar{x}^{(t)})\right\|^2\right] \lesssim \frac{1}{\sqrt{NT}} + \frac{1}{(1-\rho)^{\frac{4}{3}} T^{\frac{2}{3}}}. \tag{126}$$

Thus, it has an asymptotic convergence rate of $1/\sqrt{NT}$. Moreover, to achieve the linear speedup stage, the following condition is required:

$$\frac{1}{(1-\rho)^{\frac{4}{3}} T^{\frac{2}{3}}} \lesssim \frac{1}{\sqrt{NT}}. \tag{127}$$

The above constraint is equivalent to $T \lesssim \frac{N^3}{(1-\rho)^8}$.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.

DG LiMu Andersen and IW Park. Scaling distributed machine learning with the parameter server. In *Proceedings of the Operating Systems Design and Implementation (OSDI)*, 2014.

Sanjeev Arora, Simon Du, Sham Kakade, Yuping Luo, and Nikunj Saunshi. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pages 367–376. PMLR, 2020.

Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *Advances in neural information processing systems*, 33: 14879–14890, 2020.

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

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Tsung-Hui Chang, Mingyi Hong, and Xiangfeng Wang. Multi-agent distributed optimization via inexact consensus admm. *IEEE Transactions on Signal Processing*, 63(2):482–497, 2014.

Jianshu Chen and Ali H Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.

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

Xuxing Chen, Minhui Huang, and Shiqian Ma. Decentralized bilevel optimization. *arXiv preprint arXiv:2206.05670*, 2022.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishna Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International Conference on Machine Learning*, pages 4641–4671. PMLR, 2023a.

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

Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Accelerating gossip sgd with periodic global averaging. In *International Conference on Machine Learning (ICML)*, 2021b.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153:235–256, 2007.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

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

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

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

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pages 1568–1577. PMLR, 2018.

Hongchang Gao, Bin Gu, and My T Thai. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *International Conference on Artificial Intelligence and Statistics*, pages 9238–9281. PMLR, 2023.

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

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

Andrew Gibiansky. Bringing HPC techniques to deep learning. `https://andrew.gibiansky.com/blog/machine-learning/baidu-allreduce/`, 2017. Accessed: 2020-08-12.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

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

Minhui Huang, Dewei Zhang, and Kaiyi Ji. Achieving linear speedup in non-iid federated bilevel learning. *arXiv preprint arXiv:2302.05412*, 2023.

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

Mert Kayaalp, Stefan Vlaski, and Ali H Sayed. Dif-maml: Decentralized multi-agent meta-learning. *IEEE Open Journal of Signal Processing*, 3:71–93, 2022.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, pages 1–12, 2020.

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

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

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

Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17):4494–4506, 2019.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

Tao Lin, Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *International Conference on Machine Learning*, 2021.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35:17248–17262, 2022a.

Zhuqing Liu, Xin Zhang, Prashant Khanduri, Songtao Lu, and Jia Liu. Interact: Achieving low sample and communication complexities in decentralized bilevel learning over networks.

In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 61–70, 2022b.

Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321. IEEE, 2019.

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

Songtao Lu, Siliang Zeng, Xiaodong Cui, Mark Squillante, Lior Horesh, Brian Kingsbury, Jia Liu, and Mingyi Hong. A stochastic linearized augmented lagrangian method for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 35:30638–30650, 2022b.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pages 7111–7123. PMLR, 2021.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

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

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Youcheng Niu, Jinming Xu, Ying Sun, Yan Huang, and Li Chai. Distributed stochastic bilevel optimization: Improved complexity and heterogeneity analysis. *arXiv preprint arXiv:2312.14690*, 2023.

Rui Pan, Jipeng Zhang, Xingyuan Pan, Renjie Pi, Xiaoyu Wang, and Tong Zhang. Scalebio: Scalable bilevel optimization for llm data reweighting. *arXiv preprint arXiv:2406.19976*, 2024.

Shi Pu, Alex Olshevsky, and Ioannis Ch Paschalidis. A sharp estimate on the transient time of distributed stochastic gradient descent. *IEEE Transactions on Automatic Control*, 67 (11):5900–5915, 2021.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Ali H Sayed. Adaptive networks. *Proceedings of the IEEE*, 102(4):460–497, 2014.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.

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

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. d^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856, 2018.

Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179. PMLR, 2022.

Luis N Vicente and Paul H Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global optimization*, 5(3):291–306, 1994.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Xiaoyu Wang, Xuxing Chen, Shiqian Ma, and Tong Zhang. Fully first-order methods for decentralized bilevel optimization. *arXiv preprint arXiv:2410.19319*, 2024.

Min Wen, Chengchang Liu, Ahmed Abdelmoniem, Yipeng Zhou, and Yuedong Xu. A communication and computation efficient fully first-order method for decentralized bilevel optimization. *arXiv preprint arXiv:2410.14115*, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Quan Xiao, Han Shen, Wotao Yin, and Tianyi Chen. Alternating implicit projected sgd and its efficient variants for equality-constrained bilevel optimization. *arXiv preprint arXiv:2211.07096*, 2022.

Ran Xin, Usman A Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *arXiv preprint arXiv:2008.04195*, 2020.

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

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

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $O(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *arXiv preprint arXiv:2312.03807*, 2023a.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. *arXiv preprint arXiv:2305.19442*, 2023b.

Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Bicheng Ying, Kun Yuan, Hanbin Hu, Yiming Chen, and Wotao Yin. Bluefog: Make decentralized algorithms practical for optimization and deep learning. *arXiv preprint arXiv:2111.04287*, 2021b.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H Sayed. Exact diffusion for distributed optimization and learning—part i: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723, 2018.

Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. DecentLaM: Decentralized momentum SGD for large-batch deep training. *International Conference on Computer Vision*, 2021.

Kun Yuan, Xinmeng Huang, Yiming Chen, Xiaohan Zhang, Yingya Zhang, and Pan Pan. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. *Advances in Neural Information Processing Systems*, 35:36382–36395, 2022.

Kun Yuan, Sulaiman A Alghunaim, and Xinmeng Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized SGD. *Journal of Machine Learning Research*, 24(280):1–53, 2023.

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

Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pages 26693–26712. PMLR, 2022.

Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bi-level optimization: Foundations and applications in signal processing and machine learning. *arXiv preprint arXiv:2308.00788*, 2023b.

Shuchen Zhu, Boao Kong, Songtao Lu, Xinmeng Huang, and Kun Yuan. SPARKLE: A unified single-loop primal-dual framework for decentralized bilevel optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.