SPARKLE: A Unified Single-Loop Primal-Dual Framework for Decentralized Bilevel Optimization

Shuchen Zhu[*]
Peking University
shuchenzhu@stu.pku.edu.cn

Boao Kong[*]
Peking University
kongboao@stu.pku.edu.cn

Songtao Lu
IBM Research
songtao@ibm.com

Xinmeng Huang
University of Pennsylvania
xinmengh@sas.upenn.edu

Kun Yuan[†]
Peking University
kunyuan@pku.edu.cn

Abstract

This paper studies decentralized bilevel optimization, in which multiple agents collaborate to solve problems involving nested optimization structures with neighborhood communications. Most existing literature primarily utilizes gradient tracking to mitigate the influence of data heterogeneity, without exploring other well-known heterogeneity-correction techniques such as EXTRA or Exact Diffusion. Additionally, these studies often employ identical decentralized strategies for both upper- and lower-level problems, neglecting to leverage distinct mechanisms across different levels. To address these limitations, this paper proposes **SPARKLE**, a unified **S**ingle-loop **P**rimal-dual **A**lgo**R**ithm framewor**K** for decentra**L**ized bil**E**vel optimization. SPARKLE offers the flexibility to incorporate various heterogeneity-correction strategies into the algorithm. Moreover, SPARKLE allows for different strategies to solve upper- and lower-level problems. We present a unified convergence analysis for SPARKLE, applicable to all its variants, with state-of-the-art convergence rates compared to existing decentralized bilevel algorithms. Our results further reveal that EXTRA and Exact Diffusion are more suitable for decentralized bilevel optimization, and using mixed strategies in bilevel algorithms brings more benefits than relying solely on gradient tracking.

1 Introduction

Numerous modern machine learning tasks, such as reinforcement learning [25], meta-learning [4], adversarial learning [36], hyper-parameter optimization [19], and imitation learning [3], entail nested optimization formulations that extend beyond the traditional single-level paradigm. For instance, hyper-parameter optimization aims to identify the optimal hyper-parameters for a specific learning task in the upper level by minimizing the validation loss, achieved through training models in the lower-level process. This nested optimization structure has spurred significant attention towards Stochastic Bilevel Optimization (SBO). Since the size of data samples involved in bilevel problems has become increasingly large, this paper investigates decentralized algorithms over a network of n agents (nodes) that collaborate to solve the following distributed bilevel optimization problem:

$$\min_{x\in\mathbb{R}^p} \quad \Phi(x) = f(x, y^\star(x)) := \frac{1}{n}\sum_{i=1}^{n} f_i(x, y^\star(x)), \qquad \text{(upper-level)} \tag{1a}$$

[*]Equal Contribution

[†]Corresponding Author: Kun Yuan. Kun Yuan is also affiliated with National Engineering Labratory for Big Data Analytics and Applications, and AI for Science Institute, Beijing, China.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

$$\text{s.t.} \quad y^\star(x) = \underset{y\in\mathbb{R}^q}{\operatorname{argmin}} \Big\{ g(x,y) := \frac{1}{n}\sum_{i=1}^n g_i(x,y) \Big\}. \qquad \text{(lower-level)} \tag{1b}$$

In this formulation, each agent i holds a private upper-level objective function $f_i : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ and a strongly convex lower-level objective function $g_i : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ defined as:

$$f_i(x,y) = \mathbb{E}_{\phi\sim\mathcal{D}_{f_i}}[F_i(x,y;\phi)], \qquad g_i(x,y) = \mathbb{E}_{\xi\sim\mathcal{D}_{g_i}}[G_i(x,y;\xi)], \tag{2}$$

where $\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$ represent the local data distributions at agent i. This paper does not make any assumptions about these data distributions, implying there might be data heterogeneity across agents.

Linear speedup and transient iteration complexity. A decentralized stochastic algorithm achieves *linear speedup* if its iteration complexity decreases linearly with the network size n. Additionally, the *transient iteration complexity* refers to the number of transient iterations a decentralized algorithm must undergo to achieve the asymptotic linear speedup stage. The fewer the transient iterations, the faster the algorithm can achieve linear speedup. This paper aims to develop decentralized stochastic bilevel algorithms that can achieve linear speedup with as few transient iterations as possible.

Limitations in previous works. A significant challenge in decentralized bilevel optimization lies in accurately estimating the hyper-gradient $\nabla\Phi(x)$ through neighborhood communications. Several studies have emerged to effectively address this challenge, such as those by [9, 33, 52, 16, 21, 40, 57, 29]. However, existing works suffer from several critical limitations:

- **Stringent assumptions and inadequate convergence analysis.** Many existing studies rely on stringent assumptions to ensure convergence. For instance, references [9, 10, 33, 21, 52] assume bounded gradients, while reference [29] assumes bounded data heterogeneity (also known as bounded gradient dissimilarity). These restrictive assumptions do not arise in centralized bilevel optimization, implying their potential unnecessity. Moreover, some of these works suffer from inadequate convergence analysis, unable to clarify the transient iteration complexity [9, 33, 10] or provide a sharp estimation of the influence of network topologies [52, 21].
- **Limited exploration of various heterogeneity-correction techniques.** Several concurrent studies [16, 57, 40] have utilized Gradient Tracking (GT) [50, 13, 38] to remove the assumption of bounded data heterogeneity. However, it remains uncertain whether GT is the most suitable mechanism for decentralized bilevel optimization. Many other techniques are also useful for addressing data heterogeneity in single-level decentralized optimization, such as EXTRA [45] and Exact-Diffusion (ED) [56, 30, 54] (which is also known as D^2 [46]). Even within GT, there are variants including Adapt-Then-Combine GT (ATC-GT) [50], non-ATC-GT [38], and semi-ATC-GT [13]. It remains unexplored whether these techniques for mitigating data heterogeneity converge and even outperform GT when employed in decentralized bilevel algorithms.
- **Unknown effects of employing different upper- and lower-level update strategies.** In bilevel optimization, the challenges in solving the upper- and lower-level problems differ substantially. For instance, the upper-level problem (1a) is non-convex, whereas the lower-level problem (1b) is strongly convex. Moreover, estimating the gradient at the lower level is considerably simpler compared to estimating the hyper-gradient at the upper level. Understanding the roles of updates at each level is crucial to develop more efficient algorithms. However, most existing algorithms employ the same decentralized methods to solve both the upper- and lower-level problems. For example, references [9, 52, 29] utilize decentralized gradient descent (DGD) for updates at both levels while [57, 40, 16] leverage GT, overlooking the potential advantages of mixed strategies.

To address these limitations, several critical questions naturally arise: Should each heterogeneity-correction mechanism listed in [50, 13, 38, 56, 30, 54, 46] be explored one-by-one? Should we consider combining any two of these techniques to update the upper and lower-level problems, respectively? It is evident that examining each individual heterogeneity-correction technique, and even exploring their combinations, would involve an unbearable amount of effort.

Main results and contributions. This paper addresses all the aforementioned limitations without exhaustively exploring all heterogeneity-correction techniques. Our main results are as follows.

- **A unified decentralized bilevel framework.** To avoid examining each single heterogeneity-correction technique, we propose **SPARKLE**, a unified **S**ingle-loop **P**rimal-dual **A**lgo**R**ithm

Table 1: Comparison between different decentralized stochastic bilevel algorithms. K denotes the number of (upper-level) iterations; $1-\rho$ denotes the spectral gap of the mixing matrix (see Assumption 2); b^2 bounds the gradient dissimilarity; ε is the target stationarity such that $\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2]/K<\varepsilon$; p and q are the dimensions of the upper- and lower-level variables, reflecting per-round communication costs. Assumptions of bounded gradient, Lipschitz continuity, and bounded gradient dissimilarity are abbreviated as BG, LC, and BGD, respectively. We also list the best-known results of single-level GT, EXTRA, and ED at the bottom.

Algorithms	**Assumption**$^{\triangleright}$	**A. Rate.** $^{\diamond}$	**A. Comp.** †	**A. Comm.** ‡	**Tran. Iter.**$^{\triangleleft}$	**Loopless**
DSBO [9]	LC	$\frac{1}{\sqrt{K}}$	$\frac{1}{\varepsilon^3}$	$\left(pq\log(\frac{1}{\varepsilon})+\frac{q}{\varepsilon}\right)\frac{1}{\varepsilon^2}$	N.A.	No
MA-DSBO [10]	LC	$\frac{1}{\sqrt{K}}$	$\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon})$	$\frac{q}{\varepsilon^2}\log(\frac{1}{\varepsilon})+\frac{p}{\varepsilon^2}$	N.A.	No
SLAM [33]	LC	$\frac{1}{\sqrt{nK}}$	$\frac{1}{n\varepsilon^2}\log(\frac{1}{\varepsilon})$	$\frac{p+q}{n\varepsilon^2}$	N.A.	No
MDBO [21]	BG	$\frac{1}{\sqrt{nK}}$	$\frac{1}{n\varepsilon^2}\log(\frac{1}{\varepsilon})$	$\frac{p+q}{n\varepsilon^2}$	$\frac{n^3}{(1-\rho)^8}$	No
Gossip DSBO [52]	BG	$\frac{1}{\sqrt{nK}}$	$\frac{1}{n\varepsilon^2}\log(\frac{1}{\varepsilon})$	$\frac{q^2}{n\varepsilon^2}\log(\frac{1}{\varepsilon})+\frac{pq}{n\varepsilon^2}$	$\frac{n^3}{(1-\rho)^4}$	No
LoPA [40]*	BGD	$\frac{1}{\sqrt{K}}$	$\frac{1}{\varepsilon^2}$	$\frac{p+q}{\varepsilon^2}$	N.A.	Yes
D-SOBA [29]	BGD	$\frac{1}{\sqrt{nK}}$	$\frac{1}{n\varepsilon^2}$	$\frac{p+q}{n\varepsilon^2}$	$\max\left\{\frac{n^3}{(1-\rho)^2},\frac{n^3b^2}{(1-\rho)^4}\right\}$	Yes
SPARKLE-GT (ours)	**None**	$\boldsymbol{\frac{1}{\sqrt{nK}}}$	$\boldsymbol{\frac{1}{n\varepsilon^2}}$	$\boldsymbol{\frac{ap+q}{n\varepsilon^2}}^{\sharp}$	$\boldsymbol{\max\left\{\frac{n^3}{(1-\rho)^2},\frac{n}{(1-\rho)^{8/3}}\right\}}$	**Yes**
SPARKLE-EXTRA (ours)	**None**	$\boldsymbol{\frac{1}{\sqrt{nK}}}$	$\boldsymbol{\frac{1}{n\varepsilon^2}}$	$\boldsymbol{\frac{ap+q}{n\varepsilon^2}}^{\sharp}$	$\boldsymbol{\frac{n^3}{(1-\rho)^2}}$	**Yes**
SPARKLE-ED (ours)	**None**	$\boldsymbol{\frac{1}{\sqrt{nK}}}$	$\boldsymbol{\frac{1}{n\varepsilon^2}}$	$\boldsymbol{\frac{ap+q}{n\varepsilon^2}}^{\sharp}$	$\boldsymbol{\frac{n^3}{(1-\rho)^2}}$	**Yes**
Single-level GT [2, 28]	None	$\frac{1}{\sqrt{nK}}$	$\frac{1}{n\varepsilon^2}$	$\frac{p}{n\varepsilon^2}$	$\max\left\{\frac{n^3}{(1-\rho)^2},\frac{n}{(1-\rho)^{8/3}}\right\}$	Yes
Single-level EXTRA [2]	None	$\frac{1}{\sqrt{nK}}$	$\frac{1}{n\varepsilon^2}$	$\frac{p}{n\varepsilon^2}$	$\frac{n^3}{(1-\rho)^2}$	Yes
Single-level ED [2]	None	$\frac{1}{\sqrt{nK}}$	$\frac{1}{n\varepsilon^2}$	$\frac{p}{n\varepsilon^2}$	$\frac{n^3}{(1-\rho)^2}$	Yes

$^{\diamond}$ The convergence rate when $K\to\infty$ (smaller is better).
† The number of gradient/Jacobian/Hessian evaluations per agent to achieve ε-accuracy when $\epsilon\to 0$ (smaller is better).
‡ The communication costs per agent to achieve ε-stationarity when $\epsilon\to 0$ (smaller is better).
$^{\triangleleft}$ The transient iteration complexity to achieve linear speedup (smaller is better). "N.A." means that the algorithm cannot achieve linear speedup or the transient time cannot be accessed from existing convergence analysis.
$^{\triangleright}$ Additional assumptions beyond Assumption 1.
* LoPA solves the personalized problem, where the lower-level objectives are local to agents.
$^{\sharp}$ $a>0$ measures the relative sparsity of the mixing weights $\mathbf{W}_x,\mathbf{W}_y,\mathbf{W}_z$, which can be very small in certain cases. Here $1-\rho$ in **Tran. Iter.** denotes the smallest spectral gap of $\mathbf{W}_x,\mathbf{W}_y,\mathbf{W}_z$. See more discussions in Appendix C.2.3.

framewor**K** for decentra**L**ized bil**E**vel optimization. By specifying certain hyper-parameters, SPARKLE can be tailored to SPARKLE-EXTRA, SPARKLE-ED, and SPARKLE-GT, which employ EXTRA [45], ED [56, 30], or multiple GT variants [50, 13, 38], respectively, to facilitate the upper and lower-level problems. Additionally, SPARKLE is the *first* algorithm enabling distinct updating strategies across different levels; for example, one can utilize GT in the upper-level but ED in the lower-level, resulting in a brand new SPARKLE-GT-ED algorithm.

- **A unified and sharp analysis of various heterogeneity-correction schemes.** We provide a unified convergence analysis for SPARKLE, which immediately applies to all SPARKLE variants with distinct heterogeneity-correction techniques. The analysis does not require restrictive assumptions such as gradient boundedness used in [9, 10, 33, 21, 52] or data-heterogeneity bounded used in [29]. Moreover, our analysis demonstrates the provable superiority of SPARKLE compared to existing algorithms, as evidenced by the convergence rates listed in Table 1. Most importantly, our analysis shows that both SPARKLE-EXTRA and SPARKLE-ED outperform SPARKLE-GT (see Table 1), implying that *GT is not the best* scheme for decentralized bilevel optimization.

- **Mixing strategies outperform employing GT alone.** We demonstrate how optimization at different levels affects convergence rates. Our theoretical analysis suggests that the updating strategy at *the lower level is crucial in determining the overall performance* in decentralized bilevel algorithms. Building upon this insight, we establish that incorporating the ED or EXTRA strategy in the lower-level update phase leads to better transient iteration complexity than relying solely on the GT mechanism in both levels as proposed in [10, 16, 40], see Table 2 for more details.

- **Comparable performance with single-level algorithms.** We elucidate the comparison between bilevel and single-level stochastic decentralized optimization. On one hand, we demonstrate that the convergence performance of all our proposed algorithms is not inferior to their single-level counterparts (see the bottom part in Table 1). On the other hand, by considering specific lower-level loss functions, our bilevel results directly yield the non-asymptotic convergence of corresponding

Table 2: The transient iteration complexity of SPARKLE with mixed updating strategies at various levels. The smaller the transient iteration complexity is, the faster the algorithm will achieve its linear speedup stage. The first row and column respectively indicate the updating strategy for the upper- and lower-level problems. Please refer to Appendix B.3 for more implementation details and Appendix C.2.4 for proofs .

lower \ upper	**ED**	**EXTRA**	**GT**
ED	$\frac{n^3}{(1-\rho)^2}$	$\frac{n^3}{(1-\rho)^2}$	$\frac{n^3}{(1-\rho)^2}$
EXTRA	$\frac{n^3}{(1-\rho)^2}$	$\frac{n^3}{(1-\rho)^2}$	$\frac{n^3}{(1-\rho)^2}$
GT	$\max\left\{\frac{n^3}{(1-\rho)^2}, \frac{n}{(1-\rho)^{8/3}}\right\}$	$\max\left\{\frac{n^3}{(1-\rho)^2}, \frac{n}{(1-\rho)^{8/3}}\right\}$	$\max\left\{\frac{n^3}{(1-\rho)^2}, \frac{n}{(1-\rho)^{8/3}}\right\}$

single-level algorithms. This is the *first* result demonstrating bilevel optimization essentially subsumes the convergence of the single-level optimization.

Our main results are listed in Table 1. All SPARKLE variants achieve the state-of-the-art asymptotic rate, asymptotic gradient complexity, asymptotic communication cost, and transient iteration complexity under more relaxed assumptions compared to existing methods.

Related works. A significant challenge in decentralized bilevel optimization is accurately estimating the hyper-gradient $\nabla\Phi(x)$, necessitating solving global lower-level problems and estimating Hessian inversion. To this end, various decentralized techniques have been applied in bilevel optimization, including Neumann series in [52], JHIP oracle in [9], HIGP oracle in [10], and augmented Lagrangian-based communication in [33]. Additionally, reference [29] proposes a single-loop algorithm utilizing decentralized SOBA. To enhance algorithmic robustness against data heterogeneity, recent studies have employed Gradient Tracking (GT) in both lower- and upper-level optimization. However, existing works built upon GT suffer from several limitations. Results of [16, 9] concentrate solely on deterministic cases, while reference [40] addresses personalized problems in the lower-level, which do not require achieving global consensus in the lower-level problem. Moreover, [9, 10] introduce computationally expensive inner loops for GT steps. None of these works can establish smaller transient iteration complexity than D-SOBA for decentralized SBO, even though the latter algorithm employs no heterogeneity-correction technique.

The unified framework for single-level decentralized optimization has been extensively studied in the literature. References [1, 49, 26] propose frameworks for decentralized composite optimization in deterministic settings, while [2] investigates a framework under stochastic settings. However, none of these works can be directly applied to decentralized bilevel algorithms. Several studies [21, 57] utilize variance reduction techniques to accelerate the convergence of stochastic decentralized bilevel algorithms. Our proposed SPARKLE framework is orthogonal to variance reduction; it can also incorporate variance-reduced gradient estimation to achieve improved convergence rates. More relevant works on decentralized optimization and bilevel optimization are discussed in Appendix A.

Notations. We use lowercase letters to represent vectors and uppercase letters to represent matrices. We introduce $\text{col}\{x_1, ..., x_n\} := [x_1^\top, ..., x_n^\top]^\top \in \mathbb{R}^{pn}$ for brevity. Variables with overbar denote the average over all agents. For example, $\bar{x}^k = \sum_{i=1}^n x_i^k/n$. We denote $\overline{A} = A - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ for matrix $A \in \mathbb{R}^{n\times n}$, where $\mathbf{1}_n \in \mathbb{R}^n$ denotes the n-dimensional vector with all entries being one. For a function $f(x, y) : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$, we use $\nabla_1 f(x, y) \in \mathbb{R}^p$, $\nabla_2 f(x, y) \in \mathbb{R}^q$ to represent its partial gradients with respect to x and y, respectively. Similarly, $\nabla_{12} f(x, y) \in \mathbb{R}^{p\times q}$, $\nabla_{22} f(x, y) \in \mathbb{R}^{q\times q}$ represent the corresponding Jacobian and Hessian matrix. We use the notation $\lesssim$ to denote inequalities that hold up to constants related to the initialization of algorithms and smoothness constants.

2 SPARKLE: A unified framework for decentralized bilevel optimization

This section develops SPARKLE, a unified framework for decentralized bilevel optimization, and discusses its numerous variants by specifying certain hyper-parameters.

2.1 Three pillar subproblems in decentralized bilevel optimization.

When solving the upper-level problem (1a), it is critical to obtain the hyper-gradient $\nabla\Phi(x)$, which can be expressed as [22]

$$\nabla\Phi(x) = \nabla_1 f(x, y^\star(x)) - \nabla_{12}^2 g(x, y^\star(x)) \left[\nabla_{22}^2 g(x, y^\star(x))\right]^{-1} \nabla_2 f(x, y^\star(x)). \tag{3}$$

Evaluating this hyper-gradient is computationally expensive due to the inversion of the Hessian matrix. This evaluation becomes even more challenging over a decentralized network of collaborative agents. First, the inverse of the Hessian matrix cannot be obtained by simply averaging the local Hessian inverses due to $[\frac{1}{n}\sum_{i=1}^n \nabla_{22}^2 g_i(x, y^\star(x))]^{-1} \neq \frac{1}{n}\sum_{i=1}^n [\nabla_{22}^2 g_i(x, y^\star(x))]^{-1}$. Second, the global averaging operation cannot be realized through decentralized communication. To overcome these challenges, one can introduce an auxiliary variable $z^\star(x) := [\nabla_{22}^2 g(x, y^\star(x))]^{-1}\nabla_2 f(x, y^\star(x))$ [12], which is the solution to a quadratic problem

$$z^\star(x) = \operatorname*{argmin}_{z\in\mathbb{R}^q} \left\{ \frac{1}{2} z^\top \nabla_{22}^2 g\left(x, y^\star(x)\right) z - z^\top \nabla_2 f\left(x, y^\star(x)\right) \right\}. \tag{4}$$

Once $z^\star(x)$ is derived by solving (4), we can substitute it into (3) to achieve $\nabla\Phi(x)$.

Following this idea, solving the distributed bilevel optimization problem (1) essentially involves solving three subproblems, where $h_i(x, y^\star(x), z) := \frac{1}{2} z^\top \nabla_{22}^2 g_i(x, y^\star(x)) z - z^\top \nabla_2 f_i(x, y^\star(x))$,

$$x^\star = \operatorname*{argmin}_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n f_i(x, y^\star(x)), \qquad \text{(upper-level)} \tag{5a}$$

$$y^\star(x) = \operatorname*{argmin}_{y\in\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^n g_i(x, y), \qquad \text{(lower-level)} \tag{5b}$$

$$z^\star(x) = \operatorname*{argmin}_{z\in\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^n h_i(x, y^\star(x), z). \qquad \text{(auxiliary-level)} \tag{5c}$$

Given the variable x, one can achieve $y^\star(x)$ by solving the lower-level problem in (5b). With $y^\star(x)$ determined, $z^\star(x)$ can be obtained by solving the auxiliary-level problem in (5c). Subsequently, with $z^\star(x)$ available, one can directly compute the hyper-gradient and solve the upper-level problem in (5a) using gradient descent. This constitutes the primary methodology to solve problem (1).

A bilevel algorithm essentially solves three subproblems listed in (5), each formulated as a single-level decentralized optimization problem. Nevertheless, primary approaches may suffer from nested loops in algorithmic development. A few recent studies [12, 11, 57, 29] propose to solve each problem in (5a)-(5c) approximately with *one single* iteration, leading to practical single-loop bilevel algorithms. For example, applying a D-SGD step [43] to each of (5a)-(5c) yields the D-SOBA method [29], while further leveraging the GT technique leads to decentralized bilevel methods in [9, 16, 57, 21].

However, it is less explored whether numerous other heterogeneity-correction techniques [50, 13, 38, 56, 30, 54, 46] beyond GT can be incorporated into algorithmic design to achieve even better performance in bilevel optimization. To avoid exploring each case individually, we next introduce a general framework that unifies all these techniques for solving single-level problems.

2.2 A unified framework for decentralized single-level optimization.

In this subsection, we consider solving the single-level problem $\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n f_i(x)$ over a network of n nodes. For each k-th ($k \geq 0$) iteration, we let x_i^k denote the local x-variable maintained by the i-th agent. Furthermore, we associate the topology with a weight matrix $W = [w_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ in which $w_{ij} \in (0, 1)$ if node j is connected to node i otherwise $w_{ij} = 0$. We use bold symbols to denote stacked vectors or matrices across agents. For example, $\mathbf{x}^k = \text{col}\{x_1^k, ..., x_n^k\} \in \mathbb{R}^{pn}$ and $\mathbf{W} = W \otimes I_p$, where $\otimes$ denotes the Kronecker product operator.

A unified framework with moving average. Building on the formulation in [1, 2], we develop a unified primal-dual framework with moving average for decentralized optimization:

$$\mathbf{r}^{k+1} = (1-\theta)\mathbf{r}^k + \theta\mathbf{g}^k, \quad \mathbf{x}^{k+1} = \mathbf{C}\mathbf{x}^k - \alpha\mathbf{A}\mathbf{r}^{k+1} - \mathbf{B}\mathbf{d}^k, \quad \mathbf{d}^{k+1} = \mathbf{d}^k + \mathbf{B}\mathbf{x}^{k+1}. \tag{6}$$

Here $\mathbf{x}^k$ denotes the primal variable, $\mathbf{d}^k$ denotes the dual variable introduced to mitigate the influence of data-heterogeneity, $\mathbf{g}^k$ stacks all (stochastic) gradients evaluated at x_i^k for $1 \leq i \leq n$, $\mathbf{r}^k$ denotes the momentum introduced to boost training with coefficient $\theta \in [0, 1]$, and $\alpha > 0$ is the learning rate. Matrices $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{pn\times pn}$ are adapted from the mixing matrix $\mathbf{W}$, which determine how agents communicate with each other. See Appendix B.1 for more detailed motivations.

Framework (6) unifies various decentralized techniques in the literature. For instance, by letting $\theta = 1$ and specifying $\mathbf{A}, \mathbf{B}, \mathbf{C}$ delicately, framework (6) reduces to ED, EXTRA, and numerous GT

Algorithm 1 SPARKLE: A unified framework for decentralized stochastic bilevel optimization

Require: Initialize $\mathbf{x}^0 = \mathbf{y}^0 = \mathbf{z}^0 = \mathbf{r}^0 = \mathbf{0}$, $\mathbf{d}_x^0 = \mathbf{d}_y^0 = \mathbf{d}_z^0 = \mathbf{0}$, learning rate $\alpha_k, \beta_k, \gamma_k, \theta_k$.

for $k = 0, 1, \cdots, K-1$ **do**

$\mathbf{y}^{k+1} = \mathbf{C}_y\mathbf{y}^k - \beta_k\mathbf{A}_y\mathbf{v}^k - \mathbf{B}_y\mathbf{d}_y^k$, $\quad \mathbf{d}_y^{k+1} = \mathbf{d}_y^k + \mathbf{B}_y\mathbf{y}^{k+1}$; ▷ lower-level update

$\mathbf{z}^{k+1} = \mathbf{C}_z\mathbf{z}^k - \gamma_k\mathbf{A}_z\mathbf{p}^k - \mathbf{B}_z\mathbf{d}_z^k$, $\quad \mathbf{d}_z^{k+1} = \mathbf{d}_z^k + \mathbf{B}_z\mathbf{z}^{k+1}$; ▷ auxiliary-level update

$\mathbf{r}^{k+1} = (1-\theta_k)\mathbf{r}^k + \theta_k\mathbf{u}^k$; ▷ momentum update

$\mathbf{x}^{k+1} = \mathbf{C}_x\mathbf{x}^k - \alpha_k\mathbf{A}_x\mathbf{r}^{k+1} - \mathbf{B}_x\mathbf{d}_x^k$, $\quad \mathbf{d}_x^{k+1} = \mathbf{d}_x^k + \mathbf{B}_x\mathbf{x}^{k+1}$; ▷ upper-level update

end for

variants, see Table 3 and Appendix B.1 for more details. Framework (6) is closely related to the unified decentralized method developed in [1, 2]. The primary difference lies in the incorporation of the momentum variable $\mathbf{r}^k$, which can help improve the transient iteration complexity of the framework (6) and relax the smoothness condition for bilevel algorithms [11]. A detailed comparison between framework (6) and that proposed in [1, 2] is provided in Appendix B.2.

2.3 A unified framework for decentralized bilevel optimization.

By utilizing the unified framework (6) to approximately solve each subproblem in (5) with only *one iteration*, we achieve SPARKLE, a unified single-loop framework for decentralized bilevel optimization. In particular, we independently sample data $\xi_i^k \sim \mathcal{D}_{f_i}$, $\zeta_i^k \sim \mathcal{D}_{g_i}$ within each node at iteration k, and evaluate stochastic gradients/Jacobians/Hessians as follows

$$l_i^k = \nabla_1 F_i(x_i^k, y_i^k; \xi_i^k), \quad b_i^k = \nabla_2 F_i(x_i^k, y_i^k; \xi_i^k), \quad v_i^k = \nabla_2 G_i(x_i^k, y_i^k; \zeta_i^k),$$
$$J_i^k = \nabla_{12}^2 G_i(x_i^k, y_i^k; \zeta_i^k), \quad H_i^k = \nabla_{22}^2 G_i(x_i^k, y_i^k; \zeta_i^k).$$

Next we stack the descent directions for variables of each level as follows

$$\begin{aligned}
\text{lower-level stochstic gradient:} &\quad \mathbf{v}^k = \text{col}\{v_1^k, ..., v_n^k\},\\
\text{auxilliary-level stochstic gradient:} &\quad \mathbf{p}^k = \text{col}\{H_1^k z_1^k - b_1^k, ..., H_n^k z_n^k - b_n^k\},\\
\text{upper-level stochstic gradient:} &\quad \mathbf{u}^k = \text{col}\{l_1^k - J_1^k z_1^{k+1}, ..., l_n^k - J_n^k z_n^{k+1}\}.
\end{aligned}$$

The SPARKLE algorithm is detailed in Algorithm 1. In this algorithm, we utilize different dual variables $\mathbf{d}_s$ and communication matrices $\mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s$ for each variable $s \in \{x, y, z\}$ to optimize their respective objective functions. We use momentum $\mathbf{r}^k$ only for updating the upper-level variable, which is sufficient to enhance convergence of bilevel algorithms and relax the smoothness condition.

Versatility in decentralized strategies. SPARKLE is highly versatile, supporting various decentralized strategies by allowing the specification of different communication matrices $\mathbf{A}_s$, $\mathbf{B}_s$, and $\mathbf{C}_s$. For example, by setting $\mathbf{A}_s = \mathbf{I}$, $\mathbf{B}_s = (\mathbf{I} - \mathbf{W})^{1/2}$, and $\mathbf{C}_s = \mathbf{W}$ for any $s \in \{x, y, z\}$, SPARKLE will utilize EXTRA to update variables x, y, and z, resulting in the SPARKLE-EXTRA variant. Other variants can be achieved by setting $\mathbf{A}_s$, $\mathbf{B}_s$, and $\mathbf{C}_s$ according to Table 3. These variants can be implemented more efficiently than listed in Algorithm 1, see Appendix B.3.

Flexibility across optimization levels. SPARKLE supports different optimization and communication mechanisms for each level of (5), which can be directly achieved by choosing different $\mathbf{A}_s$, $\mathbf{B}_s$, and $\mathbf{C}_s$ matrices for each level $s \in \{x, y, z\}$. For example, SPARKLE can utilize GT to update the upper-level variable x while employing ED to update the auxiliary- and lower-level variables y and z. Throughout this paper, we denote SPARKLE using the decentralized mechanism $\mathbf{L}$ for the lower-level and auxiliary variables, and $\mathbf{U}$ for the upper-level in Algorithm 1, by SPARKLE-$\mathbf{L}$-$\mathbf{U}$, or simply SPARKLE-$\mathbf{L}$ if $\mathbf{L} = \mathbf{U}$. In addition, SPARKLE even supports utilizing different mixing matrices $\mathbf{W}_x$, $\mathbf{W}_y$, $\mathbf{W}_z$ across levels.

3 Convergence analysis

In this section, we establish the convergence properties of the SPARKLE framework and examine the influence of different decentralized techniques utilized across optimization levels.

Table 3: SPARKLE facilitates different decentralized techniques by specifying $\mathbf{A}_s, \mathbf{B}_s, \mathbf{C}_s$ for $s \in \{x, y, z\}$. We denote the stacked local variables and the associate gradients estimates by $\mathbf{s} \in \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ and $\mathbf{g}(\mathbf{s})$, respectively. The update rule refers to the specific algorithmic recursion for each level. See derivations in Appendix B.2.

Algorithms	$\mathbf{A}_s$	$\mathbf{B}_s$	$\mathbf{C}_s$	The specific update rule at the k-th iteration.
ED	$\mathbf{W}_s$	$(\mathbf{I}-\mathbf{W}_s)^{\frac{1}{2}}$	$\mathbf{W}_s$	$\mathbf{s}^{k+2} = \mathbf{W}_s\left(2\mathbf{s}^{k+1} - \mathbf{s}^k - \alpha\left(\mathbf{g}(\mathbf{s}^{k+1}) - \mathbf{g}(\mathbf{s}^k)\right)\right)$
EXTRA	$\mathbf{I}$	$(\mathbf{I}-\mathbf{W}_s)^{\frac{1}{2}}$	$\mathbf{W}_s$	$\mathbf{s}^{k+2} = \mathbf{W}_s\left(2\mathbf{s}^{k+1} - \mathbf{s}^k\right) - \alpha\left(\mathbf{g}(\mathbf{s}^{k+1}) - \mathbf{g}(\mathbf{s}^k)\right)$
ATC-GT	$\mathbf{W}_s^2$	$\mathbf{I}-\mathbf{W}_s$	$\mathbf{W}_s^2$	$\mathbf{s}^{k+1} = \mathbf{W}_s\left(\mathbf{s}^k - \alpha\mathbf{h}_s^k\right), \mathbf{h}_s^{k+1} = \mathbf{W}_s\left(\mathbf{h}_s^k + \mathbf{g}(\mathbf{s}^{k+1}) - \mathbf{g}(\mathbf{s}^k)\right)$
Semi-ATC-GT	$\mathbf{W}_s$	$\mathbf{I}-\mathbf{W}_s$	$\mathbf{W}_s^2$	$\mathbf{s}^{k+1} = \mathbf{W}_s\mathbf{s}^k - \alpha\mathbf{h}_s^k, \mathbf{h}_s^{k+1} = \mathbf{W}_s\left(\mathbf{h}_s^k + \mathbf{g}(\mathbf{s}^{k+1}) - \mathbf{g}(\mathbf{s}^k)\right)$
Non-ATC-GT	$\mathbf{I}$	$\mathbf{I}-\mathbf{W}_s$	$\mathbf{W}_s^2$	$\mathbf{s}^{k+1} = \mathbf{W}_s\mathbf{s}^k - \alpha\mathbf{h}_s^k, \mathbf{h}_s^{k+1} = \mathbf{W}_s\mathbf{h}_s^k + \mathbf{g}(\mathbf{s}^{k+1}) - \mathbf{g}(\mathbf{s}^k)$

3.1 Assumptions

Before presenting the theoretical guarantees, we first introduce the following assumptions used throughout this paper.

Assumption 1. *There exist constants $\mu_g, L_{f,0}, L_{f,1}, L_{g,1}, L_{g,2}$ such that for any $1 \le i \le n$,*

1. *$\nabla f_i, \nabla g_i, \nabla^2 g_i$ are $L_{f,1}, L_{g,1}, L_{g,2}$ Lipschitz continuous, respectively;*
2. *$\|\nabla_2 f_i\left(x, y^\star(x)\right)\| \le L_{f,0}$ for any $x \in \mathbb{R}^p$;*[3]
3. *$g_i(x, y)$ is μ_g-strongly convex with respect to y for any fixed $x \in \mathbb{R}^p$.*

Moreover, we define $L := \max\{L_{f,0}, L_{f,1}, L_{g,1}, L_{g,2}\}$ and $\kappa := L/\mu_g$.

Assumption 2. *For each $s \in \{x, y, z\}$, the corresponding mixing matrix $W_s \in \mathbb{R}^{n\times n}$ is non-negative, symmetric and doubly stochastic,* i.e.,

$$W_s = W_s^\top, \quad W_s\mathbf{1}_n = \mathbf{1}_n, \quad (W_s)_{ij} \ge 0, \quad \forall\, 1 \le i, j \le n,$$

and the corresponding communication graph is strongly-connected, i.e., *its eigenvalues satisfy $1 = \lambda_1(W_s) > \lambda_2(W_s) \ge \ldots \ge \lambda_n(W_s)$ and $\rho(W_s) := \max\{|\lambda_2(W_s)|, |\lambda_n(W_s)|\} < 1$.*

The value $1 - \rho(W_s)$ is referred to as the spectral gap in the literature [34, 53, 31] of W_s, which measures the connectivity of the communication graph. It would approach 0 for sparse networks. For example, it holds that $1 - \rho(W_s) = \Theta(1/n^2)$ for the matrix W_s induced by a ring graph.

Assumption 3. *For any $s \in \{x, y, z\}$, we assume the communication matrices A_s, B_s, C_s used in* SPARKLE *are polynomial functions of W_s. Furthermore, we assume A_s, C_s are doubly stochastic, and $\mathrm{Null}(B_s) = \mathrm{Span}\{\mathbf{1}_n\}$. In addition, we assume all eigenvalues of the augmented matrix*

$$L_s := \begin{bmatrix} \overline{C_s} - B_s^2 & B_s \\ -B_s & \overline{I_n} \end{bmatrix}$$

are strictly less than one in magnitude, where $\overline{C_s} \triangleq C_s - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $\overline{I_n} \triangleq I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$.

We remark that Assumption 3 is mild and is satisfied by all choices listed in Table 3. See more discussions in Appendix C.2.2.

Assumption 4. *We assume $\nabla F_i(x, y; \xi), \nabla G_i(x, y; \xi)$, and $\nabla^2 G_i(x, y; \xi)$ to be unbiased estimates of $\nabla f_i(x, y), \nabla g_i(x, y)$, and $\nabla^2 g_i(x, y)$ with bounded variances $\sigma_{f,1}^2, \sigma_{g,1}^2, \sigma_{g,2}^2$, respectively.*

3.2 Convergence theorem

Under the above assumptions, we establish the convergence properties as follows. Proof details can be found in Appendix C.

[3] This is more relaxed than Lipschitz continuous f_i, or bounded $\nabla_2 f_i$ in [21, 57, 33, 11].

