SLBRD: A SINGLE-LOOP AND BYZANTINE-RESILIENT DECENTRALIZED BILEVEL ALGORITHM

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized bilevel optimization has emerged as a powerful paradigm for large-scale machine learning over distributed data. However, these systems are inherently vulnerable to Byzantine agents that transmit arbitrary or adversarial updates. While extensive progress has been made on Byzantine resilience in single-level optimization, extending these robustness guarantees to the bilevel setting remains an open problem. To bridge this gap, we propose SLBRD, the first Byzantine-resilient decentralized bilevel algorithm. By adopting a fully single-loop and Hessian-inversion-free architecture, SLBRD improves communication efficiency and limits the attack surface available to adversaries. We establish its sample and transient iteration complexities, explicitly characterizing the joint effects of network topology, data heterogeneity, and Byzantine influence. We further develop the Polyak-momentum variant SLBRD-M, which removes stochastic variance from the persistent Byzantine error and improves the topology-dependent transient complexity. Experiments on hyperparameter optimization and hyper-representation tasks validate the effectiveness of the proposed methods[1].

1 INTRODUCTION

Bilevel optimization has attracted significant attention in the machine learning community due to its broad applications, including meta-learning (Bertinetto et al., 2019; Franceschi et al., 2018; Rajeswaran et al., 2019), hyperparameter optimization (Bengio, 2000; Franceschi et al., 2018; Bertrand et al., 2020), data augmentation (Cubuk et al., 2019; Rommel et al., 2022), and neural architecture search (Liu et al., 2019; Zhang et al., 2022). Meanwhile, many large-scale learning tasks are inherently distributed across multiple agents that are unwilling or unable to share their raw data, which has led to growing interest in *decentralized* bilevel optimization (Kong et al., 2024; Zhu et al., 2024; Chen et al., 2023). In this setting, n agents communicate over a peer-to-peer network and collaboratively solve the following bilevel optimization problem without relying on a central server:

$$\min_{x\in\mathbb{R}^{d_x}} \Phi(x) := \frac{1}{n}\sum_{i=1}^{n} f_i\left(x, y^\star(x)\right), \quad \text{s.t.} \quad y^\star(x) = \arg\min_{y\in\mathbb{R}^{d_y}} \frac{1}{n}\sum_{i=1}^{n} g_i(x,y), \tag{1}$$

where agent i has a local upper-level objective f_i and a local lower-level objective g_i. We assume that the aggregate lower-level objective is strongly convex in y, and refer to Φ as the hyperobjective. Each agent can access only unbiased stochastic gradient oracles for its local functions:

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

where ξ_i and ϕ_i are random samples from agent i's local upper- and lower-level data distributions $\mathcal{D}_{f_i}$ and $\mathcal{D}_{g_i}$, respectively. F_i and G_i denote the corresponding stochastic losses.

However, decentralized learning is vulnerable to faulty or adversarial participants (Armstrong, 2003; Zhang et al., 2020) arising from hardware failures, data poisoning, or attacks on the network. Byzantine workers (Xie et al., 2020; Baruch et al., 2019) may send arbitrary messages to their neighbors and prevent the honest agents from converging, even when only a δ-fraction of the n agents is Byzantine. As decentralized learning enters safety-critical applications such as autonomous driving

[1]Code is available at the anonymous repository: https://anonymous.4open.science/r/SLBRD-156C.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of the proposed methods with single-level baselines (D−Bucketing (Karimireddy et al., 2020), D−IOS (Wu et al., 2023)) and bilevel methods (SOBA (Dagréou et al., 2022), GBDSBO (Yang et al., 2022), MDBO (Gao et al., 2023), D−SOBA−SO (Kong et al., 2025), and FedBi−BR (Li et al.)) in terms of assumptions and complexity guarantees. Here, ζ^2 denotes the data-heterogeneity level for single-level objectives, while ζ_f^2 and ζ_g^2 represent the upper- and lower-level data-heterogeneity levels for bilevel objectives, respectively.

Method (single-level optimization)	Byz	SL	D	(UL) f	(LL) g	Sample Complexity	Asymptotic Error	Transient Complexity
Bucketing (Karimireddy et al., 2020)	✓	-	✗	$\mathcal{C}_L^1$	N. A.	$\tilde{\mathcal{O}}(\epsilon^{-2})$	$\mathcal{O}\left(\delta\left(\sigma^2+\zeta^2\right)\right)$	N. A.
D−IOS (Wu et al., 2023)	✓	-	✓	$\mathcal{C}_L^1$	N. A.	$\tilde{\mathcal{O}}(\epsilon^{-2})$	$\mathcal{O}\left(\delta\left(\sigma^2+\zeta^2\right)\right)$	$\mathcal{O}\left(\frac{n_h^3}{(1-\lambda-8\delta\sqrt{n_h})^6}\right)$
Method (bilevel optimization)	**Byz**	**SL**	**D**	**(UL) f**	**(LL) g**	**Sample Complexity**	**Asymptotic Error**	**Transient Complexity**
SOBA (Dagréou et al., 2022)	✗	✓	✗	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	N. A.	N. A.
GBDSBO (Yang et al., 2022)	✗	✗	✓	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\epsilon^{-2})$	N. A.	$\mathcal{O}\left(\frac{n^3G^4}{(1-\lambda)^4}\right)$
MDBO (Gao et al., 2023)	✗	✗	✓	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\epsilon^{-2})$	N. A.	$\mathcal{O}\left(\frac{n^3}{(1-\lambda)^8}\right)$
D−SOBA−SO (Kong et al., 2025)	✗	✓	✓	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\epsilon^{-2})$	N. A.	$\mathcal{O}\left(\max\left\{\frac{n^3}{(1-\lambda)^2}, \frac{n^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4}\right\}\right)$
FedBi−BR (Li et al.)	✓	✓	✗	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}\left(\delta\left(\sigma^2+\zeta_f^2+\zeta_g^2\right)\right)$	N. A.
SLBRD (Alg. 1)	✓	✓	✓	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}\left(\delta\left(\sigma^2+\zeta_f^2+\zeta_g^2\right)\right)$	$\mathcal{O}\left(\max\left\{\frac{n_h^3}{(1-\lambda)^2}, \frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4}, \frac{\delta^2 n_h^3}{(1-\lambda)^4}\right\}\right)$
SLBRD−M (Alg. 1)	✓	✓	✓	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}\left(\delta(\zeta_f^2+\zeta_g^2)\right)$	$\mathcal{O}\left(\max\left\{\frac{n_h^3}{1-\lambda}, \frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4}, \frac{\delta^2 n_h^3}{(1-\lambda)^4}\right\}\right)$

"Byz" indicates Byzantine resilience, "SL" the single-loop structure, "D" the decentralized setting, and "SC" strong convexity. $\mathcal{C}_L^{p,q}$ denotes the smoothness class of order (p, q), and $\mathcal{C}_L^1$ denotes first-order smoothness.

and healthcare (Chen et al., 2021; Kasyap et al., 2024), Byzantine resilience is essential (Blanchard et al., 2017; Yin et al., 2018).

While substantial progress has been made in developing Byzantine-resilient algorithms for single-level decentralized optimization (Blanchard et al., 2017; Yin et al., 2018; Wu et al., 2023; Farhadkhani et al., 2023), extending these techniques to the bilevel setting entails fundamental challenges. *First*, in contrast to single-level optimization, which requires consensus on only one variable, bilevel optimization involves simultaneous consensus on both the upper- and lower-level variables, coupled through $y^\star(x)$. This nested structure makes the convergence analysis significantly more intricate in the presence of Byzantine attacks. *Second*, many decentralized bilevel algorithms (Chen et al., 2025; 2023; Yang et al., 2022) rely on inner loops to estimate $y^\star(x)$ or the inverse Hessian, incurring additional communication rounds in each iteration and thereby enlarging the attack surface exposed to Byzantine adversaries. *Third*, although transient complexity has been studied in the Byzantine-free setting (Zhu et al., 2024; Kong et al., 2025), it remains unclear how it is jointly influenced by the network topology, data heterogeneity, and the fraction of Byzantine agents.

This paper presents the first study of Byzantine-resilient decentralized stochastic bilevel optimization. The contributions of this paper are summarized in Table 1 and detailed as follows:

- We propose a family of Byzantine-resilient decentralized stochastic bilevel algorithms, namely SLBRD and its momentum-enhanced variant SLBRD−M, which achieve provable convergence and robustness against Byzantine adversaries while remaining fully single-loop and free of Hessian inversion. By eliminating inner loops for estimating $y^\star(x)$ and Hessian-inverse quantities, this design reduces both communication overhead and the attack surface available to Byzantine adversaries.
- We establish the convergence rate and transient iteration complexity of SLBRD, explicitly characterizing how network topology, data heterogeneity, and the Byzantine fraction jointly affect convergence. Note that prior analyses of decentralized bilevel optimization (e.g., (Dagréou et al., 2022; Yang et al., 2022)) are restricted to the Byzantine-free setting. Moreover, our analysis relaxes the high-order smoothness assumptions commonly adopted in the bilevel literature, namely $\{f \in \mathcal{C}_L^{2,2},\ g \in \mathcal{C}_L^{3,3}\}$, to the more standard conditions $\{f \in \mathcal{C}_L^{1,1},\ g \in \mathcal{C}_L^{2,2}\}$.
- We propose SLBRD−M, a momentum-enhanced variant based on the heavy-ball method (Polyak, 1964). It attains an asymptotic error of $\mathcal{O}(\delta(\zeta_f^2+\zeta_g^2))$, matching the lower bound established in (Li et al.) and achieving optimal Byzantine resilience. It also improves the variance-dependent transient complexity from $\mathcal{O}\big(n_h^3(1-\lambda)^{-2}\big)$ to $\mathcal{O}\big(n_h^2(1-\lambda)^{-1}\big)$.
- Experiments on hyperparameter optimization and hyper-representation tasks confirm the effectiveness of the proposed algorithms on both synthetic and real-world datasets.

1.1 Related Work

Decentralized bilevel optimization. Existing decentralized bilevel methods estimate the hypergradient through inner-loop solvers or single-loop tracking. Early inner-loop approaches used JHIP,

Hessian-inverse-gradient, Neumann-series, or augmented-Lagrangian estimators (Chen et al., 2025; 2023; Yang et al., 2022; Lu et al., 2022b). Although these methods provide rigorous convergence guarantees, they incur additional communication overhead in each iteration. Subsequent work therefore developed *single-loop* methods (Kong et al., 2024; Dong et al., 2025; Gao et al., 2023), and recent studies established unified analyses with sharper transient-complexity bounds (Zhu et al., 2024; Kong et al., 2025). Other studies consider personalized lower-level objectives, differential privacy, and communication compression (Niu et al., 2025; Lu et al., 2022a; Chen & Wang, 2024; He et al., 2024). These methods generally assume benign agents and do not address Byzantine resilience.

Byzantine-resilient decentralized optimization. Byzantine-resilient optimization has been studied extensively in single-level learning (Blanchard et al., 2017; Yin et al., 2018; Alistarh et al., 2018; El-Mhamdi et al., 2021; Wu et al., 2023; Farhadkhani et al., 2023). Representative robust aggregators include Krum, coordinate-wise and geometric median, centered clipping, and bucketing (Blanchard et al., 2017; Yin et al., 2018; Acharya et al., 2022; Karimireddy et al., 2021; 2020). In decentralized settings, Byzantine resilience is obtained by aggregating messages from local neighbors, for example through trimmed means or coordinate-descent screening (Su & Vaidya, 2020; Fang et al., 2022; Yang & Bajwa, 2019). Theoretical analyses establish the lower bound $\Omega(\delta\zeta^2)$ for the asymptotic error, and Polyak's momentum can attain this bound (Karimireddy et al., 2020; Farhadkhani et al., 2023). In bilevel optimization, corruption in the lower-level tracker also perturbs the upper-level hypergradient. Existing Byzantine-resilient bilevel methods consider a centralized server (Abbas et al., 2024; Li et al.), while decentralized bilevel optimization under arbitrary Byzantine messages remains unstudied. Our work addresses this gap with a single-loop architecture and robust aggregation of the coupled upper- and lower-level updates.

2 Preliminaries

Problem statement. Consider a decentralized network of n agents indexed by $[n] = \{1, \dots, n\}$, among which at most a $\delta < 1/2$ fraction are Byzantine. Let $\mathcal{H} \subseteq [n]$ denote the set of honest agents, with cardinality $n_h := |\mathcal{H}| = (1-\delta)n$, and let $\mathcal{B} := [n] \setminus \mathcal{H}$ denotes the set of Byzantine agents. Byzantine agents do not follow the prescribed algorithm and may send arbitrary, potentially adversarial messages to their neighbors (Baruch et al., 2019; Xie et al., 2020). The objective of the honest agents is to collaboratively solve the bilevel optimization problem induced by the honest population while achieving consensus in the presence of Byzantine adversaries:

$$\min_{x\in\mathbb{R}^{d_x}} \Phi_{\mathcal{H}}(x) := \frac{1}{n_h}\sum_{i\in\mathcal{H}} f_i\big(x, y^\star(x)\big), \quad \text{s.t.} \quad y^\star(x) = \arg\min_{y\in\mathbb{R}^{d_y}} \frac{1}{n_h}\sum_{i\in\mathcal{H}} g_i(x,y), \tag{3}$$

where f_i and g_i are the local upper- and lower-level objectives of honest agent i, respectively, as introduced in (1), and $\Phi_{\mathcal{H}}$ denotes the hyperobjective associated with the honest agents.

Robust decentralized learning. The goal of robust decentralized learning is to develop algorithms that can tolerate up to a δ-fraction of Byzantine agents whose identities are unknown (Yin et al., 2018; Allouah et al., 2023). In this setting, the objective of the honest agents is to compute an ε-stationary point of the honest hyperobjective $\Phi_{\mathcal{H}}$. To formalize this notion, we adopt the following definition of (δ,ε)-Byzantine resilience from (Liu et al., 2021).

Definition 1 ((δ,ε)-Byzantine-resilient)**.** *A decentralized algorithm is said to be (δ,ε)-Byzantine-resilient if, in the presence of a δ-fraction of Byzantine attackers, it outputs $\hat{x}$ satisfying*

$$\mathbb{E}\big[\|\nabla\Phi_{\mathcal{H}}(\hat{x})\|^2\big] \le \varepsilon, \tag{4}$$

where $\Phi_{\mathcal{H}}$ denotes the honest hyperobjective defined in (3).