Theorem 1. *Under Assumptions 1 – 4, there exist proper constant step-sizes α, β, γ and momentum coefficient θ, such that the* SPARKLE *framework listed in Algorithm 1 will converge as follow:*

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[\|\nabla\Phi(\bar{x}^k)\|^2] \lesssim \frac{\kappa^5\sigma}{\sqrt{nK}} + \kappa^{\frac{16}{3}}\left(\delta_{y,1}+\delta_{z,1}\right)\frac{\sigma^{\frac{2}{3}}}{K^{\frac{2}{3}}} + \kappa^{\frac{7}{2}}\delta_{x,1}\frac{\sigma^{\frac{1}{2}}}{K^{\frac{3}{4}}}$$
$$+\left(\kappa^{\frac{26}{5}}\delta_{y,2}+\kappa^6\delta_{z,2}\right)\frac{\sigma^{\frac{2}{5}}}{K^{\frac{4}{5}}} + \left(\kappa^{\frac{16}{3}}\delta_{y,3}+\kappa^{\frac{14}{3}}\delta_{z,3}+\kappa^{\frac{8}{3}}\delta_{x,3}\right)\frac{1}{K} + \left(\kappa C_\alpha+\kappa^4 C_\theta\right)\frac{1}{K},$$

where $\sigma \triangleq \max\{\sigma_{f,1},\sigma_{g,1},\sigma_{g,2}\}$, $\{\delta_{s,i}\}_{i=1}^3$ are constants depending only on $\mathbf{W}_s,\mathbf{A}_s,\mathbf{B}_s,\mathbf{C}_s$ for $s\in\{x,y,z\}$, and C_α, C_θ are constants independent of K. See Lemma 17 for their detailed values.

In the deterministic scenario with $\sigma = 0$, SPARKLE converges at the rate $\mathcal{O}(1/K)$, see the formal theorem and derivation in Appendix C.3. This recovers the rate in [15] under even milder assumptions. Unlike reference [15], which only considers GT in the deterministic setting, SPARKLE is a unified bilevel framework for the more general stochastic setting.

Linear speedup. According to Theorem 1, SPARKLE achieves an asymptotic linear speedup as K approaches infinity, which applies to all SPARKLE variants regardless of the decentralized strategies employed and whether they are utilized at different optimization levels. Furthermore, the asymptotically dominant term $\kappa^5\sigma/(\sqrt{nK})$ matches exactly with the single-node bilevel algorithm SOBA [12] when $n = 1$, implying the tightness of Theorem 1 in terms of the asymptotic rate.

Remark 1. *We establish an upper bound for the consensus error* $\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2}{n}\right]$. *Please refer to Lemma 19 in Appendix C.2.1 for more details.*

3.3 Transient iteration complexity

With the non-asymptotic rate established in Theorem 1, we can derive the transient iteration complexity of SPARKLE as follows. The proof is in Lemma 18.

Corollary 1. *Under the same assumptions as in Theorem 1, the transient iteration complexity of* SPARKLE*—with the influence of κ and σ^2 omitted for brevity—is on the order of*

$$\max\left\{n^2\delta_x, n^3\delta_y, n^3\delta_z, n\hat{\delta}_x, n\hat{\delta}_y, n\hat{\delta}_z\right\}, \tag{8}$$

where $\delta_s,\hat{\delta}_s$ only depend $\mathbf{W}_s,\mathbf{A}_s,\mathbf{B}_s,\mathbf{C}_s$ for $s\in\{x,y,z\}$. Their values are in Lemma 18.

We obtain the transient iteration complexity of each variant of SPARKLE by applying Corollary 1.

Corollary 2. *For* SPARKLE-ED *and* SPARKLE-EXTRA*, if we choose $\mathbf{W}_y = \mathbf{W}_z$, it holds that*

$$\begin{aligned}
\delta_x &= \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-2}\right), & \delta_y &= \delta_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-2}\right),\\
\hat{\delta}_x &= \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-\frac{3}{2}}\right), & \hat{\delta}_y &= \hat{\delta}_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-2}\right).
\end{aligned} \tag{9}$$

Furthermore, if we choose $\mathbf{W}_x = \mathbf{W}_y = \mathbf{W}_z$ and denote $\rho \triangleq \rho(\mathbf{W}_x)$, the transient iteration complexity derived in (8) can be simplified as $n^3/(1-\rho)^2$.

Corollary 3. *For* SPARKLE-GT *and its variants with semi/non-ATC-GT, if we let $\mathbf{W}_y = \mathbf{W}_z$,*

$$\begin{aligned}
\delta_x &= \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-2}\right), & \delta_y &= \delta_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-2}\right),\\
\hat{\delta}_x &= \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-2}\right), & \hat{\delta}_y &= \hat{\delta}_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-\frac{8}{3}}\right).
\end{aligned}$$

Furthermore, if we let $\mathbf{W}_x = \mathbf{W}_y = \mathbf{W}_z$ and denote $\rho \triangleq \rho(\mathbf{W}_x)$, the transient iteration complexity derived in (8) can be simplified as $\max\{n^3/(1-\rho)^2, n/(1-\rho)^{8/3}\}$.

Remark 2 (*SOTA transient iterations*)**.** Comparing with algorithms listed in Table 1, all SPARKLE variants achieve smaller transient iteration complexity, implying that they can achieve linear speedup much faster than the other algorithms, especially over sparse network topologies with $1-\rho\to 0$.

Remark 3 (*GT is not the best technique for decentralized SBO*)**.** While GT is widely adopted in the literature [16, 21, 57] to facilitate decentralized SBO, a comparison of Corollary 2 and 3 reveals that both SPARKLE-EXTRA and SPARKLE-ED outperform SPARKLE-GT in terms of transient iteration complexity. This implies that EXTRA and ED are better than GT for decentralized SBO.

3.4 Different strategies across optimization levels

Corollary 1 clarifies how different update strategies for x, y, and z impact the transient iterations through constants $\{\delta_s, \hat{\delta}_s\}$ for $s \in \{x, y, z\}$. Since $\delta_y = \delta_z$ and $\hat{\delta}_y = \hat{\delta}_z$ when $\mathbf{W}_y = \mathbf{W}_z$ (Lemma 18), we naturally employ the same strategy to update y and z. The following corollary studies the utilization of both ED and GT in SPARKLE. See the transient iterations complexity of other mixed strategies in Appendix C.2.4 and Table 2.

Corollary 4. *For* SPARKLE-ED-GT *which uses ED to update y and z and GT to update x, if $\mathbf{W}_x = \mathbf{W}_y = \mathbf{W}_z$ and we denote $\rho = \rho(\mathbf{W}_x)$, it then holds that*

$$\delta_x = \delta_y = \delta_z = \mathcal{O}\left((1-\rho)^{-2}\right), \qquad \hat{\delta}_x = \hat{\delta}_y = \hat{\delta}_z = \mathcal{O}\left((1-\rho)^{-2}\right),$$

which implies that the transient iteration complexity in (8) can be simplified as $n^3/(1-\rho)^2$.

Remark 4 (*Mixed strategies outperform employing GT only*)**.** Comparing Corollary 3 and 4, we find that using ED to update y and z will lead to smaller $\hat{\delta}_y$ and $\hat{\delta}_z$, which improves the transient iteration complexity compared to employing GT only in all optimization levels (see Corollary 3).

3.5 Different topologies across optimization levels

In SPARKLE, we can utilize different topologies across levels. Theorem 1 and Corollary 1 have clarified the influence of using different topologies across levels through the constants $\{\delta_s, \hat{\delta}_s\}$ for $s \in \{x, y, z\}$. For instance, when substituting $\{\delta_s, \hat{\delta}_s\}$ established in (9) into (8), SPARKLE-ED has the following transient iteration complexity:

$$\max\{n^2(1-\rho(\mathbf{W}_x))^{-2}, n^3(1-\rho(\mathbf{W}_y))^{-2}\}$$

where $\mathbf{W}_x$ is the mixing matrix for updating x, while $\mathbf{W}_y$ is for updating y and z. As long as $(1-\rho(\mathbf{W}_x))^{-1} \lesssim \sqrt{n}(1-\rho(\mathbf{W}_y))^{-1}$ holds, SPARKLE-ED retains the transient iteration complexity of $n^3(1-\rho(\mathbf{W}_y))^{-2}$, which allows for the utilization of a sparser network topology when updating x, thereby reducing communication overheads. Consequently, the ratio a of the communication volume per round for the variables x and y can be significantly less than one. See Appendix C.2.3 for discussion on how to use different topologies across levels in other SPARKLE variants.

3.6 Recovering single-level decentralized optimization

Previous works typically study single-level and bilevel optimization separately. By taking $G_i(x, y, \xi) \equiv |y|^2/2$ and $F_i(x, y, \phi) = F_i(x, \phi)$ into (2), the decentralized SBO problem (1) reduces to stochastic single-level optimization. By setting $\mathbf{z}^k \equiv 0$, $\mathbf{y}^k \equiv 0$, $u_i^k = \nabla_1 f_i(x_i^k, \xi_i^k)$, SPARKLE reduces to the single-level framework (6), whose convergence can be naturally guaranteed by Theorem 1. Please refer to Appendix C.4 for the detailed proof and results. This is the *first* result demonstrating that bilevel optimization essentially subsumes the convergence of single-level optimization.

4 Numerical experiments

In this section, we present experiments to validate our theoretical findings. We first explore how update strategies and network structures influence the convergence of SPARKLE. Then we compare SPARKLE to the existing decentralized SBO algorithms. Additional experiments about a decentralized SBO problem with synthetic data are in Appendix D.1.

Hyper-cleaning on FashionMNIST dataset. We consider a data hyper-cleaning problem [44] on a corrupted FashionMNIST dataset [48]. Problem formulations and experimental setups can be found in Appendix D.2. Firstly, we equip SPARKLE with different decentralized strategies in different optimization levels and then compare them with D-SOBA [29], MA-DSBO-GT [10], and MDBO [21] using the corruption rate $p = 0.1, 0.2, 0.3$, respectively. As is shown in Figure 1, all the SPARKLE-based algorithms generally achieve higher test accuracy than D-SOBA, while ED and EXTRA especially outperform GT. Meanwhile, using mixed strategies (*i*.e., SPARKLE-ED-GT and SPARKLE-EXTRA-GT) achieves similar test accuracy with SPARKLE-ED and SPARKLE-EXTRA and outperform SPARKLE-GT, respectively. These observations match with the theoretical results in Corollary 2-4 and Remark 3, 4.

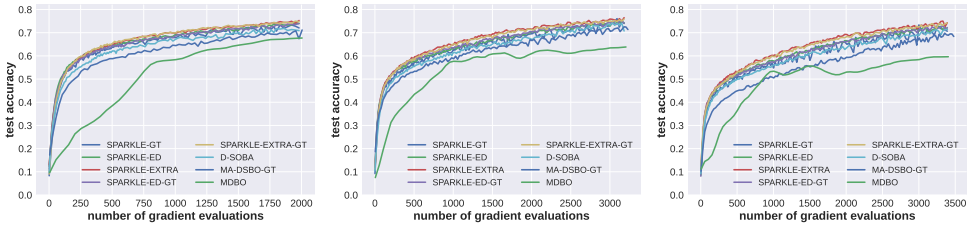


Figure 1: The test accuracy on hyper-cleaning with various SPARKLE-based algorithms using different corruption rates p. (Left: $p = 0.1$, Middle: $p = 0.2$, Right: $p = 0.3$.)

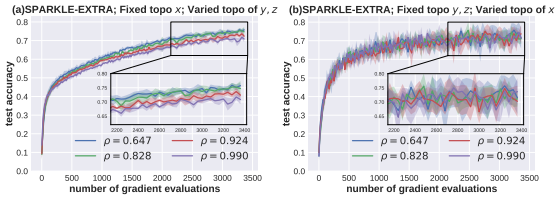


Figure 2: Test accuracy of SPARKLE-EXTRA on hyper-cleaning. (Left: fixed graph for x and varying graph for y, z; Right: fixed for y, z and varying for x)

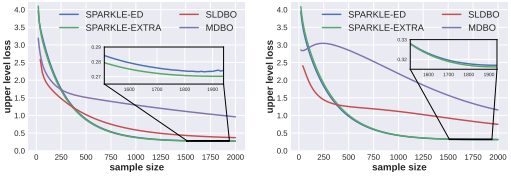


Figure 3: The upper-level loss against samples generated by one agent of different algorithms in the policy evaluation. (Left: $n = 10$, Right: $n = 20$.)

Next, we test SPARKLE-EXTRA with two communication strategies including *fixed topology for updating x and varying topology for y, z*, and *fixed topology for updating y, z and varying topology for x*. As illustrated in Figure 2, maintaining a fixed topology for x while reducing the connectivity of the topology for y and z will deteriorate the algorithmic performance. Conversely, preserving the topology for y and z while decreasing the connectivity for x has little impact on the performance. This suggests that the influence of the network topology for y and z on the algorithm dominates over the topology for x, which is consistent with our discussion in Section 3.5. We also numerically examine the influence of moving average on convergence, see discussions in Appendix D.2.

Distributed policy evaluation in reinforcement learning. We consider a multi-agent MDP problem in reinforcement learning on a distributed setting with $n \in \{10, 20\}$ agents respectively, which can be formulated as a decentralized SBO problems [52]. Here, we compare SPARKLE with existing decentralized SBO approaches including MDBO [21] and the stochastic extension of SLDBO [16] over a Ring graph. Figure 3 illustrates that SPARKLE converges faster and achieves a lower sample complexity than the other baselines, especially when $n = 20$, which shows the empirical benefits of SPARKLE in decentralized SBO algorithms with a large number of agents and sparse communication modes. More experimental details are in Appendix D.3.

Decentralized meta-learning. We investigate decentralized meta-learning on miniImageNet [47] with multiple tasks [18], formulating it as a decentralized bilevel optimization problem. This approach minimizes the validation loss with respect to shared parameters as the upper-level loss, while the training loss is managed by task-specific parameters at the lower level. Additional details about the experiment can be found in Appendix D.4. Our method, SPARKLE, is benchmarked against D-SOBA [29] and MAML [18], demonstrating a significant improvement in training accuracy.

5 Conclusions and limitations

This paper proposes SPARKLE, a unified single-loop primal-dual framework for decentralized stochastic bilevel optimization. Being highly versatile, SPARKLE can support different decentralized mechanisms and topologies across optimization levels. Moreover, all SPARKLE variants have been demonstrated to achieve state-of-the-art convergence rate compared to existing algorithms. However, SPARKLE currently supports only strongly-convex problems in the lower-level optimization. Its compatibility with generally-convex lower-level problems remains unknown. Additionally, the condition number of the lower-level problem significantly impacts the performance, as is the case with existing bilevel algorithms. We aim to address these limitations in future work.

6 Acknowledgment

The work of Shuchen Zhu, Boao Kong, and Kun Yuan is supported by Natural Science Foundation of China under Grants 92370121, 12301392, and W2441021. This work is also supported by Open Project of Key Laboratory of Mathematics and Information Networks, Ministry of Education, China. No. KF202302.

References

[1] S. A. Alghunaim, E. K. Ryu, K. Yuan, and A. H. Sayed. Decentralized proximal gradient algorithms with linear convergence rates. *IEEE Transactions on Automatic Control*, 66(6):2787–2794, 2020.

[2] S. A. Alghunaim and K. Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing*, 2022.

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

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

[5] T.-H. Chang, M. Hong, and X. Wang. Multi-agent distributed optimization via inexact consensus admm. *IEEE Transactions on Signal Processing*, 63(2):482–497, 2014.

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

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

[8] T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021.

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

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

[11] X. Chen, T. Xiao, and K. Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *arXiv preprint arXiv:2306.12067*, 2023.

[12] M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

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

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

[15] J. Dong, Z. Cao, T. Zhang, J. Ye, S. Wang, F. Feng, L. Zhao, et al. Eflops: Algorithm and system co-design for a high performance distributed training platform. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 610–622, 2020.

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

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

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

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

[20] B. Gao, Y. Yang, and Y. xiang Yuan. Lancbio: dynamic lanczos-aided bilevel optimization via krylov subspace. *arXiv preprint arXiv:2404.03331*, 2024.

[21] H. Gao, B. Gu, and M. 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.

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

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

[24] Z. Guo, Q. Hu, L. Zhang, and T. Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

[25] M. Hong, H.-T. Wai, Z. Wang, and Z. 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.

[26] D. Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2018.

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

[28] A. Koloskova, T. Lin, and S. U. Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34:11422–11435, 2021.

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

[30] Z. Li, W. Shi, and M. Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, July 2019. early acces. Also available on arXiv:1704.07807.

[31] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. 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.

[32] T. Lin, S. P. Karimireddy, S. U. Stich, and M. Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *International Conference on Machine Learning*, 2021.

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

[34] Y. Lu and C. De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pages 7111–7123. PMLR, 2021.

[35] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122. PMLR, 2015.

[36] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[37] A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[38] A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

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

[40] Y. Niu, J. Xu, Y. Sun, Y. Huang, and L. Chai. Distributed stochastic bilevel optimization: Improved complexity and heterogeneity analysis. *arXiv preprint arXiv:2312.14690*, 2023.

[41] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

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

[44] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.

[45] W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[46] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856, 2018.

[47] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[48] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[49] J. Xu, Y. Tian, Y. Sun, and G. Scutari. Distributed algorithms for composite optimization: Unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69:3555–3570, 2021.

[50] J. Xu, S. Zhu, Y. C. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *IEEE Conference on Decision and Control (CDC)*, pages 2055–2060, Osaka, Japan, 2015.

[51] J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

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

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

[54] K. Yuan, S. A. Alghunaim, B. Ying, and A. H. Sayed. On the influence of bias-correction on distributed stochastic optimization. *IEEE Transactions on Signal Processing*, 2020.

[55] K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[56] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed. Exact dffusion for distributed optimization and learning – Part I: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708 – 723, 2018.

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

Appendix for "SPARKLE: A Unified Single-Loop Primal-Dual Framework for Decentralized Bilevel Optimization"

Contents

A More related works

Bilevel optimization. Bilevel optimization presents substantial difficulties compared to single-level optimization due to its nested structure. Estimating hyper-gradient $\nabla\Phi(x)$ of the upper level involves solving lower-level problems and estimating the Hessian inverse, which requires additional calculations. Many algorithms and techniques have been proposed to solve the challenge. Approximate Implicit Differentiation (AID)-based algorithms [14, 22, 23, 27] leverage the implicit gradient form of $\nabla\Phi(x)$, which entails solving a linear system to obtain the Hessian-inverse-vector product. Similarly, [8, 25] utilize the Neumann series to handle the Hessian inverse. Iterative Differentiation (ITD)-based algorithms [19, 35, 14, 23, 27] use iterative methods solving the lower-level problem and then estimate the hyper-gradient through automatic differentiation. However, these approaches introduce inner steps, leading to extra computational overhead and memory spaces. [12] proposes a single-level algorithm called SOBA, which approximating the Hessian-inverse-vector product by solving a quadratic programming problem. A recent work [20] utilizes the Krylov subspace technique and the Lanczos process to approximate it in deterministic scenarios. For stochastic bilevel optimization, various methods have been employed to improve the convergence rate, such as momentum [7, 11] and variance reduction [51, 27, 24].

Decentralized optimization. Decentralized optimization is developed to deal with large-scale optimization problems, where datasets are distributed among multiple agents. Without a central server, each agent only gets access to its own local data and communications are limited to its neighbors in a network. Compared with centralized algorithms, decentralized ones preserve data privacy, and are more robust to contingencies in the communication network. However, due to the absence of a central server, decentralized optimization requires communication among agents, posing greater challenges for convergence, especially in the presence of severe data heterogeneity. To tackle this issue, various algorithms have emerged, such as decentralized gradient descent [39, 55], diffusion strategies [6], dual averaging [17], EXTRA [45], Exact Diffusion (a.k.a. D^2) [56, 30, 46], gradient tracking [50, 13, 38], and decentralized ADMM [5]. In stochastic scenarios, a common method for decentralized optimization is the decentralized stochastic gradient descent (DSGD), which has gained a lot of attentions recently. It has been proved to achieve linear speedup asymptotically and shares the same asymptotic rate with centralized stochastic gradient descent [31].

B More details of SPARKLE

B.1 Primal-dual deviation

Here we provide a detailed motivation of the update framework (6) for decentralized single-level algorithms. First, we rewrite the single-level distributed optimization problem in the following equivalent form:

$$\min_{x_i\in\mathbb{R}^d} f(x_1,...,x_n)=\frac{1}{n}\sum_{i=1}^{n} f_i\left(x_i\right), \quad \text{s.t. } x_1=...=x_n, \tag{10}$$

where each f_i is smooth and possibly non-convex. To simplify the notation, we assume that $d=1$ without loss of generality. Now we introduce three symmetric matrices A, B, D such that A is a doubly stochastic communication matrix with $\rho(A)<1$, and B, D satisfy $\mathrm{Null}B=\mathrm{Null}D=\mathrm{Span}\{1_n\}$. In general, B (D) determines the topology of a connected graph $\mathcal{G}_B$ ($\mathcal{G}_D$) over agents. The constraint $Bx=0$ ($Dx=0$) is equivalent to:

$$x_i=x_j \text{ if } x_i, x_j \text{ are adjacent in } \mathcal{G}_B\ (\mathcal{G}_D).$$

To simplify the derivation, we additionally assume that A, B, D are pairwise commutative. Then for $x=(x_1,...,x_n)$, we have:

$$x_1=...=x_n \Leftrightarrow Bx=0 \Leftrightarrow Dx=0 \Leftrightarrow Ax=x.$$

Therefore, (10) can be equivalently reformulated as

$$\min_{x\in\mathbb{R}^n} f(Ax), \quad \text{s.t. } Bx=0. \tag{11}$$

We construct the augmented Lagrangian function of the problem (11) as follows:

$$\mathcal{L}_\rho(x,d) = f(Ax) + \langle d, Bx\rangle + \frac{\rho}{2}\|Dx\|^2,$$

where x denotes the primal variable, d denotes the dual variable or Lagrangian multiplier associated with the consensus constraint, $\|Dx\|^2$ serves as the penalty term measuring the deviation from $Dx = 0$, or equivalently $Bx = 0$; $\rho > 0$ is the penalty coefficient. Though the introduction of matrices A, D is essentially a matter of equivalent substitution, it enhances the universality of the algorithm framework we get.

Following classical primal-dual methods, we alternately perform gradient descent on x and gradient ascent on d in the k-th iteration:

$$x^{k+1} = x^k - \alpha(A\nabla f(Ax^k) + Bd^k + \rho D^2 x^k), \quad d^{k+1} = d^k + \beta B x^{k+1},$$

where α, β denote the step-sizes. By making the change of variables

$$\hat{x}^k = Ax^k,\ \hat{d}^k = \sqrt{\frac{\alpha}{\beta}}Ad^k,\ \widehat{B} = \sqrt{\alpha\beta}B,\ \widehat{C} = I - \alpha\rho D^2,\ \widehat{A} = A^2,$$

we obtain

$$\hat{x}^{k+1} = \widehat{C}\hat{x}^k - \alpha\widehat{A}\nabla f(\hat{x}^k) - \widehat{B}\hat{d}^k, \quad \hat{d}^{k+1} = \hat{d}^k + \widehat{B}\hat{x}^{k+1}. \tag{12}$$

One should note that the definition implies that $\widehat{A}, \widehat{C}$ are doubly stochastic communication matrices under appropriate selections of α, ρ. Finally, thanks to the introduction of moving-average iteration of (12), we can obtain the framework (6) which serves as the foundation for our algorithm design. See more details in Section 2.3.

B.2 Specific instances

Relation to some existing single-level algorithm frameworks According to (12), our framework at single-level is

$$\mathbf{x}^{k+1} = \mathbf{C}\mathbf{x}^k - \alpha\mathbf{A}\mathbf{g}^k - \mathbf{B}\mathbf{d}^k, \mathbf{d}^{k+1} = \mathbf{d}^k + \mathbf{B}\mathbf{x}^{k+1}, k = 0, 1, ... \tag{13}$$

where α is the step-size, $\mathbf{g}^k$ denotes the estimated gradient at the k-th iteration, $\mathbf{d}$ serves as the dual variable.

Replacing $\mathbf{C}$ with $\mathbf{CA}$, we get UDA[1] , and equivalently, SUDA [2]:

$$\mathbf{x}^{k+1} = \mathbf{C}\mathbf{A}\mathbf{x}^k - \alpha\mathbf{A}\mathbf{g}^k - \mathbf{B}\mathbf{d}^k, \mathbf{d}^{k+1} = \mathbf{d}^k + \mathbf{B}\mathbf{x}^{k+1}, k = 0, 1, ...$$

Therefore, following SUDA, we can also recover some common state-of-the-art heterogeneity methods as follows by selecting specific $\mathbf{A}, \mathbf{B}, \mathbf{C}$. First, from (13) we get

$$\begin{aligned}\mathbf{x}^{k+2} - \mathbf{x}^{k+1} &= \mathbf{C}(\mathbf{x}^{k+1} - \mathbf{x}^k) - \alpha\mathbf{A}\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right) - \mathbf{B}\left(\mathbf{d}^{k+1} - \mathbf{d}^k\right)\\ &= \mathbf{C}\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right) - \alpha\mathbf{A}\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right) - \mathbf{B}^2\mathbf{x}^{k+1}.\end{aligned}$$

Thus, for $k \geq 0$ we have

$$\mathbf{x}^{k+2} = \left(\mathbf{I} - \mathbf{B}^2 + \mathbf{C}\right)\mathbf{x}^{k+1} - \mathbf{C}\mathbf{x}^k - \alpha\mathbf{A}\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right),$$

with $\mathbf{x}^1 = \mathbf{C}\mathbf{x}^0 - \alpha\mathbf{A}\mathbf{g}^0$.

Some specific instances We next show that how to choose $\mathbf{A}, \mathbf{B}, \mathbf{C}$ to get some common heterogeneity methods.

- ED: Taking $\mathbf{A} = \mathbf{W}, \mathbf{B} = (\mathbf{I} - \mathbf{W})^{1/2}$ and $\mathbf{C} = \mathbf{W}$, we get ED:

$$\mathbf{x}^{k+2} = \mathbf{W}\left(2\mathbf{x}^{k+1} - \mathbf{x}^k - \alpha\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\right),$$

  with $\mathbf{x}^1 = \mathbf{W}\left(\mathbf{x}^0 - \alpha\mathbf{g}^0\right)$.

- EXTRA: Taking $\mathbf{A}=\mathbf{I},\mathbf{B}=(\mathbf{I}-\mathbf{W})^{1/2}$ with $\mathbf{C}=\mathbf{W}$, we get EXTRA:
$$\mathbf{x}^{k+2}=\mathbf{W}\left(2\mathbf{x}^{k+1}-\mathbf{x}^{k}\right)-\alpha\left(\mathbf{g}^{k+1}-\mathbf{g}^{k}\right),$$
and $\mathbf{x}^1=\mathbf{W}\mathbf{x}^0-\alpha\mathbf{g}^0$.
- Adapt-then-combine gradient tracking (ATC-GT): The iteration of ATC-GT is
$$\mathbf{x}^{k+1}=\mathbf{W}\left(\mathbf{x}^{k}-\alpha\mathbf{h}^{k}\right),\mathbf{h}^{k+1}=\mathbf{W}\left(\mathbf{h}^{k}+\mathbf{g}^{k+1}-\mathbf{g}^{k}\right)$$
with $\mathbf{h}^0=\mathbf{W}\mathbf{g}^0,\mathbf{x}^0=\mathbf{W}\mathbf{x}^0$ ($x_1^0=...=x_n^0$). It follows that for $k\geq 0$
$$\mathbf{x}^{k+2}-\mathbf{W}\mathbf{x}^{k+1}=\mathbf{W}\mathbf{x}^{k+1}-\mathbf{W}^2\mathbf{x}^{k}-\alpha\mathbf{W}\left(\mathbf{h}^{k+1}-\mathbf{W}\mathbf{h}^{k}\right).$$
Then we obtain
$$\mathbf{x}^{k+2}=2\mathbf{W}\mathbf{x}^{k+1}-\mathbf{W}^2\mathbf{x}^{k}-\alpha\mathbf{W}^2\left(\mathbf{g}^{k+1}-\mathbf{g}^{k}\right),$$
with $\mathbf{x}^1=\mathbf{W}^2\mathbf{x}^0-\alpha\mathbf{W}^2\mathbf{g}^0$. Thus, we can take $\mathbf{A}=\mathbf{W^2},\mathbf{B}=(\mathbf{I}-\mathbf{W})^2,\mathbf{C}=\mathbf{W}^2$ to implement ATC-GT.
- Semi-ATC-GT: The iteration of Semi-ATC-GT is
$$\mathbf{x}^{k+1}=\mathbf{W}\left(\mathbf{x}^{k}-\alpha\mathbf{h}^{k}\right),\mathbf{h}^{k+1}=\mathbf{W}\mathbf{h}^{k}+\mathbf{g}^{k+1}-\mathbf{g}^{k}$$
with $\mathbf{h}^0=\mathbf{W}\mathbf{g}^0,\mathbf{x}^0=\mathbf{W}\mathbf{x}^0$ ($x_1^0=...=x_n^0$). Like ATC-GT, we have
$$\mathbf{x}^{k+2}=2\mathbf{W}\mathbf{x}^{k+1}-\mathbf{W}^2\mathbf{x}^{k}-\alpha\mathbf{W}\left(\mathbf{g}^{k+1}-\mathbf{g}^{k}\right),$$
with $\mathbf{x}^1=\mathbf{W}^2\mathbf{x}^0-\alpha\mathbf{W}\mathbf{g}^0$. Thus, we can take $\mathbf{A}=\mathbf{W},\mathbf{B}=(\mathbf{I}-\mathbf{W})^2,\mathbf{C}=\mathbf{W}^2$ to implement semi-ATC-GT.
- Non-ATC-GT: The iteration of Non-ATC-GT is
$$\mathbf{x}^{k+1}=\mathbf{W}\mathbf{x}^{k}-\alpha\mathbf{h}^{k},\mathbf{h}^{k+1}=\mathbf{W}\mathbf{h}^{k}+\mathbf{g}^{k+1}-\mathbf{g}^{k}$$
with $\mathbf{h}^0=\mathbf{W}\mathbf{g}^0,\mathbf{x}^0=\mathbf{W}\mathbf{x}^0$ ($x_1^0=...=x_n^0$). We have
$$\mathbf{x}^{k+2}=2\mathbf{W}\mathbf{x}^{k+1}-\mathbf{W}^2\mathbf{x}^{k}-\alpha\left(\mathbf{g}^{k+1}-\mathbf{g}^{k}\right),$$
with $\mathbf{x}^1=\mathbf{W}^2\mathbf{x}^0-\alpha\mathbf{g}^0$. Thus, we can take $\mathbf{A}=\mathbf{I},\mathbf{B}=(\mathbf{I}-\mathbf{W})^2,\mathbf{C}=\mathbf{W}^2$ to implement Non-ATC-GT.

B.3 Implementation details

Given the update method $\mathbf{L}$, we update the lower-level variable y at the k-th ($k\geq 0$) iteration as follows. For brevity, we define $y_i^{-1}=y_i^0,v_i^{-1}=0,o_i^0=\sum_{j=1}^n(W_y)_{ij}v_j^0$.

$$\begin{cases}y_i^{k+1}=\sum_{j=1}^n(W_y)_{ij}\left(2y_j^k-y_j^{k-1}-\beta_k\left(v_i^k-v_i^{k-1}\right)\right) & \text{if } \mathbf{L}=ED\\ y_i^{k+1}=\sum_{j=1}^n(W_y)_{ij}\left(2y_j^k-y_j^{k-1}\right)-\beta_k\left(v_i^k-v_i^{k-1}\right) & \text{if } \mathbf{L}=EXTRA\\ y_i^{k+1}=\sum_{j=1}^n(W_y)_{ij}\left(y_j^k-\beta_k o_j^k\right),o_i^{k+1}=\sum_{j=1}^n(W_y)_{ij}\left(o_j^k+v_i^{k+1}-v_i^k\right) & \text{if } \mathbf{L}=GT\\ \cdots & \text{others}\end{cases} \tag{14}$$

Similarly, we update the auxiliary variable z at the k-th ($k\geq 0$) iteration as follows. For brevity, we define $z_i^{-1}=z_i^0,p_i^{-1}=0,h_i^0=\sum_{j=1}^n(W_z)_{ij}p_j^0$. Note that we use the same method $\mathbf{L}$ to update z as we do for the lower-level variable y.

$$\begin{cases}z_i^{k+1}=\sum_{j=1}^n(W_z)_{ij}\left(2z_j^k-z_j^{k-1}-\gamma_k\left(p_i^k-p_i^{k-1}\right)\right) & \text{if } \mathbf{L}=ED\\ z_i^{k+1}=\sum_{j=1}^n(W_z)_{ij}\left(2z_j^k-z_j^{k-1}\right)-\gamma_k\left(p_i^k-p_i^{k-1}\right) & \text{if } \mathbf{L}=EXTRA\\ z_i^{k+1}=\sum_{j=1}^n(W_z)_{ij}\left(z_j^k-\gamma_k h_j^k\right),h_i^{k+1}=\sum_{j=1}^n(W_z)_{ij}\left(h_j^k+p_i^{k+1}-p_i^k\right) & \text{if } \mathbf{L}=GT\\ \cdots & \text{others}\end{cases} \tag{15}$$

Given the update method $\mathbf{U}$, we update the upper-level variable x at the k-th ($k\geq 0$) iteration as follows. For brevity, we define $x_i^{-1}=x_i^0,t_i^0=\sum_{j=1}^n(W_y)_{ij}r_j^1$.

$$\begin{cases} x_i^{k+1} = \sum_{j=1}^n (W_x)_{ij} \left(2x_j^k - x_j^{k-1} - \alpha_k \left(r_i^{k+1} - r_i^k\right)\right) & \text{if } \mathbf{U} = ED \\ x_i^{k+1} = \sum_{j=1}^n (W_x)_{ij} \left(2x_j^k - x_j^{k-1}\right) - \alpha_k \left(r_i^{k+1} - r_i^k\right) & \text{if } \mathbf{U} = EXTRA \\ x_i^{k+1} = \sum_{j=1}^n (W_x)_{ij} \left(x_j^k - \alpha_k t_j^k\right), t_i^{k+1} = \sum_{j=1}^n (W_x)_{ij} \left(t_j^k + r_i^{k+2} - r_i^{k+1}\right) & \text{if } \mathbf{U} = GT \\ \cdots & \text{others} \end{cases} \tag{16}$$

Then the practical implementation of SPARKLE with mixed strategies is

Algorithm 2 SPARKLE-**L**-**U**

Require: Initialize $x_i^0 = y_i^0 = z_i^0 = r_i^0 = 0$, step-sizes $\alpha_k, \beta_k, \gamma_k, \theta_k$.
  for $k = 0, 1, \cdots, K-1$, each agent i (in parallel) **do**
    Update y_i^{k+1} according to (14);
    Update z_i^{k+1} according to (15);
    $r_i^{k+1} = (1-\theta_k) r_i^k + \theta_k u_i^k$;
    Update x_i^{k+1} according to (16).
  end for

C Convergence analysis

C.1 Proof of Theorem 1

C.1.1 Notations

We use lowercase letters to represent vectors and uppercase letters to represent matrices. Stacked vectors $[x_1^\top, ..., x_n^\top]^\top$ is denoted by col$\{x_1, ..., x_n\}$ for brevity. We denote a block diagonal matrix with diagonal block $M_i (1 \le i \le l)$ by blkdiag$\{M_1, ..., M_l\}$, and a diagonal matrix with diagonal elements $d_i (1 \le i \le k)$ by diag$\{d_1, ..., d_k\}$. The Kronecker product operator is denoted by $\otimes$. For a variable v, we use v_i^k to represent its components at k-th iteration and i-th agent.

Moreover, we use an overbar *above* an iterator to denote the average over all agents. For example, $\bar{x}^k = \sum_{i=1}^n x_i^k / n$. Upright bold symbols are used to denote stacked vectors or matrices across agents. For example, $\mathbf{x}^k := \text{col}\{x_1^k, ..., x_n^k\}$, $\bar{\mathbf{x}}^k := \text{col}\{\bar{x}^k, ..., \bar{x}^k\}$ (n times), $\mathbf{W}_x := W_x \otimes I_{dim(x)}$. Denote the 2-norm of a matrix by $\|\cdot\|$.

Next, we define following σ-fields which will be used in our convergence analysis:

$$\begin{aligned} \mathcal{F}_k &= \sigma\left(\mathbf{y}^0, \ldots, \mathbf{y}^{k+1}, \mathbf{z}^0, \ldots, \mathbf{z}^{k+1}, \mathbf{x}^0, \ldots, \mathbf{x}^k, \mathbf{r}^0, \ldots, \mathbf{r}^k\right), \\ \mathcal{U}_k &= \sigma\left(\mathbf{y}^0, \ldots, \mathbf{y}^{k+1}, \mathbf{z}^0, \ldots, \mathbf{z}^k, \mathbf{x}^0, \ldots, \mathbf{x}^k, \mathbf{r}^0, \ldots, \mathbf{r}^k\right), \\ \mathcal{G}_k &= \sigma\left(\mathbf{y}^0, \ldots, \mathbf{y}^k, \mathbf{z}^0, \ldots, \mathbf{z}^k, \mathbf{x}^0, \ldots, \mathbf{x}^k, \mathbf{r}^0, \ldots, \mathbf{r}^k\right), \end{aligned}$$

and denote $\mathbb{E}[\cdot|\mathcal{F}_k]$ by $\mathbb{E}_k$, $\mathbb{E}[\cdot|\mathcal{U}_k]$ by $\widetilde{\mathbb{E}}_k$, $\mathbb{E}[\cdot|\mathcal{G}_k]$ by $\widehat{\mathbb{E}}_k$ for brevity.