Next, following prior works (Alghunaim & Yuan, 2022; Koloskova et al., 2020), we formalize the communication topology among the honest agents. Their peer-to-peer connectivity is modeled by an undirected graph, whose structure is characterized by a mixing matrix.

Assumption 1 (Mixing matrix)**.** *Let $W = [w_{ij}] \in [0,1]^{n_h\times n_h}$ denote the mixing matrix associated with the communication graph among honest agents. We assume that W is symmetric, doubly stochastic, and respects the graph sparsity pattern, i.e.,*

$$W = W^\top, \qquad W\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\top W = \mathbf{1}^\top, \qquad w_{ij} = 0 \text{ if } (i,j) \text{ is not an edge}. \tag{5}$$

In addition, $w_{ii} > 0$ for all $i \in [n_h]$. We define the lifted mixing matrix as $\mathbf{W} := W \otimes \mathbf{I}_d \in \mathbb{R}^{dn_h\times dn_h}$, where $\otimes$ denotes the Kronecker product.

Under Assumption 1, if the communication graph is connected, then W has a unique unit eigenvalue, and all remaining eigenvalues $\{\lambda_i\}_{i=2}^{n_h}$ satisfy $|\lambda_i| < 1$ (Sayed et al., 2014). We define $\lambda := \max_{i\geq 2} |\lambda_i|$ as the second-largest eigenvalue magnitude, and let $\underline{\lambda}$ denote the smallest nonzero eigenvalue magnitude of W. The spectral gap $1-\lambda \in (0,1]$ quantifies the connectivity of the network, with larger values indicating faster information mixing. For example, exact averaging yields $1-\lambda = 1$, whereas standard mixing matrices on a ring topology have $1-\lambda = \Theta(1/n_h^2)$. The spectral gap also plays a key role in characterizing the *transient complexity*, i.e., the number of iterations required for the decentralized algorithm to enter its asymptotic linear-speedup regime.

Achieving Byzantine resilience in decentralized learning requires replacing the standard averaging step with a robust aggregation rule, such as the coordinate-wise median (CWMed) (Yin et al., 2018), Krum (Blanchard et al., 2017), or Iterative Outlier Scissor (IOS) (Wu et al., 2023). To quantify the accuracy of such aggregators, we adopt the (δ,κ)-robustness criterion introduced in (Allouah et al., 2023).

Definition 2 ((δ,κ)-robustness)**.** *Let $\delta < \frac{1}{2}$ and $\kappa \geq 0$. A robust aggregation rule $\mathcal{A}(\cdot): \mathbb{R}^{d\times n} \to \mathbb{R}^d$ is said to be (δ,κ)-robust if, for any agent i with neighbor set $\mathcal{N}_i$ and any collection of vectors $\{x_i\} \cup \{x_m\}_{m\in\mathcal{N}_i} \subset \mathbb{R}^d$, it holds that*

$$\left\|\mathcal{A}\big(x_i, \{x_m\}_{m\in\mathcal{N}_i}\big) - \bar{x}_i\right\|^2 \leq \kappa \sum_{j\in\mathcal{H}_i} w_{ij} \left\|x_j - \bar{x}_i\right\|^2, \tag{6}$$

where $\bar{x}_i = \sum_{j\in\mathcal{H}_i} w_{ij} x_j$, with $\mathcal{H}_i$ denoting agent i's honest neighborhood, including agent i itself.

The (δ,κ)-robustness criterion is satisfied by many commonly used robust aggregation rules (Blanchard et al., 2017; Acharya et al., 2022; Allouah et al., 2023). In particular, the lower bound $\kappa > \frac{\delta}{1-2\delta}$ (Allouah et al., 2023) implies that no robust aggregator can exactly recover the honest weighted average. Although the precise value of κ depends on the choice of aggregation rule, it can be reduced to $\mathcal{O}(\delta)$ by incorporating pre-aggregation techniques such as NNM (Allouah et al., 2023).

3 Methodology

In this section, we derive the hypergradient estimator for the bilevel problem (3) and present the proposed Byzantine-resilient decentralized bilevel optimization algorithm.

3.1 Hypergradient Estimation

A central challenge in solving the decentralized stochastic bilevel optimization (DSBO) problem (3) is that the lower-level solution $y^\star(x)$ is defined implicitly and hence is not directly available to individual agents. As a result, the hypergradient $\nabla\Phi_{\mathcal{H}}(x)$ cannot be evaluated in a straightforward manner. With the μ_g-strong convexity of the lower-level objective (Assumption 3), $y^\star(x)$ is uniquely defined. Moreover, by the implicit function theorem (Ghadimi & Wang, 2018), the hypergradient admits the representation

$$\nabla\Phi_{\mathcal{H}}(x) = \frac{1}{n_h}\sum_{i\in\mathcal{H}} \nabla_x f_i\left(x, y^\star\right) - \nabla^2_{xy} g_{\mathcal{H}}\left(x, y^\star\right)\left[\nabla^2_{yy} g_{\mathcal{H}}\left(x, y^\star\right)\right]^{-1} \frac{1}{n_h}\sum_{i\in\mathcal{H}} \nabla_y f_i\left(x, y^\star\right), \tag{7}$$

where $y^\star := y^\star(x)$ and $g_{\mathcal{H}} := \frac{1}{n_h}\sum_{i\in\mathcal{H}} g_i$. A direct evaluation of (7) requires global second-order information of $g_{\mathcal{H}}$, which is unavailable to any individual agent in the decentralized setting. To avoid explicitly computing the full Hessian and Jacobian matrices, we follow Chen et al. (2023) and introduce the auxiliary variable $z^\star := z^\star(x)$ defined by

$$z^\star = -\Big(\frac{1}{n_h}\sum_{i\in\mathcal{H}} \nabla^2_{yy} g_i\left(x, y^\star(x)\right)\Big)^{-1}\Big(\frac{1}{n_h}\sum_{i\in\mathcal{H}} \nabla_y f_i\left(x, y^\star(x)\right)\Big), \tag{8}$$

which represents the Hessian-inverse-vector product in (7). The expression (8), however, is not directly amenable to decentralization: the inverse Hessian couples $\{\nabla^2_{yy} g_i\}_{i\in\mathcal{H}}$ nonlinearly, so $z^\star$ does not admit a finite-sum decomposition across honest agents. Following the SOBA framework (Dagréou et al., 2022), which was recently extended to the decentralized setting in (Kong et al., 2025), we characterize $z^\star$ as the unique minimizer of a strongly convex quadratic program,

$$z^\star = \operatorname*{argmin}_{z\in\mathbb{R}^{d_y}} \frac{1}{n_h}\sum_{i\in\mathcal{H}} \Big\{\frac{1}{2} z^\top \nabla^2_{yy} g_i(x, y^\star)\, z \;+\; z^\top \nabla_y f_i(x, y^\star)\Big\}. \tag{9}$$

In contrast to (8), the objective in (9) is an honest average of local quadratics, enabling each agent to track $z^\star$ by a single (stochastic) gradient step on its local summand and thereby bypassing both Hessian inversion and any inner loop for the associated linear system. Substituting (8) into (7) yields

$$\nabla\Phi_{\mathcal{H}}(x) = \frac{1}{n_h}\sum_{i\in\mathcal{H}} \underbrace{\left(\nabla_x f_i\left(x, y^\star\right) + \nabla^2_{xy} g_i\left(x, y^\star\right) z^\star\right)}_{\text{local hypergradient contribution of agent } i}. \tag{10}$$

Therefore, the global hypergradient can be expressed as the average of local contributions. This decomposition suggests that each agent only needs to estimate three quantities, namely $\nabla_x f_i(x, y^\star)$, $z^\star$, and the Jacobian-vector product $\nabla^2_{xy} g_i(x, y^\star)z^\star$. In particular, tracking the vector-valued variable $z^\star$ avoids the need to form or invert full Hessian and Jacobian matrices, thereby substantially reducing the per-iteration computational and communication costs.

3.2 The Proposed Algorithm

Based on the hypergradient decomposition, we propose `SLBRD` (**S**ingle-**L**oop **B**yzantine-**R**esilient **D**ecentralized bilevel algorithm) and its momentum-accelerated variant `SLBRD-M`, both summarized in Algorithm 1. Each honest agent i maintains three local iterates, $x_i^k \in \mathbb{R}^{d_x}$, $y_i^k \in \mathbb{R}^{d_y}$, and $z_i^k \in \mathbb{R}^{d_y}$, tracking the upper-level variable, the lower-level solution $y^\star(x)$, and the auxiliary variable $z^\star$, respectively. At iteration k, each agent draws independent samples $\phi_i^k \sim \mathcal{D}_{f_i}$ and $\varphi_i^k \sim \mathcal{D}_{g_i}$, and invokes a stochastic bilevel oracle that, using a single Hessian-vector product $\nabla^2_{yy} G_i\, z_i^k$ and a single Jacobian-vector product $\nabla^2_{xy} G_i\, z_i^k$, constructs the stochastic directions

$$D_{i,x}^k = \nabla_x F_i(x_i^k, y_i^k; \phi_i^k) + \nabla^2_{xy} G_i(x_i^k, y_i^k; \varphi_i^k)\, z_i^k, \tag{11a}$$

$$D_{i,y}^k = \nabla_y G_i(x_i^k, y_i^k; \varphi_i^k), \tag{11b}$$

$$D_{i,z}^k = \nabla_y F_i(x_i^k, y_i^k; \phi_i^k) + \nabla^2_{yy} G_i(x_i^k, y_i^k; \varphi_i^k)\, z_i^k, \tag{11c}$$

where $D_{i,x}^k$ is a stochastic estimate of the local hypergradient, $D_{i,y}^k$ is the stochastic lower-level gradient, and $D_{i,z}^k$ is the stochastic gradient of agent i's local quadratic objective in (9), thereby driving z_i^k toward the auxiliary variable $z^\star$.

Using the stochastic directions returned by the oracle, each agent performs a local update followed by robust aggregation over its neighborhood. In the base `SLBRD` algorithm, the local update is given by $v_i^{k+\frac{1}{2}} = v_i^k - \eta_v D_{i,v}^k$ for $v \in \{x, y, z\}$, where $\eta_x, \eta_y, \eta_z > 0$ are the corresponding step sizes. To further reduce variance and accelerate convergence, `SLBRD-M` incorporates Polyak's heavy-ball momentum (Polyak, 1964), whereby each agent maintains momentum variables updated according to

$$m_{i,v}^k = \beta_v\, m_{i,v}^{k-1} + (1-\beta_v)\, D_{i,v}^k, \quad v \in \{x, y, z\}, \tag{12}$$

where $\beta_x, \beta_y, \beta_z \in [0, 1)$ are momentum parameters and $m_{i,v}^{-1} = 0$. Setting $\beta_v = 0$ recovers the base `SLBRD`. In this case, the local update becomes $v_i^{k+\frac{1}{2}} = v_i^k - \eta_v\, m_{i,v}^k$. After the local update, each agent sends its intermediate iterates $v_i^{k+\frac{1}{2}}$ to its neighbors and receives their corresponding messages, which may be arbitrarily corrupted by Byzantine agents. It then applies a robust aggregation rule:

$$v_i^{k+1} = \mathcal{A}\big(v_i^{k+\frac{1}{2}},\, \{v_m^{k+\frac{1}{2}}\}_{m\in\mathcal{N}_i}\big), \quad v \in \{x, y, z\}, \tag{13}$$

where $\mathcal{N}_i$ denotes the neighborhood of agent i, which may include Byzantine agents, and $\mathcal{A}(\cdot)$ is a nonlinear aggregation rule used to robustly combine the received messages.

After aggregation, the auxiliary iterate z_i^{k+1} is clipped via $z_i^{k+1} \leftarrow \min\{1,\, \rho/\|z_i^{k+1}\|_2\}\, z_i^{k+1}$, where $\rho > 0$ is a prescribed clipping radius. This step enforces $\|z_i^{k+1}\|_2 \le \rho$ and stabilizes the subsequent Jacobian-vector product evaluation. Importantly, `SLBRD` avoids both inner loops for approximating the lower-level solution $y^\star(x)$ and explicit Hessian inversion. As a result, the method reduces communication overhead and the attack surface exposed to Byzantine adversaries.

4 Theoretical Analysis

4.1 Assumptions

We now state the standard assumptions used throughout the paper (Ghadimi & Wang, 2018; Chen et al., 2023; Kong et al., 2025).

Algorithm 1 SLBRD: Single-Loop Byzantine-Resilient Decentralized Bilevel Algorithm

Require: Initializations x_i^0, y_i^0, z_i^0 and $m_{i,x}^{-1} = m_{i,y}^{-1} = m_{i,z}^{-1} = 0$ for all $i \in \mathcal{H}$; step sizes η_x, η_y, η_z; momentum coefficients $\beta_x, \beta_y, \beta_z$; robust aggregation rule $\mathcal{A}(\cdot)$; clipping radius $\rho > 0$.

for $k = 0, 1, \ldots, K-1$ **do**
  for each honest agent $i \in \mathcal{H}$ in parallel **do**
    Draw samples $\phi_i^k \sim \mathcal{D}_{f_i}$, $\varphi_i^k \sim \mathcal{D}_{g_i}$
    Compute the stochastic directions $D_{i,x}^k, D_{i,y}^k, D_{i,z}^k$ via (11) ▷ Stochastic bilevel oracle
    $m_{i,v}^k \leftarrow \beta_v\, m_{i,v}^{k-1} + (1-\beta_v)\, D_{i,v}^k, \quad v \in \{x, y, z\}$ ▷ Momentum ($\beta_v = 0$ for SLBRD)
    $v_i^{k+\frac{1}{2}} \leftarrow v_i^k - \eta_v\, m_{i,v}^k, \quad v \in \{x, y, z\}$ ▷ Local update
    Send $v_i^{k+\frac{1}{2}}$ to the neighbors of i and receive $\{v_m^{k+\frac{1}{2}}\}_{m \in \mathcal{N}_i}$ from neighbors $\mathcal{N}_i$.
    $v_i^{k+1} \leftarrow \mathcal{A}\big(v_i^{k+\frac{1}{2}}, \{v_m^{k+\frac{1}{2}}\}_{m \in \mathcal{N}_i}\big), \quad v \in \{x, y, z\}$ ▷ Robust aggregation
    $z_i^{k+1} \leftarrow \min\big\{1,\, \rho / \|z_i^{k+1}\|_2\big\} \cdot z_i^{k+1}$ ▷ Clip the auxiliary variable
  end for
end for

Assumption 2 (Smoothness). *For all $i \in \mathcal{H}$, there exist positive constants $L_{\nabla f}$, $L_{\nabla g}$, $L_{\nabla^2 g}$, and L_f such that ∇f_i, ∇g_i, and $\nabla^2 g_i$ are $L_{\nabla f}$-, $L_{\nabla g}$-, and $L_{\nabla^2 g}$-Lipschitz continuous, respectively, and $\|\nabla_y f_i(x, y^\star(x))\| \le L_f$ for all $x \in \mathbb{R}^{d_x}$.*

The above assumption ensures the Lipschitz continuity of $\nabla\Phi$, $y^\star(x)$, $z^\star(x)$. Notably, we relax the higher-order smoothness ($f \in \mathcal{C}_L^{2,2}, g \in \mathcal{C}_L^{3,3}$) required in prior SOBA analyses (Dagréou et al., 2022), as bounding $\nabla z^\star(x)$ is algorithmically unnecessary. Furthermore, rather than assuming global Lipschitz continuity or uniform boundedness of f, we only require boundedness along the optimal trajectory.