Define

$$z^\star(x) = \left(\sum_{i=1}^n \nabla_{22}^2 g_i\left(x, y^\star(x)\right)\right)^{-1} \left(\sum_{i=1}^n \nabla_2 f_i\left(x, y^\star(x)\right)\right),$$

Then, for $k = 0, 1, \cdots$, define:

$$z_\star^{k+1} = \left(\sum_{i=1}^n \nabla_{22}^2 g_i\left(\bar{x}^k, y^\star(\bar{x}^k)\right)\right)^{-1} \left(\sum_{i=1}^n \nabla_2 f_i\left(\bar{x}^k, y^\star(\bar{x}^k)\right)\right).$$

For convenience, we define $\mathbf{x}^{-1} = \mathbf{x}^0, \mathbf{y}^{-1} = \mathbf{y}^0, y^\star(\bar{x}^{-1}) = y^\star(\bar{x}^0), z_\star^0 = z_\star^1$.

C.1.2 Basic transformations

We begin with conducting SUDA-like [2] transformations, which is fundamental of the following proofs.

Firstly, we define $\mathbf{t}^k$ to track the averaged stochastic gradients among agents as follows

$$\begin{aligned}\mathbf{t}_y^k &= \mathbf{B}_y(\mathbf{d}_y^k - \mathbf{B}_y\mathbf{y}^k) + \beta\mathbf{A}_y\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k),\\ \mathbf{t}_z^k &= \mathbf{B}_z(\mathbf{d}_z^k - \mathbf{B}_z\mathbf{z}^k) + \gamma\mathbf{A}_z\mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1}),\\ \mathbf{t}_x^k &= \mathbf{B}_x(\mathbf{d}_x^k - \mathbf{B}_x\mathbf{x}^k) + \alpha\mathbf{A}_x\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k),\end{aligned} \tag{17}$$

where

$$\begin{aligned}\mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1}) &= \mathrm{col}\left\{\nabla_{22}^2 g_i(\bar{x}^k, \bar{y}^{k+1})z_\star^k - \nabla_2 f_i(\bar{x}^k, \bar{y}^{k+1})\right\}_{i=1}^n,\\ \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k) &= \mathrm{col}\left\{\nabla_1 f_i(\bar{x}^k, y^\star(\bar{x}^k)) - \nabla_{12} g_i(\bar{x}^k, y^\star(\bar{x}^k))z_\star^{k+1}\right\}_{i=1}^n.\end{aligned}$$

Then the iteration of $\mathbf{y}, \mathbf{z}, \mathbf{x}$ in Algorithm 1 can be written as:

$$\text{iteration of } \mathbf{y}: \quad \begin{cases}\mathbf{y}^{k+1} = (\mathbf{C}_y - \mathbf{B}_y^2)\mathbf{y}^k - \mathbf{t}_y^k - \beta\mathbf{A}_y\left[\mathbf{v}^k - \nabla_2\mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right],\\ \mathbf{t}_y^{k+1} = \mathbf{t}_y^k + \mathbf{B}_y^2\mathbf{y}^k + \beta\mathbf{A}_y\left[\nabla_2\mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2\mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right],\end{cases} \tag{18}$$

$$\text{iteration of } \mathbf{z}: \quad \begin{cases}\mathbf{z}^{k+1} = (\mathbf{C}_z - \mathbf{B}_z^2)\mathbf{z}^k - \mathbf{t}_z^k - \gamma\mathbf{A}_z\left[\mathbf{p}^k - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right],\\ \mathbf{t}_z^{k+1} = \mathbf{t}_z^k + \mathbf{B}_z^2\mathbf{z}^k + \gamma\mathbf{A}_z\left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right],\end{cases} \tag{19}$$

$$\text{iteration of } \mathbf{x}: \quad \begin{cases}\mathbf{x}^{k+1} = (\mathbf{C}_x - \mathbf{B}_x^2)\mathbf{x}^k - \mathbf{t}_x^k - \alpha\mathbf{A}_x\left[\mathbf{r}^{k+1} - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right],\\ \mathbf{t}_x^{k+1} = \mathbf{t}_x^k + \mathbf{B}_x^2\mathbf{x}^k + \alpha\mathbf{A}_x\left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1}) - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right].\end{cases} \tag{20}$$

Next, we present the transformation of the matrices A, B, C. For a communication matrix W_s for the variable $s \in \{x, y, z\}$ satisfying Assumption 2, there exists an orthogonal matrix U such that:

$$W = U_s\hat{\Lambda}_s U_s^\top = \begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1} & \hat{U}_s\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & \Lambda_s\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}^\top\\ \hat{U}_s^\top\end{bmatrix},$$

where $\Lambda_s = \mathrm{diag}\{\lambda_{si}\}_{i=2}^n$, $\hat{U}_s^\top \in \mathbb{R}^{n\times(n-1)}$ satisfies $\hat{U}_s\hat{U}_s^\top = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $\mathbf{1}_n^\top\hat{U}_s = 0$. Then it follows that:

$$\mathbf{W}_s = \mathbf{U}_s\hat{\mathbf{\Lambda}}_s\mathbf{U}_s^\top = \begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}\otimes I_{dim(s)} & \hat{\mathbf{U}}_s\end{bmatrix}\begin{bmatrix}I_{dim(s)} & 0\\ 0 & \mathbf{\Lambda_s}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}^\top\otimes I_{dim(s)}\\ \hat{\mathbf{U}}_s^\top\end{bmatrix},$$

where $dim(s)$ denotes the dimension of the corresponding variable, $\mathbf{\Lambda}_s = \Lambda_s \otimes I_{dim(s)} \in \mathbb{R}^{d(n-1)\times[dim(s)\cdot(n-1)]}$, $\mathbf{U}_s \in \mathbb{R}^{[dim(s)\cdot n]\times[dim(s)\cdot n]}$ is an orthogonal matrix, and $\hat{\mathbf{U}}_s = \hat{U}_s \otimes I_{dim(s)} \in \mathbb{R}^{[dim(s)\cdot n]\times[dim(s)\cdot(n-1)]}$ satisfies:

$$\hat{\mathbf{U}}_s^\top\hat{\mathbf{U}}_s = I_{dim(s)\cdot(n-1)}, \quad \hat{\mathbf{U}}_s\hat{\mathbf{U}}_s^\top = \left[I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right]\otimes I_{dim(s)}, \quad (\mathbf{1}^\top\otimes I_{dim(s)})\hat{\mathbf{U}}_s = \mathbf{0}.$$

Now we add subscript s for $\mathbf{W}_s$. Then, as $\mathbf{A}_s, \mathbf{B}_s^2, \mathbf{C}_s$ can be expressed as a polynomial of $\mathbf{W}_s$ for $s \in \{x, y, z\}$ according to Assumption 2, we have the orthogonal decomposition:

$$\begin{aligned}\mathbf{A}_s &= \mathbf{U}_s\hat{\mathbf{\Lambda}}_{sa}\mathbf{U}_s^\top = \begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}\otimes I_{\mathrm{dim}(s)} & \hat{\mathbf{U}}_s\end{bmatrix}\begin{bmatrix}I_{\mathrm{dim}(s)} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_{sa}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}^\top\otimes I_{\mathrm{dim}(s)}\\ \hat{\mathbf{U}}_s^\top\end{bmatrix},\\ \mathbf{B}_s^2 &= \mathbf{U}_s\hat{\mathbf{\Lambda}}_{sb}^2\mathbf{U}_s^\top = \begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}\otimes I_{\mathrm{dim}(s)} & \hat{\mathbf{U}}_s\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_{sb}^2\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}^\top\otimes I_{\mathrm{dim}(s)}\\ \hat{\mathbf{U}}_s^\top\end{bmatrix},\\ \mathbf{C}_s &= \mathbf{U}_s\hat{\mathbf{\Lambda}}_{sc}\mathbf{U}_s^\top = \begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}\otimes I_{\mathrm{dim}(s)} & \hat{\mathbf{U}}_s\end{bmatrix}\begin{bmatrix}I_{\mathrm{dim}(s)} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_{sc}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbf{1}^\top\otimes I_{\mathrm{dim}(s)}\\ \hat{\mathbf{U}}_s^\top\end{bmatrix},\end{aligned} \tag{21}$$

where

$$\mathbf{\Lambda}_{sa} = \underbrace{\mathrm{diag}\{\lambda_{sa,i}\}_{i=2}^{n}}_{\Lambda_{sa}} \otimes I_{\mathrm{dim}(s)}, \quad \mathbf{\Lambda}_{sb} = \underbrace{\mathrm{diag}\{\lambda_{sb,i}\}_{i=2}^{n}}_{\Lambda_{sb}} \otimes I_{\mathrm{dim}(s)}, \quad \mathbf{\Lambda}_{sc} = \underbrace{\mathrm{diag}\{\lambda_{sc,i}\}_{i=2}^{n}}_{\Lambda_{sc}} \otimes I_{\mathrm{dim}(s)}.$$

Moreover, each $\mathbf{\Lambda}_{sb}$ is positive definite because of the null space condition in Assumption 2. Then, multiplying both sides of (18), (19) and (20) by $\mathbf{U}_y^\top, \mathbf{U}_z^\top, \mathbf{U}_x^\top$ respectively, we get:

$$\text{iter. of } \mathbf{y}: \begin{cases} \mathbf{U}_y^\top \mathbf{y}^{k+1} = (\hat{\mathbf{\Lambda}}_{yc} - \hat{\mathbf{\Lambda}}_{yb}^2)\mathbf{U}_y^\top \mathbf{y}^k - \mathbf{U}_y^\top \mathbf{t}_y^k - \beta \hat{\mathbf{\Lambda}}_{ya} \mathbf{U}_y^\top \left[\mathbf{v}^k - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right], \\ \mathbf{U}_y^\top \mathbf{t}_y^{k+1} = \mathbf{U}_y^\top \mathbf{t}_y^k + \hat{\mathbf{\Lambda}}_{yb}^2 \mathbf{U}_y^\top \mathbf{y}^k + \beta \hat{\mathbf{\Lambda}}_{ya} \mathbf{U}_y^\top \left[\nabla_2 \mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right], \end{cases} \tag{22}$$

$$\text{iter. of } \mathbf{z}: \begin{cases} \mathbf{U}_z^\top \mathbf{z}^{k+1} = (\hat{\mathbf{\Lambda}}_{zc} - \hat{\mathbf{\Lambda}}_{zb}^2)\mathbf{U}_z^\top \mathbf{z}^k - \mathbf{U}_z^\top \mathbf{t}_z^k - \gamma \hat{\mathbf{\Lambda}}_{za} \mathbf{U}_z^\top \left[\mathbf{p}^k - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right], \\ \mathbf{U}_z^\top \mathbf{t}_z^{k+1} = \mathbf{U}_z^\top \mathbf{t}_z^k + \hat{\mathbf{\Lambda}}_{zb}^2 \mathbf{U}_z^\top \mathbf{z}^k + \gamma \hat{\mathbf{\Lambda}}_{za} \mathbf{U}_z^\top \left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right]. \end{cases} \tag{23}$$

$$\text{iter. of } \mathbf{x}: \begin{cases} \mathbf{U}_x^\top \mathbf{x}^{k+1} = (\hat{\mathbf{\Lambda}}_{xc} - \hat{\mathbf{\Lambda}}_{xb}^2)\mathbf{U}_x^\top \mathbf{x}^k - \mathbf{U}_x^\top \mathbf{t}_x^k - \alpha \hat{\mathbf{\Lambda}}_{xa} \mathbf{U}^\top \left[\mathbf{r}^{k+1} - \widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^k)\right], \\ \mathbf{U}_x^\top \mathbf{t}_x^{k+1} = \mathbf{U}_x^\top \mathbf{t}_x^k + \hat{\mathbf{\Lambda}}_{xb}^2 \mathbf{U}_x^\top \mathbf{x}^k + \alpha \hat{\mathbf{\Lambda}}_{xa} \mathbf{U}_x^\top \left[\widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^{k+1}) - \widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]. \end{cases} \tag{24}$$

Then, due to Eq. (17), we have:

$$\begin{aligned} (\mathbf{1}^\top \otimes I_d)\mathbf{t}_y^k &= (\mathbf{1}^\top \otimes I_d)\left(\mathbf{B}_y(\mathbf{d}_y^k - \mathbf{B}_y \mathbf{y}^k) + \beta \mathbf{A}_y \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right) \\ &= n\beta \nabla_2 g(\bar{x}^k, \bar{y}^k). \end{aligned} \tag{25}$$

$$\begin{aligned} (\mathbf{1}^\top \otimes I_d)\mathbf{t}_z^k &= (\mathbf{1}^\top \otimes I_d)\left(\mathbf{B}_z(\mathbf{d}_z^k - \mathbf{B}_z \mathbf{z}^k) + \gamma \mathbf{A}_z \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right) \\ &= \gamma \sum_{i=1}^{n} \left[\nabla_{22}^2 g_i(\bar{x}^k, \bar{y}^{k+1}) z_\star^k - \nabla_2 f_i(\bar{x}^k, \bar{y}^{k+1})\right]. \end{aligned} \tag{26}$$

$$\begin{aligned} (\mathbf{1}^\top \otimes I_d)\mathbf{t}_x^k &= (\mathbf{1}^\top \otimes I_d)\left(\mathbf{B}_x(\mathbf{d}_x^k - \mathbf{B}_x \mathbf{x}^k) + \alpha \mathbf{A}_x \widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^k)\right) \\ &= \alpha \sum_{i=1}^{n} \left[\nabla_1 f_i(\bar{x}^k, y^\star(\bar{x}^k)) - \nabla_{12} g_i(\bar{x}^k, y^\star(\bar{x}^k)) z_\star^{k+1}\right]. \end{aligned} \tag{27}$$

Substituting (25), (26), (27) into (22), (23), (24), respectively. Then use (21) and the structure of $\hat{\mathbf{U}}_y, \hat{\mathbf{U}}_z, \hat{\mathbf{U}}_x$, we have

$$\text{iter. of } \mathbf{y}: \begin{cases} \bar{y}^{k+1} = \bar{y}^k - \beta \bar{v}^k, \\ \hat{\mathbf{U}}_y^\top \mathbf{y}^{k+1} = (\mathbf{\Lambda}_{yc} - \mathbf{\Lambda}_{yb}^2)\hat{\mathbf{U}}_y^\top \mathbf{y}^k - \hat{\mathbf{U}}_y^\top \mathbf{t}_y^k - \beta \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\mathbf{v}^k - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right], \\ \hat{\mathbf{U}}_y^\top \mathbf{t}_y^{k+1} = \hat{\mathbf{U}}_y^\top \mathbf{t}_y^k + \mathbf{\Lambda}_{yb}^2 \hat{\mathbf{U}}_y^\top \mathbf{y}^k + \beta \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\nabla_2 \mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right], \end{cases}$$

$$\text{iter. of } \mathbf{z}: \begin{cases} \bar{z}^{k+1} = \bar{z}^k - \gamma \bar{p}^k, \\ \hat{\mathbf{U}}_z^\top \mathbf{z}^{k+1} = (\mathbf{\Lambda}_{zc} - \mathbf{\Lambda}_{zb}^2)\hat{\mathbf{U}}_z^\top \mathbf{z}^k - \hat{\mathbf{U}}_z^\top \mathbf{t}_z^k - \gamma \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\mathbf{p}^k - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right], \\ \hat{\mathbf{U}}_z^\top \mathbf{t}_z^{k+1} = \hat{\mathbf{U}}_z^\top \mathbf{t}_z^k + \mathbf{\Lambda}_{zb}^2 \hat{\mathbf{U}}_z^\top \mathbf{z}^k + \gamma \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right], \end{cases}$$

$$\text{iter. of } \mathbf{x}: \begin{cases} \bar{x}^{k+1} = \bar{x}^k - \alpha \bar{r}^{k+1}, \\ \hat{\mathbf{U}}_x^\top \mathbf{x}^{k+1} = (\hat{\mathbf{\Lambda}}_{xc} - \hat{\mathbf{\Lambda}}_{xb}^2)\hat{\mathbf{U}}_x^\top \mathbf{x}^k - \hat{\mathbf{U}}_x^\top \mathbf{t}_x^k - \alpha \hat{\mathbf{\Lambda}}_{xa} \hat{\mathbf{U}}_x^\top \left[\mathbf{r}^{k+1} - \widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^k)\right], \\ \hat{\mathbf{U}}_x^\top \mathbf{t}_x^{k+1} = \hat{\mathbf{U}}_x^\top \mathbf{t}_x^k + \hat{\mathbf{\Lambda}}_{xb}^2 \hat{\mathbf{U}}_x^\top \mathbf{x}^k + \alpha \hat{\mathbf{\Lambda}}_{xa} \hat{\mathbf{U}}_x^\top \left[\widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^{k+1}) - \widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]. \end{cases}$$

The above three equations are equivalent to:

$$\begin{aligned}\begin{bmatrix} \hat{\mathbf{U}}_y^\top \mathbf{y}^{k+1} \\ \mathbf{\Lambda}_{yb}^{-1}\hat{\mathbf{U}}_y^\top \mathbf{t}_y^{k+1} \end{bmatrix} =& \begin{bmatrix} \mathbf{\Lambda}_{yc}-\mathbf{\Lambda}_{yb}^2 & -\mathbf{\Lambda}_{yb} \\ \mathbf{\Lambda}_{yb} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{U}}_y^\top \mathbf{y}^{k} \\ \mathbf{\Lambda}_{yb}^{-1}\hat{\mathbf{U}}_y^\top \mathbf{t}_k^{y} \end{bmatrix} \\ & - \beta \begin{bmatrix} \mathbf{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top \left[\mathbf{v}^k - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right] \\ \mathbf{\Lambda}_{yb}^{-1}\mathbf{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top \left[\nabla_2 \mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right] \end{bmatrix},\end{aligned} \tag{28}$$

$$\begin{aligned}\begin{bmatrix} \hat{\mathbf{U}}_z^\top \mathbf{z}^{k+1} \\ \mathbf{\Lambda}_{zb}^{-1}\hat{\mathbf{U}}_z^\top \mathbf{t}_z^{k+1} \end{bmatrix} =& \begin{bmatrix} \mathbf{\Lambda}_{zc}-\mathbf{\Lambda}_{zb}^2 & -\mathbf{\Lambda}_{zb} \\ \mathbf{\Lambda}_{zb} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{U}}_z^\top \mathbf{z}^{k} \\ \mathbf{\Lambda}_{zb}^{-1}\hat{\mathbf{U}}_z^\top \mathbf{t}_k^{z} \end{bmatrix} \\ & - \gamma \begin{bmatrix} \mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top \left[\mathbf{p}^k - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right] \\ \mathbf{\Lambda}_{zb}^{-1}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top \left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right] \end{bmatrix},\end{aligned} \tag{29}$$

$$\begin{aligned}\begin{bmatrix} \hat{\mathbf{U}}_x^\top \mathbf{x}^{k+1} \\ \mathbf{\Lambda}_{xb}^{-1}\hat{\mathbf{U}}_x^\top \mathbf{t}_x^{k+1} \end{bmatrix} =& \begin{bmatrix} \mathbf{\Lambda}_{xc}-\mathbf{\Lambda}_{xb}^2 & -\mathbf{\Lambda}_{xb} \\ \mathbf{\Lambda}_{xb} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{U}}_x^\top \mathbf{x}^{k} \\ \mathbf{\Lambda}_{xb}^{-1}\hat{\mathbf{U}}_x^\top \mathbf{t}_k^{x} \end{bmatrix} \\ & - \alpha \begin{bmatrix} \mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top \left[\mathbf{r}^{k+1} - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right] \\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top \left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1}) - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right] \end{bmatrix}.\end{aligned} \tag{30}$$

For $\mathbf{s} \in \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$, define:

$$\mathbf{e}_s^k = \begin{bmatrix} \hat{\mathbf{U}}_s^\top \mathbf{s}^k \\ \mathbf{\Lambda}_{sb}^{-1}\hat{\mathbf{U}}_s^\top \mathbf{t}_s^k \end{bmatrix}, \quad \mathbf{M}_s = \begin{bmatrix} \mathbf{\Lambda}_{sc}-\mathbf{\Lambda}_{sb}^2 & -\mathbf{\Lambda}_{sb} \\ \mathbf{\Lambda}_{sb} & \mathbf{I} \end{bmatrix}.$$

Then (28), (29), (30) are respectively equivalent to:

$$\begin{aligned}\mathbf{e}_y^{k+1} &= \mathbf{M}_y \mathbf{e}_y^k - \beta \begin{bmatrix} \mathbf{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top \left[\mathbf{v}^k - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right] \\ \mathbf{\Lambda}_{yb}^{-1}\mathbf{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top \left[\nabla_2 \mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right] \end{bmatrix}, \\ \mathbf{e}_z^{k+1} &= \mathbf{M}_z \mathbf{e}_z^k - \gamma \begin{bmatrix} \mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top \left[\mathbf{p}^k - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right] \\ \mathbf{\Lambda}_{zb}^{-1}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top \left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right] \end{bmatrix}, \\ \mathbf{e}_x^{k+1} &= \mathbf{M}_x \mathbf{e}_k^x - \alpha \begin{bmatrix} \mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top \left[\mathbf{r}^{k+1} - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right] \\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top \left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1}) - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right] \end{bmatrix}.\end{aligned}$$

Assumption 2, 3 imply that all eigenvalues of

$$\begin{aligned}&\begin{bmatrix} \mathrm{diag}\{0, \Lambda_{sc}-\Lambda_{sb}^2\} & -\mathrm{diag}\{0, \Lambda_{sb}\} \\ \mathrm{diag}\{0, \Lambda_{sb}\} & \mathrm{diag}\{0, 1, ..., 1\} \end{bmatrix} \\ &= \begin{bmatrix} U_s^\top & \\ & U_s^\top \end{bmatrix} \begin{bmatrix} C_s - \frac{1}{n}1_n 1_n^\top - B_s^2 & -B_s \\ B_s & I_n - \frac{1}{n}1_n 1_n^\top \end{bmatrix} \begin{bmatrix} U_s & \\ & U_s \end{bmatrix}\end{aligned} \tag{31}$$

are strictly less than one in magnitude. Thus by symmetrically exchanging columns and rows of the matrix, we know that equivalently, all eigenvalues of

$$\begin{bmatrix} \Lambda_{sc}-\Lambda_{sb}^2 & -\Lambda_{sb} \\ \Lambda_{sb} & I_{n-1} \end{bmatrix} \text{ and } \mathbf{M}_s = \begin{bmatrix} \mathbf{\Lambda}_{sc}-\mathbf{\Lambda}_{sb}^2 & -\mathbf{\Lambda}_{sb} \\ \mathbf{\Lambda}_{sb} & I_{n-1} \otimes I_{dim(s)} \end{bmatrix} \tag{32}$$

are strictly less than one in magnitude,.

Then according to Lemma 3, for $s \in \{x, y, z\}$, $\mathbf{M}_s$ has the similarity transformation:

$$\mathbf{M}_s = \mathbf{O}_s \mathbf{\Gamma}_s \mathbf{O}_s^{-1},$$

where $\mathbf{O}_s$ is invertible and $\|\mathbf{\Gamma}_s\| < 1$. Moreover, we define $\hat{\mathbf{e}}_s^k = \mathbf{O}_s^{-1}\mathbf{e}_s^k$. It yields

$$\hat{\mathbf{e}}_y^{k+1} = \mathbf{\Gamma}_y \hat{\mathbf{e}}_y^k - \beta \mathbf{O}_y^{-1} \begin{bmatrix} \mathbf{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top \left[\mathbf{v}^k - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right] \\ \mathbf{\Lambda}_{yb}^{-1}\mathbf{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top \left[\nabla_2 \mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k)\right] \end{bmatrix}, \tag{33}$$

$$\hat{\mathbf{e}}_z^{k+1} = \mathbf{\Gamma}_y \hat{\mathbf{e}}_z^k - \gamma \mathbf{O}_z^{-1} \begin{bmatrix} \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\mathbf{p}^k - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right] \\ \mathbf{\Lambda}_{zb}^{-1} \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\right] \end{bmatrix}, \tag{34}$$

$$\hat{\mathbf{e}}_x^{k+1} = \mathbf{\Gamma}_x \hat{\mathbf{e}}_x^k - \alpha \mathbf{O}_x^{-1} \begin{bmatrix} \mathbf{\Lambda}_{xa} \hat{\mathbf{U}}_x^\top \left[\mathbf{r}^{k+1} - \widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^k)\right] \\ \mathbf{\Lambda}_{xb}^{-1} \mathbf{\Lambda}_{xa} \hat{\mathbf{U}}_x^\top \left[\widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^{k+1}) - \widetilde{\nabla} \mathbf{\Phi}(\bar{\mathbf{x}}^k)\right] \end{bmatrix}. \tag{35}$$

Then, for $\mathbf{s} \in \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$, the consensus errors between different agents have the upper bound of:

$$\left\|\mathbf{s}^k - \bar{\mathbf{s}}^k\right\|^2 = \|\hat{\mathbf{U}}_s^\top \mathbf{s}^k\|^2 \leq \|\mathbf{e}_s^k\|^2 \leq \|\mathbf{O}_s\|^2 \|\hat{\mathbf{e}}_s^k\|^2. \tag{36}$$

Thus, we can define:

$$\Delta_k = \kappa^2 \|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2 + \kappa^2 \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^{k+1}\|^2 + \|\mathbf{O}_z\|^2 \|\hat{\mathbf{e}}_z^{k+1}\|^2$$

to measure the consensus error during the iteration.

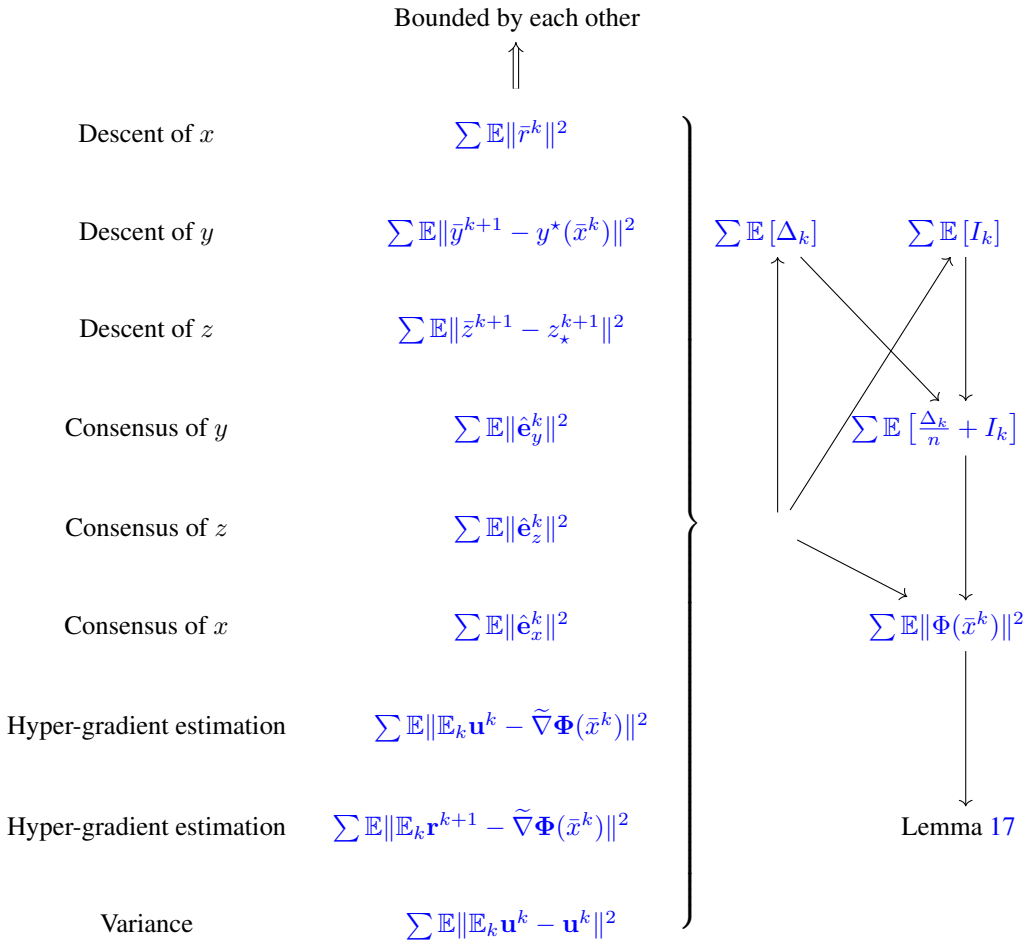
We also define

$$I_k = \|\bar{z}^{k+1} - z_\star^{k+1}\|^2 + \kappa^2 \|\bar{y}^{k+1} - y^\star(\bar{x}^k)\|^2,$$

to measure the estimation accuracy of the lower- and auxiliary-level problems.

C.1.3 Proof sketch

Before proceeding with the formal proof, we first present the structure of the proof in Appendix C.



C.1.4 Technical lemmas

Lemma 1. *Suppose Assumptions 1 hold, we know $\nabla\Phi(x)$,$\widetilde{\nabla}\mathbf{\Phi}(x)$, $z^\star(x)$ and $y^\star(x)$ defined above are $L_{\nabla\Phi}$, $\widetilde{L}$, $L_{z^\star}$, $L_{y^\star}$- Lipschitz continuous respectively with the constants satisfying:*

$$
\begin{aligned}
L_{\nabla\Phi} &\leq L_{f,1} + \frac{2L_{f,1}L_{g,1} + L_{g,2}L_{f,0}}{\mu_g} + \frac{2L_{g,1}L_{f,0}L_{g,2} + L_{g,1}^2 L_{f,1}}{\mu_g^2} + \frac{L_{g,2}L_{g,1}^2 L_{f,0}}{\mu_g^3}, \\
\widetilde{L} &\leq L_{f,1} + \frac{2L_{f,1}L_{g,1} + L_{g,2}L_{f,0}}{\mu_g} + \frac{2L_{g,1}L_{f,0}L_{g,2} + L_{g,1}^2 L_{f,1}}{\mu_g^2} + \frac{L_{g,2}L_{g,1}^2 L_{f,0}}{\mu_g^3}, \\
L_{y^\star} &\leq \frac{L_{g,1}}{\mu_g}, \\
L_{z^\star} &\leq \sqrt{1+L_{y^\star}^2}\left(\frac{L_{f,1}}{\mu_g} + \frac{L_{f,0}L_{g,2}}{\mu_g^2}\right).
\end{aligned}
$$

And we also have:

$$
\|z^\star(x)\| \leq \frac{L_{f,0}}{\mu_g}, \quad \forall x \in \mathbb{R}^p.
$$

Proof. See Lemma 2.2 in [22] and Lemma B.2 in [11]. □

Lemma 2. *Suppose that $g(x)$ is μ-strongly convex and L-smooth. Then for any x and $0 < \alpha < \frac{2}{\mu+L}$, we have*

$$
\|x - \alpha\nabla g(x) - x^\star\| \leq (1 - \alpha\mu)\|x - x^\star\|,
$$

where $x^\star = \operatorname{argmin} g(x)$.

Proof. See Lemma 10 in [41]. □

Lemma 3. *Given diagonal matrices $A, B, C, D \in \mathbb{R}^{(n-1)\times(n-1)}$, and*

$$
\mathbf{M} = \begin{bmatrix} A \otimes I_d & B \otimes I_d \\ C \otimes I_d & D \otimes I_d \end{bmatrix}.
$$

Suppose that the eigenvalues of $\mathbf{M}$ are strictly less than one in magnitude. Then there exist an invertible matrix $\mathbf{O}$ and a matrix $\mathbf{\Gamma}$ with $\|\mathbf{\Gamma}\| < 1$, such that $\mathbf{M}$ has the similarity transformation:

$$
\mathbf{M} = \mathbf{O}\mathbf{\Gamma}\mathbf{O}^{-1}.
$$

Proof. See Lemma 1 in [2]. □

Remark 5. *Asserting the existence of $\mathbf{\Gamma}$ with $\|\mathbf{\Gamma}\| < 1$, Lemma 3 only guarantees the convergence of* SPARKLE. *However, to obtain a precise non-asymptotic convergence rate, one must construct appropriate $\mathbf{O}$ and $\mathbf{\Gamma}$. See more details in Appendix C.2.2.*

C.1.5 Descent lemmas for the upper-level

In this subsection, we estimate the upper bound of the errors induced by the moving average in hyper-gradient estimation, as well as the upper bound of $\|\nabla\Phi(x)\|^2$ based on I_k, Δ_k.

Lemma 4. *Suppose Assumptions 1- 4 hold. We have:*

$$
\begin{aligned}
\left\|\mathbb{E}_k\bar{u}^k - \nabla\Phi(\bar{x}^k)\right\|^2 \leq& \frac{20}{n}L^2(\Delta_k + nI_k), \\
\left\|\mathbb{E}_k\mathbf{u}^k - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2 \leq& 20L^2(\Delta_k + nI_k).
\end{aligned} \tag{38}
$$

Proof. Cauchy Schwartz inequality implies that:

$$
\begin{aligned}
&\left\|\mathbb{E}_k \mathbf{u}^k - \widetilde{\nabla}\mathbf{\Phi}\left(\bar{\mathbf{x}}^k\right)\right\|^2 \\
\leq & 5\sum_{i=1}^n \left\|\nabla_1 f_i\left(x_i^k, y_i^{k+1}\right) - \nabla_1 f_i\left(\bar{x}^k, \bar{y}^{k+1}\right)\right\|^2 + 5\sum_{i=1}^n \left\|\nabla_1 f_i\left(\bar{x}^k, \bar{y}^{k+1}\right) - \nabla_1 f_i\left(\bar{x}^k, y^\star(\bar{x}^k)\right)\right\|^2 \\
&+ 5\sum_{i=1}^n \left\|\nabla_{12}^2 g_i\left(x_i^k, y_i^{k+1}\right)\left(z_i^{k+1} - z_\star^{k+1}\right)\right\|^2 \\
&+ 5\sum_{i=1}^n \left\|\left(\nabla_{12}^2 g_i\left(x_i^k, y_i^{k+1}\right) - \nabla_{12}^2 g_i\left(\bar{x}^k, \bar{y}^{k+1}\right)\right) z_\star^{k+1}\right\|^2 \\
&+ 5\sum_{i=1}^n \left\|\left(\nabla_{12}^2 g_i\left(\bar{x}^k, \bar{y}^{k+1}\right) - \nabla_{12}^2 g_i\left(\bar{x}^k, y^\star(\bar{x}^k)\right)\right) z_\star^{k+1}\right\|^2 \\
\leq & 10\left(L_{f,1}^2 + \kappa^2 L_{f,0}^2\right)\left(\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2 + \left\|\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1}\right\|^2 + \left\|\bar{\mathbf{y}}^{k+1} - \mathbf{y}^\star(\bar{\mathbf{x}}^k)\right\|^2\right) \\
&+ 10L_{g,1}^2\left(\left\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^{k+1}\right\|^2 + \left\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}_\star^{k+1}\right\|^2\right) \\
\leq & 20L^2\left(\Delta_k + nI_k\right).
\end{aligned}
$$

For the term $\left\|\mathbb{E}_k \bar{u}^k - \nabla\Phi\left(\bar{x}^k\right)\right\|^2$, we have:

$$
\left\|\mathbb{E}_k \bar{u}^k - \nabla\Phi(\bar{x}^k)\right\|^2 \leq \frac{1}{n}\left\|\mathbb{E}_k \mathbf{u}^k - \widetilde{\nabla}\mathbf{\Phi}\left(\bar{\mathbf{x}}^k\right)\right\|^2 \leq 20L^2\left(\frac{\Delta_k}{n} + I_k\right).
$$

□

Lemma 5. *Suppose that Assumptions 1- 4 hold. We have*

$$
\begin{aligned}
&n^2\sum_{k=0}^K \mathbb{E}\left[\|\bar{u}^k - \mathbb{E}_k[\bar{u}^k]\|^2\right] = \sum_{k=0}^K \mathbb{E}\left[\|\mathbf{u}^k - \mathbb{E}_k[\mathbf{u}^k]\|^2\right] \\
\leq & 9\sigma_{g,2}^2\sum_{k=0}^K\left(\mathbb{E}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^{k+1}\|^2 + \mathbb{E}\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}_\star^{k+1}\|^2\right) + 3(K+1)n\left(\sigma_{f,1}^2 + 3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right).
\end{aligned} \tag{39}
$$

Proof. For $k \geq 0$, Cauchy Schwartz inequality implies that

$$
\begin{aligned}
&\frac{1}{3}\mathbb{E}_k\left[\|\mathbf{u}^k - \mathbb{E}_k[\mathbf{u}^k]\|^2\right] \\
\leq & \mathbb{E}_k\left[\sum_{i=1}^n \|\nabla_1 f_i(x_i^k, y_i^{k+1}, \xi_i^k) - \nabla_1 f_i(x_i^k, y_i^{k+1})\|^2\right] \\
&+ \mathbb{E}_k\left[\sum_{i=1}^n \left\|\left(\nabla_{12} g_i(x_i^k, y_i^{k+1}, \zeta_i^k) - \nabla_{12} g_i(x_i^k, y_i^{k+1})\right) z_i^{k+1}\right\|^2\right] \\
\leq & n\sigma_{f,1}^2 + \sigma_{g,2}^2\|\mathbf{z}^{k+1}\|^2 \\
\leq & n\sigma_{f,1}^2 + 3\sigma_{g,2}^2\left(\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^{k+1}\|^2 + \|\bar{\mathbf{z}}^{k+1} - \mathbf{z}_\star^{k+1}\|^2 + \|\mathbf{z}_\star^{k+1}\|^2\right) \\
\leq & n\sigma_{f,1}^2 + 3\sigma_{g,2}^2\left(\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^{k+1}\|^2 + \|\bar{\mathbf{z}}^{k+1} - \mathbf{z}_\star^{k+1}\|^2 + n\frac{L_{f,0}^2}{\mu_g^2}\right).
\end{aligned}
$$

Then taking expectation and summation on both sides, we get

$$
\begin{aligned}
&\sum_{k=0}^K \mathbb{E}\left[\|\mathbf{u}^k - \mathbb{E}_k[\mathbf{u}^k]\|^2\right] \\
\leq & 9\sigma_{g,2}^2\sum_{k=0}^K\left(\mathbb{E}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^{k+1}\|^2 + \mathbb{E}\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}_\star^{k+1}\|^2\right) + 3(K+1)n\left(\sigma_{f,1}^2 + 3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right).
\end{aligned}
$$

Since samples among agents are independent, it follows that

$$\sum_{k=0}^{K}\mathbb{E}_k\left[\|\bar{u}^k-\mathbb{E}_k[\bar{u}^k]\|^2\right]=\frac{1}{n^2}\sum_{k=0}^{K}\mathbb{E}_k\left[\|\mathbf{u}^k-\mathbb{E}_k[\mathbf{u}^k]\|^2\right].$$

Taking expectations, we get the conclusion. □

Lemma 6. *Suppose that Assumptions 1- 4, and Lemmas 4, 5 hold. If*

$$\alpha\le\frac{1}{2L_{\nabla\Phi}},\tag{40}$$

we have

$$\begin{aligned}&\frac{1}{4}\sum_{k=0}^{K}\mathbb{E}\left\|\bar{r}^{k+1}\right\|^2\\ \le&\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha}+10\left(L^2+\frac{\theta\sigma_{g,2}^2}{n}\right)\sum_{k=0}^{K}\mathbb{E}\left(\frac{\Delta_k}{n}+I_k\right)+\frac{3\theta}{n}(K+1)\left(\sigma_{f,1}^2+2\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right).\end{aligned}\tag{41}$$

Proof. The $L_{\nabla\Phi}$-smoothness of Φ indicates that

$$\begin{aligned}&\mathbb{E}_k[\Phi\left(\bar{x}^{k+1}\right)]-\Phi(\bar{x}^k)\\ \le&\left\langle\nabla\Phi(\bar{x}^k),\left(-\alpha\mathbb{E}_k[\bar{r}^{k+1}]\right)\right\rangle+\frac{L_{\nabla\Phi}\alpha^2}{2}\mathbb{E}_k\|\bar{r}^{k+1}\|^2\\ =&\left\langle\nabla\Phi(\bar{x}^k)-\mathbb{E}_k[\bar{u}^k],-\alpha\mathbb{E}_k[\bar{r}^{k+1}]\right\rangle+\frac{L_{\nabla\Phi}}{2}\alpha^2\mathbb{E}_k\|\bar{r}^{k+1}\|^2-\alpha\left\langle\mathbb{E}_k[\bar{u}^k],\mathbb{E}_k[\bar{r}^{k+1}]\right\rangle.\end{aligned}$$

Then, due to $\mathbb{E}_k[\bar{u}^k]=\theta^{-1}(\mathbb{E}_k[\bar{r}^{k+1}]-(1-\theta)\bar{r}^k)$, we have:

$$\begin{aligned}&\mathbb{E}_k[\Phi\left(\bar{x}^{k+1}\right)]-\Phi(\bar{x}^k)\\ \le&\frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)-\mathbb{E}_k[\bar{u}^k]\|^2+\frac{\alpha}{2}\|\mathbb{E}_k[\bar{r}^{k+1}]\|^2\\ &+\frac{L_{\nabla\Phi}}{2}\alpha^2\mathbb{E}_k\|\bar{r}^{k+1}\|^2-\alpha\left\langle\mathbb{E}_k[\bar{u}^k],\mathbb{E}_k[\bar{r}^{k+1}]\right\rangle\\ =&\frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)-\mathbb{E}_k[\bar{u}^k]\|^2+(-\frac{\alpha}{2}+\frac{L_{\nabla\Phi}}{2}\alpha^2)\mathbb{E}_k\|\bar{r}^{k+1}\|^2-\frac{\alpha(1-\theta)}{2\theta}\|\mathbb{E}_k[\bar{r}^{k+1}]-\bar{r}^k\|^2\\ &+\frac{\alpha(1-\theta)}{2\theta}\left(\|\bar{r}^k\|^2-\mathbb{E}_k\|\bar{r}^{k+1}\|^2\right)+\frac{\alpha}{2\theta}\mathbb{E}_k\|\bar{r}^{k+1}-\mathbb{E}_k[\bar{r}^{k+1}]\|^2\\ \le&\frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)-\mathbb{E}_k[\bar{u}^k]\|^2+(-\frac{\alpha}{2}+\frac{L_{\nabla\Phi}}{2}\alpha^2)\mathbb{E}_k\|\bar{r}^{k+1}\|^2+\frac{\alpha\theta}{2}\mathbb{E}_k\|\bar{u}^k-\mathbb{E}_k[\bar{u}^k]\|^2\\ &+\frac{\alpha(1-\theta)}{2\theta}\left(\|\bar{r}^k\|^2-\mathbb{E}_k\|\bar{r}^{k+1}\|^2\right),\end{aligned}$$

where the first equality uses $2\left\langle\bar{r}^k,\mathbb{E}_k[\bar{r}^{k+1}]\right\rangle=\|\bar{r}^k\|^2+\|\mathbb{E}_k[\bar{r}^{k+1}]\|^2-\|\bar{r}^k-\mathbb{E}_k[\bar{r}^{k+1}]\|^2$ and $\mathbb{E}_k\|\bar{r}^{k+1}\|^2=\|\mathbb{E}_k[\bar{r}^{k+1}]\|^2+\mathbb{E}_k\|\bar{r}^{k+1}-\mathbb{E}_k[\bar{r}^{k+1}]\|^2$.

Taking expectation and summation, and using $\alpha\le\frac{1}{2L_{\nabla\Psi}}$, we get

$$\begin{aligned}&\inf\Phi-\Phi(\bar{x}_0)\\ \le&\frac{\alpha}{2}\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)-\mathbb{E}_k[\bar{u}^k]\|^2-\frac{\alpha}{4}\sum_{k=0}^{K}\mathbb{E}\|\bar{r}^{k+1}\|^2+\frac{\alpha\theta}{2}\sum_{k=0}^{K}\mathbb{E}\left[\mathbb{E}_k\|\bar{u}^k-\mathbb{E}_k[\bar{u}^k]\|^2\right].\end{aligned}\tag{42}$$

Since samples of different agents are independent, we have

$$\mathbb{E}_k\|\bar{u}^k-\mathbb{E}_k[\bar{u}^k]\|^2=\frac{1}{n^2}\mathbb{E}_k\|\mathbf{u}^k-\mathbb{E}_k[\mathbf{u}^k]\|^2.$$

Combining it with the conclusion of Lemma 4 and 5, we get from (42) that

$$
\begin{aligned}
&\frac{\alpha}{4}\sum_{k=0}^{K}\mathbb{E}\left\|\bar{r}^{k+1}\right\|^2 \\
\leq&\Phi(\bar{x}_0)-\inf\Phi+\frac{\alpha}{2}\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)-\mathbb{E}_k[\bar{u}^k]\|^2+\frac{\alpha\theta}{2}\sum_{k=0}^{K}\mathbb{E}\left[\mathbb{E}_k\|\bar{u}^k-\mathbb{E}_k[\bar{u}^k]\|^2\right] \\
\leq&\Phi(\bar{x}_0)-\inf\Phi+10\alpha\left(L^2+\frac{\theta\sigma_{g,2}^2}{n}\right)\sum_{k=0}^{K}\mathbb{E}\left(\frac{\Delta_k}{n}+I_k\right)+\frac{3\alpha\theta}{n}(K+1)\left(\sigma_{f,1}^2+2\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right).
\end{aligned}
$$

□

Lemma 7. *Suppose that Assumptions 1- 4 hold, then we have*

$$
\begin{aligned}
&\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right] \\
\leq&\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2+2\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbb{E}_k[\mathbf{u}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right] \\
&+\frac{2\widetilde{L}^2(1-\theta)^2}{\theta^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\right\|^2\right]+(1-\theta)\theta\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbf{u}^k-\mathbb{E}_k[\mathbf{u}^k]\right\|^2\right].
\end{aligned}
$$

Proof. We define $\mathbf{u}^{-1}=\mathbf{0}$ for brevity. From the definition of $\mathbb{E}_k$, we have :

$$
\begin{aligned}
&\mathbb{E}_{k-1}\left[\left\|\mathbf{r}^k-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})\right\|^2\right] \\
=&\mathbb{E}_{k-1}\left[\left\|\mathbb{E}_{k-1}[\mathbf{r}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})\right\|^2\right]+\mathbb{E}_{k-1}\left[\left\|\mathbf{r}^k-\mathbb{E}_{k-1}[\mathbf{r}^k]\right\|^2\right] \\
=&\mathbb{E}_{k-1}\left[\left\|\mathbb{E}_{k-1}[\mathbf{r}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})\right\|^2\right]+\theta^2\mathbb{E}_{k-1}\left[\left\|\mathbf{u}^{k-1}-\mathbb{E}_{k-1}[\mathbf{u}^{k-1}]\right\|^2\right].
\end{aligned}
\tag{43}
$$

Jensen's inequality implies that

$$
\begin{aligned}
&\mathbb{E}_k\left[\left\|\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right] \\
\leq&(1-\theta)\mathbb{E}_k\left[\left\|\mathbf{r}^k-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})\right\|^2\right] \\
&+\theta\mathbb{E}_k\left[\left\|\left(\mathbb{E}_k[\mathbf{u}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right)+\theta^{-1}(1-\theta)\left(\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right)\right\|^2\right] \\
\leq&(1-\theta)\mathbb{E}_k\left[\left\|\mathbf{r}^k-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})\right\|^2\right]+2\theta\mathbb{E}_k\left[\left\|\mathbb{E}_k[\mathbf{u}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right] \\
&+\frac{2(1-\theta)^2}{\theta}\mathbb{E}_k\left[\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right].
\end{aligned}
\tag{44}
$$

Substituting (43) into (44) , and taking expectation and summation on both sides, we get:

$$
\begin{aligned}
&\theta\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbb{E}_{k-1}[\mathbf{r}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})\right\|^2\right] \\
\leq&\mathbb{E}\left[\left\|\mathbf{r}^0-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{-1})\right\|^2\right]-\mathbb{E}\left[\left\|\mathbb{E}_K[\mathbf{r}^{K+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right]+2\theta\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbb{E}_k[\mathbf{u}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right] \\
&+\frac{2(1-\theta)^2}{\theta}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k-1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right]+(1-\theta)\theta^2\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbf{u}^k-\mathbb{E}_k[\mathbf{u}^k]\right\|^2\right].
\end{aligned}
$$

Finally, note that $\mathbf{x}^{-1} = \mathbf{x}^0$, $\mathbf{r}^0 = \mathbf{0}$, and $\mathbb{E}_{-1} = \mathbb{E}_0$. Subtracting $\theta\mathbb{E}\left[\left\|\mathbb{E}_{-1}[\mathbf{r}^0] - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{-1})\right\|^2\right] = \theta\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2$ from both sides of this equation, we get:

$$\begin{aligned}
&\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbb{E}_k[\mathbf{r}^{k+1}] - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right] \\
\leq&\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2 + 2\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbb{E}_k[\mathbf{u}^k] - \widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right] \\
&+ \frac{2\widetilde{L}^2(1-\theta)^2}{\theta^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\right\|^2\right] + (1-\theta)\theta\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbf{u}^k - \mathbb{E}_k[\mathbf{u}^k]\right\|^2\right].
\end{aligned}$$

□

Lemma 8 (Descent lemma)**.** *Suppose that Assumptions 1- 4 and Lemmas 4, 5 hold. If*

$$\frac{\alpha^2}{\theta^2}(1-\theta) \leq \frac{1}{32L_{\nabla\Phi}^2},\ \alpha \leq \frac{1}{10L_{\nabla\Phi}}, \tag{45}$$

then we have

$$\begin{aligned}
\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2 \lesssim&\frac{\Phi(\bar{x}_0) - \inf\Phi}{\alpha} + \left(L^2 + \left(\theta(1-\theta) + L_{\nabla\Phi}\alpha\theta^2\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n} + I_k\right] \\
&+ (K+1)\left(\theta(1-\theta) + L_{\nabla\Phi}\alpha\theta^2\right)(\sigma_{f,1}^2 + \kappa^2\sigma_{g,2}^2) + \frac{(1-\theta)^2}{\theta}\|\nabla\Phi(\bar{x}^0)\|^2.
\end{aligned}$$

Proof. The $L_{\nabla\Phi}$-smoothness of Φ indicates that

$$\begin{aligned}
&\mathbb{E}_k[\Phi(\bar{x}^{k+1})] - \Phi(\bar{x}^k) \\
\leq&\left\langle\nabla\Phi(\bar{x}^k), -\alpha\mathbb{E}_k[\bar{r}^{k+1}]\right\rangle + \frac{L_{\nabla\Phi}\alpha^2}{2}\mathbb{E}_k\left\|\bar{r}^{k+1}\right\|^2 \\
=&-\alpha\left\langle\nabla\Phi(\bar{x}^k), \mathbb{E}_k[\bar{r}^{k+1}] - \nabla\Phi(\bar{x}^k)\right\rangle - \alpha\|\nabla\Phi(\bar{x}^k)\|^2 + \frac{L_{\nabla\Phi}}{2}\alpha^2\mathbb{E}_k\left\|\bar{r}^{k+1}\right\|^2 \\
\leq&-\frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 + \frac{\alpha}{2}\|\mathbb{E}_k[\bar{r}^{k+1}] - \nabla\Phi(\bar{x}^k)\|^2 + \frac{L_{\nabla\Phi}}{2}\alpha^2\mathbb{E}_k\left\|\bar{r}^{k+1}\right\|^2.
\end{aligned}$$

Taking expectation and summation on both sides, we get:

$$\begin{aligned}
&\sum_{k=0}^{K}\alpha\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2 \\
\leq&2(\Phi(\bar{x}^0) - \inf\Phi) + \sum_{k=0}^{K}\alpha\mathbb{E}\|\mathbb{E}_k[\bar{r}^{k+1}] - \nabla\Phi(\bar{x}^k)\|^2 + \sum_{k=0}^{K}L_{\nabla\Phi}\alpha^2\mathbb{E}\left\|\bar{r}^{k+1}\right\|^2.
\end{aligned} \tag{46}$$

Define auxiliary series m^k as:

$$m^0 = \bar{r}^0 = 0, m^{k+1} = (1-\theta)m^k + \theta\nabla\Phi(\bar{x}^k).$$

Note that

$$\begin{aligned}
\mathbb{E}_k\left\|\bar{r}^{k+1}\right\|^2 &= \left\|\mathbb{E}_k\bar{r}^{k+1}\right\|^2 + \mathbb{E}_k\left\|\bar{r}^{k+1} - \mathbb{E}_k\bar{r}^{k+1}\right\|^2 \\
&\leq 2\|\mathbb{E}_k[\bar{r}^{k+1}] - \nabla\Phi(\bar{x}^k)\|^2 + 2\|\nabla\Phi(\bar{x}^k)\|^2 + \theta^2\mathbb{E}_k\|\bar{u}^k - \mathbb{E}_k\bar{u}^k\|^2.
\end{aligned} \tag{47}$$

Then using the Jenson's Inequality, we get:

$$\begin{aligned}
\|\mathbb{E}_k\bar{r}^{k+1} - m^{k+1}\|^2 &= \|(1-\theta)(\bar{r}^k - m^k) + \theta(\mathbb{E}_k\bar{u}^k - \nabla\Phi(\bar{x}^k))\|^2 \\
&\leq (1-\theta)\|\bar{r}^k - m^k\|^2 + \theta\|\mathbb{E}_k\bar{u}^k - \nabla\Phi(\bar{x}^k)\|^2.
\end{aligned}$$

It follows that for $k \geq 0$

$$\begin{aligned}
&\mathbb{E}\|\mathbb{E}_k \bar{r}^{k+1} - m^{k+1}\|^2 \\
\leq &(1-\theta)\mathbb{E}\|\mathbb{E}_{k-1}[\bar{r}^k] - m^k\|^2 + (1-\theta)\theta^2\mathbb{E}\|\bar{u}^{k-1} - \mathbb{E}_{k-1}\bar{u}^{k-1}\|^2 + \theta\mathbb{E}\|\mathbb{E}_k \bar{u}^k - \nabla\Phi(\bar{x}^k)\|^2,
\end{aligned}$$

where for brevity we define $\bar{u}^{-1} = 0$.

Taking the summation on both sides from $k = 0$ to K, we get

$$\begin{aligned}
\sum_{k=0}^{K} \theta\mathbb{E}\|\mathbb{E}_k \bar{r}^{k+1} - m^{k+1}\|^2 &\leq \sum_{k=0}^{K-1} \theta\mathbb{E}\|\mathbb{E}_k \bar{r}^{k+1} - m^{k+1}\|^2 + \mathbb{E}\|\mathbb{E}_K \bar{r}^{K+1} - m^{K+1}\|^2 \\
&\leq \sum_{k=0}^{K} \theta\mathbb{E}\|\mathbb{E}_k \bar{u}^k - \nabla\Phi(\bar{x}^k)\|^2 + \sum_{k=0}^{K-1}(1-\theta)\theta^2\mathbb{E}\|\bar{u}^k - \mathbb{E}_k\bar{u}^k\|^2.
\end{aligned} \tag{48}$$

On the other hand, due to the definition of m^k and Jenson's Inequality, we have:

$$\begin{aligned}
\|m^{k+1} - \nabla\Phi(\bar{x}^k)\|^2 &= \|(1-\theta)(m^k - \nabla\Phi(\bar{x}^k))\|^2 \\
&= (1-\theta)^2\|m^k - \nabla\Phi(\bar{x}^{k-1}) + \nabla\Phi(\bar{x}^{k-1}) - \nabla\Phi(\bar{x}^k)\|^2 \\
&\leq (1-\theta)\|m^k - \nabla\Phi(\bar{x}^{k-1})\|^2 + \frac{(1-\theta)^2}{\theta} L_{\nabla\Phi}^2\alpha^2\|\bar{r}^k\|^2.
\end{aligned}$$

Taking the summation, we get

$$\begin{aligned}
\sum_{k=0}^{K} \theta\|m^{k+1} - \nabla\Phi(\bar{x}^k)\|^2 &\leq \|m^0 - \nabla\Phi(\bar{x}^{-1})\|^2 + \sum_{k=0}^{K} \frac{(1-\theta)^2}{\theta} L_{\nabla\Phi}^2\alpha^2\|\bar{r}^k\|^2 \\
&= (1-\theta)^2\|\nabla\Phi(\bar{x}^0)\|^2 + \sum_{k=0}^{K} \frac{(1-\theta)^2}{\theta} L_{\nabla\Phi}^2\alpha^2\|\bar{r}^k\|^2.
\end{aligned} \tag{49}$$

Combining (48) and (49), we obtain:

$$\begin{aligned}
&\sum_{k=0}^{K} \theta\mathbb{E}\|\mathbb{E}_k \bar{r}^{k+1} - \nabla\Phi(\bar{x}^k)\|^2 \\
\leq &2\sum_{k=0}^{K} \theta\mathbb{E}\|\mathbb{E}_k \bar{r}^{k+1} - m^{k+1}\|^2 + 2\sum_{k=0}^{K} \theta\|m^{k+1} - \nabla\Phi(\bar{x}^k)\|^2 \\
\leq &2\sum_{k=0}^{K} \theta\mathbb{E}\|\mathbb{E}_k \bar{u}^k - \nabla\Phi(\bar{x}^k)\|^2 + 2\sum_{k=0}^{K-1}(1-\theta)\theta^2\mathbb{E}\|\bar{u}^k - \mathbb{E}_k\bar{u}^k\|^2 \\
&+ 2\sum_{k=0}^{K} \frac{(1-\theta)^2}{\theta} L_{\nabla\Phi}^2\alpha^2\mathbb{E}\|\bar{r}^k\|^2 + 2(1-\theta)^2\|\nabla\Phi\left(\bar{x}^0\right)\|^2 \\
\leq &2\sum_{k=0}^{K} \theta\mathbb{E}\|\mathbb{E}_k \bar{u}^k - \nabla\Phi(\bar{x}^k)\|^2 + 2\sum_{k=0}^{K-1}\left(1 + 2\frac{1-\theta}{\theta} L_{\nabla\Phi}^2\alpha^2\right)(1-\theta)\theta^2\mathbb{E}\|\bar{u}^k - \mathbb{E}_k\bar{u}^k\|^2 \\
&+ 2(1-\theta)^2\|\nabla\Phi\left(\bar{x}^0\right)\|^2 + 2\sum_{k=0}^{K-1} \frac{(1-\theta)^2}{\theta} L_{\nabla\Phi}^2\alpha^2\left(2\mathbb{E}\|\mathbb{E}_k[\bar{r}^{k+1}] - \nabla\Phi(\bar{x}^k)\|^2 + 2\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\right),
\end{aligned} \tag{50}$$

where the last inequality uses (47).

(45) indicates that $4\frac{1-\theta}{\theta} L_{\nabla\Phi}^2\alpha^2 \leq \frac{\theta}{8}$. Subtracting

$$2\sum_{k=0}^{K-1} \frac{(1-\theta)^2}{\theta} L_{\nabla\Phi}^2\alpha^2 \cdot 2\mathbb{E}\|\mathbb{E}_k[\bar{r}^{k+1}] - \nabla\Phi(\bar{x}^k)\|^2$$

from both sides of (50), we have:

$$\begin{aligned}
&\sum_{k=0}^{K}\theta\mathbb{E}\|\mathbb{E}_k\bar{r}^{k+1}-\nabla\Phi(\bar{x}^k)\|^2\\
\leq&4\sum_{k=0}^{K}\theta\mathbb{E}\|\mathbb{E}_k\bar{u}^k-\nabla\Phi(\bar{x}^k)\|^2+8\sum_{k=0}^{K-1}(1-\theta)\theta^2\mathbb{E}\|\bar{u}^k-\mathbb{E}_k\bar{u}^k\|^2+4(1-\theta)^2\|\nabla\Phi(\bar{x}^0)\|^2\\
&+\frac{\theta}{4}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2.
\end{aligned}\tag{51}$$

Substituting (47), (51) into (46), we get:

$$\begin{aligned}
&\sum_{k=0}^{K}\alpha\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\\
\leq&2(\Phi(\bar{x}^0)-\inf\Phi)+\sum_{k=0}^{K}(\alpha+2L_{\nabla\Phi}\alpha^2)\mathbb{E}\|\mathbb{E}_k[\bar{r}^{k+1}]-\nabla\Phi(\bar{x}^k)\|^2+\sum_{k=0}^{K}2L_{\nabla\Phi}\alpha^2\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\\
&+\sum_{k=0}^{K}2L_{\nabla\Phi}\alpha^2\theta^2\mathbb{E}_k\|\bar{u}^k-\mathbb{E}_k\bar{u}^k\|^2\\
\leq&2(\Phi(\bar{x}_0)-\inf\Phi)+5\alpha\sum_{k=0}^{K}\mathbb{E}\|\mathbb{E}_k\bar{u}^k-\nabla\Phi(\bar{x}^k)\|^2+5\frac{\alpha}{\theta}(1-\theta)^2\|\nabla\Phi\left(\bar{x}^0\right)\|^2\\
&+\sum_{k=0}^{K}\left(10\frac{\alpha}{\theta}(1-\theta)+2L_{\nabla\Phi}\alpha^2\right)\theta^2\mathbb{E}\|\bar{u}^k-\mathbb{E}_k\bar{u}^k\|^2+\frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2,
\end{aligned}\tag{52}$$

where the last inequality uses $\alpha\leq\frac{1}{10L_{\nabla\Phi}}$.

Subtracting $\frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2$ from both sides of (52), and substituting (38), (39) into it, we get:

$$\begin{aligned}
&\frac{1}{2}\sum_{k=0}^{K}\alpha\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\\
\leq&2(\Phi(\bar{x}^0)-\inf\Phi)+100\alpha L^2\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}+I_k\right]+5\frac{\alpha}{\theta}(1-\theta)^2\|\nabla\Phi\left(\bar{x}^0\right)\|^2\\
&+\frac{\left(10\frac{\alpha}{\theta}(1-\theta)+2L_{\nabla\Phi}\alpha^2\right)}{n^2}\theta^2\cdot9\sigma_{g,2}^2\sum_{k=0}^{K}\left(\mathbb{E}\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\|^2+\mathbb{E}\|\bar{\mathbf{z}}^{k+1}-\mathbf{z}_\star^{k+1}\|^2\right)\\
&+\frac{\left(10\frac{\alpha}{\theta}(1-\theta)+2L_{\nabla\Phi}\alpha^2\right)}{n^2}\theta^2\cdot3(K+1)n\left(\sigma_{f,1}^2+3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
\leq&2(\Phi(\bar{x}_0)-\inf\Phi)+\left(100\alpha L^2+9\left(10\alpha\theta(1-\theta)+2L_{\nabla\Phi}\alpha^2\theta^2\right)\frac{\sigma_{g,2}^2}{n}\right)\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}+I_k\right]\\
&+3(K+1)\left(10\alpha(1-\theta)+2L_{\nabla\Phi}\alpha^2\theta\right)\frac{\theta}{n}\left(\sigma_{f,1}^2+3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
&+5\frac{\alpha}{\theta}(1-\theta)^2\|\nabla\Phi\left(\bar{x}^0\right)\|^2.
\end{aligned}$$

Finally, multiplying $\frac{2}{\alpha}$ on both sides, we get:

$$\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2 \lesssim \frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha}+\left(L^2+\left(\theta(1-\theta)+L_{\nabla\Phi}\alpha\theta^2\right)\frac{\sigma_{g,2}^2}{n}\right)\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}+I_k\right] + \frac{K+1}{n}\left(\theta(1-\theta)+L_{\nabla\Phi}\alpha\theta^2\right)(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{(1-\theta)^2}{\theta}\|\nabla\Phi(\bar{x}^0)\|^2.$$

□

C.1.6 Descent lemmas for the lower- and auxiliary-level

The following lemmas present the error analysis of the estimation of $y^\star(\bar{x}^k)$ and $z_\star^k$, i.e., the term I_k:

Lemma 9 (Estimation error of $y^\star(x)$)**.** *Suppose Assumptions 1- 4hold, and:*

$$\beta \le \frac{\mu_g}{32L_{g,1}^2}. \tag{53}$$

Then we have the estimation error of $y^\star$:

$$\begin{aligned}
&\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\sum_{k=0}^{K}\mathbb{E}[\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2]\\
\le&\frac{4}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\sum_{k=1}^{K}\frac{6\alpha^2L_{y^\star}^2}{\beta^2\mu_g^2}\mathbb{E}\|\bar{r}^k\|^2+\sum_{k=1}^{K}\frac{6}{\mu_g^2}L_{g,1}^2\mathbb{E}\left[\frac{\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^k\|^2}{n}\right]\\
&+\frac{4K\beta\sigma_{g,1}^2}{n\mu_g},
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{y}^{k+1}-\bar{y}^k\|^2\right]\\
\le&\frac{\beta^2L_{g,1}^2}{n}\left(4+\frac{48L_{g,1}^2}{\mu_g^2}\right)\sum_{k=1}^{K}\mathbb{E}\left(\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^k\|^2\right)+\frac{48\alpha^2L_{g,1}^2}{\mu_g^2}L_{y^\star}^2\sum_{k=1}^{K}\mathbb{E}\|\bar{r}^k\|^2\\
&+\frac{3(K+1)\beta^2}{n}\sigma_{g,1}^2+\frac{32\beta L_{g,1}^2}{\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2.
\end{aligned} \tag{54}$$

Proof. For each $k\ge 0$, due to the independence of samples, we have:

$$\begin{aligned}
&\widehat{\mathbb{E}}_k[\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2]=\widehat{\mathbb{E}}_k[\|\bar{y}^k-\beta\bar{v}^k-y^\star(\bar{x}^k)\|^2]\\
=&\widehat{\mathbb{E}}_k\left[\left\|\bar{y}^k-\beta\frac{1}{n}\sum_{i=1}^{n}\nabla_2 g_i(x_i^k,y_i^k)-y^\star(\bar{x}^k)+\beta\frac{1}{n}\sum_{i=1}^{n}\left(\nabla_2 g_i(x_i^k,y_i^k)-v_i^k\right)\right\|^2\right]\\
\le&\left\|\bar{y}^k-\beta\frac{1}{n}\sum_{i=1}^{n}\nabla_2 g_i(\bar{x}^k,\bar{y}^k)-y^\star(\bar{x}^k)+\beta\frac{1}{n}\sum_{i=1}^{n}(\nabla_2 g_i(\bar{x}^k,\bar{y}^k)-\nabla_2 g_i(x_i^k,y_i^k))\right\|^2+\beta^2\frac{\sigma_{g,1}^2}{n}.
\end{aligned}$$

Then,

$$
\begin{aligned}
&\widehat{\mathbb{E}}_k[\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2]\\
\leq&\left(1+\frac{\beta\mu_g}{2}\right)\left\|\bar{y}^k-\beta\frac{1}{n}\sum_{i=1}^n\nabla_2 g_i(\bar{x}^k,\bar{y}^k)-y^\star(\bar{x}^k)\right\|^2\\
&+\beta^2\left(1+\frac{2}{\beta\mu_g}\right)\left\|\frac{1}{n}\sum_{i=1}^n(\nabla_2 g_i(\bar{x}^k,\bar{y}^k)-\nabla_2 g_i(x_i^k,y_i^k))\right\|^2+\beta^2\frac{\sigma_{g,1}^2}{n}\\
\leq&\left(1+\frac{\beta\mu_g}{2}\right)(1-\beta\mu_g)^2\|\bar{y}^k-y^\star(\bar{x}^k)\|^2\\
&+\beta^2\left(1+\frac{2}{\beta\mu_g}\right)L_{g,1}^2\left(\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2}{n}\right)+\beta^2\frac{\sigma_{g,1}^2}{n}\\
\leq&(1-\beta\mu_g)\left[\left(1+\frac{\beta\mu_g}{2}\right)\|\bar{y}^k-y^\star(\bar{x}^{k-1})\|^2+\left(1+\frac{2}{\beta\mu_g}\right)\|y^\star(\bar{x}^k)-y^\star(\bar{x}^{k-1})\|^2\right]\\
&+\beta^2\left(1+\frac{2}{\beta\mu_g}\right)L_{g,1}^2\left(\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2}{n}\right)+\beta^2\frac{\sigma_{g,1}^2}{n}\\
\leq&\left(1-\frac{\beta\mu_g}{2}\right)\|\bar{y}^k-y^\star(\bar{x}^{k-1})\|^2+\frac{3}{\beta\mu_g}L_{y^\star}^2\|\bar{x}^k-\bar{x}^{k-1}\|^2\\
&+\frac{3\beta}{\mu_g}L_{g,1}^2\left(\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2}{n}\right)+\beta^2\frac{\sigma_{g,1}^2}{n},
\end{aligned}
$$

where the first and the third inequality is due to the Jenson's inequality, the second inequality holds according to Lemma 2 and the fact that $\beta\leq\frac{\mu_g}{32L_{g,1}^2}\leq\frac{1}{3(\mu_g+L_{g,1})}$, and the last inequality uses $\beta\mu_g\leq\frac{1}{3}$. Taking the summation and expectation on the both sides, we get:

$$
\begin{aligned}
&\sum_{k=0}^K\frac{\beta\mu_g}{2}\mathbb{E}[\|\bar{y}^k-y^\star(\bar{x}^{k-1})\|^2]+\mathbb{E}[\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2]\\
\leq&\mathbb{E}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\sum_{k=0}^K\mathbb{E}\left[\frac{3\alpha^2}{\beta\mu_g}L_{y^\star}^2\|\bar{r}^k\|^2+\frac{3\beta}{\mu_g}L_{g,1}^2\left(\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2}{n}\right)+\beta^2\frac{\sigma_{g,1}^2}{n}\right].
\end{aligned}
$$

Using (36) and the fact that $\mathbf{x}^0,\mathbf{y}^0$ is consensual, it follows that:

$$
\begin{aligned}
\sum_{k=0}^{K+1}\frac{\beta\mu_g}{2}\mathbb{E}[\|\bar{y}^k-y^\star(\bar{x}^{k-1})\|^2]\leq&2\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\sum_{k=1}^K\frac{3\alpha^2}{\beta\mu_g}L_{y^\star}^2\mathbb{E}\|\bar{r}^k\|^2\\
&+\sum_{k=1}^K\frac{3\beta}{\mu_g}L_{g,1}^2\mathbb{E}\left[\frac{\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^k\|^2}{n}\right]+\frac{2K\beta^2\sigma_{g,1}^2}{n}.
\end{aligned}
\tag{55}
$$

On the other hand,

$$
\begin{aligned}
&\widehat{\mathbb{E}}_k\left[\|\bar{y}^{k+1}-\bar{y}^k\|^2\right]\\
\leq&\beta^2\left\|\frac{1}{n}\sum_{i=1}^n\nabla_2 g_i(x_i^k,y_i^k)\right\|^2+\frac{\beta^2}{n}\sigma_{g,1}^2\\
\leq&2\beta^2\left(\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla_2 g_i(x_i^k,y_i^k)-\nabla_2 g_i(\bar{x}^k,\bar{y}^k)\right)\right\|^2+\left\|\nabla_2 g(\bar{x}^k,\bar{y}^k)-\nabla_2 g(\bar{x}^k,y^\star(\bar{x}^k))\right\|^2\right)\\
&+\frac{\beta^2}{n}\sigma_{g,1}^2\\
\leq&\frac{2\beta^2L_{g,1}^2}{n}\left(\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2+\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2+2\|\bar{\mathbf{y}}^{k+1}-\mathbf{y}^\star(\bar{\mathbf{x}}^k)\|^2+2\|\bar{\mathbf{y}}^{k+1}-\bar{\mathbf{y}}^k\|^2\right)+\frac{\beta^2}{n}\sigma_{g,1}^2,
\end{aligned}
$$

where the second inequality uses $\nabla_2 g(\bar{x}^k, y^\star(\bar{x}^k)) = 0$.