Assumption 3 (μ_g-strong convexity). *For all $i \in \mathcal{H}$, $g_i(x, \cdot)$ is μ_g-strongly convex for every $x \in \mathbb{R}^{d_x}$.*

Assumption 4 (Bounded stochastic noise). *For all $i \in \mathcal{H}$, the stochastic first- and second-order oracles of f_i and g_i are unbiased with variance bounded by σ^2.*

Assumption 5 (Bounded gradient dissimilarity). *There exist constants $\zeta_f, \zeta_g \ge 0$ such that for all $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$,*

$$\frac{1}{n_h} \sum_{i \in \mathcal{H}} \|\nabla_\alpha f_i(x, y) - \nabla_\alpha f_{\mathcal{H}}(x, y)\|^2 \le \zeta_f^2, \quad \alpha \in \{x,\, y\},$$

and the analogous bound with constant ζ_g^2 holds for the lower-level derivatives $\nabla_y g_i$, $\nabla_{xy}^2 g_i$, and $\nabla_{yy}^2 g_i$ relative to their honest averages, where $f_{\mathcal{H}} := \frac{1}{n_h} \sum_{i \in \mathcal{H}} f_i$ and $g_{\mathcal{H}} := \frac{1}{n_h} \sum_{i \in \mathcal{H}} g_i$.

4.2 Convergence Results

We next present the convergence guarantees for SLBRD and its momentum-enhanced variant SLBRD−M. The full proofs are deferred to Appendices D and E, respectively.

Theorem 1 (Convergence of SLBRD). *Suppose Assumptions 1–5 hold, and the robust aggregator $\mathcal{A}(\cdot)$ satisfies Definition 2 with parameter κ. Set the step sizes as $\eta_x = \Theta\big(1/\sqrt{n_h K}\big)$, $\eta_y = c_y \eta_x$, $\eta_z = c_z \eta_x$ for problem-dependent constants $c_y, c_z > 0$. Then the iterates of SLBRD satisfy*

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\big[\|\nabla \Phi_{\mathcal{H}}(\bar{x}^k)\|^2\big] \le\ & \underbrace{\mathcal{O}\Big(\frac{1}{\sqrt{n_h K}}\Big)}_{\text{dominant term}} + \underbrace{\mathcal{O}\Big(\frac{n_h \sigma^2}{K(1-\lambda)} + \frac{n_h(\zeta_f^2 + \zeta_g^2)}{K(1-\lambda)^2}\Big)}_{\text{topology-related terms}} \\ & + \underbrace{\mathcal{O}\Big(\frac{\kappa\, n_h \sigma^2}{K(1-\lambda)^2} + \frac{\kappa\, n_h (\zeta_f^2 + \zeta_g^2)}{K(1-\lambda)^2}\Big)}_{\text{Byzantine-induced terms}} + \underbrace{\mathcal{O}\Big(\kappa\big(\sigma^2 + \zeta_f^2 + \zeta_g^2\big)\Big)}_{\text{asymptotic error}}, \end{aligned} \tag{14}$$

where $\bar{x}^k := \frac{1}{n_h} \sum_{i \in \mathcal{H}} x_i^k$ denotes the average iterate over the honest agents.

Remark 1 (Convergence of SLBRD)**.** The leading term $\mathcal{O}(1/\sqrt{n_h K})$ in Theorem 1 shows that SLBRD achieves linear speedup with respect to the number of honest agents n_h. Moreover, when $\kappa = \mathcal{O}(\delta)$, the non-vanishing asymptotic error reduces to $\mathcal{O}\big(\delta(\sigma^2 + \zeta_f^2 + \zeta_g^2)\big)$. Compared with the fundamental lower bound $\Omega\big(\delta(\zeta_f^2 + \zeta_g^2)\big)$ established in (Li et al.), this asymptotic error contains an additional variance-dependent term $\mathcal{O}(\delta\sigma^2)$. This gap motivates the momentum-enhanced variant SLBRD−M, which removes the effect of stochastic variance in the asymptotic error and attains the lower bound.

Corollary 1 (Transient time of SLBRD)**.** *Under the conditions of Theorem 1, suppose $\mathcal{A}(\cdot)$ satisfies the (δ, κ)-robustness criterion with $\kappa = \mathcal{O}(\delta)$, then the transient time τ of SLBRD satisfies*

$$\tau = \mathcal{O}\left(\max\left\{\frac{n_h^3}{(1-\lambda)^2}, \frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4}, \frac{\delta^2 n_h^3}{(1-\lambda)^4}\right\}\right). \tag{15}$$

Remark 2 (Transient complexity of SLBRD)**.** In contrast to prior works on Byzantine-resilient decentralized learning (Farhadkhani et al., 2023), which focus primarily on asymptotic complexity, Corollary 1 explicitly characterizes how network topology, data heterogeneity, the Byzantine fraction, and the nested bilevel structure jointly affect the transient performance of SLBRD. In particular, sparse communication topologies, corresponding to a small spectral gap $1-\lambda$, significantly amplify the effect of heterogeneity across both the upper and lower levels. Moreover, even in the homogeneous setting where $\zeta_f = \zeta_g = 0$, a large Byzantine fraction δ further aggravates the effect of poor connectivity, worsening the topology-dependent transient complexity from $\mathcal{O}\big((1-\lambda)^{-2}\big)$ to $\mathcal{O}\big((1-\lambda)^{-4}\big)$.

Theorem 2 (Convergence of SLBRD−M)**.** *Under the same conditions as Theorem 1, by choosing the momentum coefficients such that $1-\beta_v = \Theta(\eta_v L_{\beta_v})$ for all $v \in \{x, y, z\}$, where $L_{\beta_x}, L_{\beta_y}, L_{\beta_z} > 0$ are problem-dependent constants, the iterates generated by SLBRD−M in Algorithm 1 satisfy:*

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla\Phi_{\mathcal{H}}(\bar{x}^k)\|^2\big] \le{}& \underbrace{\mathcal{O}\Big(\frac{1}{\sqrt{n_h K}}\Big)}_{\text{dominant term}} + \underbrace{\mathcal{O}\Big(\frac{n_h^{3/2}\sigma^2}{K^{3/2}(1-\lambda)} + \frac{n_h(\zeta_f^2+\zeta_g^2)}{K(1-\lambda)^2}\Big)}_{\text{topology-related terms}} \\ &+ \underbrace{\mathcal{O}\Big(\frac{\kappa\, n_h^{3/2}\sigma^2}{K^{3/2}(1-\lambda)^2} + \frac{\kappa\, n_h\left(\zeta_f^2+\zeta_g^2\right)}{K(1-\lambda)^2}\Big)}_{\text{Byzantine-induced terms}} + \underbrace{\mathcal{O}\Big(\kappa\left(\zeta_f^2+\zeta_g^2\right)\Big)}_{\text{asymptotic error}},\end{aligned} \tag{16}$$

where $\bar{x}^k := \frac{1}{n_h}\sum_{i\in\mathcal{H}} x_i^k$ is the honest iterate average.

Remark 3 (Convergence of SLBRD−M)**.** Comparing (16) with (14), momentum yields two key improvements. First, the variance-dependent transient term is improved from $\mathcal{O}\big(n_h\sigma^2/(K(1-\lambda))\big)$ to $\mathcal{O}\big(n_h^{3/2}\sigma^2/(K^{3/2}(1-\lambda))\big)$, leading to faster convergence in the stochastic regime. Second, and more importantly, the stochastic variance term σ^2 is completely removed from the non-vanishing Byzantine-induced error, which is reduced from $\mathcal{O}\big(\kappa(\sigma^2+\zeta_f^2+\zeta_g^2)\big)$ to $\mathcal{O}\big(\kappa(\zeta_f^2+\zeta_g^2)\big)$. In particular, when $\kappa = \mathcal{O}(\delta)$, the asymptotic error becomes $\mathcal{O}\big(\delta(\zeta_f^2+\zeta_g^2)\big)$, which matches the lower bound in (Li et al.). Therefore, SLBRD−M achieves the optimal asymptotic error under Byzantine attack.

Corollary 2 (Transient time of SLBRD−M)**.** *Under the conditions of Theorem 2, suppose $\mathcal{A}(\cdot)$ satisfies the (δ, κ)-robustness criterion with $\kappa = \mathcal{O}(\delta)$, then the transient time τ of SLBRD−M satisfies*

$$\tau = \mathcal{O}\left(\max\left\{\frac{n_h^2}{1-\lambda}, \frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4}, \frac{\delta^2 n_h^3}{(1-\lambda)^4}\right\}\right). \tag{17}$$

Remark 4 (Transient complexity of SLBRD−M)**.** Compared with Corollary 1 for SLBRD, Corollary 2 shows that momentum strictly improves the topology-dependent transient complexity from $\mathcal{O}\big((1-\lambda)^{-2}\big)$ to $\mathcal{O}\big((1-\lambda)^{-1}\big)$. This improvement is most pronounced when the heterogeneity- and Byzantine-dependent terms are of lower order. Consequently, the acceleration effect of SLBRD−M is most significant in regimes with mild heterogeneity and a low Byzantine fraction.

5 Experiments

We evaluate the proposed methods on hyperparameter optimization and hyper-representation tasks over decentralized networks in the presence of Byzantine agents. As baselines, we consider two

Table 2: Test accuracy (%) on the synthetic **hyperparameter optimization** task with heterogeneous data. Column blocks correspond to aggregators Krum, Median, TrMean, and IOS. "Worst" reports the worst accuracy across all attacks. **Bold** marks the best and underline the second best in each column.

Algorithm	ALIE	SF	IPM	RN	Worst	ALIE	SF	IPM	RN	Worst
MA-DSBO	$70.73_{\pm 1.02}$	$64.50_{\pm 0.29}$	$69.33_{\pm 2.42}$	$70.60_{\pm 0.51}$	64.50	$72.83_{\pm 2.33}$	$69.13_{\pm 1.25}$	$70.83_{\pm 2.05}$	$72.73_{\pm 0.37}$	69.13
MA-DSBO-M	$75.33_{\pm 2.08}$	$81.27_{\pm 0.50}$	$82.23_{\pm 0.90}$	$96.03_{\pm 0.52}$	75.33	$82.90_{\pm 1.53}$	$80.67_{\pm 2.74}$	$80.53_{\pm 1.81}$	$82.60_{\pm 1.10}$	80.53
GBDSBO	$65.47_{\pm 1.39}$	$65.37_{\pm 2.22}$	$68.90_{\pm 1.10}$	$71.60_{\pm 1.22}$	65.37	$69.87_{\pm 0.95}$	$67.90_{\pm 1.81}$	$69.60_{\pm 2.45}$	$73.47_{\pm 0.62}$	67.90
GBDSBO-M	$78.50_{\pm 0.29}$	$86.30_{\pm 0.33}$	$87.50_{\pm 1.69}$	$88.50_{\pm 0.57}$	78.50	$81.60_{\pm 0.51}$	$88.67_{\pm 0.76}$	$87.50_{\pm 0.08}$	$89.43_{\pm 1.31}$	81.60
SLBRD	$76.77_{\pm 2.49}$	$\mathbf{93.07_{\pm 0.29}}$	$\mathbf{96.87_{\pm 0.58}}$	$\mathbf{96.73_{\pm 0.50}}$	76.77	$76.30_{\pm 2.33}$	$65.50_{\pm 2.40}$	$87.83_{\pm 1.31}$	$90.83_{\pm 1.43}$	65.50
SLBRD-M	$\mathbf{94.83_{\pm 0.62}}$	$90.87_{\pm 0.90}$	$96.00_{\pm 0.94}$	$96.07_{\pm 0.78}$	**90.87**	$\mathbf{94.50_{\pm 0.49}}$	$\mathbf{92.87_{\pm 1.02}}$	$\mathbf{96.00_{\pm 0.80}}$	$\mathbf{96.50_{\pm 0.99}}$	**92.87**
Algorithm	ALIE	SF	IPM	RN	Worst	ALIE	SF	IPM	RN	Worst
MA-DSBO	$69.77_{\pm 3.39}$	$68.90_{\pm 1.71}$	$70.20_{\pm 0.86}$	$72.73_{\pm 1.01}$	68.90	$72.17_{\pm 1.35}$	$65.20_{\pm 2.19}$	$71.23_{\pm 1.16}$	$70.90_{\pm 1.57}$	65.20
MA-DSBO-M	$75.77_{\pm 2.70}$	$79.30_{\pm 1.10}$	$82.43_{\pm 1.75}$	$82.97_{\pm 1.20}$	75.77	$74.17_{\pm 1.84}$	$80.43_{\pm 0.19}$	$82.20_{\pm 0.92}$	$95.83_{\pm 1.05}$	74.17
GBDSBO	$66.33_{\pm 0.96}$	$68.27_{\pm 2.78}$	$70.50_{\pm 1.08}$	$71.53_{\pm 1.01}$	66.33	$65.47_{\pm 1.39}$	$65.87_{\pm 0.39}$	$69.50_{\pm 0.29}$	$70.67_{\pm 1.44}$	65.47
GBDSBO-M	$78.77_{\pm 0.21}$	$89.23_{\pm 1.02}$	$88.33_{\pm 0.09}$	$88.30_{\pm 1.07}$	78.77	$78.63_{\pm 0.25}$	$87.00_{\pm 0.43}$	$87.53_{\pm 0.45}$	$89.93_{\pm 0.56}$	78.63
SLBRD	$70.97_{\pm 2.36}$	$93.23_{\pm 0.54}$	$\mathbf{97.20_{\pm 0.70}}$	$96.93_{\pm 0.88}$	70.97	$77.10_{\pm 2.73}$	$\mathbf{93.60_{\pm 0.51}}$	$\mathbf{96.80_{\pm 0.33}}$	$\mathbf{97.00_{\pm 0.73}}$	77.10
SLBRD-M	$\mathbf{94.07_{\pm 0.45}}$	$\mathbf{95.53_{\pm 0.34}}$	$96.80_{\pm 0.79}$	$\mathbf{97.17_{\pm 0.77}}$	**94.07**	$\mathbf{94.87_{\pm 0.59}}$	$82.53_{\pm 2.46}$	$96.07_{\pm 0.58}$	$96.77_{\pm 0.63}$	**82.53**

Table 3: Test accuracy (%) on the MNIST **hyper-representation** task with heterogeneous agents. Column blocks correspond to aggregators Krum, Median, TrMean, and IOS. "Worst" reports the worst accuracy across all attacks. **Bold** marks the best and underline the second best in each column.