Note that $\beta^2 \leq \frac{\mu_g^2}{32L_{g,1}^4} \leq \frac{1}{8L_{g,1}^2}$. Subtracting $2\beta^2 L_{g,1}^2 \|\bar{y}^{k+1} - \bar{y}^k\|$ on both sides, and taking expectation and summation, we get:

$$
\begin{aligned}
&\sum_{k=0}^{K} \mathbb{E}\left[\|\bar{y}^{k+1} - \bar{y}^k\|^2\right] \\
\leq &\sum_{k=0}^{K} \left[\frac{4\beta^2 L_{g,1}^2}{n} \mathbb{E}\left(\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + \|\mathbf{y}^k - \bar{\mathbf{y}}^k\|^2\right) + \frac{8\beta^2 L_{g,1}^2}{n} \mathbb{E}\|\bar{\mathbf{y}}^{k+1} - \mathbf{y}^\star(\bar{\mathbf{x}}^k)\|^2 + \frac{2\beta^2}{n}\sigma_{g,1}^2\right] \\
\leq &\frac{4\beta^2 L_{g,1}^2}{n} \sum_{k=1}^{K} \mathbb{E}\left(\|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_k^x\|^2 + \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_k^y\|^2\right) + \frac{8\beta^2 L_{g,1}^2}{n} \sum_{k=0}^{K} \mathbb{E}\|\bar{\mathbf{y}}^{k+1} - \mathbf{y}^\star(\bar{\mathbf{x}}^k)\|^2 \\
&+ \frac{2(K+1)\beta^2}{n}\sigma_{g,1}^2 \\
\leq &\frac{4\beta^2 L_{g,1}^2}{n} \sum_{k=1}^{K} \mathbb{E}\left(\|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2 + \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^k\|^2\right) + \frac{2(K+1)\beta^2}{n}\sigma_{g,1}^2 \\
&+ 8\beta^2 L_{g,1}^2 \left(\frac{4}{\beta\mu_g}\|\bar{y}_0 - y^\star(\bar{x}^0)\|^2 + \sum_{k=1}^{K} \frac{6\alpha^2}{\beta^2\mu_g^2} L_{y^\star}^2 \mathbb{E}\|\bar{r}^k\|^2\right) \\
&+ 8\beta^2 L_{g,1}^2 \left(\sum_{k=1}^{K} \frac{6}{\mu_g^2} L_{g,1}^2 \mathbb{E}\left[\frac{\|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2 + \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^k\|^2}{n}\right] + \frac{4K\beta\sigma_{g,1}^2}{n\mu_g}\right) \\
\leq &\frac{\beta^2 L_{g,1}^2}{n}\left(4 + \frac{48L_{g,1}^2}{\mu_g^2}\right) \sum_{k=1}^{K} \mathbb{E}\left(\|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2 + \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^k\|^2\right) + \frac{48\alpha^2 L_{g,1}^2}{\mu_g^2} L_{y^\star}^2 \sum_{k=1}^{K} \mathbb{E}\|\bar{r}^k\|^2 \\
&+ \frac{3(K+1)\beta^2}{n}\sigma_{g,1}^2 + \frac{32\beta L_{g,1}^2}{\mu_g}\|\bar{y}^0 - y^\star(\bar{x}^0)\|^2.
\end{aligned}
$$

where the second inequality holds since $\mathbf{x}^0, \mathbf{y}^0$ are consensual, the third inequality uses (55), and the last inequality holds since $\beta \leq \frac{\mu_g}{32L_{g,1}^2}$. □

Lemma 10 (Estimation error of $z^\star(x)$)**.** *Suppose that Assumptions 1- 4hold, and*

$$
\gamma < \min\left\{\frac{1}{\mu_g}, \frac{nL_{g,1}^2}{\mu_g^2\sigma_{g,2}^2}, \frac{n\mu_g}{36\sigma_{g,2}^2}\right\}. \tag{56}
$$

We have:

$$
\begin{aligned}
&\sum_{k=0}^{K+1} \mathbb{E}\|\bar{z}^k - z_\star^k\|^2 \\
\leq &\sum_{k=0}^{K} \frac{9\alpha^2 L_{z^\star}^2}{\gamma^2\mu_g^2} \mathbb{E}\|\bar{r}^k\|^2 \\
&+ 72\kappa^2 \sum_{k=1}^{K} \mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2 + \|\mathbf{O}_z\|^2\|\hat{\mathbf{e}}_z^k\|^2}{n}\right] + 72\kappa^2 \sum_{k=0}^{K} \mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2}{n}\right] \\
&+ \sum_{k=0}^{K} 72\kappa^4 \mathbb{E}\left[\|\bar{y}^{k+1} - y^\star(\bar{x}^k)\|^2\right] + \frac{3\|z_\star^1\|^2}{\mu_g\gamma} + \frac{6(K+1)\gamma}{\mu_g n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2} + \sigma_{f,1}^2\right).
\end{aligned}
$$

Proof. For each $k \geq 0$, note that $z_\star^k = \nabla_{22}^2 g(\bar{x}^k, y^\star(\bar{x}^k))^{-1} \nabla_2 f_2(\bar{x}^k, y^\star(\bar{x}^k))$, we have:

$$
\widetilde{\mathbb{E}}_k[\bar{z}^{k+1}] - z_\star^{k+1} = \bar{z}^k - \frac{\gamma}{n}\sum_{i=1}^{n}\left(\nabla_{22}^2 g_i(x_i^k, y_i^{k+1}) z_i^k - \nabla_2 f_i(x_i^k, y_i^{k+1})\right) - z_\star^{k+1}
$$

$$= \left[I - \frac{\gamma}{n} \sum_{i=1}^{n} \nabla_{22}^2 g_i(x_i^k, y_i^{k+1}) \right] (\bar{z}^k - z_\star^{k+1}) + \frac{\gamma}{n} \sum_{i=1}^{n} \left[\nabla_2 f_i(x_i^k, y_i^{k+1}) - \nabla_2 f_i(\bar{x}^k, y^\star(\bar{x}^k)) \right]$$
$$+ \frac{\gamma}{n} \sum_{i=1}^{n} \nabla_{22}^2 g_i(x_i^k, y_i^{k+1})(\bar{z}^k - z_i^k) + \frac{\gamma}{n} \sum_{i=1}^{n} \left[\nabla_{22}^2 g_i(\bar{x}^k, y^\star(\bar{x}^k)) - \nabla_{22}^2 g_i(x_i^k, y_i^{k+1}) \right] z_\star^{k+1}.$$

Then we have:

$$\begin{aligned}
&\left\| \widetilde{\mathbb{E}}_k[\bar{z}^{k+1}] - z_\star^{k+1} \right\|^2 \\
\leq &(1+\gamma\mu_g) \left\| \left[I - \frac{\gamma}{n} \sum_{i=1}^{n} \nabla_{22}^2 g_i(x_i^k, y_i^{k+1}) \right] (\bar{z}^k - z_\star^{k+1}) \right\|^2 \\
&+ 3\gamma^2 \left(1 + \frac{1}{\gamma\mu_g}\right) \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\nabla_2 f_i(x_i^k, y_i^{k+1}) - \nabla_2 f_i(\bar{x}^k, y^\star(\bar{x}^k)) \right] \right\|^2 \\
&+ 3\gamma^2 \left(1 + \frac{1}{\gamma\mu_g}\right) \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\nabla_{22}^2 g_i(\bar{x}^k, y^\star(\bar{x}^k)) - \nabla_{22}^2 g_i(x_i^k, y_i^{k+1}) \right] z_\star^{k+1} \right\|^2 \\
&+ 3\gamma^2 \left(1 + \frac{1}{\gamma\mu_g}\right) \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{22}^2 g_i(x_i^k, y_i^{k+1})(\bar{z}^k - z_i^k) \right\|^2 \\
\leq &(1+\gamma\mu_g)(1-\gamma\mu_g)^2 \|\bar{z}^k - z_\star^{k+1}\|^2 \\
&+ \frac{6\gamma}{\mu_g} \left(L_{g,1}^2 \frac{\|\mathbf{z}^k - \bar{\mathbf{z}}^k\|}{n} + \left(\frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2} + L_{f,1}^2 \right) \left(\frac{\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2}{n} + \frac{\|\mathbf{y}^{k+1} - \mathbf{y}^\star(\bar{x}^k)\|^2}{n} \right) \right) \\
\leq &(1-\gamma\mu_g) \left(1 + \frac{\gamma\mu_g}{2}\right) \|\bar{z}^k - z_\star^k\|^2 + \left(1 + \frac{2}{\gamma\mu_g}\right) \|z_\star^k - z_\star^{k+1}\|^2 + \frac{6\gamma}{\mu_g} L_{g,1}^2 \frac{\|\mathbf{z}^k - \bar{\mathbf{z}}^k\|}{n} \\
&+ \frac{12\gamma}{\mu_g} \left(\frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2} + L_{f,1}^2 \right) \left(\frac{\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2}{n} + \frac{\|\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1}\|^2}{n} + \|\bar{y}^{k+1} - y^\star(\bar{x}^k)\|^2 \right) \\
\leq &(1 - \frac{\gamma\mu_g}{2}) \|\bar{z}^k - z_\star^k\|^2 + \frac{3\alpha^2 L_{z^\star}^2}{\gamma\mu_g} \|\bar{r}^k\|^2 + \frac{6\gamma}{\mu_g} L_{g,1}^2 \frac{\|\mathbf{z}^k - \bar{\mathbf{z}}^k\|}{n} \\
&+ \frac{12\gamma}{\mu_g} \left(\frac{L_{g,2}^2 L_{f,0}^2}{\mu_g^2} + L_{f,1}^2 \right) \left(\frac{\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2}{n} + \frac{\|\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1}\|^2}{n} + \|\bar{y}^{k+1} - y^\star(\bar{x}^k)\|^2 \right)
\end{aligned}$$

where the first and third inequality uses Jensen's inequality and Cauchy Schwartz inequality, the second inequality holds due to Assumption 1 and $\gamma\mu_g < 1$, the last inequality holds since $z^\star(x)$ is $L_{z^\star}$ Lipschitz continuous.

Moreover, the independence of samples implies that

$$\begin{aligned}
\widetilde{\mathbb{E}}_k \left\| \bar{z}^{k+1} - \widetilde{\mathbb{E}}_k[\bar{z}^{k+1}] \right\|^2 &= \gamma^2 \widetilde{\mathbb{E}}_k \left\| \frac{1}{n} \sum_{i=1}^{n} (H_i^k - \widetilde{\mathbb{E}}_k[H_i^k]) z_i^k + \frac{1}{n} \sum_i (b_i^k - \widetilde{\mathbb{E}}_k[b_i^k]) \right\|^2 \\
&\leq \frac{2\gamma^2}{n} \left(\sigma_{g,2}^2 \frac{\|\mathbf{z}^k\|^2}{n} + \sigma_{f,1}^2 \right) \\
&\leq \frac{2\gamma^2}{n} \left(3\sigma_{g,2}^2 \left(\frac{\|\mathbf{z}^k - \bar{\mathbf{z}}_k\|^2}{n} + \|\bar{z}^k - z_\star^k\|^2 + \frac{L_{f,0}^2}{\mu_g^2} \right) + \sigma_{f,1}^2 \right).
\end{aligned}$$

As γ satisfies

$$\frac{6\sigma_{g,2}^2\gamma^2}{n} \leq \frac{6\gamma L_{g,1}^2}{\mu_g^2}, \quad \frac{6\sigma_{g,2}^2\gamma^2}{n} \leq \frac{\gamma\mu_g}{6},$$

we get:

$$\begin{aligned}
&\widetilde{\mathbb{E}}_k[\|\bar{z}^{k+1}-z_\star^{k+1}\|^2]=\widetilde{\mathbb{E}}_k\|\widetilde{\mathbb{E}}_k[\bar{z}^{k+1}]-z_\star^{k+1}\|^2+\widetilde{\mathbb{E}}_k\|\bar{z}^{k+1}-\widetilde{\mathbb{E}}_k[\bar{z}^{k+1}]\|^2\\
\leq&\left(1-\frac{\gamma\mu_g}{3}\right)\|\bar{z}^k-z_\star^k\|^2+\frac{3\alpha^2L_{z^\star}^2}{\gamma\mu_g}\|\bar{r}^k\|^2+\frac{12\gamma}{\mu_g}L_{g,1}^2\frac{\|\mathbf{z}^k-\bar{\mathbf{z}}^k\|^2}{n}+\frac{2\gamma^2}{n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)\\
&+\frac{12\gamma}{\mu_g}\left(\frac{L_{g,2}^2L_{f,0}^2}{\mu_g^2}+L_{f,1}^2\right)\left(\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\|^2}{n}+\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2\right).
\end{aligned}$$

Taking expectation and summation on both sides, we get

$$\begin{aligned}
&\sum_{k=0}^K\frac{\gamma\mu_g}{3}\mathbb{E}\|\bar{z}^k-z_\star^k\|^2+\mathbb{E}\|\bar{z}^{K+1}-z_\star^{K+1}\|^2\\
\leq&\mathbb{E}\|\bar{z}^0-z_\star^0\|^2+\sum_{k=0}^K\left[\frac{3\alpha^2L_{z^\star}^2}{\gamma\mu_g}\mathbb{E}\|\bar{r}^k\|^2+\frac{12\gamma}{\mu_g}L_{g,1}^2\frac{\mathbb{E}\|\mathbf{z}^k-\bar{\mathbf{z}}^k\|^2}{n}+\frac{2\gamma^2}{n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)\right]\\
&+\sum_{k=0}^K\frac{12\gamma}{\mu_g}\left(\frac{L_{g,2}^2L_{f,0}^2}{\mu_g^2}+L_{f,1}^2\right)\mathbb{E}\left[\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\|^2}{n}+\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2\right].
\end{aligned}$$

It follows that

$$\begin{aligned}
&\sum_{k=0}^{K+1}\mathbb{E}\|\bar{z}^k-z_\star^k\|^2\\
\leq&\sum_{k=0}^K\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}\mathbb{E}\|\bar{r}^k\|^2\\
&+72\kappa^2\sum_{k=1}^K\mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_z\|^2\|\hat{\mathbf{e}}_z^k\|^2}{n}\right]+72\kappa^2\sum_{k=0}^K\mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2}{n}\right]\\
&+\sum_{k=0}^K72\kappa^4\mathbb{E}\left[\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2\right]+\frac{3\|z_\star^1\|^2}{\mu_g\gamma}+\frac{6(K+1)\gamma}{\mu_gn}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right),
\end{aligned}$$

since $z_\star^0=z_\star^1$ and $\mathbf{z}^0$ is consensual. □

Then, we combine the results in Lemmas 9, 10 and give an upper bound of $\mathbb{E}[I_k]$:

Lemma 11. *Suppose that Lemmas 9 and 10 hold. Then we have:*

$$\begin{aligned}
\sum_{k=-1}^K\mathbb{E}[I_k]\leq&\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\sum_{k=0}^K\mathbb{E}\|\bar{r}^k\|^2+510\kappa^4\sum_{k=0}^K\mathbb{E}\left[\frac{\Delta_k}{n}\right]+\frac{3\|z_\star^1\|^2}{\mu_g\gamma}\\
&+\frac{6(K+1)\gamma}{\mu_gn}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)+73\kappa^4\left(\frac{4}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{n\mu_g}\beta\right).
\end{aligned}\tag{57}$$

Remark 6. *Here $I_{-1}=\|\bar{z}^0-z_\star^0\|^2+\kappa^2\|\bar{y}^0-y^\star(\bar{x}^{-1})\|^2$. The aim of introducing this term is to simplify the subsequent proofs of other lemmas.*

Proof. Lemma 10 implies that:

$$
\begin{aligned}
&\sum_{k=0}^{K+1}\mathbb{E}\|\bar{z}^k - z_\star^k\|^2 \\
\leq&\sum_{k=0}^{K}\frac{9\alpha^2 L_{z^\star}^2}{\gamma^2\mu_g^2}\mathbb{E}\|\bar{r}^k\|^2 \\
&+72\kappa^2\sum_{k=1}^{K}\mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_z\|^2\|\hat{\mathbf{e}}_z^k\|^2}{n}\right]+72\kappa^2\sum_{k=0}^{K}\mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2}{n}\right] \\
&+\sum_{k=0}^{K}72\kappa^4\mathbb{E}\left[\|\bar{y}^{k+1}-y^\star(\bar{x}^k)\|^2\right]+\frac{3\|z_\star^1\|^2}{\mu_g\gamma}+\frac{6(K+1)\gamma}{\mu_g n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)
\end{aligned}
$$

Then, using Lemma 9, we have:

$$
\begin{aligned}
&\sum_{k=-1}^{K}\mathbb{E}[I_k]=\sum_{k=0}^{K+1}\mathbb{E}\|\bar{z}^k-z_\star^k\|^2+\kappa^2\sum_{k=0}^{K+1}\mathbb{E}\|\bar{y}^k-y^\star(\bar{x}^{k-1})\|^2 \\
\leq&\sum_{k=0}^{K}\frac{9\alpha^2 L_{z^\star}^2}{\gamma^2\mu_g^2}\mathbb{E}\|\bar{r}^k\|^2 \\
&+72\kappa^2\sum_{k=1}^{K}\mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_z\|^2\|\hat{\mathbf{e}}_z^k\|^2}{n}\right]+72\kappa^2\sum_{k=0}^{K}\mathbb{E}\left[\frac{\kappa^2\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2}{n}\right] \\
&+\frac{6(K+1)\gamma}{\mu_g n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)+73\kappa^4\left[\frac{4}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\sum_{k=1}^{K}\frac{6\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\mathbb{E}\|\bar{r}^k\|^2\right] \\
&+\frac{3\|z_\star^1\|^2}{\mu_g\gamma}+73\kappa^4\left[\sum_{k=1}^{K}\frac{6}{\mu_g^2}L_{g,1}^2\mathbb{E}\left[\frac{\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^k\|^2}{n}\right]+\frac{4K\sigma_{g,1}^2}{n\mu_g}\beta\right] \\
\leq&\sum_{k=0}^{K}\left(\frac{9\alpha^2 L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\mathbb{E}\|\bar{r}^k\|^2+\sum_{k=0}^{K}510\kappa^4\mathbb{E}\left[\frac{\Delta_k}{n}\right]+\frac{3\|z_\star^1\|^2}{\mu_g\gamma} \\
&+\frac{6(K+1)\gamma}{\mu_g n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)+73\kappa^4\left(\frac{4}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{n\mu_g}\beta\right).
\end{aligned}
$$

□

C.1.7 Consensus error analysis

In this subsection we aim to bound the consensus errors of y, z, x (*i*.e. the terms $\|\hat{\mathbf{e}}_y^k\|^2$, $\|\hat{\mathbf{e}}_z^k\|^2$, and $\|\hat{\mathbf{e}}_x^k\|^2$).

Lemma 12 (Consensus error of y)**.** *Suppose that Assumptions 1- 4 hold, and*

$$
\beta^2\leq\frac{(1-\|\mathbf{\Gamma}_y\|)^2}{8L_{g,1}^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{O}_y\|^2\|\mathbf{\Lambda}_{ya}\|^2}. \tag{58}
$$

We have

$$
\begin{aligned}
\sum_{k=0}^{K+1}\mathbb{E}\|\hat{\mathbf{e}}_y^k\|^2\leq&3\sum_{k=0}^{K}\frac{\beta^2\|\mathbf{O}_y^{-1}\|^2}{(1-\|\mathbf{\Gamma}_y\|)^2}\|\mathbf{\Lambda}_{yb}^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2L_{g,1}^2\mathbb{E}\left[\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2+\|\bar{\mathbf{y}}^{k+1}-\bar{\mathbf{y}}^k\|^2\right] \\
&+\frac{\|\mathbf{O}_x\|^2}{3\|\mathbf{O}_y\|^2}\sum_{k=0}^{K}\mathbb{E}\|\hat{\mathbf{e}}_x^k\|^2+\frac{3(K+1)\beta^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}.
\end{aligned} \tag{59}
$$

Proof. Firstly, the term $\|\hat{\mathbf{e}}_y^{k+1}\|^2$ can be deformed as

$$
\begin{aligned}
&\|\hat{\mathbf{e}}_y^{k+1}\|^2 \\
=&\left\|\boldsymbol{\Gamma}_y\hat{\mathbf{e}}_y^k-\beta\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\widehat{\mathbb{E}}_k[\mathbf{v}^k]-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\\ \boldsymbol{\Lambda}_{yb}^{-1}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\nabla_2\mathbf{g}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+1})-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\end{bmatrix}-\beta\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\mathbf{v}^k-\widehat{\mathbb{E}}_k[\mathbf{v}^k]\right]\\ \mathbf{0}\end{bmatrix}\right\|^2\\
=&\left\|\boldsymbol{\Gamma}_y\hat{\mathbf{e}}_y^k-\beta\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\widehat{\mathbb{E}}_k[\mathbf{v}^k]-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\\ \boldsymbol{\Lambda}_{yb}^{-1}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\nabla_2\mathbf{g}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+1})-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\end{bmatrix}\right\|^2\\
&+\beta^2\left\|\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\mathbf{v}^k-\widehat{\mathbb{E}}_k[\mathbf{v}^k]\right]\\ \mathbf{0}\end{bmatrix}\right\|^2-2\left\langle\boldsymbol{\Gamma}_y\hat{\mathbf{e}}_y^k,\beta\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\mathbf{v}^k-\widehat{\mathbb{E}}_k[\mathbf{v}^k]\right]\\ \mathbf{0}\end{bmatrix}\right\rangle\\
&+2\beta^2\left\langle\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\widehat{\mathbb{E}}_k[\mathbf{v}^k]-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\\ \boldsymbol{\Lambda}_{yb}^{-1}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\nabla_2\mathbf{g}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+1})-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\end{bmatrix},\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\mathbf{v}^k-\widehat{\mathbb{E}}_k[\mathbf{v}^k]\right]\\ \mathbf{0}\end{bmatrix}\right\rangle
\end{aligned}
\tag{60}
$$

due to Eq.(33). Then, for the first term in the right-hand side of (60), we have:

$$
\begin{aligned}
&\widehat{\mathbb{E}}_k\left[\left\|\boldsymbol{\Gamma}_y\hat{\mathbf{e}}_y^k-\beta\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\widehat{\mathbb{E}}_k[\mathbf{v}^k]-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\\ \boldsymbol{\Lambda}_{yb}^{-1}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\nabla_2\mathbf{g}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+1})-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\end{bmatrix}\right\|^2\right]\\
\le&\|\boldsymbol{\Gamma}_y\|\|\hat{\mathbf{e}}_y^k\|^2+\frac{\beta^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\boldsymbol{\Gamma}_y\|}\widehat{\mathbb{E}}_k\left[\left\|\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\widehat{\mathbb{E}}_k[\mathbf{v}^k]-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\\ \boldsymbol{\Lambda}_{yb}^{-1}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\nabla_2\mathbf{g}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+1})-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\right]\end{bmatrix}\right\|^2\right]\\
\le&\|\boldsymbol{\Gamma}_y\|\|\hat{\mathbf{e}}_y^k\|^2+\frac{\beta^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\boldsymbol{\Gamma}_y\|}\cdot\|\boldsymbol{\Lambda}_{ya}\|^2\|\nabla_2\mathbf{g}(\mathbf{x}^k,\mathbf{y}^k)-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\|^2\\
&+\frac{\beta^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\boldsymbol{\Gamma}_y\|}\|\boldsymbol{\Lambda}_{yb}^{-1}\|^2\|\boldsymbol{\Lambda}_{ya}\|^2\widehat{\mathbb{E}}_k\left[\|\nabla_2\mathbf{g}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+1})-\nabla_2\mathbf{g}(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^k)\|^2\right]\\
\le&\|\boldsymbol{\Gamma}_y\|\|\hat{\mathbf{e}}_y^k\|^2+\frac{\beta^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\boldsymbol{\Gamma}_y\|}\cdot\|\boldsymbol{\Lambda}_{ya}\|^2L_{g,1}^2\left(\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2+\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2\right)\\
&+\frac{\beta^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\boldsymbol{\Gamma}_y\|}\|\boldsymbol{\Lambda}_{yb}^{-1}\|^2\|\boldsymbol{\Lambda}_{ya}\|^2L_{g,1}^2\widehat{\mathbb{E}}_k\left[\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2+\|\bar{\mathbf{y}}^{k+1}-\bar{\mathbf{y}}^k\|^2\right],
\end{aligned}
\tag{61}
$$

where the first inequality uses the Jenson's inequality, the second inequality hold since $\|\hat{\mathbf{U}}_y^\top\|\le 1$. For the second term, we have:

$$
\widehat{\mathbb{E}}_k\left[\left\|\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\mathbf{v}^k-\widehat{\mathbb{E}}_k[\mathbf{v}^k]\right]\\ \mathbf{0}\end{bmatrix}\right\|^2\right]\le\|\mathbf{O}_y^{-1}\|^2\|\boldsymbol{\Lambda}_{ya}\|^2n\sigma_{g,1}^2. \tag{62}
$$

For the third them, we have:

$$
\widehat{\mathbb{E}}_k\left[\left\langle\boldsymbol{\Gamma}_y\hat{\mathbf{e}}_y^k,\beta\mathbf{O}_y^{-1}\begin{bmatrix}\boldsymbol{\Lambda}_{ya}\hat{\mathbf{U}}_y^\top\left[\mathbf{v}^k-\widehat{\mathbb{E}}_k[\mathbf{v}^k]\right]\\ \mathbf{0}\end{bmatrix}\right\rangle\right]=0. \tag{63}
$$

Next, for the last term, we have:

$$
\begin{aligned}
&\widehat{\mathbb{E}}_k \left\langle \mathbf{O}_y^{-1} \begin{bmatrix} \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\widehat{\mathbb{E}}_k[\mathbf{v}^k] - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k) \right] \\ \mathbf{\Lambda}_{yb}^{-1} \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\nabla_2 \mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k) \right] \end{bmatrix}, \mathbf{O}_y^{-1} \begin{bmatrix} \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\mathbf{v}^k - \widehat{\mathbb{E}}_k[\mathbf{v}^k] \right] \\ \mathbf{0} \end{bmatrix} \right\rangle \\
\leq & \frac{1}{2} \|\mathbf{O}_y^{-1}\|^2 \widehat{\mathbb{E}}_k \left\| \begin{bmatrix} \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\widehat{\mathbb{E}}_k[\mathbf{v}^k] - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k) \right] \\ \mathbf{\Lambda}_{yb}^{-1} \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\nabla_2 \mathbf{g}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+1}) - \nabla_2 \mathbf{g}(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^k) \right] \end{bmatrix} \right\|^2 \\
& + \frac{1}{2} \|\mathbf{O}_y^{-1}\|^2 \widehat{\mathbb{E}}_k \left\| \begin{bmatrix} \mathbf{\Lambda}_{ya} \hat{\mathbf{U}}_y^\top \left[\mathbf{v}^k - \widehat{\mathbb{E}}_k[\mathbf{v}^k] \right] \\ \mathbf{0} \end{bmatrix} \right\|^2 \\
\leq & \frac{1}{2} \beta^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \widehat{\mathbb{E}}_k \left[\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + \|\mathbf{y}^k - \bar{\mathbf{y}}^k\|^2 \right] \\
& + \frac{1}{2} \beta^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{yb}^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \widehat{\mathbb{E}}_k \left[\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \|\bar{\mathbf{y}}^{k+1} - \bar{\mathbf{y}}^k\|^2 \right] \\
& + \frac{1}{2} \beta^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 n \sigma_{g,1}^2.
\end{aligned} \tag{64}
$$

Taking expectations on both sides of (60), and plugging (61), (62), (63), (64) into it, we obtain:

$$
\begin{aligned}
& \mathbb{E} \left[\|\hat{\mathbf{e}}_y^{k+1}\|^2 \right] \\
\leq & 2 \frac{\beta^2 \|\mathbf{O}_y^{-1}\|^2}{1 - \|\mathbf{\Gamma}_y\|} \|\mathbf{\Lambda}_{yb}^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \mathbb{E} \left[\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \|\bar{\mathbf{y}}^{k+1} - \bar{\mathbf{y}}^k\|^2 \right] + \|\mathbf{\Gamma}_y\| \mathbb{E} \|\hat{\mathbf{e}}_y^k\|^2 \\
& + 2 \frac{\beta^2 \|\mathbf{O}_y^{-1}\|^2}{1 - \|\mathbf{\Gamma}_y\|} \| \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \mathbb{E} \left[\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + \|\mathbf{y}^k - \bar{\mathbf{y}}^k\|^2 \right] + 2 \beta^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 n \sigma_{g,1}^2.
\end{aligned}
$$

Taking summation over k and using $\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \leq \|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2$, $\|\mathbf{y}^k - \bar{\mathbf{y}}^k\|^2 \leq \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^k\|^2$, we get:

$$
\begin{aligned}
& (1 - \|\mathbf{\Gamma}_y\|) \sum_{k=0}^K \mathbb{E} \left[\|\hat{\mathbf{e}}_y^k\|^2 \right] \\
\leq & 2 \sum_{k=0}^K \frac{\beta^2 \|\mathbf{O}_y^{-1}\|^2}{1 - \|\mathbf{\Gamma}_y\|} \|\mathbf{\Lambda}_{yb}^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \mathbb{E} \left[\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \|\bar{\mathbf{y}}^{k+1} - \bar{\mathbf{y}}^k\|^2 \right] \\
& + \mathbb{E} \|\hat{\mathbf{e}}_y^0\|^2 - \mathbb{E} \|\hat{\mathbf{e}}_y^{k+1}\|^2 + 2 \sum_{k=0}^K \frac{\beta^2 \|\mathbf{O}_y^{-1}\|^2}{1 - \|\mathbf{\Gamma}_y\|} \| \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \mathbb{E} \left[\|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2 + \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^k\|^2 \right] \\
& + 2(K+1) \beta^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 n \sigma_{g,1}^2 \\
\leq & 2 \sum_{k=0}^K \frac{\beta^2 \|\mathbf{O}_y^{-1}\|^2}{1 - \|\mathbf{\Gamma}_y\|} \|\mathbf{\Lambda}_{yb}^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \mathbb{E} \left[\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \|\bar{\mathbf{y}}^{k+1} - \bar{\mathbf{y}}^k\|^2 \right] \\
& + \mathbb{E} \|\hat{\mathbf{e}}_y^0\|^2 - \mathbb{E} \|\hat{\mathbf{e}}_y^{k+1}\|^2 + \frac{1 - \|\mathbf{\Gamma}_y\|}{4 \|\mathbf{O}_y\|^2} \sum_{k=0}^K \mathbb{E} \left[\|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2 + \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^k\|^2 \right] \\
& + 2(K+1) \beta^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 n \sigma_{g,1}^2,
\end{aligned}
$$

where the last inequality uses $\beta^2 \leq \dfrac{(1 - \|\mathbf{\Gamma}_y\|)^2}{8 L_{g,1}^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{O}_y\|^2 \|\mathbf{\Lambda}_{ya}\|^2}$.

It follows that

$$
\begin{aligned}
\sum_{k=0}^{K+1} \mathbb{E} \left[\|\hat{\mathbf{e}}_y^k\|^2 \right] \leq & 3 \sum_{k=0}^K \frac{\beta^2 \|\mathbf{O}_y^{-1}\|^2}{(1 - \|\mathbf{\Gamma}_y\|)^2} \|\mathbf{\Lambda}_{yb}^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 L_{g,1}^2 \mathbb{E} \left[\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \|\bar{\mathbf{y}}^{k+1} - \bar{\mathbf{y}}^k\|^2 \right] \\
& + \frac{\|\mathbf{O}_x\|^2}{3 \|\mathbf{O}_y\|^2} \sum_{k=0}^K \mathbb{E} \left[\|\hat{\mathbf{e}}_x^k\|^2 \right] + \frac{3(K+1) \beta^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2}{1 - \|\mathbf{\Gamma}_y\|} n \sigma_{g,1}^2 + \frac{2 \mathbb{E} \|\hat{\mathbf{e}}_y^0\|^2}{1 - \|\mathbf{\Gamma}_y\|}.
\end{aligned}
$$

□

Lemma 13 (Consensus error of z). *Suppose that Assumptions 1- 4 hold, and γ satisfies*

$$\frac{6\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{O}_z\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\cdot(2L^2+2(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)\leq\frac{1-\|\mathbf{\Gamma}_z\|}{4}. \tag{65}$$

We have

$$\begin{aligned}
&\sum_{k=0}^{K+1}\mathbb{E}\left[\|\hat{\mathbf{e}}_z^k\|^2\right]\\
\leq&\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{\mathbf{z}}^k-\mathbf{z}_\star^k\|^2\right]+\frac{2\mathbb{E}[\|\hat{\mathbf{e}}_z^0\|^2]}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{8\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^{K}\mathbb{E}\left[\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+L_{g,1}^2L_{z^\star}^2\right)\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2\right]\\
&+\frac{8\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^{K}\mathbb{E}\left[\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\|\bar{\mathbf{y}}^{k+2}-\bar{\mathbf{y}}^{k+1}\|^2\right]\\
&+16(K+1)n\gamma^2\frac{\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+\frac{\kappa^2}{3\|\mathbf{O}_z\|^2}\sum_{k=0}^{K}\mathbb{E}(\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2).
\end{aligned}\tag{66}$$

Proof. Firstly, Eq. (34) implies that:

$$\begin{aligned}
\hat{\mathbf{e}}_z^{k+1}=&\mathbf{\Gamma}_z\hat{\mathbf{e}}_z^k-\gamma\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\\ \mathbf{\Lambda}_{zb}^{-1}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+2})-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\end{bmatrix}\\
&+\gamma\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k\right]\\ 0\end{bmatrix}.
\end{aligned}$$

Then using Cauchy Schwartz inequality, we get

$$\begin{aligned}
&\|\hat{\mathbf{e}}_z^{k+1}\|^2\\
\leq&\left\|\mathbf{\Gamma}_z\hat{\mathbf{e}}_z^k-\gamma\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\\ \mathbf{\Lambda}_{zb}^{-1}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+2})-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\end{bmatrix}\right\|^2\\
&+\gamma^2\left\|\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k\right]\\ 0\end{bmatrix}\right\|^2-2\left\langle\mathbf{\Gamma}_z\hat{\mathbf{e}}_z^k,\gamma\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k\right]\\ 0\end{bmatrix}\right\rangle\\
&+\gamma^2\left\|\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\\ \mathbf{\Lambda}_{zb}^{-1}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+2})-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\end{bmatrix}\right\|^2\\
&+\gamma^2\left\|\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k\right]\\ 0\end{bmatrix}\right\|^2
\end{aligned}\tag{67}$$

To obtain the upper bound of the right-hand side of the above equation, we first estimate some individual terms in it as follows. Note that:

$$
\begin{aligned}
&\widetilde{\mathbb{E}}_k\|\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\|^2\\
=&\sum_{i=1}^n\widetilde{\mathbb{E}}_k\left\|\nabla^2_{22}g_i(x_i^k,y_i^{k+1})z_i^k-\nabla_2 f_i(x_i^k,y_i^{k+1})-\left(\nabla^2_{22}g_i(\bar{x}^k,\bar{y}^{k+1})z_k^\star-\nabla_2 f_i(\bar{x}^k,\bar{y}^{k+1})\right)\right\|^2\\
\le&3\sum_{i=1}^n\widetilde{\mathbb{E}}_k\left\|\nabla^2_{22}g_i(x_i^k,y_i^{k+1})(z_i^k-z_\star^k)\right\|^2+3\sum_{i=1}^n\widetilde{\mathbb{E}}_k\left\|(\nabla^2_{22}g_i(x_i^k,y_i^{k+1})-\nabla^2_{22}g_i(\bar{x}^k,\bar{y}^{k+1}))z_\star^k\right\|^2\\
&+3\sum_{i=1}^n\widetilde{\mathbb{E}}_k\left\|\nabla_2 f_i(\bar{x}^k,\bar{y}^{k+1})-\nabla_2 f_i(x_i^k,y_i^{k+1})\right\|^2\\
\le&6L_{g,1}^2(\|\mathbf{z}^k-\bar{\mathbf{z}}^k\|^2+\|\bar{\mathbf{z}}^k-\mathbf{z}_\star^k\|^2)+3\left(L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+L_{f,1}^2\right)\left(\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2+\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\|^2\right)
\end{aligned}
\tag{68}
$$

and

$$
\begin{aligned}
&\widetilde{\mathbb{E}}_k\|\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+2})-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\|^2\\
=&\sum_{i=1}^n\widetilde{\mathbb{E}}_k\|\nabla^2_{22}g_i(\bar{x}^{k+1},\bar{y}^{k+2})z_\star^{k+1}-\nabla_2 f_i(\bar{x}^{k+1},\bar{y}^{k+2})-\nabla^2_{22}g_i(\bar{x}^k,\bar{y}^{k+1})z_\star^k+\nabla_2 f_i(\bar{x}^k,\bar{y}^{k+1})\|^2\\
\le&3\sum_{i=1}^n\widetilde{\mathbb{E}}_k\|(\nabla^2_{22}g_i(\bar{x}^{k+1},\bar{y}^{k+2})-\nabla^2_{22}g_i(\bar{x}^k,\bar{y}^{k+1}))z_\star^{k+1}\|^2\\
&+3\sum_{i=1}^n\widetilde{\mathbb{E}}_k\|\nabla^2_{22}g_i(\bar{x}^k,\bar{y}^{k+1})(z_\star^{k+1}-z_\star^k)\|^2+3\sum_{i=1}^n\widetilde{\mathbb{E}}_k\|\nabla_2 f_i(\bar{x}^{k+1},\bar{y}^{k+2})-\nabla_2 f_i(\bar{x}^k,\bar{y}^{k+1})\|^2\\
\le&3\widetilde{\mathbb{E}}_k\left[\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)(\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2+\|\bar{\mathbf{y}}^{k+2}-\bar{\mathbf{y}}^{k+1}\|^2)+L_{g,1}^2L_{z^\star}^2\|\bar{\mathbf{x}}^k-\bar{\mathbf{x}}^{k-1}\|^2\right].
\end{aligned}
\tag{69}
$$

Then we present the bound of the right-hand side of (67). For the first term, we have the following evaluations:

$$
\begin{aligned}
&\widetilde{\mathbb{E}}_k\left[\left\|\mathbf{\Gamma}_z\hat{\mathbf{e}}_z^k-\gamma\mathbf{O}_z^{-1}\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\\ \mathbf{\Lambda}_{zb}^{-1}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+2})-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\end{bmatrix}\right\|\right]\\
\le&\|\mathbf{\Gamma}_z\|\|\hat{\mathbf{e}}_z^k\|^2+\frac{\gamma^2\|\mathbf{O}_z^{-1}\|^2}{1-\|\mathbf{\Gamma}_z\|}\widetilde{\mathbb{E}}_k\left[\left\|\begin{bmatrix}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\\ \mathbf{\Lambda}_{zb}^{-1}\mathbf{\Lambda}_{za}\hat{\mathbf{U}}_z^\top\left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+2})-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\right]\end{bmatrix}\right\|^2\right]\\
\le&\|\mathbf{\Gamma}_z\|\|\hat{\mathbf{e}}_z^k\|^2+\frac{\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\widetilde{\mathbb{E}}_k\left[\|\widetilde{\mathbb{E}}_k[\mathbf{p}^k]-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\|^2\right]\\
&+\frac{\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{1-\|\mathbf{\Gamma}_z\|}\widetilde{\mathbb{E}}_k\left[\|\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1},\bar{\mathbf{y}}^{k+2})-\mathbf{p}^k(\bar{\mathbf{x}}^k,\bar{\mathbf{y}}^{k+1})\|^2\right]
\end{aligned}
\tag{70}
$$

where the first inequality use Jensen's inequality and the second inequality use $\|\hat{\mathbf{U}}_z^\top\|\le 1$.

For the second term, since $\mathbf{z}^k, \mathbf{y}^{k+1} \in \mathcal{U}_k$, we have:

$$\begin{aligned}
&\widetilde{\mathbb{E}}_k \left\| \mathbf{O}_z^{-1} \begin{bmatrix} \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k \right] \\ 0 \end{bmatrix} \right\|^2 \\
\leq & \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \widetilde{\mathbb{E}}_k \left[\left\| \widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k \right\|^2 \right] \\
\leq & 2\|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 (\|\mathbf{z}^k\|^2 \sigma_{g,2}^2 + n\sigma_{f,1}^2) \\
\leq & 6\|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \left(\left(\|\mathbf{z}^k - \bar{\mathbf{z}}^k\|^2 + \|\bar{\mathbf{z}}^k - \mathbf{z}_\star^k\|^2 + n \frac{L_{f,1}^2}{\mu_g^2} \right) \sigma_{g,2}^2 + n\sigma_{f,1}^2 \right).
\end{aligned} \tag{71}$$

For the third term, we have:

$$2\widetilde{\mathbb{E}}_k \left\langle \mathbf{\Gamma}_z \hat{\mathbf{e}}_z^k, \gamma \mathbf{O}_z^{-1} \begin{bmatrix} \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k \right] \\ 0 \end{bmatrix} \right\rangle = 0, \tag{72}$$

since $\hat{\mathbf{e}}_z^k \in \mathcal{U}_k$.