Algorithm	ALIE	SF	IPM	RN	Worst	ALIE	SF	IPM	RN	Worst
MA-DSBO	$88.96_{\pm 0.28}$	$88.34_{\pm 0.34}$	$88.32_{\pm 0.33}$	$88.33_{\pm 0.31}$	88.32	$89.09_{\pm 0.04}$	$86.91_{\pm 0.46}$	$86.91_{\pm 0.43}$	$88.06_{\pm 0.04}$	86.91
MA-DSBO-M	$92.76_{\pm 0.20}$	$91.04_{\pm 1.54}$	$92.10_{\pm 0.46}$	$92.20_{\pm 0.46}$	91.04	$92.88_{\pm 0.07}$	$90.23_{\pm 0.24}$	$90.23_{\pm 0.30}$	$\mathbf{92.73_{\pm 0.19}}$	90.23
GBDSBO	$66.31_{\pm 21.95}$	$82.32_{\pm 0.35}$	$83.74_{\pm 0.56}$	$87.78_{\pm 1.38}$	66.31	$86.06_{\pm 1.77}$	$82.53_{\pm 1.87}$	$82.44_{\pm 1.56}$	$86.28_{\pm 0.79}$	82.44
GBDSBO-M	$87.75_{\pm 0.07}$	$81.70_{\pm 0.94}$	$83.30_{\pm 0.92}$	$87.39_{\pm 1.14}$	81.70	$82.47_{\pm 2.60}$	$80.78_{\pm 2.11}$	$80.62_{\pm 1.94}$	$84.56_{\pm 0.41}$	80.62
SLBRD	$92.48_{\pm 0.35}$	$92.35_{\pm 0.06}$	$92.38_{\pm 0.07}$	$93.10_{\pm 0.09}$	92.35	$92.97_{\pm 0.00}$	$85.90_{\pm 1.28}$	$85.83_{\pm 1.33}$	$91.30_{\pm 0.25}$	85.83
SLBRD-M	$\mathbf{93.78_{\pm 2.19}}$	$\mathbf{95.27_{\pm 1.21}}$	$\mathbf{94.63_{\pm 0.46}}$	$\mathbf{95.60_{\pm 0.27}}$	**93.78**	$\mathbf{94.83_{\pm 0.25}}$	$\mathbf{93.45_{\pm 0.68}}$	$\mathbf{93.08_{\pm 0.76}}$	$92.34_{\pm 0.10}$	**92.34**
Algorithm	ALIE	SF	IPM	RN	Worst	ALIE	SF	IPM	RN	Worst
MA-DSBO	$89.22_{\pm 0.00}$	$85.57_{\pm 0.76}$	$85.55_{\pm 0.77}$	$89.03_{\pm 0.06}$	85.55	$89.08_{\pm 0.14}$	$89.35_{\pm 0.04}$	$89.34_{\pm 0.02}$	$89.34_{\pm 0.04}$	89.08
MA-DSBO-M	$92.86_{\pm 0.20}$	$89.72_{\pm 0.58}$	$89.77_{\pm 0.60}$	$93.07_{\pm 0.06}$	89.72	$92.91_{\pm 0.11}$	$92.88_{\pm 0.16}$	$92.91_{\pm 0.14}$	$93.04_{\pm 0.10}$	92.88
GBDSBO	$50.88_{\pm 37.35}$	$76.35_{\pm 0.67}$	$53.22_{\pm 30.71}$	$86.56_{\pm 0.46}$	50.88	$88.32_{\pm 0.25}$	$58.73_{\pm 35.23}$	$84.84_{\pm 0.67}$	$88.56_{\pm 0.25}$	58.73
GBDSBO-M	$88.32_{\pm 0.45}$	$77.38_{\pm 2.33}$	$71.33_{\pm 3.92}$	$87.52_{\pm 0.61}$	71.33	$48.98_{\pm 39.16}$	$59.14_{\pm 34.69}$	$84.89_{\pm 0.72}$	$88.62_{\pm 0.32}$	48.98
SLBRD	$92.66_{\pm 0.08}$	$86.81_{\pm 0.81}$	$86.76_{\pm 0.74}$	$93.63_{\pm 0.03}$	86.76	$93.11_{\pm 0.14}$	$93.30_{\pm 0.04}$	$93.30_{\pm 0.06}$	$93.92_{\pm 0.01}$	93.11
SLBRD-M	$\mathbf{94.07_{\pm 0.18}}$	$\mathbf{90.68_{\pm 0.75}}$	$\mathbf{90.61_{\pm 0.60}}$	$\mathbf{96.68_{\pm 0.04}}$	**90.61**	$\mathbf{94.35_{\pm 0.14}}$	$\mathbf{94.56_{\pm 0.10}}$	$\mathbf{95.98_{\pm 0.14}}$	$\mathbf{96.56_{\pm 0.01}}$	**94.35**

representative sub-loop decentralized bilevel algorithms, namely MA-DSBO (Chen et al., 2023) and GBDSBO (Yang et al., 2022), together with their momentum-enhanced variants MA-DSBO-M and GBDSBO-M. To comprehensively assess robustness, we evaluate these methods under four representative Byzantine attack models: ALIE (Baruch et al., 2019), Sign Flipping (SF), Inner Product Manipulation (IPM) (Xie et al., 2020), and Random Noise (RN). For each attack setting, we equip each method with four well-established robust aggregation rules: Krum (Blanchard et al., 2017), Median, Trimmed Mean (Yin et al., 2018), and Iterative Outlier Scissor (IOS) (Wu et al., 2023). Detailed experimental settings are deferred to Appendix A.

5.1 Decentralized Hyperparameter Optimization

We formulate hyperparameter optimization as the following bilevel problem:

$$\min_{\lambda} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_v^i|} \sum_{\phi \in \mathcal{D}_v^i} L_i\big(\lambda, \theta^\star(\lambda); \phi\big), \text{ s.t. } \theta^\star(\lambda) = \operatorname*{argmin}_{\theta} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_\tau^i|} \sum_{\varphi \in \mathcal{D}_\tau^i} L_i(\lambda, \theta; \varphi) + R(\lambda, \theta), \tag{18}$$

where $\mathcal{H}$ denotes the set of honest agents, $\mathcal{D}_\tau^i$ and $\mathcal{D}_v^i$ represent the local training and validation sets of agent i, respectively, and $R(\lambda, \theta)$ is a dataset-dependent regularizer. We evaluate (18) on two benchmark tasks: a synthetic dataset with agent-dependent covariance $\mathcal{N}(0, i^2 I)$ (Chen et al., 2023), and MNIST (LeCun, 1998) with an ℓ_2-regularized logistic regression model under Dirichlet partitioning with $\alpha = 0.5$. In all experiments, we set the Byzantine fraction to $\delta = 0.3$, use SGD with batch size 64, and run the algorithms for 1000 communication rounds on the synthetic task and 2000 rounds on MNIST. Additional details are provided in Appendix A.1.

Table 2 summarizes the test accuracy on the synthetic task, while the corresponding results on MNIST are deferred to Appendix A.2.2. Overall, the proposed methods SLBRD and SLBRD-M consistently outperform the sub-loop baselines across both datasets. In contrast, the sub-loop baselines, MA-DSBO and GBDSBO, are substantially more vulnerable to Byzantine perturbations, often suffering accuracy drops of more than 20% under SF and IPM. On the synthetic task, SLBRD-M achieves the best

worst-case accuracy under every robust aggregator, with particularly clear gains under the variance-exploiting ALIE attack. On MNIST, the two proposed variants perform comparably, and for each robust aggregator the best worst-case accuracy is attained by one of them.

5.2 Decentralized Hyper-representation

Following Franceschi et al. (2018), we consider the following hyper-representation bilevel problem:

$$\min_{\lambda} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_v^i|} \sum_{\phi \in \mathcal{D}_v^i} f_i(\lambda, \theta^\star(\lambda); \phi), \quad \text{s.t.} \quad \theta^\star(\lambda) = \operatorname*{argmin}_{\theta} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_\tau^i|} \sum_{\varphi \in \mathcal{D}_\tau^i} g_i(\lambda, \theta; \varphi),$$

where $\mathcal{D}_\tau^i$ and $\mathcal{D}_v^i$ denote agent i's training and validation sets. The last fully connected layer is the lower-level variable θ, and the remaining network parameters form the upper-level variable λ. We use a two-layer MLP with 200 hidden units on MNIST for 500 communication rounds and a three-convolutional-layer CNN on CIFAR-10 (Krizhevsky et al., 2009) for 1000 rounds. Both settings use Byzantine fraction $\delta = 0.3$ and batch size 64.

We conduct hyper-representation experiments on MNIST and CIFAR-10 under the four robust aggregation rules and attack models above. The test accuracy on MNIST is reported in Table 3, while CIFAR-10 results and convergence curves are deferred to Appendices A.2.3 and A.2.4. Overall, the proposed single-loop methods `SLBRD` and `SLBRD-M` demonstrate consistently stronger robustness than the sub-loop baselines across both datasets. On MNIST, `SLBRD` and `SLBRD-M` achieve the highest test accuracy in nearly every aggregator–attack configuration, whereas the sub-loop baselines exhibit noticeable performance degradation, particularly under SF and ALIE. On CIFAR-10, this advantage becomes even more evident. The sub-loop baseline `GBDSBO` experiences substantial accuracy degradation, along with its momentum variant, and occasionally collapses to chance-level accuracy under certain configurations. In contrast, the proposed single-loop methods remain stable throughout training and consistently attain higher test accuracy.

5.3 Ablation Studies

We conduct ablation studies on the MNIST hyper-representation task to further assess the robustness of `SLBRD` and `SLBRD-M` across different Byzantine fractions, network sizes, and levels of data heterogeneity. Complete results are reported in Appendix A.3.

(1) Byzantine fraction. We vary the Byzantine fraction as $\delta \in \{0.1, 0.2, 0.3, 0.4\}$ while keeping all other settings fixed. As δ increases, both `SLBRD` and `SLBRD-M` only exhibit gradual performance degradation and remain effective even when 40% of the agents are adversarial. Across the entire range of δ, the proposed methods consistently outperform the baselines under all attack models.

(2) Network size. We fix $\delta = 0.3$ and vary the number of agents as $n \in \{10, 20, 30\}$. Both proposed methods converge reliably across all tested network sizes and maintain a clear advantage over the sub-loop baselines. This indicates that the robustness gain of the proposed methods is not tied to a specific network scale.

(3) Data heterogeneity. We further assess robustness to non-IID data by partitioning MNIST across agents using a Dirichlet distribution with $\alpha \in \{0.1, 0.5, 1\}$, where a smaller α indicates stronger data heterogeneity. Across all settings, `SLBRD` and `SLBRD-M` consistently achieve superior performance over the sub-loop baselines and exhibit stable behavior under varying levels of heterogeneity.

6 Conclusion

We proposed `SLBRD` and its momentum variant `SLBRD-M`, the first Byzantine-resilient decentralized bilevel algorithms with provable convergence under standard smoothness assumptions. The proposed single-loop, Hessian-inversion-free architecture reduces communication overhead and limits the attack surface exposed to Byzantine adversaries. Our analysis characterizes the effects of network topology, data heterogeneity, and Byzantine corruption on both transient and asymptotic performance. The momentum variant removes stochastic variance from the asymptotic error and attains the corresponding lower-bound dependence on Byzantine fraction and honest-client heterogeneity. Experiments on hyperparameter optimization and hyper-representation tasks demonstrate the effectiveness of both methods. Extending the framework and its convergence analysis to more general convex lower-level objectives remains an important direction for future work.

AI USE STATEMENT

We used generative AI tools solely for language polishing to improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted edits and take full responsibility for the final content of this work.

REFERENCES

Momin Abbas, Yi Zhou, Nathalie Baracaldo, Horst Samulowitz, Parikshit Ram, and Theodoros Salonidis. Byzantine-resilient bilevel federated learning. In *2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 1–5. IEEE, 2024.

Anish Acharya, Abolfazl Hashemi, Prateek Jain, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Robust training in high dimensions via block coordinate geometric median descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 11145–11168. PMLR, 2022.

Sulaiman A Alghunaim and Kun Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing*, 70:3264–3279, 2022.

Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. *Advances in neural information processing systems*, 31, 2018.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1232–1300. PMLR, 2023.

Joe Armstrong. *Making reliable distributed systems in the presence of software errors*. PhD thesis, 2003.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8): 1889–1900, 2000.

Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxnZh0ct7.

Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pp. 810–821. PMLR, 2020.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

Jin-Hua Chen, Min-Rong Chen, Guo-Qiang Zeng, and Jia-Si Weng. Bdfl: A byzantine-fault-tolerance decentralized federated learning method for autonomous vehicle. *IEEE Transactions on Vehicular Technology*, 70(9):8639–8652, 2021.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishnakumar Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International conference on machine learning*, pp. (to appear). PMLR, 2023.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024.

Xuxing Chen, Minhui Huang, and Shiqian Ma. Decentralized bilevel optimization: X. chen et al. *Optimization Letters*, 19(7):1249–1313, 2025.

Ziqin Chen and Yongqiang Wang. Locally differentially private decentralized stochastic bilevel optimization with guaranteed convergence accuracy. In *Forty-first International Conference on Machine Learning*, 2024.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=wlEOsQ917F.

Youran Dong, Shiqian Ma, Junfeng Yang, and Chao Yin. A single-loop algorithm for decentralized bilevel optimization. *Mathematics of Operations Research*, 2025.

El Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in neural information processing systems*, 34:25044–25057, 2021.

Cheng Fang, Zhixiong Yang, and Waheed U Bajwa. Bridge: Byzantine-resilient decentralized gradient descent. *IEEE Transactions on Signal and Information Processing over Networks*, 8: 610–626, 2022.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pp. 6246–6283. PMLR, 2022.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. Robust collaborative learning with linear gradient overhead. In *International Conference on Machine Learning*, pp. 9761–9813. PMLR, 2023.

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

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

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

Eduard Gorbunov, Samuel Horváth, Peter Richtárik, and Gauthier Gidel. Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top. *arXiv preprint arXiv:2206.00529*, 2022.

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

Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via clippedgossip. *arXiv preprint arXiv:2202.01545*, 2022.

Yutong He, Jie Hu, Xinmeng Huang, Songtao Lu, Bin Wang, and Kun Yuan. Distributed bilevel optimization with communication compression. *arXiv preprint arXiv:2405.18858*, 2024.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. *arXiv preprint arXiv:2006.09365*, 2020.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pp. 5311–5319. PMLR, 2021.

Harsh Kasyap, Ugur Ilker Atmaca, and Carsten Maple. Privacy-preserving personalised federated learning financial fraud detection. In *International Conference on AI and the Digital Economy (CADE 2024)*, volume 2024, pp. 87–88. IET, 2024.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pp. 5381–5393. PMLR, 2020.

Boao Kong, Shuchen Zhu, Songtao Lu, Xinmeng Huang, and Kun Yuan. Decentralized bilevel optimization over graphs: Loopless algorithmic update and transient iteration complexity. *arXiv e-prints*, pp. arXiv–2402, 2024.

Boao Kong, Shuchen Zhu, Songtao Lu, Xinmeng Huang, and Kun Yuan. Decentralized bilevel optimization: A perspective from transient iteration complexity. *Journal of Machine Learning Research*, 26(240):1–64, 2025.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yangnan Li, Shenghui Song, and Xuanyu Cao. Single-loop byzantine-resilient federated bilevel optimization. In *The Fourteenth International Conference on Learning Representations*.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Shuo Liu, Nirupam Gupta, and Nitin H Vaidya. Approximate byzantine fault-tolerance in distributed optimization. In *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, pp. 379–389, 2021.

Yuchen Liu, Chen Chen, Lingjuan Lyu, Fangzhao Wu, Sai Wu, and Gang Chen. Byzantine-robust learning on heterogeneous data via gradient splitting. In *International Conference on Machine Learning*, pp. 21404–21425. PMLR, 2023.

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

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

Tomoya Murata, Kenta Niwa, Takumi Fukami, and Iifan Tyou. Simple minimax optimal byzantine robust algorithm for nonconvex objectives with uniform gradient heterogeneity. In *The Twelfth International Conference on Learning Representations*, 2024.

Youcheng Niu, Jinming Xu, Ying Sun, Yan Huang, and Li Chai. Distributed stochastic bilevel optimization: Improved complexity and heterogeneity analysis. *Journal of Machine Learning Research*, 26(76):1–58, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Jie Peng, Weiyu Li, and Qing Ling. Byzantine-robust decentralized stochastic optimization over static and time-varying networks. *Signal Processing*, 183:108020, 2021.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

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

Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort. Cadda: Class-wise automatic differentiable data augmentation for eeg signals. In *ICLR 2022-International Conference on Learning Representations*, 2022.

Ali H Sayed et al. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014.

Lili Su and Nitin H Vaidya. Byzantine-resilient multiagent optimization. *IEEE Transactions on Automatic Control*, 66(5):2227–2233, 2020.

Zhaoxian Wu, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE transactions on signal processing*, 71:3179–3195, 2023.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pp. 261–270. PMLR, 2020.

Shuoguang Yang, Xuezhou Zhang, and Mengdi Wang. Decentralized gossip-based stochastic bilevel optimization over communication networks. *Advances in neural information processing systems*, 35:238–252, 2022.

Zhixiong Yang and Waheed U Bajwa. Byrdie: Byzantine-resilient distributed coordinate descent for decentralized learning. *IEEE Transactions on Signal and Information Processing over Networks*, 5 (4):611–627, 2019.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pp. 5650–5659. Pmlr, 2018.

Xuezhou Zhang, Xiaojin Zhu, and Laurent Lessard. Online data poisoning attacks. In *Learning for Dynamics and Control*, pp. 201–210. PMLR, 2020.

Yihua Zhang, Yuguang Yao, Parikshit Ram, Pu Zhao, Tianlong Chen, Mingyi Hong, Yanzhi Wang, and Sijia Liu. Advancing model pruning via bi-level optimization. In *Advances in Neural Information Processing Systems*, 2022.

Shuchen Zhu, Boao Kong, Songtao Lu, Xinmeng Huang, and Kun Yuan. Sparkle: a unified single-loop primal-dual framework for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 37:62912–62987, 2024.

Appendix for "SLBRD: A Single-Loop and Byzantine-Resilient Decentralized Bilevel Algorithm"

A Experiment 14

B Notations and Technical Lemmas 26

C Convergence Analysis: Proof of Theorem 1 (SLBRD) 27

D Convergence Analysis: Proof of Theorem 2 (SLBRD-M) 49

A Experiment

A.1 Implementation Details

This subsection provides comprehensive implementation details for our experimental evaluation. We first describe the computational environment and software framework used for all experiments. We then introduce the datasets employed in our evaluation, including both synthetic data for controlled analysis and real-world benchmark datasets. Subsequently, we detail the four Byzantine attack strategies used to evaluate robustness, and finally describe the four robust aggregation mechanisms integrated with the proposed methods.

A.1.1 Computational Environment

All methods were implemented using the PyTorch framework (Paszke et al., 2019) and executed on NVIDIA A100 GPUs with 40GB of memory or V100 GPUs with 32GB of memory. This hardware configuration ensures sufficient computational resources for training deep neural networks in the decentralized setting.

A.1.2 Decentralized Hyperparameter Optimization

We evaluate the proposed methods on both synthetic and real-world datasets to demonstrate their effectiveness across different problem settings and data distributions.

Synthetic Data. Following Chen et al. (2023), we instantiate (18) as a decentralized ℓ_2-regularized logistic regression problem:

$$\min_{\lambda\in\mathbb{R}^{200}} \frac{1}{n_h}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{D}_v^i|}\sum_{(x_{i,e},y_{i,e})\in\mathcal{D}_v^i} L\big(y_{i,e}\, x_{i,e}^\top \theta^\star(\lambda)\big)\,,$$

$$\text{s.t.}\quad \theta^\star(\lambda)=\operatorname*{argmin}_{\theta\in\mathbb{R}^{200}} \frac{1}{n_h}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{D}_\tau^i|}\sum_{(x_{i,e},y_{i,e})\in\mathcal{D}_\tau^i} L\big(y_{i,e}\, x_{i,e}^\top \theta\big)+\tfrac{1}{2}\sum_{s=1}^{200} e^{\lambda_s}\,\theta_s^2,$$

where $L(u)=\log(1+e^{-u})$ is the logistic loss, $\mathcal{H}$ is the set of honest agents with $n_h=|\mathcal{H}|$, $x_{i,e}\in\mathbb{R}^{200}$ and $y_{i,e}\in\mathbb{R}$ are the feature vector and label of a sample, λ_s,θ_s are the s-th entries of λ and θ, and $\mathcal{D}_\tau^i,\mathcal{D}_v^i$ are the training and validation datasets of agent i.

To simulate heterogeneous data distributions across agents, for each agent i, the features $x_{i,e}\in\mathbb{R}^{200}$ are drawn from a normal distribution $\mathcal{N}\left(0,i^2 I\right)$, which introduces heterogeneity through different variances. Given a fixed ground-truth vector $\theta^\star\in\mathbb{R}^{200}$, the label $y_{i,e}$ is generated as $y_{i,e}=\operatorname{sign}\big(x_{i,e}^\top\theta^\star+0.1\,\varepsilon\big)$, where $\varepsilon\sim\mathcal{N}(0,1)$ is an additive scalar noise. The algorithm was executed for 1000 communication rounds, with each agent randomly selecting 64 training samples in every round. The test dataset contains 20,000 samples, with 1,000 samples randomly selected for each round.

MNIST. Following Grazzi et al. (2020), we instantiate (18) as the decentralized ℓ_2-regularized multinomial logistic regression problem:

$$\begin{aligned} &\min_{\lambda \in \mathbb{R}^p} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_v^i|} \sum_{(x_{i,e}, y_{i,e}) \in \mathcal{D}_v^i} L\big(\theta^\star(\lambda)\, x_{i,e},\, y_{i,e}\big), \\ &\text{s.t.} \quad \theta^\star(\lambda) = \operatorname*{argmin}_{\theta \in \mathbb{R}^{c \times p}} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_\tau^i|} \sum_{(x_{i,e}, y_{i,e}) \in \mathcal{D}_\tau^i} L\big(\theta\, x_{i,e},\, y_{i,e}\big) + \frac{1}{cp} \sum_{r=1}^{c} \sum_{s=1}^{p} e^{\lambda_s}\, \theta_{rs}^2, \end{aligned} \tag{19}$$

where $L(\cdot, \cdot)$ is the cross-entropy loss, $c = 10$ and $p = 784$ denote the number of classes and features, λ_s is the s-th entry of $\lambda \in \mathbb{R}^p$, θ_{rs} is the (r, s)-th element of $\theta \in \mathbb{R}^{c \times p}$, and $\mathcal{D}_\tau^i, \mathcal{D}_v^i$ are the training and validation datasets of honest agent $i \in \mathcal{H}$. On both datasets, we use $n = 10$ agents with Byzantine fraction $\delta = 0.3$ (i.e., $f = 3$ Byzantine and $n - f = 7$ honest workers); the MNIST data are partitioned across agents via a Dirichlet distribution with concentration $\alpha = 0.5$ to induce heterogeneity. Complete hyperparameter settings for this task are summarized in Table 4.

Table 4: Implementation details for the hyperparameter optimization task.

Details	**Synthetic**	**MNIST**
Model architecture	$\theta \in \mathbb{R}^{200}, \lambda \in \mathbb{R}^{200}$	$\theta \in \mathbb{R}^{10 \times 784}, \lambda \in \mathbb{R}^{784}$
Learning rate (η)	$\{\eta_y, \eta_z, \eta_x\} = \{0.1, 0.02, 0.01\}$	$\{\eta_y, \eta_z, \eta_x\} = \{0.05, 0.005, 0.005\}$
Batch size (b)	$b = 64$	$b = 64$
Momentum (β)	$\beta \in \{0.9, 0.9, 0.9\}$	$\beta \in \{0.9, 0.9, 0.9\}$
Communication rounds	1000	2000
Number of Byzantine workers	$f = 3$	$f = 3$
Number of honest workers	$n - f = 7$	$n - f = 7$

A.1.3 Decentralized hyper-representation

Following the formulation presented in Franceschi et al. (2018), the hyper-representation problem can be described as:

$$\begin{aligned} &\min_{\lambda} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_v^i|} \sum_{\phi \in \mathcal{D}_v^i} f_i(\lambda, \theta^\star(\lambda); \phi) \\ &\text{s.t.} \quad \theta^\star(\lambda) = \arg\min_{\theta} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}_\tau^i|} \sum_{\varphi \in \mathcal{D}_\tau^i} g_i(\lambda, \theta; \varphi), \end{aligned}$$

where $\mathcal{H}$ denotes the index set of honest agents with $n_h = |\mathcal{H}|$, and $\mathcal{D}_\tau^i$ and $\mathcal{D}_v^i$ denote the training and validation datasets of agent i, respectively. The last fully connected layer is treated as the lower-level variable θ, and the remaining parameters form the upper-level variable λ. We conduct experiments on MNIST using a 2-layer multilayer perceptron and on CIFAR-10 using a CNN, where the last fully connected layer serves as the lower-level variable θ and the remaining parameters as the upper-level variable λ.

In our experiments, we test two datasets: MNIST (LeCun, 1998) and CIFAR-10 (Krizhevsky et al., 2009). For MNIST, we use a 2-layer multilayer perceptron (MLP) with 200 hidden units. For CIFAR-10, we employ a convolutional neural network (CNN) with three convolutional blocks (with 64, 128, and 256 channels), followed by two fully connected layers ($4096 \to 256 \to 10$). In both models, we designate the last fully connected layer as the lower-level variable θ, while the remaining parameters are treated as the upper-level variable λ. On both datasets, we use $n = 10$ agents with Byzantine fraction $\delta = 0.3$ (i.e., $f = 3$ Byzantine and $n - f = 7$ honest workers). Details of the model architectures and training configurations are summarized in Table 5.

Table 5: Implementation details for the hyper-representation task.

Details	**MNIST**	**CIFAR-10**
Model type	2-layer MLP	CNN
Model architecture	Input: 784, Hidden: 200, Output: 10	3 Conv blocks $(64, 128, 256)$ + FC: $4096 \to 256 \to 10$
Learning rate (η)	$\{\eta_y, \eta_z, \eta_x\} = \{0.01, 0.01, 0.01\}$	$\{\eta_y, \eta_z, \eta_x\} = \{0.1, 0.05, 0.03\}$
Batch size (b)	$b = 64$	$b = 64$
Momentum (β)	$\beta \in \{0.5, 0.5, 0.5\}$	$\beta \in \{0.5, 0.5, 0.5\}$
Communication rounds	500	1000
Number of Byzantine workers	$f = 3$	$f = 3$
Number of honest workers	$n - f = 7$	$n - f = 7$

A.1.4 Byzantine Attack Strategies

To comprehensively evaluate the Byzantine robustness of the proposed methods, we consider four representative gradient-based attacks that cover different adversarial strategies. In our experiments, the Byzantine workers execute A Little is Enough (ALIE) (Baruch et al., 2019), Fall of Empires (also known as Inner Product Manipulation, IPM) (Xie et al., 2020), Sign-Flipping (SF), and Random Noise (RN) attacks.