Next, for the last two terms, we have:

$$\begin{aligned}
&\widetilde{\mathbb{E}}_k \left[\left\| \mathbf{O}_z^{-1} \begin{bmatrix} \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1}) \right] \\ \mathbf{\Lambda}_{zb}^{-1} \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1}) \right] \end{bmatrix} \right\|^2 \right] \\
\leq & \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \widetilde{\mathbb{E}}_k \left[\|\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\|^2 \right] \\
& + \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \|\mathbf{\Lambda}_{zb}^{-1}\|^2 \widetilde{\mathbb{E}}_k \left[\|\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\|^2 \right].
\end{aligned} \tag{73}$$

$$\widetilde{\mathbb{E}}_k \left[\left\| \mathbf{O}_z^{-1} \begin{bmatrix} \mathbf{\Lambda}_{za} \hat{\mathbf{U}}_z^\top \left[\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k \right] \\ 0 \end{bmatrix} \right\|^2 \right] \leq \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \widetilde{\mathbb{E}}_k \left[\|\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k\|^2 \right]. \tag{74}$$

Taking the expectation $\widetilde{\mathbb{E}}_k$ on both sides of (67) and plugging (70), (71), (72), (73) and (74) into it, we obtain:

$$\begin{aligned}
&\widetilde{\mathbb{E}}_k \left[\|\hat{\mathbf{e}}_z^{k+1}\|^2 \right] \\
\leq & \|\mathbf{\Gamma}_z\| \|\hat{\mathbf{e}}_z^k\|^2 + \frac{2\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2}{1 - \|\mathbf{\Gamma}_z\|} \widetilde{\mathbb{E}}_k \left[\|\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\|^2 \right] \\
& + \frac{2\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \|\mathbf{\Lambda}_{zb}^{-1}\|^2}{1 - \|\mathbf{\Gamma}_z\|} \widetilde{\mathbb{E}}_k \left[\|\mathbf{p}^{k+1}(\bar{\mathbf{x}}^{k+1}, \bar{\mathbf{y}}^{k+2}) - \mathbf{p}^k(\bar{\mathbf{x}}^k, \bar{\mathbf{y}}^{k+1})\|^2 \right] \\
& + 2\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \widetilde{\mathbb{E}}_k \|\widetilde{\mathbb{E}}_k[\mathbf{p}^k] - \mathbf{p}^k\|^2 \\
\leq & \|\mathbf{\Gamma}_z\| \|\hat{\mathbf{e}}_z^k\|^2 + 12n\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \left(\frac{L_{f,0}^2}{\mu_g^2} \sigma_{g,2}^2 + \sigma_{f,1}^2 \right) \\
& + \frac{12\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2}{1 - \|\mathbf{\Gamma}_z\|} (L_{g,1}^2 + (1 - \|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2) \widetilde{\mathbb{E}}_k \left[\|\mathbf{O}_z\|^2 \|\hat{\mathbf{e}}_z^k\|^2 + \|\bar{\mathbf{z}}^k - \mathbf{z}_\star^k\|^2 \right] \\
& + \frac{6\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2}{1 - \|\mathbf{\Gamma}_z\|} \left(L_{g,2}^2 \frac{L_{f,0}^2}{\mu_g^2} + L_{f,1}^2 \right) \widetilde{\mathbb{E}}_k \left[\|\mathbf{O}_x\|^2 \|\hat{\mathbf{e}}_x^k\|^2 + \|\mathbf{O}_y\|^2 \|\hat{\mathbf{e}}_y^{k+1}\|^2 \right] \\
& + \frac{6\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{zb}^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2}{1 - \|\mathbf{\Gamma}_z\|} \left(L_{f,1}^2 + L_{g,2}^2 \frac{L_{f,0}^2}{\mu_g^2} \right) \widetilde{\mathbb{E}}_k \left[\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 + \|\bar{\mathbf{y}}^{k+2} - \bar{\mathbf{y}}^{k+1}\|^2 \right] \\
& + \frac{6\gamma^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{zb}^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2}{1 - \|\mathbf{\Gamma}_z\|} L_{g,1}^2 L_{z^\star}^2 \widetilde{\mathbb{E}}_k \left[\|\bar{\mathbf{x}}^k - \bar{\mathbf{x}}^{k-1}\|^2 \right],
\end{aligned}$$

where the second inequality uses (36), (68), (69), and (71).

Thanks to

$$\frac{6\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{O}_z\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\cdot(2L^2+2(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)\leq\frac{1-\|\mathbf{\Gamma}_z\|}{4},$$

we have:

$$\begin{aligned}
&\widetilde{\mathbb{E}}_k\left[\|\hat{\mathbf{e}}_z^{k+1}\|^2\right]\\
\leq&\frac{1+3\|\mathbf{\Gamma}_z\|}{4}\|\hat{\mathbf{e}}_z^k\|^2+\frac{12\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\widetilde{\mathbb{E}}_k\left[\|\bar{\mathbf{z}}^k-\mathbf{z}_\star^k\|^2\right]\\
&+12n\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+\frac{1-\|\mathbf{\Gamma}_z\|}{4\|\mathbf{O}_z\|^2}\kappa^2(\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2)\\
&+\frac{6\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\widetilde{\mathbb{E}}_k\left[\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2+\|\bar{\mathbf{y}}^{k+2}-\bar{\mathbf{y}}^{k+1}\|^2\right]\\
&+\frac{6\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}L_{g,1}^2L_{z^\star}^2\widetilde{\mathbb{E}}_k\left[\|\bar{\mathbf{x}}^k-\bar{\mathbf{x}}^{k-1}\|^2\right].
\end{aligned}$$

Taking summation and expectation on both sides, we get:

$$\begin{aligned}
&\frac{3}{4}(1-\|\mathbf{\Gamma}_z\|)\sum_{k=0}^K\mathbb{E}\left[\|\hat{\mathbf{e}}_z^k\|^2\right]\\
\leq&\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2-\mathbb{E}\left[\|\hat{\mathbf{e}}_z^{k+1}\|^2\right]\\
&+\frac{12\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=0}^K\mathbb{E}\left[\|\bar{\mathbf{z}}^k-\mathbf{z}_\star^k\|^2\right]\\
&+\frac{1-\|\mathbf{\Gamma}_z\|}{4\|\mathbf{O}_z\|^2}\kappa^2\sum_{k=0}^K\mathbb{E}(\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2)\\
&+12n\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+\frac{6\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=0}^K\mathbb{E}\left[\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+L_{g,1}^2L_{z^\star}^2\right)\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2\right]\\
&+\frac{6\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=0}^K\mathbb{E}\left[\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\|\bar{\mathbf{y}}^{k+2}-\bar{\mathbf{y}}^{k+1}\|^2\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{k=0}^{K+1}\mathbb{E}\left[\|\hat{\mathbf{e}}_z^k\|^2\right]\\
\leq&\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^K\mathbb{E}\left[\|\bar{\mathbf{z}}^k-\mathbf{z}_\star^k\|^2\right]+\frac{2\mathbb{E}[\|\hat{\mathbf{e}}_z^0\|^2]}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{8\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^K\mathbb{E}\left[\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+L_{g,1}^2L_{z^\star}^2\right)\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2\right]\\
&+\frac{8\gamma^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^K\mathbb{E}\left[\left(L_{f,1}^2+L_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\|\bar{\mathbf{y}}^{k+2}-\bar{\mathbf{y}}^{k+1}\|^2\right]
\end{aligned}$$

$$+ 16(K+1)n\gamma^2 \frac{\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)$$
$$+ \frac{\kappa^2}{3\|\mathbf{O}_z\|^2}\sum_{k=0}^{K}\mathbb{E}(\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2).$$

□

Lemma 14 (Consensus error of x)**.** *Suppose that Assumptions 1- 4 and Lemmas 4, 5, and 7 hold. We have*

$$\begin{aligned}
&\sum_{k=0}^{K+1}\mathbb{E}\|\hat{\mathbf{e}}_x^k\|^2\\
\leq&\frac{\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+\frac{6n\alpha^2\theta(K+1)\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\left(\sigma_{f,1}^2+3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
&+\frac{\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}\left[\Delta_k+nI_k\right]\\
&+\frac{2\widetilde{L}^2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left(\|\mathbf{\Lambda}_{xb}^{-1}\|^2+\frac{2(1-\theta)^2}{\theta^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\right\|^2\right].
\end{aligned}\tag{75}$$

Proof. Firstly, the term $\|\hat{\mathbf{e}}_x^{k+1}\|^2$ can be deformed as

$$\begin{aligned}
&\|\hat{\mathbf{e}}_x^{k+1}\|^2\\
=&\left\|\mathbf{\Gamma}_x\hat{\mathbf{e}}_x^k-\alpha\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbf{r}^{k+1}-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\end{bmatrix}\right\|^2\\
=&\left\|\mathbf{\Gamma}_x\hat{\mathbf{e}}_x^k-\alpha\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\end{bmatrix}\right\|^2\\
&+\alpha^2\mathbb{E}_k\left[\left\|\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\right]\\ 0\end{bmatrix}\right\|^2\right]\\
&-2\left\langle\mathbf{\Gamma}_x\hat{\mathbf{e}}_x^k,\alpha\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\right]\\ 0\end{bmatrix}\right\rangle\\
&+2\alpha^2\left\langle\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\end{bmatrix},\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\right]\\ 0\end{bmatrix}\right\rangle
\end{aligned}\tag{76}$$

due to Eq. (35).

Then, for the first term of the right-hand side of (76), we use Jensen's Inequality and get:

$$
\begin{aligned}
&\mathbb{E}_k\left[\left\|\mathbf{\Gamma}\hat{\mathbf{e}}_x^k-\alpha\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\end{bmatrix}\right\|^2\right]\\
\leq&\|\mathbf{\Gamma}_x\|\|\hat{\mathbf{e}}_x^k\|^2+\frac{\alpha^2\|\mathbf{O}_x^{-1}\|^2}{1-\|\mathbf{\Gamma}_x\|}\mathbb{E}_k\left[\left\|\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\end{bmatrix}\right\|^2\right]\\
\leq&\|\mathbf{\Gamma}_x\|\|\hat{\mathbf{e}}_x^k\|^2+\frac{\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\mathbb{E}_k\left[\left\|\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right],\\
&+\frac{\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}{1-\|\mathbf{\Gamma}_x\|}\mathbb{E}_k\left[\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right].
\end{aligned}
\tag{77}
$$

For the second term in the right-hand side of (76), we have:

$$
\begin{aligned}
\mathbb{E}_k\left\|\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\right]\\ 0\end{bmatrix}\right\|^2\leq&\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\mathbb{E}_k\|\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\|^2\\
=&\theta^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\mathbb{E}_k\|\mathbb{E}_k[\mathbf{u}^k]-\mathbf{u}^k\|^2.
\end{aligned}
\tag{78}
$$

Like (72), we have:

$$
\mathbb{E}_k\left[\left\langle\mathbf{\Gamma}_x\hat{\mathbf{e}}_x^k,\alpha\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\right]\\ 0\end{bmatrix}\right\rangle\right]=0.
\tag{79}
$$

Next, for the last term, we have:

$$
\begin{aligned}
&2\alpha^2\mathbb{E}_k\left\langle\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\\ \mathbf{\Lambda}_{xb}^{-1}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right]\end{bmatrix},\mathbf{O}_x^{-1}\begin{bmatrix}\mathbf{\Lambda}_{xa}\hat{\mathbf{U}}_x^\top\left[\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\right]\\ 0\end{bmatrix}\right\rangle\\
\leq&\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\mathbb{E}_k\left[\|\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\|^2+\|\mathbb{E}_k[\mathbf{r}^{k+1}]-\mathbf{r}^{k+1}\|^2\right]\\
&+\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\mathbb{E}_k\left[\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\|^2\right].
\end{aligned}
\tag{80}
$$

Taking the expectation on both sides of (76), and plugging (77), (78), (79), (80) into it, we obtain:

$$
\begin{aligned}
&\mathbb{E}_k\|\hat{\mathbf{e}}_x^{k+1}\|^2\\
\leq&\|\mathbf{\Gamma}_x\|\|\hat{\mathbf{e}}_x^k\|^2+2\alpha^2\theta^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\mathbb{E}_k\|\mathbb{E}_k[\mathbf{u}^k]-\mathbf{u}^k\|^2\\
&+\frac{2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\mathbb{E}_k\left[\|\mathbb{E}_k[\mathbf{r}^{k+1}]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\|^2+\|\mathbf{\Lambda}_{xb}^{-1}\|^2\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^{k+1})-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\|^2\right].
\end{aligned}
$$

Taking expectation and summation on both sides, we obtain:

$$
\begin{aligned}
&(1-\|\mathbf{\Gamma}_x\|)\sum_{k=0}^{K}\mathbb{E}\|\hat{\mathbf{e}}_x^k\|^2\\
\leq&\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2-\mathbb{E}\|\hat{\mathbf{e}}_x^{k+1}\|^2+\frac{2\widetilde{L}^2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2\\
&+\frac{2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2+2\sum_{k=0}^{K}\mathbb{E}\left[\left\|\mathbb{E}_k[\mathbf{u}^k]-\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^k)\right\|^2\right]\right)\\
&+\frac{2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\cdot\frac{2\widetilde{L}^2(1-\theta)^2}{\theta^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\right\|^2\right]\\
&+\left[2\alpha^2\theta^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2+\frac{2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\theta(1-\theta)\right]\sum_{k=0}^{K}\mathbb{E}\|\mathbb{E}_k[\mathbf{u}^k]-\mathbf{u}^k\|^2\\
\leq&\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2-\mathbb{E}\|\hat{\mathbf{e}}_x^{k+1}\|^2+\frac{2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+6(K+1)n\alpha^2\theta\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\left(\sigma_{f,1}^2+3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
&+\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}\left[\Delta_k+nI_k\right]\\
&+\frac{2\widetilde{L}^2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\|\mathbf{\Lambda}_{xb}^{-1}\|^2+\frac{2(1-\theta)^2}{\theta^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\right\|^2\right].
\end{aligned}
$$

where the first inequality uses $\widetilde{L}$- Lipschitz continuity of $\widetilde{\nabla}\mathbf{\Phi}$, Lemma 7, and the second inequality uses Lemma 4 , Lemma 5 and

$$\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\|^2\leq\Delta_k,\quad\|\bar{\mathbf{z}}^{k+1}-\mathbf{z}_\star^{k+1}\|^2\leq nI_k.$$

Hence we get

$$
\begin{aligned}
&\sum_{k=0}^{K+1}\mathbb{E}\|\hat{\mathbf{e}}_x^k\|^2\\
\leq&\frac{\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+\frac{6n\alpha^2\theta(K+1)\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\left(\sigma_{f,1}^2+3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
&+\frac{\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}\left[\Delta_k+nI_k\right]\\
&+\frac{2\widetilde{L}^2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left(\|\mathbf{\Lambda}_{xb}^{-1}\|^2+\frac{2(1-\theta)^2}{\theta^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\right\|^2\right].
\end{aligned}
$$

□

The following lemma gather the consensus analysis of x, y, z together:

Lemma 15. *Take*

$$\eta_1=\frac{3\kappa^2\beta^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2}{(1-\|\mathbf{\Gamma}_y\|)^2}\|\mathbf{\Lambda}_{yb}^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2L_{g,1}^2+16\gamma^2L^2\left(2\kappa^2+L_{z^\star}^2\right)\frac{\|\mathbf{O}_z\|^2\|\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}$$

$$+4\kappa^2\widetilde{L}^2\left(1+\frac{(1-\theta)^2}{\theta^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}\right)\alpha^2\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2},$$

$$\eta_2 = 3\kappa^2 L_{g,1}^2\beta^2\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_y\|)^2}+16L^2\left(2\kappa^2+L_{z^\star}^2\right)\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}.$$

Suppose that Assumptions 1- 4 and Lemmas 12, 13, 14 hold, and α, β satisfy

$$\begin{aligned}\alpha^2 &\le \frac{(1-\|\mathbf{\Gamma}_x\|)^2}{24\kappa^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2\left[80L^2+18\theta(1-\|\mathbf{\Gamma}_x\|)\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right]},\\ \eta_2\beta^2 &\le \frac{1}{1248L_{g,1}^2}.\end{aligned} \tag{82}$$

We have:

$$\begin{aligned}&\frac{1}{4}\sum_{k=0}^{K}\mathbb{E}[\Delta_k] \\ \le&(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^{k+1}\|^2+\frac{32L_{g,1}^2\eta_2\beta}{\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+3(K+1)\eta_2\beta^2\sigma_{g,1}^2\\ &+\frac{\kappa^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2\alpha^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}[nI_k]\\ &+\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\cdot\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=-1}^{K}\mathbb{E}[nI_k]\\ &+\frac{3\kappa^2\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{2\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{2\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}\\ &+\frac{\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{2\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\ &+16(K+1)n\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\ &+24(K+1)n\kappa^2\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right).\end{aligned} \tag{83}$$

Proof. Adding (59), (66) and (75) together, we get:

$$\begin{aligned}&\kappa^2\|\mathbf{O}_x\|^2\sum_{k=0}^{K+1}\mathbb{E}[\|\hat{\mathbf{e}}_x^k\|^2]+\kappa^2\|\mathbf{O}_y\|^2\sum_{k=0}^{K+1}\mathbb{E}[\|\hat{\mathbf{e}}_y^k\|^2]+\|\mathbf{O}_z\|^2\sum_{k=0}^{K+1}\mathbb{E}[\|\hat{\mathbf{e}}_z^k\|^2]\\ \le&3\kappa^2\sum_{k=0}^{K}\frac{\beta^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2}{(1-\|\mathbf{\Gamma}_y\|)^2}\|\mathbf{\Lambda}_{yb}^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2L_{g,1}^2\mathbb{E}\left[\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2+\|\bar{\mathbf{y}}^{k+1}-\bar{\mathbf{y}}^k\|^2\right]\\ &+\frac{\kappa^2\|\mathbf{O}_x\|^2}{3}\sum_{k=0}^{K}\mathbb{E}\left[\|\hat{\mathbf{e}}_x^k\|^2\right]+\frac{3\kappa^2(K+1)\beta^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{2\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}\\ &+\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}\|)\sigma_{g,2}^2)\|\mathbf{O}_z^{-1}\|^2\|\mathbf{O}_z\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{\mathbf{z}}^k-\mathbf{z}_\star^k\|^2\right]\\ &+\frac{\kappa^2}{3}\sum_{k=0}^{K}\mathbb{E}(\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\|\mathbf{O}_y\|^2\|\hat{\mathbf{e}}_y^{k+1}\|^2)\\ &+\frac{8\gamma^2(2\kappa^2+L_{z^\star}^2)L^2\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\|^2\right]\\ &+\frac{16\gamma^2\kappa^2L^2\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\sum_{k=0}^{K}\mathbb{E}\left[\|\bar{\mathbf{y}}^{k+2}-\bar{\mathbf{y}}^{k+1}\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&+16(K+1)n\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)+\frac{2\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{2\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+\frac{6n\kappa^2\alpha^2\theta(K+1)\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\left(\sigma_{f,1}^2+3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
&+\frac{\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}[\Delta_k+nI_k]\\
&+\frac{2\kappa^2\widetilde{L}^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left(\|\mathbf{\Lambda}_{xb}^{-1}\|^2+\frac{2(1-\theta)^2}{\theta^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k\right\|^2\right]\\
\leq&\frac{3}{4}\sum_{k=0}^{K+1}\mathbb{E}\left(\kappa^2\|\mathbf{O}_x\|^2\|\hat{\mathbf{e}}_x^k\|^2+\kappa^2\|\mathbf{O}_y\|^2\|\|\hat{\mathbf{e}}_y^k\|^2+\|\mathbf{O}_z\|^2\|\|\hat{\mathbf{e}}_z^k\|^2\right)\\
&+(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^{k+1}\|^2+\eta_2\left(\frac{32L_{g,1}^2\beta}{\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+3(K+1)\beta^2\sigma_{g,1}^2\right)\\
&+\frac{\kappa^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2\alpha^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}[nI_k]\\
&+\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\cdot\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=-1}^{K}\mathbb{E}[nI_k]\\
&+\frac{3\kappa^2\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{2\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{2\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{2\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+16(K+1)n\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+24(K+1)n\kappa^2\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right).
\end{aligned}
$$

where the second inequality uses (54) and

$$
\eta_2L_{g,1}^2\beta^2\cdot 52+\frac{\kappa^2\alpha^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\leq\frac{1}{12}.
$$

Hence:

$$
\begin{aligned}
&\frac{1}{4}\sum_{k=0}^{K}\mathbb{E}[\Delta_k]\\
\leq&(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^{k+1}\|^2+\frac{32L_{g,1}^2\eta_2\beta}{\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+3(K+1)\eta_2\beta^2\sigma_{g,1}^2\\
&+\frac{\kappa^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2\alpha^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}[nI_k]\\
&+\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\cdot\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=-1}^{K}\mathbb{E}[nI_k]\\
&+\frac{3\kappa^2\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{2\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{2\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{2\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]
\end{aligned}
$$

$$+ 16(K+1)n\gamma^2 \frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)$$
$$+ 24(K+1)n\kappa^2\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right).$$

□

C.1.8 Proof of the main theorem

Before giving the final result of the convergence analysis, we present the following Lemma that combines the results in the analysis of I_k and Δ_k:

Lemma 16. *Suppose that Assumptions 1- 4 and Lemmas 6, 11, 15 hold. If $\alpha, \beta, \gamma, \theta$ satisfy*

$$\begin{aligned}
\alpha^2 &\leq \frac{(1-\|\mathbf{\Gamma}_x\|)^2}{16\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\left[80L^2+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)(1-\|\mathbf{\Gamma}_x\|)\sigma_{g,2}^2\right]\cdot 2040\kappa^6},\\
\beta^2\eta_2 &\leq \frac{1}{1024L_{g,1}^2},\\
\gamma^2 &\leq \frac{(1-\|\mathbf{\Gamma}_z\|)^2}{256\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)\cdot 2040\kappa^4},
\end{aligned} \tag{84}$$

and

$$40\left[4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\left(L^2+\frac{\theta\sigma_{g,2}^2}{n}\right)\leq\frac{1}{4080\kappa^4}, \tag{85}$$

then we have:

$$\begin{aligned}
\sum_{k=0}^{K}\mathbb{E}[\Delta_k+nI_k] \lesssim& \kappa^4(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2\left(\frac{n(\Phi(\bar{x}_0)-\inf\Phi)}{\alpha}+\theta(K+1)(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\right)\\
&+\left(\frac{\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\left(\frac{n(\Phi(\bar{x}_0)-\inf\Phi)}{\alpha}+\theta(K+1)(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\right)\\
&+\frac{\kappa^6\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\kappa^4\eta_2\beta^2(K+1)\sigma_{g,1}^2\\
&+\frac{\kappa^6\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{\kappa^4\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{\kappa^6\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}\\
&+\frac{\kappa^6\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\cdot\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\\
&+(K+1)\kappa^4\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2n}{1-\|\mathbf{\Gamma}_z\|}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+(K+1)\kappa^6\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2n}{1-\|\mathbf{\Gamma}_x\|}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+\frac{(K+1)\gamma}{\mu_g}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{n\|z_\star^1\|^2}{\mu_g\gamma}+\kappa^4\left(\frac{\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2}{\beta\mu_g}+\frac{K\sigma_{g,1}^2}{\mu_g}\beta\right).
\end{aligned}$$

Proof. Combining (57) and (83), we obtain

$$
\begin{aligned}
&\sum_{k=0}^{K}\mathbb{E}[\Delta_k]+\frac{1}{1020\kappa^4}\sum_{k=-1}^{K}\mathbb{E}[nI_k]\\
\leq&4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^{k+1}\|^2+\frac{128L_{g,1}^2\eta_2\beta}{\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+12(K+1)\eta_2\beta^2\sigma_{g,1}^2\\
&+4\frac{\kappa^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2\alpha^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}[nI_k]\\
&+\frac{4\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\cdot\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=-1}^{K}\mathbb{E}[nI_k]\\
&+\frac{12\kappa^2\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{8\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{8\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{4\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{8\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+64(K+1)n\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+96(K+1)n\kappa^2\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^k\|^2+\frac{1}{2}\sum_{k=0}^{K}\mathbb{E}\left[\Delta_k\right]+\frac{1}{1020\kappa^4}\cdot\frac{3n\|z_\star^1\|^2}{\mu_g\gamma}\\
&+\frac{(K+1)}{1020\kappa^4}\frac{6\gamma}{\mu_g}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)+\frac{73\kappa^4}{1020\kappa^4}\left(\frac{4}{\beta\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{\mu_g}\beta\right).
\end{aligned}
\tag{86}
$$

Subtracting the term

$$
\begin{aligned}
&4\frac{\kappa^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2\alpha^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{80L^2}{1-\|\mathbf{\Gamma}_x\|}+18\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\sigma_{g,2}^2\right)\sum_{k=0}^{K}\mathbb{E}[nI_k]\\
&+\frac{4\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\frac{16\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)}{1-\|\mathbf{\Gamma}_z\|}\sum_{k=-1}^{K}\mathbb{E}[nI_k]+\frac{1}{2}\sum_{k=0}^{K}\mathbb{E}[\Delta_k]
\end{aligned}
$$

from both sides of (86) and using the restriction of α,γ in (84), we can get:

$$
\begin{aligned}
&\frac{1}{2040\kappa^4}\left(\sum_{k=0}^{K}\mathbb{E}[\Delta_k]+\sum_{k=-1}^{K}\mathbb{E}[nI_k]\right)\\
\leq&4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^{k+1}\|^2+\frac{128L_{g,1}^2\eta_2\beta}{\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+12(K+1)\eta_2\beta^2\sigma_{g,1}^2\\
&+\frac{12\kappa^2\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{8\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{8\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{4\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{8\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+64(K+1)n\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+96(K+1)n\kappa^2\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^k\|^2+\frac{(K+1)}{1020\kappa^4}\cdot\frac{6\gamma}{\mu_g}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)\\
&+\frac{1}{1020\kappa^4}\cdot\frac{3n\|z_\star^1\|^2}{\mu_g\gamma}+\frac{1}{1020\kappa^4}\cdot 73\kappa^4\left(\frac{4}{\beta\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{\mu_g}\beta\right)\\
\leq&\left[4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{r}}^{k+1}\|^2\\
&+12(K+1)\eta_2\beta^2\sigma_{g,1}^2+\frac{12\kappa^2\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{8\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}\\
&+\frac{8\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{4\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{8\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+64(K+1)n\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+96(K+1)n\kappa^2\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+(K+1)\frac{18\gamma}{\mu_g\cdot 1020\kappa^4}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+\frac{1}{1020\kappa^4}\cdot\frac{3n\|z_\star^1\|^2}{\mu_g\gamma}+\frac{1}{1020\kappa^4}\cdot 73\kappa^4\left(\frac{8}{\beta\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{\mu_g}\beta\right),
\end{aligned}
$$

where the second inequality holds since

$$
\frac{128L_{g,1}^2\eta_2\beta}{\mu_g}\leq\frac{1}{8\beta\mu_g}.
$$

Then taking (41) into the concern, we know:

$$
\begin{aligned}
&\frac{1}{2040\kappa^4}\left(\sum_{k=0}^{K}\mathbb{E}[\Delta_k]+\sum_{k=-1}^{K}\mathbb{E}[nI_k]\right)\\
\leq&40\left[4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\left(L^2+\frac{\theta\sigma_{g,2}^2}{n}\right)\sum_{k=0}^{K}\mathbb{E}[\Delta_k+nI_k]\\
&+4\left[4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\frac{n(\Phi(\bar{x}_0)-\inf\Phi)}{\alpha}\\
&+12\theta(K+1)\left[4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\left(\sigma_{f,1}^2+2\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
&+12(K+1)\eta_2\beta^2\sigma_{g,1}^2+\frac{12\kappa^2\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{8\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}\\
&+\frac{8\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{4\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{8\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left[\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\right]\\
&+64(K+1)n\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+96(K+1)n\kappa^2\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+(K+1)\frac{18\gamma}{\mu_g\cdot 1020\kappa^4}\left(\frac{L_{f,0}^2}{\mu_g^2}\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+\frac{1}{1020\kappa^4}\cdot\frac{3n\|z_\star^1\|^2}{\mu_g\gamma}+\frac{1}{1020\kappa^4}\cdot 73\kappa^4\left(\frac{8}{\beta\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{\mu_g}\beta\right),
\end{aligned}
$$

Since

$$40\left[4(\eta_1+48\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\frac{1}{1020\kappa^4}\left(\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\left(L^2+\frac{\theta\sigma_{g,2}^2}{n}\right)\leq\frac{1}{4080\kappa^4},$$

it follows that

$$\begin{aligned}
&\sum_{k=0}^{K}\mathbb{E}[\Delta_k+nI_k]\\
\lesssim&\left[\kappa^4(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\left(\frac{\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\frac{n(\Phi(\bar{x}_0)-\inf\Phi)}{\alpha}\\
&+\theta(K+1)\left[\kappa^4(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\left(\frac{\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right](\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+\frac{\kappa^6\beta^2(K+1)\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\kappa^4\eta_2\beta^2(K+1)\sigma_{g,1}^2+\frac{\kappa^6\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}\\
&+\frac{\kappa^4\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{\kappa^6\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{\kappa^6\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\frac{1-\theta}{\theta}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\\
&+\left[\kappa^4\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2n}{1-\|\mathbf{\Gamma}_z\|}+\frac{\gamma}{\mu_g}\right](K+1)(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+(K+1)\kappa^6\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2n}{1-\|\mathbf{\Gamma}_x\|}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+\frac{n\|z_\star^1\|^2}{\mu_g\gamma}+\kappa^4\left(\frac{1}{\beta\mu_g}\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+\frac{K\sigma_{g,1}^2}{\mu_g}\beta\right).
\end{aligned}$$

Then, we finish the proof of this lemma. □

Finally, we can give the proof of Lemma 17, which is a detailed version of Theorem 1:

Lemma 17 (Detailed version of Theorem 1)**.** *Suppose that Assumptions 1- 4 hold. Then there exist constant step-sizes α, β, γ, θ, such that*

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\\
\lesssim&\frac{\kappa^5\sigma}{\sqrt{nK}}+\kappa^{\frac{16}{3}}\left[\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}+\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\right]\frac{\sigma^{\frac{2}{3}}}{K^{\frac{2}{3}}}\\
&+\kappa^{\frac{7}{2}}\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{2}}\frac{\sigma^{\frac{1}{2}}}{K^{\frac{3}{4}}}\\
&+\left[\kappa^{\frac{26}{5}}\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_y\|)^2}\right)^{\frac{1}{5}}+\kappa^6\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_z\|)^2}\right)^{\frac{1}{5}}\right]\frac{\sigma^{\frac{2}{5}}}{K^{\frac{4}{5}}}\\
&+\left[\kappa^{\frac{16}{3}}\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}+\kappa^{\frac{14}{3}}\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\right.\\
&\left.+\kappa^{\frac{8}{3}}\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{3}}\right]\frac{1}{K}+\left(\kappa C_\alpha+\kappa^4C_\theta\right)\frac{1}{K},
\end{aligned}$$

where $\sigma=\max\{\sigma_{f,1},\sigma_{g,1},\sigma_{g,2}\}$, C_α, C_θ are defined as:

$$\begin{aligned}
C_\alpha=&L_{\nabla\Phi}+\kappa^3\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|L}{1-\|\mathbf{\Gamma}_x\|}+\kappa^3L\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{2}}+\kappa^4\left(\frac{L_{g,1}^2}{\mu_g}+\frac{\sigma_{g,1}^2}{n\mu_g}\right)\\
&+\kappa^4\frac{\|\mathbf{O}_y\|\|\mathbf{O}_y^{-1}\|\|\mathbf{\Lambda}_{ya}\|L_{g,1}}{1-\|\mathbf{\Gamma}_y\|}+\kappa^{\frac{9}{2}}L_{g,1}\left(\frac{\|\mathbf{O}_y\|\|\mathbf{O}_y^{-1}\|\|\mathbf{\Lambda}_{ya}\|\|\mathbf{\Lambda}_{yb}^{-1}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{2}}+\kappa^4\left(\mu_g+\frac{\mu_g^2\sigma_{g,1}^2}{nL_{g,1}^2}\right)
\end{aligned}$$

$$+\kappa^6\frac{\|\mathbf{O}_z\|\|\mathbf{O}_z^{-1}\|\|\mathbf{\Lambda}_{za}\|\sqrt{L^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2}}{1-\|\mathbf{\Gamma}_z\|}+\kappa^{\frac{11}{2}}L\left(\frac{\|\mathbf{O}_z\|\|\mathbf{O}_z^{-1}\|\|\mathbf{\Lambda}_{za}\|\|\mathbf{\Lambda}_{zb}^{-1}\|}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{2}},$$

$$C_\theta=\frac{\sigma_{g,2}^2}{nL_{g,1}^2}+\frac{\sigma_{g,2}^2}{L^2}+1$$

Proof. Take $L_1=L^2+\left(\theta(1-\theta)+L_{\nabla\Phi}\alpha\theta^2\right)\frac{\sigma_{g,2}^2}{n}$ and use the conclusion of Lemmas 8 and 16, we get:

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\\
\lesssim&\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha(K+1)}+\frac{1}{n}\left(\theta(1-\theta)+L_{\nabla\Phi}\alpha\theta^2\right)(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{(1-\theta)^2}{\theta(K+1)}\|\nabla\Phi\left(\bar{x}^0\right)\|^2\\
&+L_1\left[\kappa^4(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\left(\frac{\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\left(\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha(K+1)}+\frac{\theta}{n}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\right)\\
&+L_1\frac{\kappa^6\beta^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\sigma_{g,1}^2+L_1\kappa^4\eta_2\beta^2\frac{\sigma_{g,1}^2}{n}\\
&+\frac{L_1}{K+1}\left[\frac{\kappa^6\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{n(1-\|\mathbf{\Gamma}_y\|)}+\frac{\kappa^4\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{n(1-\|\mathbf{\Gamma}_z\|)}+\frac{\kappa^6\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{n(1-\|\mathbf{\Gamma}_x\|)}\right]\\
&+L_1\frac{\kappa^6\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(K+1)(1-\|\mathbf{\Gamma}_x\|)^2}\cdot\frac{1-\theta}{\theta}\cdot\frac{\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2}{n}\\
&+L_1\left[\kappa^4\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{\gamma}{n\mu_g}\right](\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+L_1\kappa^6\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+L_1\frac{\|z_\star^1\|^2}{(K+1)\mu_g\gamma}+L_1\kappa^4\left(\frac{1}{\beta\mu_g(K+1)}\|\bar{y}_0-y^\star(\bar{x}^0)\|^2+\frac{\sigma_{g,1}^2}{n\mu_g}\beta\right).
\end{aligned}\tag{87}$$

Define:

$$\begin{aligned}
\zeta_0^y=&\frac{1}{n}\sum_{i=1}^{n}\|\nabla_2g_i(\bar{x}_0,\bar{y}_0)-\nabla_2g(\bar{x}_0,\bar{y}_0)\|^2,\\
\zeta_0^z=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla_{22}^2g_i(\bar{x}_0,\bar{y}_1)-\nabla_{22}^2g(\bar{x}_0,\bar{y}_1)\|^2\|z_\star^1\|^2+\|\nabla_2g_i(\bar{x}_0,\bar{y}_1)-\nabla_2g(\bar{x}_0,\bar{y}_1)\|^2\right],\\
\zeta_0^x=&\frac{1}{n}\sum_{i=1}^{n}\|\nabla_1f_i(\bar{x}_0,y^\star(\bar{x}_0))-\nabla_1f(\bar{x}_0,y^\star(\bar{x}_0))\|^2\\
&+\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{12}^2g_i(\bar{x}_0,y^\star(\bar{x}_0))-\nabla_{12}^2g(\bar{x}_0,y^\star(\bar{x}_0))\|^2\|z_\star^1\|^2,\\
\hat{\zeta}_0=&\frac{1}{n}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2.
\end{aligned}$$