We formalize these attacks as follows. Let a^k denote the attack direction at iteration k, and let $\eta \geq 0$ be a fixed scalar. At each step k, every Byzantine worker sends to its neighbors a manipulated vector $B^k = \bar{s}^k + \eta a^k$, where $\bar{s}^k$ is an estimate of the true gradient (or momentum). In our gradient descent (GD) experiments, we compute this estimate as $\bar{s}^k = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} g_i^k$, where g_i^k is the gradient computed by honest worker w_i at iteration k. The specific attack strategies are defined as follows:

- **ALIE (A Little is Enough):** The adversarial direction is chosen as $a^k = \sigma^k$, where σ^k is the coordinate-wise standard deviation of $\bar{s}^k$. This attack exploits the statistical properties of honest gradients to craft subtle perturbations that can evade detection.
- **IPM (Inner Product Manipulation):** The adversary sets $a^k = -\bar{s}^k$. In this case, each Byzantine worker sends $(1 - \eta)\bar{s}^k$ to its neighbors, effectively reducing the magnitude of the aggregated gradient.
- **SF (Sign-Flipping):** This attack also uses $a^k = -\bar{s}^k$, but fixes $\eta = 2$. Thus, Byzantine workers send $B^k = a^k = -\bar{s}^k$ at every step, directly flipping the sign of the honest gradient estimate.
- **RN (Random Noise):** The adversarial direction a^k is drawn from a Gaussian distribution, $a^k \sim \mathcal{N}(\mu, \sigma^2)$, so that each Byzantine worker injects random Gaussian noise into its reported vector. Unlike the other three attacks, RN is non-adaptive to the honest gradients and models oblivious or malfunctioning agents.

A.1.5 Robust Aggregation Mechanisms

To demonstrate the versatility and effectiveness of the proposed methods, we evaluate their compatibility and performance when integrated with four established robust aggregation mechanisms: Krum (Blanchard et al., 2017), Median, Trimmed Mean (Yin et al., 2018), and Iterative Outlier Scissor (IOS) (Wu et al., 2023). These aggregators represent different design philosophies for Byzantine-robust aggregation, ranging from selection-based (Krum) to filtering-based (Trimmed Mean, IOS) and optimization-based (Median) approaches.

- **Krum.** Given input vectors $x_1, \ldots, x_n$, Krum selects one of them based on its proximity to its closest neighbors while ignoring the f farthest vectors. For each $i \in [n]$, let $i_1, \ldots, i_{n-1}$ be the indices of the other vectors sorted by distance from x_i, and define the set of its $n - f - 1$ nearest neighbors as $C_i = \{i_1, \ldots, i_{n-f-1}\}$. Krum then chooses the vector whose neighbors are collectively the closest:

$$\mathrm{Krum}(x_1, \ldots, x_n) = x_{i^*}, \quad i^* \in \arg\min_{i \in [n]} \sum_{j \in C_i} \|x_i - x_j\|^2.$$

- **Median.** Given vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, the multivariate median is defined as the point $\mathrm{Med}(x_1, \ldots, x_n) := \arg\min_{x \in \mathbb{R}^d} \sum_{i=1}^n \|x - x_i\|_2$. This estimator minimizes the sum of Euclidean distances to all input vectors. Although robust, it is computationally expensive in high-dimensional settings.

- **Trimmed Mean.** Given $x_1, \ldots, x_n \in \mathbb{R}^d$ and a trimming parameter f, let $x_{(1)}, \ldots, x_{(n)}$ denote the vectors sorted by their distance to the mean, with $\|x_{(1)} - \bar{x}\|_2 \leq \cdots \leq \|x_{(n)} - \bar{x}\|_2$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The trimmed mean removes the f farthest and f closest vectors and averages the remaining:

$$\mathrm{TM}(x_1, \ldots, x_n) = \frac{1}{n-2f} \sum_{j=f+1}^{n-f} x_{(j)}.$$

- **Iterative Outlier Scissor (IOS).** Given vectors $x_1, \ldots, x_n$ with weights $\{w_m\}$ and an estimate q of Byzantine inputs, IOS iteratively prunes outliers. Starting from the full set $\mathcal{U}^{(0)} = [n]$, at each iteration $i = 0, \ldots, q-1$, it computes the weighted average $x_{\mathrm{avg}}^{(i)}$ and removes the farthest vector:

$$m^{(i)} = \arg\max_{m \in \mathcal{U}^{(i)}} \|x_m - x_{\mathrm{avg}}^{(i)}\|_2, \quad \mathcal{U}^{(i+1)} = \mathcal{U}^{(i)} \setminus \{m^{(i)}\}.$$

  After q removals, it returns the weighted average of the remaining trusted set $\mathcal{U}^{(q)}$.

A.2 Additional Experiment Results

This subsection complements the summarized results in Section 5 by providing the complete set of training curves for each (aggregator, attack) combination. For both tasks, we evaluate every method under all four robust aggregators—Krum, Median, Trimmed Mean (TrMean), and Iterative Outlier Scissor (IOS)—and all four Byzantine attacks—Sign-Flipping (SF), A Little is Enough (ALIE), Inner Product Manipulation (IPM), and Random Noise (RN).

A.2.1 Hyperparameter Optimization on Synthetic Data

Figures 1–4 report the test accuracy curves on the synthetic hyperparameter optimization task under each of the four aggregators. Across all four aggregators and all four attacks (SF, ALIE, IPM, and RN), `SLBRD` and `SLBRD-M` attain the highest final accuracy and the most stable learning dynamics, whereas the non-momentum baselines `MA-DSBO` and `GBDSBO` exhibit pronounced oscillations and a clear accuracy gap under the adaptive attacks ALIE, IPM, and SF; under the non-adaptive RN attack the gap narrows, but the proposed methods still match or exceed the strongest momentum baselines under every aggregator.

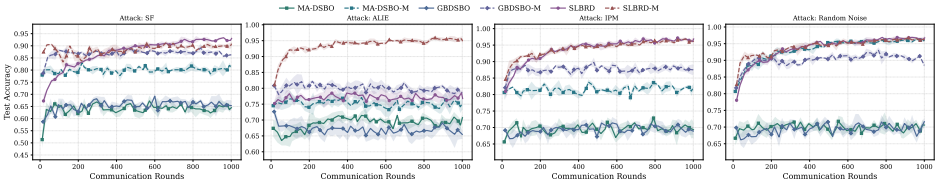

Figure 1: Test accuracy of decentralized bilevel optimization algorithms on the synthetic hyperparameter optimization task using Krum aggregation under SF, ALIE, IPM, and RN attacks.

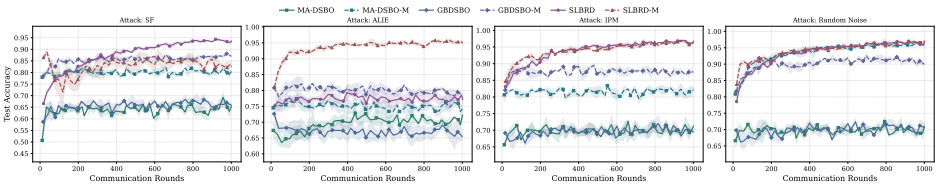

Figure 2: Test accuracy of decentralized bilevel optimization algorithms on the synthetic hyperparameter optimization task using IOS aggregation under SF, ALIE, IPM, and RN attacks.

Table 6: Test accuracy (%) on the MNIST hyperparameter optimization task with heterogeneous data. Column blocks correspond to aggregators Krum, Median, TrMean, and IOS. "Worst" reports the worst accuracy across all attacks. **Bold** marks the best and underline the second best in each column.

Algorithm	ALIE	SF	IPM	RN	Worst	ALIE	SF	IPM	RN	Worst
MA-DSBO	$82.30_{\pm 1.34}$	$60.40_{\pm 14.07}$	$62.50_{\pm 13.91}$	$83.37_{\pm 1.84}$	60.40	$81.30_{\pm 1.98}$	$76.77_{\pm 0.49}$	$76.63_{\pm 0.45}$	$78.07_{\pm 2.30}$	76.63
MA-DSBO-M	$82.17_{\pm 1.34}$	$71.93_{\pm 15.09}$	$83.20_{\pm 1.78}$	$83.33_{\pm 1.81}$	71.93	$81.17_{\pm 2.05}$	$76.90_{\pm 0.28}$	$76.80_{\pm 0.36}$	$77.97_{\pm 2.53}$	76.80
GBDSBO	$85.47_{\pm 0.59}$	$56.10_{\pm 10.83}$	$56.43_{\pm 7.32}$	$86.77_{\pm 0.68}$	56.10	$84.93_{\pm 0.68}$	$80.87_{\pm 0.82}$	$80.63_{\pm 0.78}$	$83.27_{\pm 2.27}$	80.63
GBDSBO-M	$85.97_{\pm 0.61}$	$64.97_{\pm 17.15}$	$63.57_{\pm 21.24}$	$87.23_{\pm 0.63}$	63.57	$84.70_{\pm 1.13}$	$82.37_{\pm 0.66}$	$82.33_{\pm 0.61}$	$84.20_{\pm 2.18}$	82.33
SLBRD	$87.03_{\pm 2.61}$	$90.10_{\pm 0.24}$	$87.77_{\pm 0.78}$	$87.93_{\pm 0.79}$	87.03	$\mathbf{87.67_{\pm 1.64}}$	$\mathbf{90.17_{\pm 0.41}}$	$87.50_{\pm 0.64}$	$\mathbf{88.70_{\pm 0.57}}$	**87.50**
SLBRD-M	$\mathbf{87.57_{\pm 2.45}}$	$\mathbf{91.07_{\pm 0.38}}$	$\mathbf{90.13_{\pm 0.29}}$	$\mathbf{90.37_{\pm 0.33}}$	**87.57**	$85.00_{\pm 1.84}$	$90.00_{\pm 0.08}$	$\mathbf{87.77_{\pm 0.54}}$	$87.73_{\pm 1.96}$	85.00
Algorithm	**ALIE**	**SF**	**IPM**	**RN**	**Worst**	**ALIE**	**SF**	**IPM**	**RN**	**Worst**
MA-DSBO	$83.00_{\pm 1.56}$	$76.80_{\pm 0.85}$	$76.60_{\pm 0.88}$	$78.53_{\pm 1.84}$	76.60	$82.27_{\pm 1.27}$	$49.93_{\pm 2.17}$	$60.67_{\pm 16.29}$	$83.43_{\pm 1.87}$	49.93
MA-DSBO-M	$82.93_{\pm 1.73}$	$76.90_{\pm 1.07}$	$76.87_{\pm 1.05}$	$78.87_{\pm 1.77}$	76.87	$82.23_{\pm 1.30}$	$71.93_{\pm 15.16}$	$83.20_{\pm 1.78}$	$83.30_{\pm 1.78}$	71.93
GBDSBO	$85.27_{\pm 1.05}$	$81.57_{\pm 1.17}$	$81.30_{\pm 0.75}$	$82.53_{\pm 2.93}$	81.30	$85.43_{\pm 0.34}$	$57.33_{\pm 7.83}$	$55.73_{\pm 6.47}$	$86.73_{\pm 0.58}$	55.73
GBDSBO-M	$\mathbf{85.67_{\pm 0.97}}$	$83.13_{\pm 0.95}$	$83.07_{\pm 1.10}$	$84.00_{\pm 1.93}$	83.07	$85.90_{\pm 0.62}$	$63.47_{\pm 16.49}$	$69.40_{\pm 13.77}$	$87.23_{\pm 0.63}$	63.47
SLBRD	$84.93_{\pm 3.69}$	$90.53_{\pm 0.48}$	$88.50_{\pm 0.86}$	$\mathbf{89.27_{\pm 0.68}}$	**84.93**	$86.83_{\pm 2.53}$	$90.30_{\pm 0.16}$	$87.73_{\pm 0.77}$	$87.90_{\pm 0.73}$	86.83
SLBRD-M	$83.43_{\pm 4.75}$	$\mathbf{91.23_{\pm 0.49}}$	$\mathbf{88.53_{\pm 0.95}}$	$88.93_{\pm 0.78}$	83.43	$\mathbf{87.60_{\pm 2.49}}$	$\mathbf{91.07_{\pm 0.38}}$	$\mathbf{90.13_{\pm 0.29}}$	$\mathbf{90.37_{\pm 0.33}}$	**87.60**

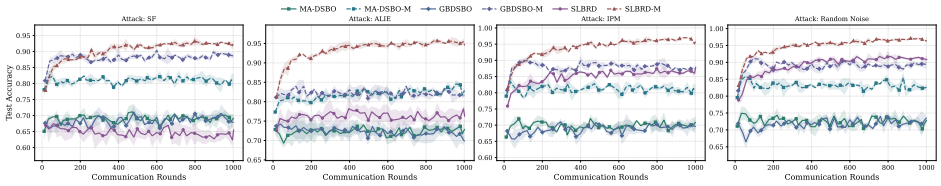

Figure 3: Test accuracy of decentralized bilevel optimization algorithms on the synthetic hyperparameter optimization task using Median aggregation under SF, ALIE, IPM, and RN attacks.

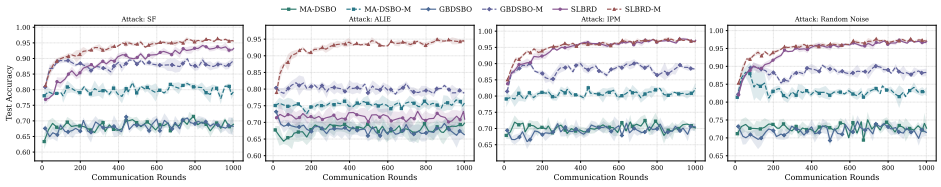

Figure 4: Test accuracy of decentralized bilevel optimization algorithms on the synthetic hyperparameter optimization task using Trimmed Mean aggregation under SF, ALIE, IPM, and RN attacks.

A.2.2 Hyperparameter Optimization on MNIST Dataset

Figures 5–8 and Table 6 show the test accuracy curves on the MNIST hyperparameter optimization task. SLBRD and SLBRD-M reach the highest accuracy in nearly every (aggregator, attack) pair, whereas MA-DSBO and GBDSBO plateau at substantially lower levels, particularly under the SF and IPM attacks, and their momentum variants only partially close the gap. Compared with the synthetic case, the absolute accuracy gap narrows on the smoother real-data objective, but the qualitative ranking between the proposed algorithms and the baselines is preserved.

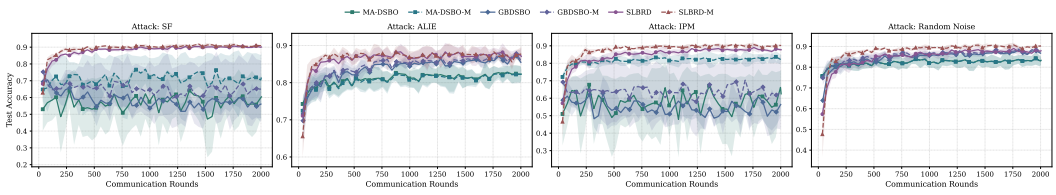

Figure 5: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyperparameter optimization task using Krum aggregation under SF, ALIE, IPM, and RN attacks.