Then we take:

$$\begin{aligned}
\alpha_1=&\kappa^{-4}\sqrt{\frac{n}{K\sigma^2}},\\
\alpha_{x,2}=&\left(\frac{(1-\|\mathbf{\Gamma}_x\|)^2}{\kappa^{10}K\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\sigma^2}\right)^{\frac{1}{4}}\\
\alpha_{y,2}=&\left(\frac{1-\|\mathbf{\Gamma}_y\|}{\kappa^{13}K\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\sigma_{g,1}^2}\right)^{\frac{1}{3}},
\end{aligned}\tag{88}$$

$$
\begin{aligned}
\alpha_{y,3} &= \left(\frac{n(1-\|\mathbf{\Gamma}_y\|)^2}{\kappa^{21} K \|\mathbf{O}_y\|^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 \|\mathbf{\Lambda}_{yb}^{-1}\|^2 \sigma_{g,1}^2} \right)^{\frac{1}{5}}, \\
\alpha_{z,2} &= \left(\frac{1-\|\mathbf{\Gamma}_z\|}{\kappa^{13} K \|\mathbf{O}_z\|^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \sigma^2} \right)^{\frac{1}{3}}, \\
\alpha_{z,3} &= \left(\frac{n(1-\|\mathbf{\Gamma}_z\|)^2}{\kappa^{25} K \|\mathbf{O}_z\|^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \|\mathbf{\Lambda}_{zb}^{-1}\|^2 \sigma_{g,1}^2} \right)^{\frac{1}{5}}, \\
\alpha_{yb,2} &= \left(\frac{1-\|\mathbf{\Gamma}_y\|}{\kappa^{13} \|\mathbf{O}_y\|^2 \|\mathbf{O}_y^{-1}\|^2 \|\mathbf{\Lambda}_{ya}\|^2 \|\mathbf{\Lambda}_{yb}^{-1}\|^2 \zeta_0^y} \right)^{\frac{1}{3}}, \\
\alpha_{zb,2} &= \left(\frac{1-\|\mathbf{\Gamma}_z\|}{\kappa^{11} \|\mathbf{O}_z\|^2 \|\mathbf{O}_z^{-1}\|^2 \|\mathbf{\Lambda}_{za}\|^2 \|\mathbf{\Lambda}_{zb}^{-1}\|^2 \zeta_0^z} \right)^{\frac{1}{3}}, \\
\alpha_{xb,2} &= \left(\frac{1-\|\mathbf{\Gamma}_x\|}{\kappa^{5} \|\mathbf{O}_x\|^2 \|\mathbf{O}_x^{-1}\|^2 \|\mathbf{\Lambda}_{xa}\|^2 \|\mathbf{\Lambda}_{xb}^{-1}\|^2 \zeta_0^x} \right)^{\frac{1}{3}}, \\
\theta_1 &= \left(\frac{n\kappa^2 \hat{\zeta_0}}{K\sigma^2} \right)^{\frac{1}{2}}, \\
\theta_2 &= \kappa^3 \alpha_{x,2},
\end{aligned}
$$

and

$$
\begin{aligned}
\theta =& \left(C_\theta + \frac{1}{\theta_1} + \frac{1}{\theta_2} \right)^{-1}, \\
\alpha =& \Theta \left(C_\alpha + \frac{\sqrt{1-\theta}}{\theta} \kappa^3 + \frac{1}{\alpha_1} + \frac{1}{\alpha_{y,2}} + \frac{1}{\alpha_{y,3}} + \frac{1}{\alpha_{z,3}} + \frac{1}{\alpha_{yb,2}} + \frac{1}{\alpha_{zb,2}} + \frac{1}{\alpha_{xb,2}} + \frac{1}{\alpha_{z,2}} + \frac{1}{\alpha_{x,2}} \right)^{-1}, \\
\beta =& \Theta \left(\kappa^4 \alpha \right), \\
\gamma =& \Theta \left(\kappa^4 \alpha \right),
\end{aligned}
\tag{89}
$$

It yields $L_1 = \Theta(L^2)$, and (45), (53), (56), (40), (58), (65), (82), (84), and (85) hold. It implies that the restrictions on the step-sizes $\alpha, \beta, \gamma, \theta$ in all previous lemma conditions hold. Thus all previous lemmas hold. We obtain:

$$
\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\\
\lesssim&\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha K}+\frac{\theta}{n}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{\hat{\zeta}_0}{\theta K}+\frac{\kappa^6\beta^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\sigma_{g,1}^2+\eta_2\kappa^4\beta^2\frac{\sigma_{g,1}^2}{n}\\
&+\frac{1}{K}\left[\frac{\kappa^6\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{n(1-\|\mathbf{\Gamma}_y\|)}+\frac{\kappa^4\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{n(1-\|\mathbf{\Gamma}_z\|)}+\frac{\kappa^6\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{n(1-\|\mathbf{\Gamma}_x\|)}\right]\\
&+\left[\kappa^4\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{\gamma}{n\mu_g}\right](\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+\kappa^6\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+\frac{\|z_\star^1\|^2}{(K+1)\mu_g\gamma}+\kappa^4\left(\frac{1}{\beta\mu_g(K+1)}\|\bar{y}_0-y^\star(\bar{x}^0)\|^2+\frac{\sigma_{g,1}^2}{n\mu_g}\beta\right)\\
\lesssim&\frac{\theta}{n}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{1}{\theta K}+\frac{\kappa}{\alpha K}+\frac{\kappa^9\sigma_{g,1}^2}{n}\alpha+\frac{\kappa^{14}\alpha^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\sigma_{g,1}^2\\
&+\left(\kappa^{10}\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_y\|)^2}+\kappa^{14}\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\right)\kappa^{12}\alpha^4\frac{\sigma_{g,1}^2}{n}\\
&+\alpha^2\frac{\kappa^{14}\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{K(1-\|\mathbf{\Gamma}_y\|)}+\alpha^2\frac{\kappa^{12}\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{K(1-\|\mathbf{\Gamma}_z\|)}\\
&+\alpha^2\frac{\kappa^6\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{K(1-\|\mathbf{\Gamma}_x\|)}\\
&+\left[\kappa^{12}\alpha^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}+\kappa^6\alpha^2\theta\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}+\frac{\kappa^5\alpha}{n}\right](\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
\lesssim&\frac{\theta_1}{n}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{\kappa^4}{\theta_1 K}+\frac{C_\theta\kappa^4}{K}+\frac{\kappa}{\alpha_1 K}+\frac{\kappa^9\sigma_{g,1}^2}{n}\alpha_1+\frac{\kappa^5\alpha_1}{n}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)\\
&+\frac{\kappa^{14}\alpha_{y,2}^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\sigma_{g,1}^2+\frac{\kappa}{\alpha_{y,2}K}\\
&+\kappa^{10}\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_y\|)^2}\kappa^{12}\alpha_{y,3}^4\frac{\sigma_{g,1}^2}{n}+\frac{\kappa}{\alpha_{y,3}K}\\
&+\kappa^{14}\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\kappa^{12}\alpha_{z,3}^4\frac{\sigma_{g,1}^2}{n}+\frac{\kappa}{\alpha_{z,3}K}\\
&+\alpha_{yb,2}^2\frac{\kappa^{14}\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{K(1-\|\mathbf{\Gamma}_y\|)}+\frac{\kappa}{\alpha_{yb,2}K}\\
&+\alpha_{zb,2}^2\frac{\kappa^{12}\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{K(1-\|\mathbf{\Gamma}_z\|)}+\frac{\kappa}{\alpha_{zb,2}K}\\
&+\alpha_{xb,2}^2\frac{\kappa^6\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{K(1-\|\mathbf{\Gamma}_x\|)}+\frac{\kappa}{\alpha_{xb,2}K}\\
&+\kappa^{12}\alpha_{z,2}^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{\kappa}{\alpha_{z,2}K}\\
&+\kappa^6\alpha_{x,2}^2\theta_2\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2)+\frac{\kappa}{\alpha_{x,2}K}+\frac{\kappa^4}{\theta_2 K},
\end{aligned}
$$

where the last inequality uses (89).

Finally, substituting (88) and (89) into the last inequality, we can get:

$$
\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla\Phi(\bar{x}^k)\|^2\\
\lesssim&\frac{\kappa^5\sigma}{\sqrt{nK}}+\kappa^{\frac{16}{3}}\left[\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}+\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\right]\frac{\sigma^{\frac{2}{3}}}{K^{\frac{2}{3}}}\\
&+\kappa^{\frac{7}{2}}\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{2}}\frac{\sigma^{\frac{1}{2}}}{K^{\frac{3}{4}}}\\
&+\left[\kappa^{\frac{26}{5}}\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_y\|)^2}\right)^{\frac{1}{5}}+\kappa^6\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_z\|)^2}\right)^{\frac{1}{5}}\right]\frac{\sigma^{\frac{2}{5}}}{K^{\frac{4}{5}}}\\
&+\left[\kappa^{\frac{16}{3}}\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}+\kappa^{\frac{14}{3}}\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\right.\\
&\left.+\kappa^{\frac{8}{3}}\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{3}}\right]\frac{1}{K}+\left(\kappa C_\alpha+\kappa^4 C_\theta\right)\frac{1}{K},
\end{aligned}
$$

where $\sigma=\max\{\sigma_{f,1},\sigma_{g,1},\sigma_{g,2}\}$. □

Remark 7. *From the proof of Lemma 17, the impact of the moving average technique on variance reduction becomes evident. The term $\frac{\theta}{n}\sigma^2$ absorb $\alpha^2\eta_1\sigma^2$, which includes the high order term $\alpha^4\sigma^2$. Additionally, compared to y,z, the quadratic term related to σ^2 of x has an extra term θ multiplied in the numerator ($\alpha^2\theta\sigma^2$). These details reduce the impacts of noise to terms related to x, confirming the conclusion that terms related to y,z dominate the rate in precious sections. Notably, taking $\theta<1$ is indispensable our proof. If we take $\theta=1$, there would be a constant term $\frac{1}{n}\sigma^2$ in the convergence rate (see the first inequality of (87)), since the coefficient $\alpha^2/\beta^2+\alpha^2/\gamma^2=\mathcal{O}(1)$. This would not guarantee the convergence of* SPARKLE.

C.2 Analysis of consensus error and transient iteration complexity

From Lemma 17, we can immediately obtain the transient time complexity of Algorithm 1. Here we omit the impacts of the condition number κ.

Lemma 18. *The transient time complexity of Algorithm 1 has an upper bound of:*

$$
\begin{aligned}
\max\Bigg\{&n^3\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right)^2\|\mathbf{\Lambda}_{ya}\|^2, n^3\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2}{1-\|\mathbf{\Gamma}_z\|}\right)^2\|\mathbf{\Lambda}_{za}\|^2,\\
&n^2\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}\right)^2\|\mathbf{\Lambda}_{xa}\|^2, n\left(\frac{\|\mathbf{O}_y\|\|\mathbf{O}_y^{-1}\|\|\mathbf{\Lambda}_{yb}^{-1}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{4}{3}}\|\mathbf{\Lambda}_{ya}\|,\\
&n\left(\frac{\|\mathbf{O}_z\|\|\mathbf{O}_z^{-1}\|\|\mathbf{\Lambda}_{zb}^{-1}\|}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{4}{3}}\|\mathbf{\Lambda}_{za}\|, n\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{2}{3}},\\
&n\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}, n\Bigg\}.
\end{aligned}
\tag{90}
$$

Proof. According to lemma 17, SPARKLE achieves linear speedup if:

$$
\frac{1}{\sqrt{nK}}\gtrsim\left[\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}+\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\right]\frac{1}{K^{\frac{2}{3}}}
$$

$$
\begin{aligned}
&+\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{2}}\frac{1}{K^{\frac{3}{4}}}\\
&+\left[\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_y\|)^2}\right)^{\frac{1}{5}}+\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_z\|)^2}\right)^{\frac{1}{5}}\right]\frac{1}{K^{\frac{4}{5}}}\\
&+\left[\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}+\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\right.\\
&\left.+\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{3}}\right]\frac{1}{K}+\left(C_\alpha+C_\theta\right)\frac{1}{K}.
\end{aligned}
$$

It holds when K satisfies:

$$
\begin{aligned}
\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}\frac{1}{K^{\frac{2}{3}}}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\frac{1}{K^{\frac{2}{3}}}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_y\|)^2}\right)^{\frac{1}{5}}\frac{1}{K^{\frac{4}{5}}}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{2}}\frac{1}{K^{\frac{3}{4}}}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{n(1-\|\mathbf{\Gamma}_z\|)^2}\right)^{\frac{1}{5}}\frac{1}{K^{\frac{4}{5}}}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}\frac{1}{K}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\frac{1}{K}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{3}}\frac{1}{K}&\lesssim\frac{1}{\sqrt{nK}},\\
\left(C_\alpha+C_\theta\right)\frac{1}{K}&\lesssim\frac{1}{\sqrt{nK}}.
\end{aligned}
$$

Then we get:

$$
\begin{aligned}
K\gtrsim\max\Bigg\{&n^3\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right)^2\|\mathbf{\Lambda}_{ya}\|^2,n^3\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2}{1-\|\mathbf{\Gamma}_z\|}\right)^2\|\mathbf{\Lambda}_{za}\|^2,\\
&n^2\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}\right)^2\|\mathbf{\Lambda}_{xa}\|^2,n\left(\frac{\|\mathbf{O}_y\|\|\mathbf{O}_y^{-1}\|\|\mathbf{\Lambda}_{yb}^{-1}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{4}{3}}\|\mathbf{\Lambda}_{ya}\|,\\
&n\left(\frac{\|\mathbf{O}_z\|\|\mathbf{O}_z^{-1}\|\|\mathbf{\Lambda}_{zb}^{-1}\|}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{4}{3}}\|\mathbf{\Lambda}_{za}\|,n\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{2}{3}},\\
&n\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}{1-\|\mathbf{\Gamma}_x\|},n\Bigg\}.
\end{aligned}
$$

□

C.2.1 Consensus Error

Lemma 19. *Suppose that Assumptions 1- 4 hold. Then there exist constant step-sizes $\alpha, \beta, \gamma, \theta$, such that Lemma 17 holds and*

$$\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2}{n}\right]$$
$$\lesssim_K \frac{n}{K}\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right),$$

where $\lesssim_K$ denotes the the asymptotic rate when $K\to\infty$.

Proof. Suppose α, β, γ, and θ satisfy the constraints given in (88) and (89), which ensures that Theorem 1 (Lemma 17) holds.

For clarity, we define the constants:

$$c_1=\frac{9\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{438\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2,\quad c_2=10\left(L^2+\frac{\theta\sigma_{g,2}^2}{n}\right).$$

Then there exist α, β, γ, and θ that satisfy the constraints in (88) and (89), and also:

$$c_1\le 0.01L^{-2},\quad c_2\le 11L^2. \tag{91}$$

We take such values for step-sizes in the following proof.

We proceed by substituting (41) into (57), yielding:

$$\begin{aligned}\sum_{k=-1}^{K}\mathbb{E}[I_k]\le&4c_1\left(\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha}+c_2\sum_{k=0}^{K-1}\mathbb{E}\left(\frac{\Delta_k}{n}+I_k\right)+\frac{3\theta}{n}K\left(\sigma_{f,1}^2+2\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\right)\\&+510\kappa^4\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}\right]+\frac{3\|z_\star^1\|^2}{\mu_g\gamma}\\&+\frac{6(K+1)\gamma}{\mu_g n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)+73\kappa^4\left(\frac{4}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{n\mu_g}\beta\right).\end{aligned}$$

Subtracting $4c_1c_2\sum_{k=0}^{K-1}\mathbb{E}[I_k]$ from both sides, we get:

$$\begin{aligned}\sum_{k=-1}^{K}\mathbb{E}[I_k]\lesssim&\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha}+\frac{\theta}{n}K\left(\sigma_{f,1}^2+\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)+\kappa^4\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}\right]+\frac{\|z_\star^1\|^2}{\mu_g\gamma}\\&+\frac{K\gamma}{\mu_g n}\left(\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)+\kappa^4\left(\frac{1}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\frac{K\sigma_{g,1}^2}{n\mu_g}\beta\right).\end{aligned}$$

Substituting (57) into (41), we obtain:

$$\begin{aligned}&\frac{1}{4}\sum_{k=0}^{K}\mathbb{E}\left\|\bar{r}^{k+1}\right\|^2\\\le&\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha}+c_2\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}\right]+c_2c_1\sum_{k=0}^{K}\mathbb{E}\|\bar{r}^k\|^2\\&+c_2\left[510\kappa^4\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}\right]+\frac{3\|z_\star^1\|^2}{\mu_g\gamma}+\frac{6(K+1)\gamma}{\mu_g n}\left(3\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)\right.\\&\left.+73\kappa^4\left(\frac{4}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\frac{4K\sigma_{g,1}^2}{n\mu_g}\beta\right)\right]\\&+\frac{3\theta}{n}(K+1)\left(\sigma_{f,1}^2+2\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right).\end{aligned}$$

Subtracting $c_2c_1\sum_{k=0}^K\mathbb{E}\|\bar{r}^k\|^2$ from both sides, we get

$$\begin{aligned}
&\sum_{k=0}^K\mathbb{E}\left\|\bar{r}^{k+1}\right\|^2\\
\lesssim&\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha}+\kappa^4\sum_{k=0}^K\mathbb{E}\left[\frac{\Delta_k}{n}\right]+\frac{\theta}{n}K\left(\sigma_{f,1}^2+\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}\right)\\
&+\frac{\|z_\star^1\|^2}{\mu_g\gamma}+\frac{K\gamma}{\mu_g n}\left(\sigma_{g,2}^2\frac{L_{f,0}^2}{\mu_g^2}+\sigma_{f,1}^2\right)+\kappa^4\left(\frac{1}{\beta\mu_g}\|\bar{y}^0-y^\star(\bar{x}^0)\|^2+\frac{K\sigma_{g,1}^2}{n\mu_g}\beta\right).
\end{aligned}$$

Taking

$$\eta_3=\left(\frac{\kappa^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{O}_x\|^2\|\mathbf{\Lambda}_{xa}\|^2\alpha^2}{(1-\|\mathbf{\Gamma}_x\|)^2}+\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\cdot\frac{\gamma^2(L_{g,1}^2+(1-\|\mathbf{\Gamma}_z\|)\sigma_{g,2}^2)}{1-\|\mathbf{\Gamma}_z\|}\right),$$

and combining previous results with (83), we obtain

$$\begin{aligned}
&\sum_{k=0}^K\mathbb{E}\left[\Delta_k\right]\\
\lesssim&(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2\sum_{k=0}^K\mathbb{E}\|\bar{\mathbf{r}}^{k+1}\|^2+\kappa\eta_2\beta\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+K\eta_2\beta^2\sigma_{g,1}^2+\eta_3\sum_{k=-1}^K\mathbb{E}[nI_k]\\
&+\frac{\kappa^2\beta^2K\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2+\frac{\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}\\
&+\frac{\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}+\frac{\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{\theta(1-\|\mathbf{\Gamma}_x\|)^2}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\\
&+Kn\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\kappa^2\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+Kn\kappa^2\alpha^2\theta\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left(\kappa^2\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
\lesssim&\left[(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\eta_3\right]\cdot\kappa^4\sum_{k=0}^K\mathbb{E}\left[\Delta_k\right]+\kappa\eta_2\beta\|\bar{\mathbf{y}}^0-\mathbf{y}^\star(\bar{x}^0)\|^2+K\eta_2\beta^2\sigma_{g,1}^2\\
&+n\left[(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2+\eta_3\right]\left[\frac{1}{\alpha}+\frac{\theta}{n}K\left(\sigma_{f,1}^2+\kappa^2\sigma_{g,2}^2\right)+\frac{1}{\mu_g\gamma}+\frac{K\gamma}{\mu_g n}\left(\kappa^2\sigma_{g,2}^2+\sigma_{f,1}^2\right)\right.\\
&\left.+\kappa^4\left(\frac{1}{\beta\mu_g}+\frac{K\sigma_{g,1}^2}{n\mu_g}\beta\right)\right]+\frac{\kappa^2\beta^2K\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}n\sigma_{g,1}^2\\
&+\frac{\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{\theta(1-\|\mathbf{\Gamma}_x\|)^2}\left\|\widetilde{\nabla}\mathbf{\Phi}(\bar{\mathbf{x}}^0)\right\|^2\\
&+\frac{\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{1-\|\mathbf{\Gamma}_y\|}+\frac{\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{1-\|\mathbf{\Gamma}_z\|}+\frac{\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{1-\|\mathbf{\Gamma}_x\|}\\
&+Kn\gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\kappa^2\sigma_{g,2}^2+\sigma_{f,1}^2\right)\\
&+Kn\kappa^2\alpha^2\theta\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left(\kappa^2\sigma_{g,2}^2+\sigma_{f,1}^2\right).
\end{aligned}$$

(88) and (89) imply that

$$\begin{aligned}
&\eta_1\lesssim\kappa^2+\kappa^2\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2},\quad\eta_2\lesssim\kappa^2,\\
&(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2\lesssim\kappa^{-4},\quad\eta_3\lesssim\kappa^{-4}
\end{aligned}$$

where η_1, η_2 are defined in Lemma 15.

Then taking $\alpha, \beta, \gamma, \theta$ such that (88), (89), (91) hold and $\kappa^4[(\eta_1 + \kappa^2 L_{y^\star}^2 \eta_2)\alpha^2 + \eta_3]$ is a sufficiently small constant, we can derive the following result:

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}\right] \\
\lesssim&\frac{\kappa\eta_2\beta}{K} + \eta_2\beta^2\frac{\sigma_{g,1}^2}{n} \\
&+ \left[(\eta_1 + \kappa^2 L_{y^\star}^2\eta_2)\alpha^2 + \eta_3\right]\left[\frac{1}{\alpha K} + \frac{\theta}{n}\left(\sigma_{f,1}^2 + \kappa^2\sigma_{g,2}^2\right) + \frac{1}{\mu_g\gamma K} + \frac{\gamma}{\mu_g n}\left(\kappa^2\sigma_{g,2}^2 + \sigma_{f,1}^2\right)\right] \\
&+ \left[(\eta_1 + \kappa^2 L_{y^\star}^2\eta_2)\alpha^2 + \eta_3\right]\kappa^4\left(\frac{1}{\beta\mu_g K} + \frac{\sigma_{g,1}^2}{n\mu_g}\beta\right) + \frac{\kappa^2\beta^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\sigma_{g,1}^2 \\
&+ \frac{\kappa^2\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{\theta K(1-\|\mathbf{\Gamma}_x\|)^2} + \frac{\kappa^2\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{(1-\|\mathbf{\Gamma}_y\|)Kn} + \frac{\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{(1-\|\mathbf{\Gamma}_z\|)Kn} + \frac{\kappa^2\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{(1-\|\mathbf{\Gamma}_x\|)Kn} \\
&+ \gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\kappa^2\sigma_{g,2}^2 + \sigma_{f,1}^2\right) + \kappa^2\alpha^2\theta\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left(\kappa^2\sigma_{g,2}^2 + \sigma_{f,1}^2\right) \\
\lesssim&\frac{\kappa^5\eta_2\alpha}{K} + \kappa^{10}\alpha^2\frac{\sigma_{g,1}^2}{n} + \frac{\kappa}{K}\left[(\eta_1 + \kappa^2 L_{y^\star}^2\eta_2)\alpha + \frac{\eta_3}{\alpha}\right] \\
&+ \left[(\eta_1 + \kappa^2 L_{y^\star}^2\eta_2)\alpha^2 + \eta_3\right]\left[\frac{\theta}{n}\left(\sigma_{f,1}^2 + \kappa^2\sigma_{g,2}^2\right) + \frac{\kappa^5\alpha}{n}\left(\kappa^2\sigma_{g,2}^2 + \sigma_{f,1}^2\right) + \kappa^9\frac{\sigma_{g,1}^2}{n}\alpha\right] \\
&+ \frac{\kappa^2\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\sigma_{g,1}^2\kappa^8\alpha^2 + \frac{\kappa^{-1}\alpha\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{K(1-\|\mathbf{\Gamma}_x\|)^2} \\
&+ \alpha^2\frac{\kappa^{10}\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{K(1-\|\mathbf{\Gamma}_y\|)} + \alpha^2\frac{\kappa^8\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{K(1-\|\mathbf{\Gamma}_z\|)} \\
&+ \alpha^2\frac{\kappa^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{K(1-\|\mathbf{\Gamma}_x\|)} \\
&+ \kappa^8\alpha^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|}\left(\kappa^2\sigma_{g,2}^2 + \sigma_{f,1}^2\right) \\
&+ \kappa^2\alpha^2\theta\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\left(\kappa^2\sigma_{g,2}^2 + \sigma_{f,1}^2\right).
\end{aligned}
$$

From (88) and (89), we can determine the asymptotic orders for α, β, γ and θ when $K \to \infty$

$$
\alpha = \mathcal{O}\left(\kappa^{-4}\sqrt{\frac{n}{K\sigma^2}}\right), \quad \beta = \mathcal{O}\left(\sqrt{\frac{n}{K\sigma^2}}\right), \quad \gamma = \mathcal{O}\left(\sqrt{\frac{n}{K\sigma^2}}\right), \quad \theta = \mathcal{O}\left(\kappa\sqrt{\frac{n}{K\sigma^2}}\right).
$$

Then we get

$$
\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\frac{\Delta_k}{n}\right] \lesssim_K \frac{\kappa^2 n}{K}\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|} + \frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right),
$$

where $\lesssim_K$ denotes the the asymptotic rate when $K \to \infty$.

Then using (36) and the definition of Δ_k, we get

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\frac{\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2}{n} + \frac{\|\mathbf{y}^k - \bar{\mathbf{y}}^k\|^2}{n}\right] \\
\lesssim_K&\frac{n}{K}\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2}{1-\|\mathbf{\Gamma}_z\|} + \frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right).
\end{aligned}
$$

In particular, the corresponding result of SPARKLE variants that using EXTRA, ED or GT is

$$\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\frac{\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|^2}{n}+\frac{\|\mathbf{y}^k-\bar{\mathbf{y}}^k\|^2}{n}\right]\lesssim_K \frac{n}{K}\left(\frac{1}{1-\rho_y}+\frac{1}{1-\rho_z}\right),$$

where ρ_y, ρ_z are spectrum gaps of relevant mixing matrices.

□

C.2.2 Essential matrix norms for analysis

Common heterogeneity-correction algorithms, including ED, EXTRA and GT, satisfy Assumption 3, according to transformations (31), (32) and discussions in [2, Appendix B.2]. Then Lemma 3 ensures that $\|\mathbf{\Gamma}\|<1$. From Lemma 18, the transient time complexity depends on the coefficients $\|\mathbf{O}\|^2$, $\|\mathbf{O}^{-1}\|^2$, $\|\mathbf{\Lambda}_a\|^2$, $\|\mathbf{\Lambda}_b^{-1}\|^2$, and $\|\mathbf{\Gamma}\|^2$. The solution of these matrices is constructive. Table 4 presents the upper bounds of these coefficients with different communication modes. Please refer to [2, Appendix B.2] for more details about the construction of these matrices and the computation of relevant norms. It is required that W is positive definite for ED, EXTRA, and we denote the smallest nonzero eigenvalue of W by $\underline{\rho}$. $\underline{\rho}$ can view as a constant. Otherwise we replace $\mathbf{W}$ with $t\mathbf{I}+(1-t)\mathbf{W}$ for some constant $t\in(0,1)$ (e.g. $t=1/2$).

Substituting values of $\|\mathbf{O}_s\|, \|\mathbf{O}_s^{-1}\|, \|\mathbf{\Lambda}_{sa}\|, \|\mathbf{\Lambda}_{sb}^{-1}\|, \|\mathbf{\Gamma}_s\|$ into (90) , we obtain the explicit transient iteration complexity for some specific examples of Algorithm 1, which are listed in Table 2. Note that all GT variants exhibit the same transient iteration complexity.

Table 4: Upper bounds of coefficients for different heterogeneity-correction modes in Lemma 18, where notation $\mathcal{O}$ is omitted for $\|\mathbf{O}\|$ and $\|\mathbf{O}^{-1}\|$.

Mode	A	B	C	$\|\mathbf{O}\|$	$\|\mathbf{O}^{-1}\|$	$\|\mathbf{\Lambda}_a\|$	$\|\mathbf{\Lambda}_b^{-1}\|$	$\|\mathbf{\Gamma}\|$
ED	$\mathbf{W}$	$(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}$	$\mathbf{W}$	1	$\underline{\rho}^{-\frac{1}{2}}$	ρ	$(1-\rho)^{-\frac{1}{2}}$	$\sqrt{\rho}$
EXTRA	$\mathbf{I}$	$(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}$	$\mathbf{W}$	1	$\underline{\rho}^{-\frac{1}{2}}$	1	$(1-\rho)^{-\frac{1}{2}}$	$\sqrt{\rho}$
ATC-GT	$\mathbf{W}^2$	$\mathbf{I}-\mathbf{W}$	$\mathbf{W}^2$	1	1	ρ^2	$(1-\rho)^{-1}$	$\frac{1+\rho}{2}$
Semi-ATC-GT	$\mathbf{W}$	$\mathbf{I}-\mathbf{W}$	$\mathbf{W}^2$	1	1	ρ	$(1-\rho)^{-1}$	$\frac{1+\rho}{2}$
Non-ATC-GT	$\mathbf{I}$	$\mathbf{I}-\mathbf{W}$	$\mathbf{W}^2$	1	1	1	$(1-\rho)^{-1}$	$\frac{1+\rho}{2}$

C.2.3 Theoretical gap between upper-level and lower-level

Note that $\|\mathbf{\Lambda}_{sa}\|\le 1$. We rewrite the upper bound of the transient iteration complexity in Lemma 18 as

$$\max\{n^3\delta_y, n^3\delta_z, n^2\delta_x, n\hat{\delta}_y, n\hat{\delta}_z, n\hat{\delta}_x\} \tag{92}$$

where

$$\begin{aligned}
&\delta_y=\left(\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2}{1-\|\mathbf{\Gamma}_y\|}\right)^2\|\mathbf{\Lambda}_{ya}\|^2, \delta_z=\left(\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2}{1-\|\mathbf{\Gamma}_z\|}\right)^2\|\mathbf{\Lambda}_{za}\|^2, \delta_x=\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|}{1-\|\mathbf{\Gamma}_z\|}\|\mathbf{\Lambda}_{za}\|\right)^2,\\
&\hat{\delta}_y=\left(\frac{\|\mathbf{O}_y\|\|\mathbf{O}_y^{-1}\|\|\mathbf{\Lambda}_{yb}^{-1}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{4}{3}}, \hat{\delta}_z=\left(\frac{\|\mathbf{O}_z\|\|\mathbf{O}_z^{-1}\|\|\mathbf{\Lambda}_{zb}^{-1}\|}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{4}{3}},\\
&\hat{\delta}_x=\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{2}{3}}+\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}.
\end{aligned} \tag{93}$$

Suppose that we use the same communication matrices and heterogeneity-correction methods for updating x, y, z, *i*.e.

$$\begin{aligned}
&\|\mathbf{O}_x\|=\|\mathbf{O}_y\|=\|\mathbf{O}_z\|, \|\mathbf{O}_x^{-1}\|=\|\mathbf{O}_y^{-1}\|=\|\mathbf{O}_z^{-1}\|, \|\mathbf{\Gamma}_x\|=\|\mathbf{\Gamma}_y\|=\|\mathbf{\Gamma}_z\|,\\
&\|\mathbf{\Lambda}_{xa}\|=\|\mathbf{\Lambda}_{ya}\|=\|\mathbf{\Lambda}_{za}\|, \|\mathbf{\Lambda}_{xb}^{-1}\|=\|\mathbf{\Lambda}_{yb}^{-1}\|=\|\mathbf{\Lambda}_{zb}^{-1}\|.
\end{aligned}$$

Then we have

$$\delta_x \lesssim \delta_y = \delta_z,\ \hat{\delta}_x \lesssim \hat{\delta}_y = \hat{\delta}_z, \tag{94}$$

Now we fix the update strategies for y, z. (94) implies that we can appropriately increase $\delta_x, \hat{\delta}_x$ while keeping the transient iteration complexity (92) unchanged (at most scaled by a constant factor). For example, we can use a moderately sparser communication network for updating x than y, z. We illustrate this point with three examples: SPARKLE-ED, SPARKLE-EXTRA and SPARKLE-GT (variants), where y, z share the same communication matrix $\mathbf{W}_y$.

- SPARKLE-ED, SPARKLE-EXTRA: From Table 4, we have

  $$\delta_x = \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-2}\right), \delta_y = \delta_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-2}\right),$$
  $$\hat{\delta}_x = \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-\frac{3}{2}}\right), \hat{\delta}_y = \hat{\delta}_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-2}\right).$$

  Substituting these values into (92), we get the transient iteration complexity is bounded by

  $$\max\left\{n^2(1-\rho(\mathbf{W}_x))^{-2}, n^3(1-\rho(\mathbf{W}_y))^{-2}\right\}$$

  SPARKLE-ED will keep the transient iteration complexity $n^3(1-\rho(\mathbf{W}_y))^{-2}$ (the dominated term) if

  $$(1-\rho(\mathbf{W}_x))^{-1} \lesssim \sqrt{n}(1-\rho(\mathbf{W}_y))^{-1}. \tag{95}$$

- SPARKLE-GT variants: Results in Table 4 imply that

  $$\delta_x = \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-2}\right), \delta_y = \delta_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-2}\right),$$
  $$\hat{\delta}_x = \mathcal{O}\left((1-\rho(\mathbf{W}_x))^{-2}\right), \hat{\delta}_y = \hat{\delta}_z = \mathcal{O}\left((1-\rho(\mathbf{W}_y))^{-\frac{8}{3}}\right).$$

  Following the same argument as before, we have the following upper bound of the transient iteration complexity of SPARKLE-GT

  $$\max\left\{n^2(1-\rho(\mathbf{W}_x))^{-2}, n^3(1-\rho(\mathbf{W}_y))^{-2}, n(1-\rho(\mathbf{W}_y))^{-\frac{8}{3}}\right\}.$$

  we get the constraints of the spectral gap $1-\rho(\mathbf{W}_x)$ that maintains the transient iteration complexity $\max\left\{n^3(1-\rho(\mathbf{W}_y))^{-2}, n(1-\rho(\mathbf{W}_y))^{-\frac{8}{3}}\right\}$:

  $$(1-\rho(\mathbf{W}_x))^{-1} \lesssim \max\left\{\sqrt{n}(1-\rho(\mathbf{W}_y))^{-1}, n^{-1/2}(1-\rho(\mathbf{W}_y))^{-\frac{4}{3}}\right\}. \tag{96}$$

Denote the communication times per agent of $\mathbf{W}_x, \mathbf{W}_y$ by c_x, c_y respectively. For example, we have $c_x = 2$, $c_y = n-1$ when taking Ring Graph for x (*i.e.* $[\mathbf{W}_x]_{ij} \neq 0$ iff $|i-j| \in \{0, 1, n-1\}$), and Complete Graph for y (*i.e.* $\mathbf{W}_y = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$).

Then for each agent, the communication cost per round is $\mathcal{O}(c_x p + c_y q)$. If we take $a = c_x/c_y$ to measure the relative sparsity of the two communication matrices, and consider $c_y = \mathcal{O}(1)$, then for each agent, the communication cost per round is $\mathcal{O}(ap+q)$. (95) and (96) theoretically provide the range of the sparsity (connectivity) degree of $\mathbf{W}_x$ relative to $\mathbf{W}_y$. From (95) and (96), we can set $a \ll 1$, while maintaining the transient iteration complexity for SPARKLE-GT, SPARKLE-ED, SPARKLE-EXTRA.

C.2.4 The transient iteration complexities of some specific examples in SPARKLE.

Now we compute the transient iteration complexities of each SPARKLE-**L**-**U** algorithm, where $\mathbf{L}, \mathbf{U} \in \{\text{GT (variants)}, \text{ED}, \text{EXTRA}\}$. For brevity, here we assume that $\mathbf{W}_x = \mathbf{W}_y = \mathbf{W}_z$, use the same heterogeneity-correction method to y, z, and denote the spectral gap $1-\rho(\mathbf{W}_x)$ by $1-\rho$.

Substituting the results in Table 4 into (92) and (93), we get

$$\delta_x = \mathcal{O}\left(\frac{1}{(1-\rho)^2}\right), \delta_y = \delta_z = \mathcal{O}\left(\frac{1}{(1-\rho)^2}\right)$$

for any $\mathbf{L}, \mathbf{U} \in \{\text{GT (variants)}, \text{ED}, \text{EXTRA}\}$,

$$\hat{\delta}_x = \mathcal{O}\left(\frac{1}{(1-\rho)^2}\right), \mathcal{O}\left(\frac{1}{(1-\rho)^{3/2}}\right), \mathcal{O}\left(\frac{1}{(1-\rho)^{3/2}}\right)$$

for $\mathbf{U} = \{\text{GT (variants)}, \text{ED}, \text{EXTRA}\}$ respectively, and

$$\hat{\delta}_y = \hat{\delta}_z = \mathcal{O}\left(\frac{1}{(1-\rho)^{8/3}}\right), \mathcal{O}\left(\frac{1}{(1-\rho)^2}\right), \mathcal{O}\left(\frac{1}{(1-\rho)^2}\right)$$

for $\mathbf{L} = \{\text{GT (variants)}, \text{ED}, \text{EXTRA}\}$ respectively.

Combining the above results, we can directly obtain Table 2, the transient iteration complexities of SPARKLE with mixed heterogeneity-correction techniques in different levels.

C.3 Convergence analysis in deterministic scenarios

The following lemma gives the convergence rate of Algorithm 1 without a moving average when there is no sample noise:

Lemma 20. *Suppose that Assumptions 1- 4 hold. If $\sigma^2 = 0$, then there exist α, β, γ and $\theta = 1$ such that*

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\Phi(\bar{x}^k)\|^2 \\
\lesssim&\left(\frac{\kappa^{16}\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}\frac{1}{K} + \left(\frac{\kappa^{14}\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\frac{1}{K} \\
&+\left(\frac{\kappa^{8}\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{3}}\frac{1}{K} + \widetilde{C}_\alpha\frac{1}{K}.
\end{aligned}$$

where $\widetilde{C}_\alpha$ is a series of overheads which is defined below.