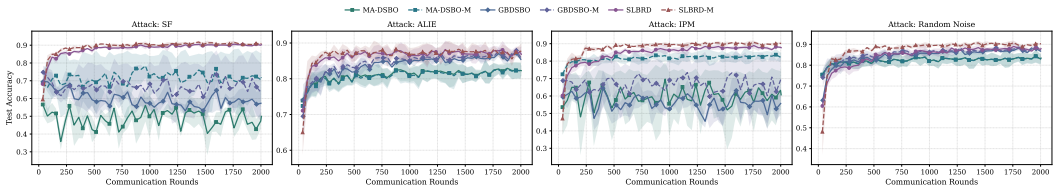

Figure 6: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyperparameter optimization task using IOS aggregation under SF, ALIE, IPM, and RN attacks.

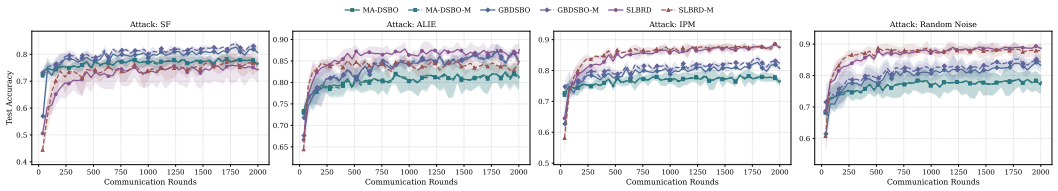

Figure 7: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyperparameter optimization task using Median aggregation under SF, ALIE, IPM, and RN attacks.

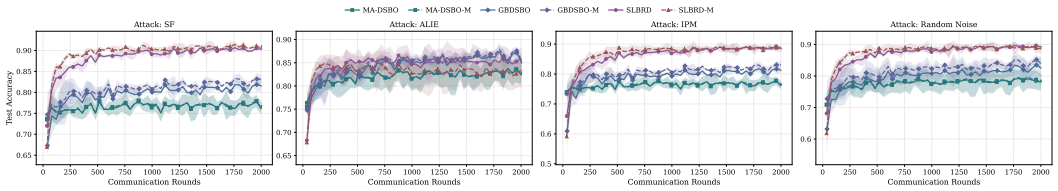

Figure 8: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyperparameter optimization task using Trimmed Mean aggregation under SF, ALIE, IPM, and RN attacks.

A.2.3 Decentralized Hyper-representation on MNIST Dataset

We train a 2-layer MLP (200 hidden units) on MNIST, treating the last fully connected layer as the lower-level variable θ and the remaining parameters as the upper-level variable λ. The final test accuracy across all 16 (aggregator, attack) combinations is reported in Table 3 of the main text, and Figures 9-12 report the corresponding test accuracy curves. `SLBRD-M` attains the highest accuracy in nearly every column, and `SLBRD` takes the second-best position in about half of the remaining ones. The non-momentum baseline `GBDSBO` incurs the largest accuracy degradation, particularly under the SF and IPM attacks, while `MA-DSBO` trails the proposed methods consistently though with less catastrophic drops; their momentum variants only partially mitigate these gaps. The curves further indicate that `SLBRD-M` enjoys both faster convergence and smaller oscillations than the baselines throughout training.

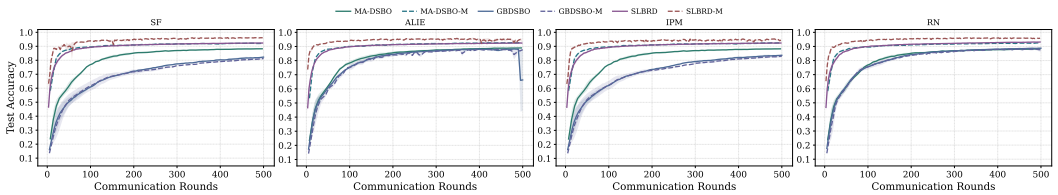

Figure 9: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyper-representation task using Krum aggregation under SF, ALIE, IPM, and RN attacks.

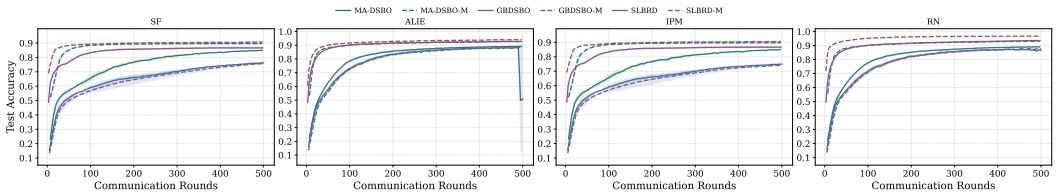

Figure 10: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyper-representation task using Trimmed Mean aggregation under SF, ALIE, IPM, and RN attacks.

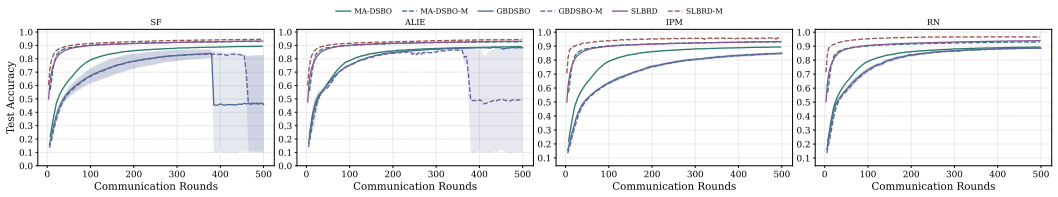

Figure 11: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyper-representation task using IOS aggregation under SF, ALIE, IPM, and RN attacks.

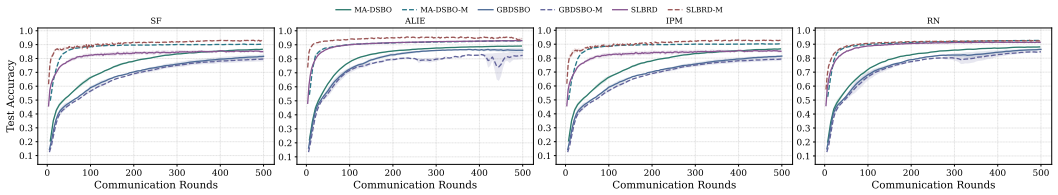

Figure 12: Test accuracy of decentralized bilevel optimization algorithms on the MNIST hyper-representation task using Median aggregation under SF, ALIE, IPM, and RN attacks.

A.2.4 Decentralized Hyper-representation on CIFAR-10 Dataset

On the more challenging CIFAR-10 task, we train a CNN with three convolutional blocks (64, 128, 256 channels) followed by two fully connected layers ($4096 \to 256 \to 10$); the last fully connected layer serves as the lower-level variable θ and the remaining parameters form the upper-level variable λ. Figures 13–16 present the test accuracy curves under the four aggregators and four attacks. The performance gap between the proposed methods and the baselines widens significantly relative to MNIST: the `GBDSBO` family occasionally collapses to near-random accuracy under certain (aggregator, attack) configurations, and the `MA-DSBO` family suffers substantial accuracy drops across all four attacks (SF, ALIE, IPM, and RN), whereas `SLBRD` and `SLBRD-M` together sustain stable training dynamics and attain the highest final accuracy across all 16 (aggregator, attack) combinations, with each variant taking the lead on complementary subsets of settings. This confirms that the proposed bias-corrected hypergradient estimator, combined with Polyak's momentum, is essential for Byzantine-resilient bilevel optimization on harder real-data objectives.

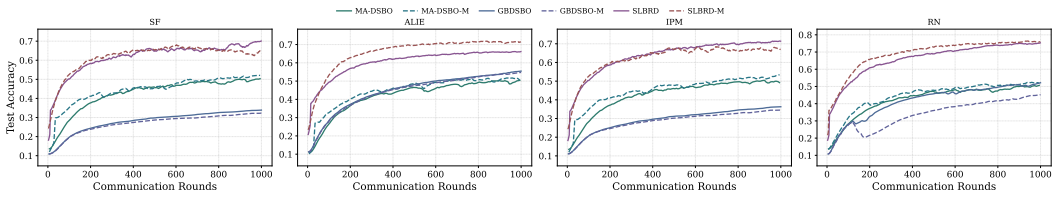

Figure 13: Test accuracy of decentralized bilevel optimization algorithms on the CIFAR10 hyper-representation task using Krum aggregation under SF, ALIE, IPM, and RN attacks.

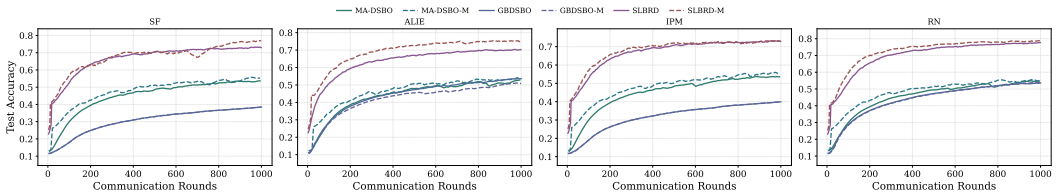

Figure 14: Test accuracy of decentralized bilevel optimization algorithms on the CIFAR10 hyper-representation task using Trimmed Mean aggregation under SF, ALIE, IPM, and RN attacks.

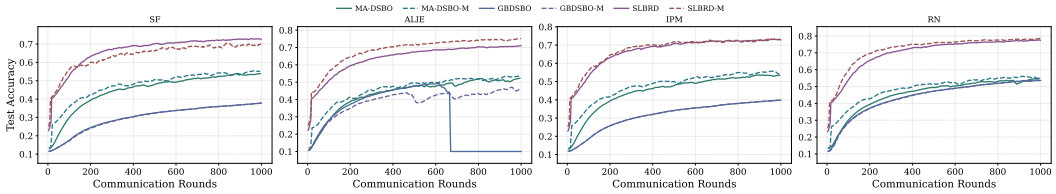

Figure 15: Test accuracy of decentralized bilevel optimization algorithms on the CIFAR10 hyper-representation task using IOS aggregation under SF, ALIE, IPM, and RN attacks.

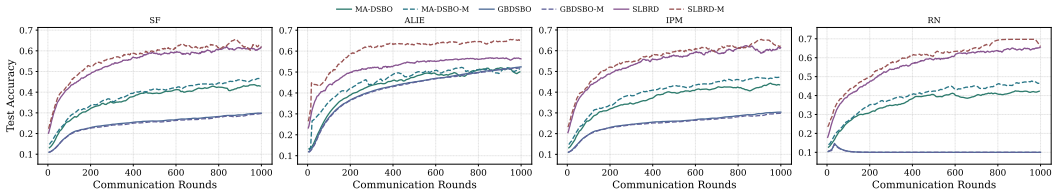

Figure 16: Test accuracy of decentralized bilevel optimization algorithms on the CIFAR10 hyper-representation task using Median aggregation under SF, ALIE, IPM, and RN attacks.

A.3 Ablation Studies

In this subsection, we conduct ablation studies on the MNIST hyper-representation task to analyze the impact of key factors on the performance of `SLBRD` and `SLBRD-M`. Specifically, we investigate: (i) the effect of the Byzantine attacker fraction on convergence and robustness, (ii) the influence of the number of participating agents on system performance, and (iii) the impact of data heterogeneity, controlled by the Dirichlet concentration parameter $\alpha \in \{0.1, 0.5, 1\}$, on the algorithms' effectiveness. Unless stated otherwise, all other settings follow the default MNIST hyper-representation configuration described above.

A.3.1 Effect of Byzantine Attacker Fraction

We investigate the sensitivity of the proposed methods to the proportion of Byzantine agents in the network. Specifically, we vary the Byzantine fraction $\delta \in \{0.1, 0.2, 0.3, 0.4\}$ on the MNIST hyper-representation task while keeping all other settings fixed. For each fraction, we evaluate all four robust aggregators (Median, Krum, Trimmed Mean, and IOS) under four attack strategies: ALIE, SF, IPM, and Random Noise. The results are presented in Figures 17–20.

As expected, increasing the Byzantine fraction generally degrades performance across all aggregators and attacks. However, `SLBRD` and `SLBRD-M` maintain reasonable accuracy even at $\delta = 0.4$, where up to 40% of the agents are adversarial. Among the aggregators, Krum shows the highest sensitivity at larger fractions, while Trimmed Mean and IOS exhibit the most graceful degradation as the attacker fraction increases, with Median lying in between. Across all attack types, the ALIE attack proves the most challenging at higher Byzantine fractions due to its ability to craft perturbations that closely mimic honest gradients, whereas SF and IPM are more effectively mitigated by the robust aggregators.

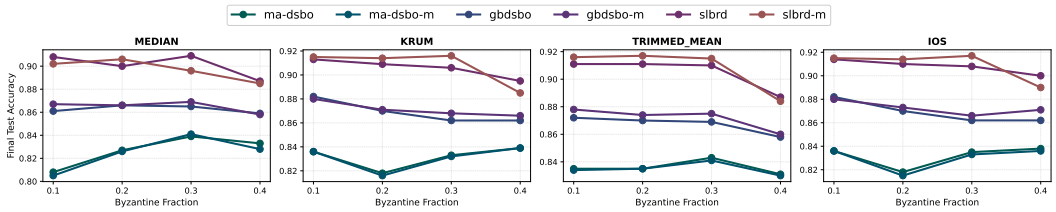

Figure 17: Effect of Byzantine attacker fraction under the ALIE attack on MNIST. Each subplot corresponds to a different robust aggregator. The Byzantine fraction varies from 0.1 to 0.4.

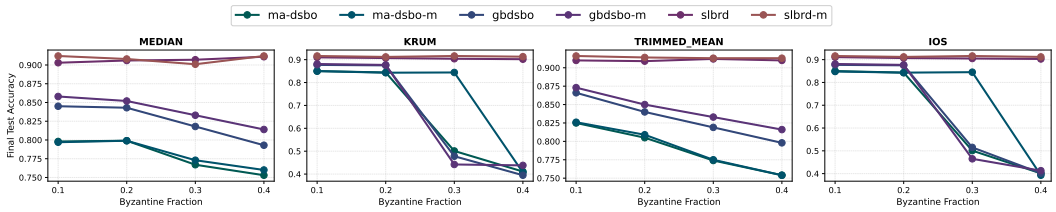

Figure 18: Effect of Byzantine attacker fraction under the SF attack on MNIST. Each subplot corresponds to a different robust aggregator. The Byzantine fraction varies from 0.1 to 0.4.