Proof. Note that $\sigma^2 = 0$ implies that $L_1 = \Theta(L^2)$ when $\alpha = \mathcal{O}(L_{\nabla\Phi}^{-1})$. Thus (87) implies that:

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\Phi(\bar{x}^k)\|^2 \\
\lesssim&\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha(K+1)} \\
&+L^2\left[\kappa^4(\eta_1+\kappa^2L_{y^\star}^2\eta_2)\alpha^2 + \left(\frac{\alpha^2L_{z^\star}^2}{\gamma^2\mu_g^2}+\frac{\kappa^4\alpha^2}{\beta^2\mu_g^2}L_{y^\star}^2\right)\right]\left(\frac{\Phi(\bar{x}_0)-\inf\Phi}{\alpha(K+1)}\right) \\
&+L^2\frac{\kappa^6\|\mathbf{O}_y\|^2\mathbb{E}\|\hat{\mathbf{e}}_y^0\|^2}{n(K+1)(1-\|\mathbf{\Gamma}_y\|)} + L^2\frac{\kappa^4\|\mathbf{O}_z\|^2\mathbb{E}\|\hat{\mathbf{e}}_z^0\|^2}{n(K+1)(1-\|\mathbf{\Gamma}_z\|)} + L^2\frac{\kappa^6\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{n(K+1)(1-\|\mathbf{\Gamma}_x\|)} \\
&+L^2\frac{\|z_\star^1\|^2}{\mu_g\gamma(K+1)} + \frac{L^2\kappa^4}{K+1}\frac{1}{\beta\mu_g}\|\bar{y}_0 - y^\star(\bar{x}^0)\|^2.
\end{aligned} \tag{97}$$

Then we aim to choose the stepsize α, β, γ. Define:

$$\begin{aligned}
\widetilde{C}_\alpha =& L_{\nabla\Phi} + \kappa^3 \frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|L}{1-\|\mathbf{\Gamma}_x\|} + \kappa^3 L \left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{2}} \\
&+ \kappa^4 \frac{L_{g,1}^2}{\mu_g} + \kappa^4 \frac{\|\mathbf{O}_y\|\|\mathbf{O}_y^{-1}\|\|\mathbf{\Lambda}_{ya}\|L_{g,1}}{1-\|\mathbf{\Gamma}_y\|} + \kappa^4 L_{g,1} \left(\frac{\kappa\|\mathbf{O}_y\|\|\mathbf{O}_y^{-1}\|\|\mathbf{\Lambda}_{ya}\|\|\mathbf{\Lambda}_{yb}^{-1}\|}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{2}} \\
&+ \kappa^6 L \frac{\|\mathbf{O}_z\|\|\mathbf{O}_z^{-1}\|\|\mathbf{\Lambda}_{za}\|}{1-\|\mathbf{\Gamma}_z\|} + \kappa^{\frac{11}{2}} L \left(\frac{\|\mathbf{O}_z\|\|\mathbf{O}_z^{-1}\|\|\mathbf{\Lambda}_{za}\|\|\mathbf{\Lambda}_{zb}^{-1}\|}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{2}}, \\
\widetilde{\alpha}_{yb,2} =& \left(\frac{1-\|\mathbf{\Gamma}_y\|}{\kappa^{13}\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}\right)^{\frac{1}{3}}, \\
\widetilde{\alpha}_{zb,2} =& \left(\frac{1-\|\mathbf{\Gamma}_z\|}{\kappa^{11}\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}\right)^{\frac{1}{3}}, \\
\widetilde{\alpha}_{xb,2} =& \left(\frac{1-\|\mathbf{\Gamma}_x\|}{\kappa^{5}\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}\right)^{\frac{1}{3}}.
\end{aligned}$$

Then there exist

$$\alpha = \Theta\left(\widetilde{C}_\alpha + \widetilde{\alpha}_{xb,2}^{-1} + \widetilde{\alpha}_{yb,2}^{-1} + \widetilde{\alpha}_{zb,2}^{-1}\right)^{-1}, \beta = \Theta\left(\kappa^4\alpha\right), \gamma = \Theta\left(\kappa^4\alpha\right)$$

such that (45), (53), (56), (40), (58), (65), (82), and (84) hold. Then all previous lemmas hold.

Then from (97) we have:

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\Phi(\bar{x}^k)\|^2 \\
\lesssim& \frac{\kappa}{\alpha K} + \frac{\alpha^2\kappa^{14}\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{K(1-\|\mathbf{\Gamma}_y\|)} \\
&+ \frac{\alpha^2\kappa^{12}\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{K(1-\|\mathbf{\Gamma}_z\|)} + \frac{\alpha^2\kappa^{6}\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{K(1-\|\mathbf{\Gamma}_x\|)} \\
\lesssim& \left(\frac{\kappa^{16}\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2\zeta_0^y}{1-\|\mathbf{\Gamma}_y\|}\right)^{\frac{1}{3}}\frac{1}{K} + \left(\frac{\kappa^{14}\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2\zeta_0^z}{1-\|\mathbf{\Gamma}_z\|}\right)^{\frac{1}{3}}\frac{1}{K} \\
&+ \left(\frac{\kappa^{8}\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{3}}\frac{1}{K} + \widetilde{C}_\alpha\frac{1}{K}.
\end{aligned}$$

□

C.4 Degenerating to single-level algorithms

We consider the bilevel problem with the following upper- and lower-level loss function on the i-th agent:

$$F_i(x,y,\phi) = F_i(x,\phi), \quad G_i(x,y,\xi) \equiv \frac{\|y\|^2}{2}.$$

Actually, this optimization problem with respect to x is single-level, since we have $\mathbf{z}^k \equiv 0$, $\mathbf{y}^k \equiv 0$, $u_i^k = \nabla_1 f_i(x_i^k, \xi_i^k)$ by induction. By taking $\theta = 1$, we get the following single-level algorithm framework for decentralized stochastic single-level algorithm. As we discuss in previous sections, it can recover various heterogeneity-correction algorithms, including GT, EXTRA and ED, by selecting specific $\mathbf{A}_x, \mathbf{B}_x, \mathbf{C}_x$.

Algorithm 3 SPARKLE: degenerating to single-level decentralized stochastic algorithms

Require: Initialize $\mathbf{x}^0 = \mathbf{0}$, $\mathbf{d}_x^0 = \mathbf{0}$, learning rate α_k.
  for $k = 0, 1, \cdots, K-1$ **do**
    $\mathbf{x}^{k+1} = \mathbf{C}_x\mathbf{x}^k - \alpha_k\mathbf{A}_x\mathbf{u}^k - \mathbf{B}_x\mathbf{d}_x^k$, $\mathbf{d}_x^{k+1} = \mathbf{d}_x^k + \mathbf{B}_x\mathbf{x}^{k+1}$;
  end for

In this case, we have $z_k^\star \equiv 0$, $y_k^\star \equiv 0$. Notice that $L_{y^\star} = 0$, $L_{z^\star} = 0$. It gives

$$\eta_2 = \mathcal{O}\left(\beta^2\frac{\|\mathbf{O}_y\|^2\|\mathbf{O}_y^{-1}\|^2\|\mathbf{\Lambda}_{ya}\|^2\|\mathbf{\Lambda}_{yb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_y\|)^2} + \gamma^2\frac{\|\mathbf{O}_z\|^2\|\mathbf{O}_z^{-1}\|^2\|\mathbf{\Lambda}_{za}\|^2\|\mathbf{\Lambda}_{zb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_z\|)^2}\right),$$
$$\eta_1 = \mathcal{O}\left(\eta_2 + \alpha^2\left(1+\frac{(1-\theta)^2}{\theta^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\right).$$

If we take

$$\alpha \lesssim \min\left\{1, \frac{1-\|\mathbf{\Gamma}_x\|}{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|}, \left(\frac{1-\|\mathbf{\Gamma}_x\|}{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}\right)^{\frac{1}{2}}\right\}$$

and $\theta = 1$, $\beta \to 0$, $\gamma \to 0$, then (45), (53), (56), (40), (58), (65), (82), and (84) hold. Thus all previous lemmas hold. Then (87) transforms into

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\Phi(\bar{x}^k)\|^2\\
\lesssim&\frac{f(\bar{x}_0)-\inf f}{\alpha(K+1)} + \frac{1}{n}\left(\theta(1-\theta)+\alpha\theta^2\right)\sigma_{f,1}^2 + \frac{(1-\theta)^2}{\theta(K+1)}\|\nabla f\left(\bar{x}^0\right)\|^2\\
&+\eta_1\alpha^2\left(\frac{f(\bar{x}_0)-\inf f}{\alpha(K+1)} + \frac{\theta}{n}\sigma_{f,1}^2\right) + \frac{\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{n(K+1)(1-\|\mathbf{\Gamma}_x\|)}\\
&+\frac{\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{n(1-\|\mathbf{\Gamma}_x\|)^2(K+1)}\left[\frac{1-\theta}{\theta}\sum_{i=1}^{n}\left\|\nabla f_i(\bar{x}^0)\right\|^2\right]\\
&+\frac{1}{n}\left[\alpha^2\theta\left(\theta+\frac{1-\theta}{1-\|\mathbf{\Gamma}_x\|}\right)\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2 n}{1-\|\mathbf{\Gamma}_x\|}\right]\sigma_{f,1}^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\Phi(\bar{x}^k)\|^2\\
\lesssim&\frac{f(\bar{x}_0)-\inf f}{\alpha(K+1)} + \frac{\alpha\sigma_{f,1}^2}{n} + \left(\alpha^4\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2}{(1-\|\mathbf{\Gamma}_x\|)^2}\right)\left(\frac{f(\bar{x}_0)-\inf f}{\alpha(K+1)} + \frac{1}{n}\sigma_{f,1}^2\right)\\
&+\frac{\|\mathbf{O}_x\|^2\mathbb{E}\|\hat{\mathbf{e}}_x^0\|^2}{n(K+1)(1-\|\mathbf{\Gamma}_x\|)} + \alpha^2\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\sigma_{f,1}^2\\
\lesssim&\frac{f(\bar{x}_0)-\inf f}{\alpha(K+1)} + \frac{\alpha\sigma_{f,1}^2}{n} + \alpha^2\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2}{1-\|\mathbf{\Gamma}_x\|}\sigma_{f,1}^2\\
&+\frac{\alpha^2\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{(K+1)(1-\|\mathbf{\Gamma}_x\|)} + \frac{\alpha^4\|\mathbf{O}\|_x^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\sigma_{f,1}^2}{n(1-\|\mathbf{\Gamma}_x\|)^2}.
\end{aligned} \tag{98}$$

Like (88), we take

$$C_0 = 1 + \frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|}{1-\|\mathbf{\Gamma}_x\|} + \left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xa}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{2}},$$

$$\alpha_1 = \sqrt{\frac{n}{K\sigma_{f,1}^2}}, \quad \alpha_2 = \left(\frac{1-\|\mathbf{\Gamma}_x\|}{K\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\sigma_{f,1}^2}\right)^{\frac{1}{3}},$$

$$\alpha_3 = \left(\frac{1-\|\mathbf{\Gamma}_x\|}{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}\right)^{\frac{1}{3}},$$

$$\alpha_4 = \left(\frac{n(1-\|\mathbf{\Gamma}_x\|)^2}{K\|\mathbf{O}\|_x^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\sigma_{f,1}^2}\right)^{\frac{1}{5}},$$

$$\alpha = \Theta\left(C_0 + \frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \frac{1}{\alpha_3} + \frac{1}{\alpha_4}\right)^{-1}.$$

Substituting these values into (98), we get

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\Phi(\bar{x}^k)\|^2 \lesssim \frac{\sigma_{f,1}}{\sqrt{nK}} + \left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\sigma_{f,1}^2}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{1}{3}} K^{-2/3} + \frac{C_0}{K}$$
$$+ \left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\zeta_0^x}{(1-\|\mathbf{\Gamma}_x\|)}\right)^{\frac{1}{3}} \frac{1}{K} + \left(\frac{\|\mathbf{O}\|_x^2\|\mathbf{O}_x^{-1}\|^2\|\mathbf{\Lambda}_{xa}\|^2\|\mathbf{\Lambda}_{xb}^{-1}\|^2\sigma_{f,1}^2}{n(1-\|\mathbf{\Gamma}_x\|)^2}\right)^{\frac{1}{5}} K^{-4/5}.$$

Like Lemma 18, we get the transient iterating complexity for Algorithm 3 is

$$\left\{n^3\left(\frac{\|\mathbf{O}_x\|^2\|\mathbf{O}_x^{-1}\|^2}{1-\|\mathbf{\Gamma}_x\|}\right)^2\|\mathbf{\Lambda}_{xa}\|^2, n\left(\frac{\|\mathbf{O}_x\|\|\mathbf{O}_x^{-1}\|\|\mathbf{\Lambda}_{xb}^{-1}\|}{1-\|\mathbf{\Gamma}_x\|}\right)^{\frac{4}{3}}\|\mathbf{\Lambda}_{xa}\|, n\right\}.$$

Substituting the value of relevant norms in Table 4, we get the transient iteration complexity for GT, EXTRA, ED are

$$\mathcal{O}\left(\max\left\{\frac{n^3}{(1-\rho)^2}, \frac{n}{(1-\rho)^{8/3}}\right\}\right), \mathcal{O}\left(\frac{n^3}{(1-\rho)^2}\right), \mathcal{O}\left(\frac{n^3}{(1-\rho)^2}\right)$$

respectively, where $\rho := \rho(\mathbf{W}_x)$. These upper bounds are the same as the state-of-the-art results shown in Table 1. It indicates that our analysis accurately captures the impacts of updates at each level on the convergence results.

D Experimental details

In this section, we provide the details of our numerical experiments discussed in Section 4. We also provide addition experimental results which are not mentioned in the main text due to the space limitation. For all GT variants, we focus on one typical representative, ATC-GT, in our experiments, which we denote as GT for brevity. All experiments described in this section were run on an NVIDIA A100 server.

D.1 Synthetic bilevel optimization

Here, we consider problem (1) whose upper- and lower level loss functions on the i-th agents ($1 \le i \le N$) are denoted as:

$$f_i(x,y) = \mathbb{E}_{A_i,b_i}\left[\|A_i y - b_i\|^2\right],$$
$$g_i(x,y) = \mathbb{E}_{A_i,b_i}\left[\|A_i y - x\|^2 + C_r\|y\|^2\right],$$

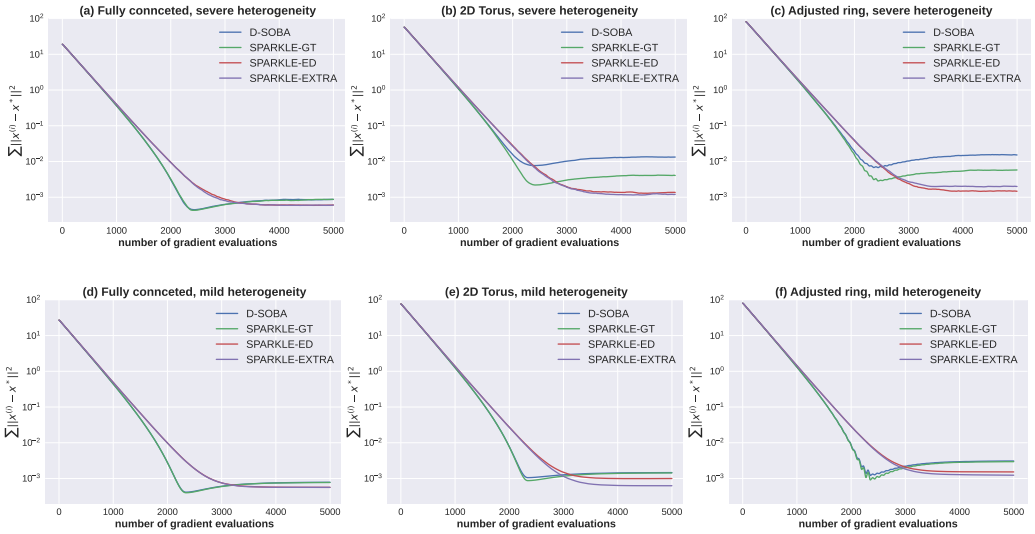


Figure 4: The estimation error of D-SOBA, SPARKLE-GT, SPARKLE-ED, and SPARKLE-EXTRA under different networks and data heterogeneity.

where $x \in \mathbb{R}^D, y \in \mathbb{R}^K$ and C_r denotes a fixed regularization parameter. For each agent i, we firstly generate the local solution y_i^*, x_i^* as $y_i^* = y^* + \zeta_i$ and $x_i^* = A^* b^* + \xi_i$, where $x^* \sim \mathcal{N}(0, I_K)$ is a randomly generated vector, each element of A^* is independently sampled from $\mathcal{N}(0, 9)$. The observation (A_i, b_i) on agent i is generated in a streaming manner by $A_i = A^* + \phi_i$, $b_i = x_i^* + \psi_i$, in which each element of $\phi_i \in \mathbb{R}^{K \times D}$ and $\psi_i \in \mathbb{R}^D$ are independently generated by $\mathcal{N}(0, \sigma_g^2)$. The terms $\xi_i \sim \mathcal{N}(0, \sigma_h^2 I_K)$ and $\zeta_i \sim \mathcal{N}(0, \sigma_h^2 I_D)$ control the heterogeneity of data distributions across different agents.

We set $D = 20, K = 10, \sigma_g = 0.001, C_r = 0.001$. Then we set $\sigma_h = 0.5$ to represent severe heterogeneity across agents and $\sigma_h = 0.1$ for mild heterogeneity. We run D-SOBA, SPARKLE-GT, SPARKLE-ED, and SPARKLE-EXTRA over Ring, 2D-Torus [37], and fully connected networks with $N = 64$ agents. The moving-average term $\theta = 0.1$ and the step-size at the t-th iteration are $\alpha_t = \beta_t = \gamma_t = 1/(500 + 0.01t)$. The batch size is 10.

Fig. 4 illustrates the averaged estimation error $\sum_{i=1}^N \left\| x_i^{(t)} - x^* \right\|^2$ of the mentioned algorithms with different communication topology and data heterogeneity. It is observed that SPARKLE with ED, EXTRA, GT achieve better convergence performances with decentralized communication networks. Meanwhile, SPARKLE-ED and SPARKLE-EXTRA are more robust to data heterogeneity and the sparsity of network topology than SPARKLE-GT. All the results are consistent with our theoretical results.

D.2 Hyper-cleaning on FashionMNIST dataset

Here, we consider a data hyper-clean problem [44] on FashionMNIST dataset [48]. The FashionMNIST dataset consists of 60000 images for training and 10000 images for testing and we randomly split 50000 training images into a training set and the other 10000 images into a validation set.

The data hyper-cleaning problem aims to train a classifier from a corrupted dataset, in which the label of each training data is replaced by a random class number with a probability p (*i*.e. the corruption rate). It can be considered as a stochastic bilevel problem (1) whose upper- and lower-level loss functions on the i-th agents ($1 \leq i \leq n$) are formulated as:

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

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

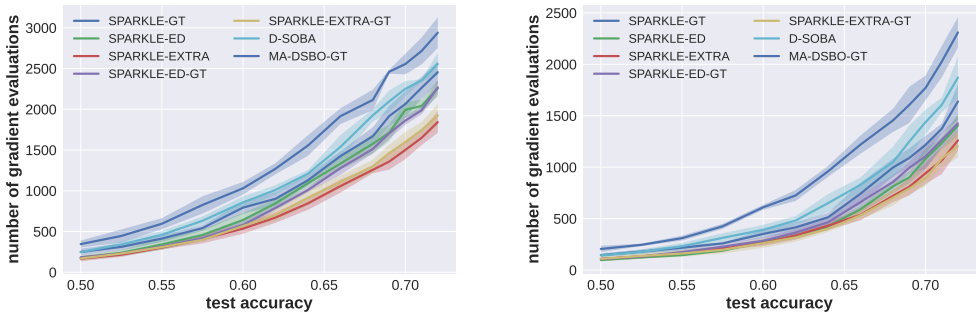


Figure 5: Hypergradient evaluation times for required test accuracy in hyper-cleaning problem. (Left: $p = 0.2$; Right: $p = 0.3$)

where ϕ denotes a training model while y denotes its parameters, L denotes the cross-entropy loss function and $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. $\mathcal{D}_{tr}^{(i)}$ and $\mathcal{D}_{val}^{(i)}$ denotes the training and validation set of the i-th agent, respectively. $C > 0$ is a fixed regularization parameter.

Data generation and experiment settings. In this experiment, we let ϕ be a two-layer MLP network with a 300-dim hidden layer and ReLU activation while y denotes its parameters. For $1 \leq i \leq 10$, we sample a probability distribution $\mathcal{P}_i$ randomly by Dirichlet distribution with parameters $\alpha = 0.1$. The training and validation images with label i are sent to different agents according the probability distribution $\mathcal{P}_i$. Then $\mathcal{D}_{tr}^{(i)}$ and $\mathcal{D}_{val}^{(i)}$ are generated sufficiently heterogeneous [32]. We set $C = 0.001$. The batch size is set to 50.

Convergence performances with different corruption rates. We set the moving-average term $\theta_k = 0.2$ and run D-SOBA [29], MA-DSBO-GT [10], MDBO [21] SPARKLE-GT, SPARKLE-ED, SPARKLE-EXTRA, SPARKLE-ED-GT, and SPARKLE-EXTRA-GT on an Adjusted Ring graph with $n = 10$ agents and $p = 0.1, 0.2, 0.3$ separately. The step-sizes for all the algorithms are set to $\alpha_k = \beta_k = \gamma_k = 0.03$ and the term η in MDBO is set to 0.5. The weight matrix of Adjust Ring $W = [w_{ij}]_{n\times n}$ satisfies:

$$w_{ij} = \begin{cases} a, & \text{if } j = i, \\ \dfrac{1-a}{2}, & \text{if } (j-i)\%n = \pm 1, \\ 0, & \text{else.} \end{cases}$$

Moreover, we run SPARKLE with ED in the lower level and auxiliary variable and gradient tracking in the upper level (i.e. SPARKLE-ED-GT) as well as SPARKLE with EXTRA in the lower level and auxiliary variable and gradient tracking in the upper level (i.e. SPARKLE-EXTRA-GT) and compare their test accuracy with the other four algorithms.

Figure 1 shows that SPARKLE-ED and SPARKLE-EXTRA outperforms in different cases than SPARKLE-GT. Meanwhile, SPARKLE-EXTRA, SPARKLE-EXTRA-GT achieve similar test accuracy, as do those for SPARKLE-ED and SPARKLE-ED-GT, which matches our theoretical results in transient iteration analysis. Figure 5 presents the times of gradient evaluation for different test accuracies of these algorithms at $p = 0.2, 0.3$, demonstrating similar results.

Influence of network topology. We set the corruption rate $p = 0.3$, the step sizes $\alpha_k = \beta_k = \gamma_k = 0.02$, and the moving-average term $\theta_k = 0.2$. Then we run SPARKLE-EXTRA and SPARKLE-EXTRA-GT on a network containing $n = 10$ nodes with different topologies in the following two cases:

- **Fixed upper, varied lower**: x communicates through a five-peer graph; y, z communicate through different adjusted rings with $\rho = 0.647, 0.828, 0.924, 0.990$.
- **Fixed lower, varied upper**: y, z communicate through a five-peer graph; x communicates through different adjusted rings with $\rho = 0.647, 0.828, 0.924, 0.990$.

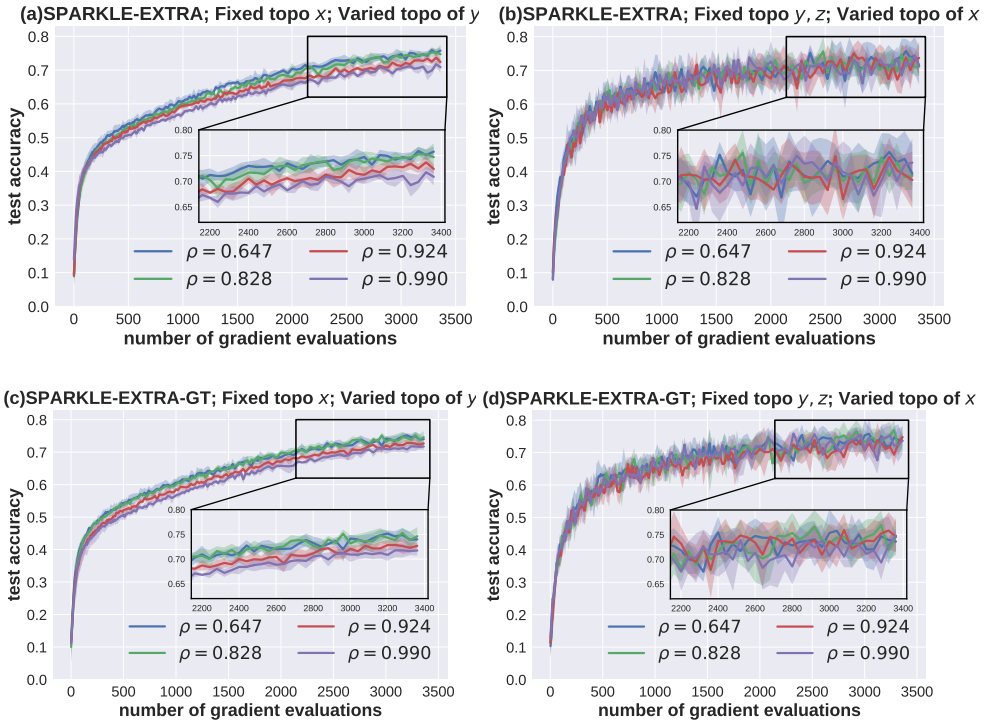


Figure 6: The average test accuracy of SPARKLE-EXTRA and SPARKLE-EXTRA-GT on hyper-cleaning with different communicating strategy of x, y, z.

Table 5: Mean and standard deviation of the average test accuracy of last 40 iterations during 10 trials with different moving-average terms

Algorithm	$\theta = 0.05$	$\theta = 0.2$	$\theta = 0.3$
SPARKLE-**GT**	$\mathbf{0.7080 \pm 0.0215}$	0.7045 ± 0.0126	0.7064 ± 0.0113
SPARKLE-**ED**	0.7096 ± 0.0074	$\mathbf{0.7113 \pm 0.0047}$	0.7110 ± 0.0081
SPARKLE-**EXTRA**	0.7190 ± 0.0103	$\mathbf{0.7277 \pm 0.0090}$	0.7243 ± 0.0028
SPARKLE-**ED-GT**	0.7064 ± 0.0063	$\mathbf{0.7178 \pm 0.0037}$	0.7162 ± 0.0041
SPARKLE-**EXTRA-GT**	0.7198 ± 0.0051	$\mathbf{0.7262 \pm 0.0058}$	0.7247 ± 0.0048

The weight matrix of five-peer graph $W = [w_{ij}]_{n\times n}$ satisfies:

$$w_{ij} = \begin{cases} 0.2, & \text{if } (j-i)\%n = 0, \pm 1, \pm 2, \\ 0, & \text{else.} \end{cases}$$

Figure 6 shows the average test accuracy of both SPARKLE-EXTRA and SPARKLE-EXTRA-GT over 10 trials. It indicates that the test accuracy decays with increasing spectral gap of topologies related to y, z while the topology of x is fixed during the whole iterations. However, such convergence gap becomes milder when the topologies of y, z are fixed and that of x varies. This phenomenon supports our theoretical findings, which suggest that the transient iteration complexity is more sensitive to the network topologies of y, z than to that of x.

Influence of moving-average iteration on convergence. Moreover, for $\theta_t = 0.05, 0.2, 0.3$, we run SPARKLE-GT, SPARKLE-ED, SPARKLE-EXTRA, SPARKLE-ED-GT, and SPARKLE-EXTRA-GT on an Adjusted Ring graph with $n = 10$ agents, $\alpha_k = \beta_k = \gamma_k = 0.03$ and $p = 0.3$ for 3000 iterations. We obtain the average test accuracy of the last 40 iterations over 10 trials, and present the mean and standard deviation during the different trials in Table 5. We can observe that most algorithms achieve the highest test accuracy when $\theta = 0.2$, which may prove that a suitable θ can benefit the test accuracy in hyper-cleaning problems.

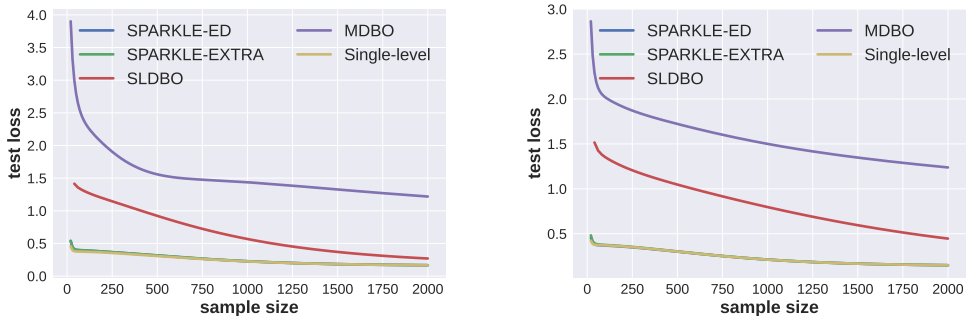


Figure 7: The test loss against samples generated by one agent of different algorithms in the policy evaluation. (Left: $n = 20$, Right: $n = 10$.)

Table 6: The average training loss of the last 500 iterations for 10 independent trials in the distributed policy evaluation.

Algorithm	$N = 10$	$N = 20$
SPARKLE-**ED**	$0.2781 \pm 1.09 \times 10^{-3}$	$0.3198 \pm 3.21 \times 10^{-3}$
SPARKLE-**EXTRA**	$0.2743 \pm 0.88 \times 10^{-3}$	$0.3207 \pm 2.94 \times 10^{-3}$
MDBO	$1.0408 \pm 4.51 \times 10^{-3}$	$1.3293 \pm 8.38 \times 10^{-3}$
SLDBO	$0.4132 \pm 1.18 \times 10^{-3}$	$0.8374 \pm 2.47 \times 10^{-3}$
Single-level ED	$0.2948 \pm 0.92 \times 10^{-3}$	$0.3164 \pm 3.12 \times 10^{-3}$

D.3 Distributed policy evaluation in reinforcement learning

Following the result of [52], we consider a multi-agent MDP problem in reinforcement learning on a distributed setting with n agents. Denote $\mathcal{S}$ as the state space. Suppose that the value function in each state $s \in \mathcal{S}$ is a linear function $V(s) = \phi_s^\top x$, where $\phi_s \in \mathbb{R}^m$ is a feature and $x \in \mathbb{R}^m$ is a parameter. To obtain the optimal solution x^*, we consider the following Bellman minimization problem:

$$\min_{x \in \mathbb{R}^m} \quad F(x) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{2|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(\phi_s^\top x - \mathbb{E}_{s'} \left[r^i(s, s') + \gamma \phi_{s'}^\top x \middle| s \right] \right)^2 \right]$$

where $r^i(s, s')$ denotes the reward incurred from transition s to s' on the i-th agent, $\gamma \in (0, 1)$ denotes the discount factor. The expectation is taken over all random transitions from state s to s'. It can be viewed as a bilevel optimization problem with the following upper- and lower-level loss:

$$f_i(x, y) = \frac{1}{2|\mathcal{S}|} \sum_{s \in \mathcal{S}} (\phi_s^\top x - y_s)^2,$$
$$g_i(x, y) = \sum_{s \in \mathcal{S}} \left(y_s - \mathbb{E}_{s'} \left[r^i(s, s') + \gamma \phi_{s'}^\top x \middle| s \right] \right)^2,$$

where $y = (y_1, \cdots, y_{|\mathcal{S}|})^\top \in \mathbb{R}^{|\mathcal{S}|}$. In our experiment, we set the number of states $|\mathcal{S}| = 200$ and $m = 10$. For each $s \in \mathcal{S}$, we generate its feature $\phi_s \sim U[0, 1]^m$. The non-negative transition probabilities are generated randomly and standardized to satisfy $\sum_{s' \in \mathcal{S}} p_{s,s'} = 1$. The mean reward $\bar{r}^i(s, s')$ are independently generated from the uniform distribution $U[0, 1]$. In each iteration, the stochastic reward $r^i(s, s') \sim \mathcal{N}(\bar{r}^i(s, s'), 0.02^2)$.

For $n = 10, 20$, we run SPARKLE-ED and SPARKLE-EXTRA as well as existing decentralized SBO algorithms MDBO [21] and SLDBO [16] (here we use the stochastic gradient instead of deterministic gradient) over a Ring graph. For MDBO, the number of Hessian-inverse estimation iterations is set to 5. The step sizes are 0.03 for all methods. Figure 3 illustrates the upper-level loss against samples generated by one agent for 10 independent trials. Table 6 shows the average training loss of the last 500 iterations for 10 independent trials of the four decentralized SBO algorithms as well as single-level ED [56] (For bilevel algorithms, *training loss* means the upper-level loss here).

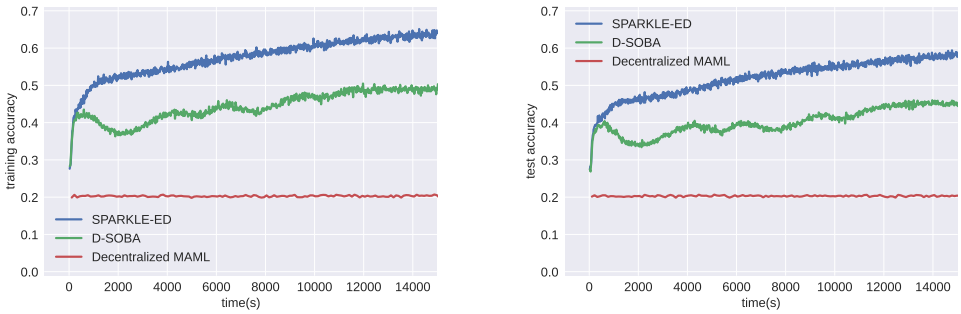


Figure 8: The accuracy on training and testing set of different algorithms for the meta-learning problem.

Both Figure 3 and Table 6 demonstrate that SPARKLE-ED and SPARKLE-EXTRA converge faster than other methods.

Finally, we create a fixed "test set" with 10000 sample generated from $\mathcal{S}$. Figure 7 shows the loss on the test set of SPARKLE-ED, SPARKLE-EXTRA, SLDBO, MDBO and single-level ED algorithm, demonstrating the superior performance of SPARKLE compared to other decentralized SBO algorithms.

D.4 Decentralized meta-learning

We consider a meta-learning problem as described in [18]. There are R tasks $\{\mathcal{T}_s, s = 1, \cdots, R\}$. Each task $\mathcal{T}_s$ has its own loss function $L(x, y_s, \xi)$, where ξ_s represents a stochastic sample drawn from the data distribution $\mathcal{D}_s$, y_s denotes the task-specific parameters and x denotes the global parameters shared by all the tasks. In meta-learning problem, we aim to find the parameters $(x^*, y_1^*, \cdots, y_R^*)$ that minimizes the loss function across all R tasks, *i*.e.,

$$\min_{x, y_1, \cdots, y_R} l(x, y_1, \cdots, y_R) = \frac{1}{R} \sum_{s=1}^{R} \mathbb{E}_{\xi \sim \mathcal{D}_s} \left[L(x, y_s, \xi) \right]. \tag{102}$$

The problem (102) can be formulated as a decentralized SBO problem with heterogeneous data distributions across N nodes. For $i = 1, 2, \cdots, N$, let $\mathcal{D}_{s,i}^{\text{train}}$ and $\mathcal{D}_{s,i}^{\text{val}}$ denote the training and validation datasets for the s-th task $\mathcal{T}_s$ received by node i respectively. We can then address the meta-learning problem by minimizing (1), with the upper- and lower-level loss functions defined as:

$$f_i(x, y) = \frac{1}{R} \sum_{s=1}^{R} \mathbb{E}_{\xi \sim \mathcal{D}_{s,i}^{\text{val}}} \left[L(x, y_s, \xi) \right],$$

$$g_i(x, y) = \frac{1}{R} \sum_{s=1}^{R} \left[\mathbb{E}_{\xi \sim \mathcal{D}_{s,i}^{\text{train}}} \left[L(x, y_s, \xi) \right] + \mathcal{R}(y_s) \right],$$

where L denotes the cross-entropy loss and $\mathcal{R}(y_s) = C_r \|y_s\|^2$ is a strongly convex regularization function.

In this experiment, we compare SPARKLE-ED with D-SOBA [29] and MAML [18] in a decentralized communication setting over a 5-way 5-shot task across a network of $N = 8$ nodes connected by Ring graph. The dataset used is miniImageNet [47], derived from ImageNet [42], which comprises 100 classes, each containing 600 images of size 84×84. We set $R = 2000$ and partition these classes into 64 for training, 16 for validation, and 20 for testing. For the training and validation classes, the data is split according to a Dirichlet distribution with parameter $\alpha = 0.1$ [32]. We utilize a four-layer CNN with four convolution blocks, where each block sequentially consists of a 3×3 convolution with 32 filters, batch normalizationm, ReLU activation, and 2×2 max pooling. The batch size is 32 , and $C_r = 0.001$. The parameters of the last linear layer are designated as task-specific, while the other parameters are shared globally. For SPARKLE and D-SOBA, the step-sizes are

$\beta = \gamma = 0.1$ and $\alpha = 0.01$. For MAML, the inner step-size is 0.1 and the outer step-size is 0.001, and the number of inner-loop steps as 3. For all algorithms, the task number is set to 32. And we only repeat the experiment only once due to the time limitation. Figure 8 shows the average accuracy on the training dataset for all nodes, as well as the test accuracy of the three algorithms. We observe that SPARKLE-ED outperforms other algorithms, demonstrating the efficiency of SPARKLE in decentralized meta-learning problems.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Refer to Abstract and Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to Section Conclusions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Refer to Section Assumptions for our assumptions, and Appendix for detailed proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Refer to Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We may consider making data and code openly accessible when it is deemed necessary.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Refer to Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We show error bars in experiments where we consider them essential.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification:Our research conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no such risk in the paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We comply with the licenses of existing assets used in the paper and provide necessary references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.