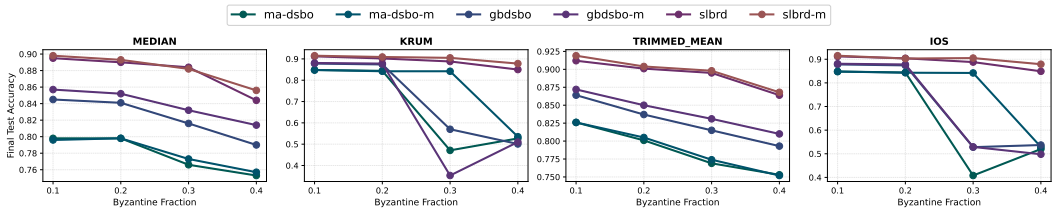

Figure 19: Effect of Byzantine attacker fraction under the IPM attack on MNIST. Each subplot corresponds to a different robust aggregator. The Byzantine fraction varies from 0.1 to 0.4.

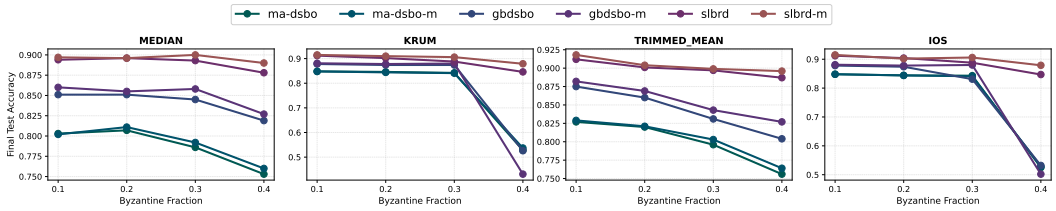

Figure 20: Effect of Byzantine attacker fraction under the Random Noise attack on MNIST. Each subplot corresponds to a different robust aggregator. The Byzantine fraction varies from 0.1 to 0.4.

A.3.2 Effect of the Number of Participating Agents

We investigate how the number of participating agents affects the convergence behavior and robustness of SLBRD on MNIST. Specifically, we vary the total number of agents $n \in \{10, 20, 30\}$ while maintaining a fixed Byzantine fraction of $\delta = 0.3$ (i.e., 30% of agents are adversarial) and keeping the network topology structure consistent across all configurations. For each network size, we evaluate all four robust aggregators (Median, Krum, Trimmed Mean, and IOS) under four attack strategies: IPM, Random Noise, ALIE, and SF. The results are presented in Figures 21–24.

The results demonstrate that both SLBRD and SLBRD-M maintain robust performance across all network sizes, with larger networks generally improving convergence speed and final accuracy due to

better statistical estimates and more effective Byzantine filtering. Notably, SLBRD and SLBRD−M exhibit stable convergence even with only 10 agents, demonstrating effectiveness in resource-constrained settings. Among the aggregators, Trimmed Mean and IOS show the most consistent improvement with increasing network size, while performance gains are most pronounced under the ALIE attack where additional honest gradients help distinguish subtle adversarial perturbations.

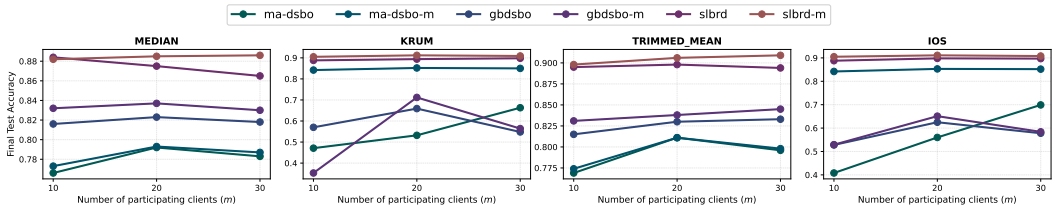

Figure 21: Effect of the number of participating agents under the IPM attack on MNIST. Each subplot corresponds to a different robust aggregator. The number of agents varies from 10 to 30.

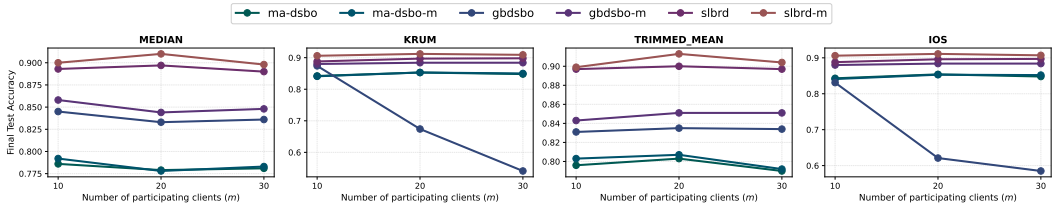

Figure 22: Effect of the number of participating agents under the Random Noise attack on MNIST. Each subplot corresponds to a different robust aggregator. The number of agents varies from 10 to 30.

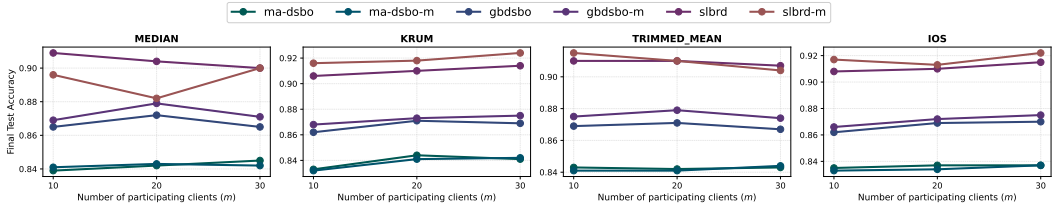

Figure 23: Effect of the number of participating agents under the ALIE attack on MNIST. Each subplot corresponds to a different robust aggregator. The number of agents varies from 10 to 30.

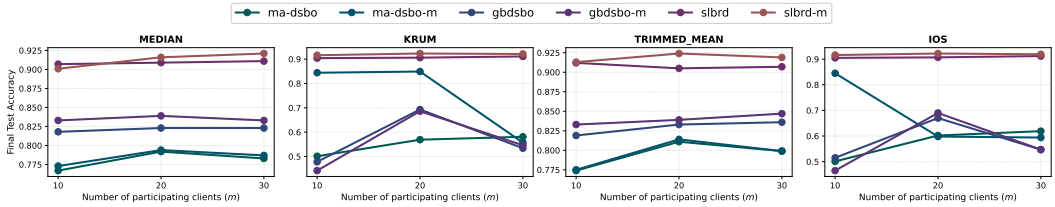

Figure 24: Effect of the number of participating agents under the SF attack on MNIST. Each subplot corresponds to a different robust aggregator. The number of agents varies from 10 to 30.

A.3.3 Effect of Data Heterogeneity

We investigate how data heterogeneity across agents affects the convergence and robustness of SLBRD and SLBRD−M on MNIST. We partition the training data across agents using a Dirichlet distribution

over class proportions with concentration parameter α; smaller α induces stronger heterogeneity (more skewed local class mixes), while larger α approaches more balanced allocations. Specifically, we set $\alpha \in \{0.1, 0.5, 1\}$ while keeping all other settings fixed (10 agents, Byzantine fraction $\delta = 0.3$). For each α, we evaluate all four robust aggregators (Median, Krum, Trimmed Mean, and IOS) under four attack strategies: SF, ALIE, Random Noise, and IPM. The results are presented in Figures 25–28.

The results show that SLBRD and SLBRD−M remain stable across the chosen α values, with stronger heterogeneity (smaller α) generally making aggregation harder and modestly affecting convergence or final accuracy compared with milder heterogeneity at $\alpha = 1$. Among the aggregators, Krum tends to be the most sensitive to severe heterogeneity, especially under the ALIE attack, whereas Trimmed Mean and IOS exhibit more consistent behavior across α, with Median in between. Under the Random Noise attack, differences across α are comparatively muted, consistent with random perturbations being easier to separate from honest updates regardless of moderate distribution shift.

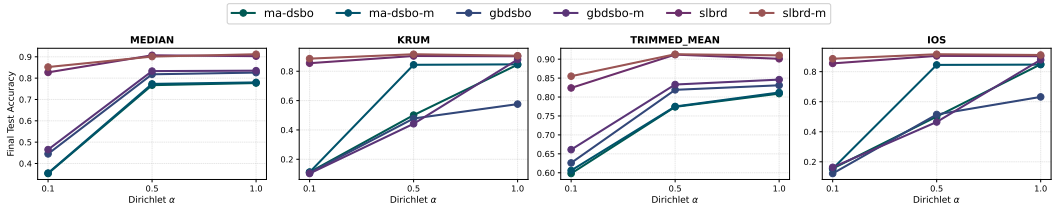

Figure 25: Effect of data heterogeneity (Dirichlet $\alpha \in \{0.1, 0.5, 1\}$) under the SF attack on MNIST. Each subplot corresponds to a different robust aggregator.

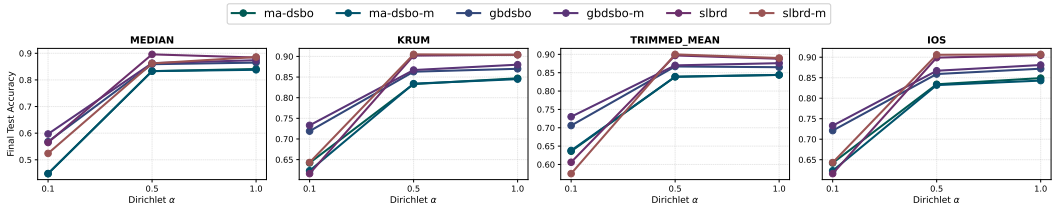

Figure 26: Effect of data heterogeneity (Dirichlet $\alpha \in \{0.1, 0.5, 1\}$) under the ALIE attack on MNIST. Each subplot corresponds to a different robust aggregator.

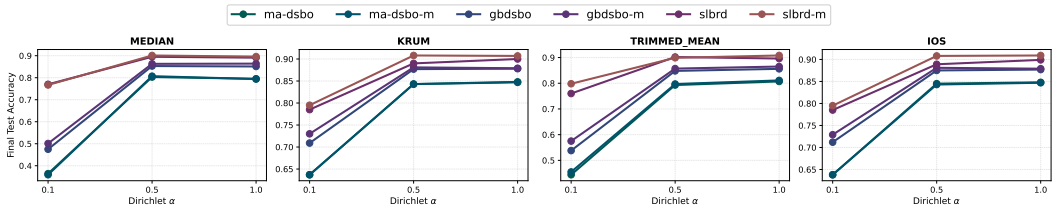

Figure 27: Effect of data heterogeneity (Dirichlet $\alpha \in \{0.1, 0.5, 1\}$) under the Random Noise attack on MNIST. Each subplot corresponds to a different robust aggregator.

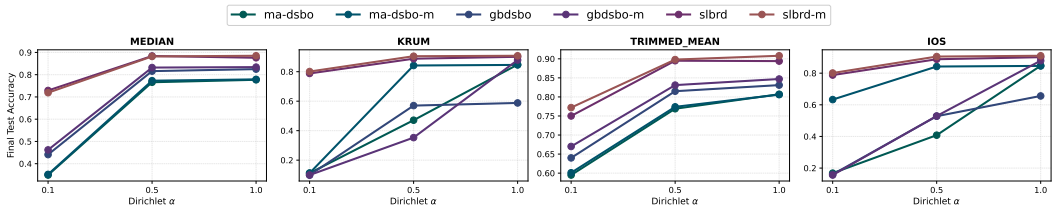

Figure 28: Effect of data heterogeneity (Dirichlet $\alpha \in \{0.1, 0.5, 1\}$) under the IPM attack on MNIST. Each subplot corresponds to a different robust aggregator.

B Additional Related Work

Decentralized bilevel optimization. A core difficulty in decentralized bilevel optimization lies in estimating the hypergradient $\nabla\Phi(x)$ with peer-to-peer communication. Agents must collaboratively track the coupled lower-level solution and approximate the Hessian-inverse term induced by implicit differentiation. Early works addressed these challenges through inner-loop procedures. Chen et al. (2025) developed a Jacobian-Hessian-Inverse-Product (JHIP) estimator for hypergradient tracking. `MA-DSBO` (Chen et al., 2023) employed a Hessian-inverse-gradient-product subroutine. Yang et al. (2022) adopted a truncated Neumann-series estimator, and `SLAM` (Lu et al., 2022b) used an augmented-Lagrangian inner solver. Although these methods enjoyed rigorous convergence guarantees, they typically incurred substantial communication overhead per iteration. Subsequent works therefore focused on *single-loop* methods (Kong et al., 2024; Dong et al., 2025; Gao et al., 2023). More recent studies developed unified frameworks and established sharper transient iteration-complexity bounds (Zhu et al., 2024; Kong et al., 2025). Other works considered personalized lower-level objectives (Niu et al., 2025; Lu et al., 2022a), differential privacy (Chen & Wang, 2024), and communication compression (He et al., 2024). These studies generally assume benign agents, leaving Byzantine-resilient decentralized bilevel optimization underexplored.

Byzantine-resilient decentralized optimization. Byzantine-resilient optimization has progressed rapidly in recent years (Blanchard et al., 2017; Yin et al., 2018; Alistarh et al., 2018; El-Mhamdi et al., 2021; Wu et al., 2023; Farhadkhani et al., 2023). In the server-based setting, robust aggregators include Krum (Blanchard et al., 2017), coordinate-wise and geometric median (Yin et al., 2018; Acharya et al., 2022), centered clipping (Karimireddy et al., 2021), and bucketing (Karimireddy et al., 2020). These methods are often combined with momentum (Karimireddy et al., 2021; Farhadkhani et al., 2022) or variance reduction (Gorbunov et al., 2022) to handle data heterogeneity (Allouah et al., 2023; Liu et al., 2023; Murata et al., 2024). In the decentralized setting, Byzantine resilience is achieved by robustly aggregating messages from local neighbors, for example through coordinate-wise trimmed mean (Su & Vaidya, 2020; Fang et al., 2022) or coordinate-descent screening (Yang & Bajwa, 2019). Representative stochastic-gradient methods include decentralized RSA (Peng et al., 2021), Self-Centered Clipping (SCC) (He et al., 2022), and Iterative Outlier Scissor (IOS) (Wu et al., 2023). On the theoretical side, Karimireddy et al. (2020) established the lower bound $\Omega(\delta\zeta^2)$ on the asymptotic error, where δ is the Byzantine fraction and ζ^2 quantifies data heterogeneity. Farhadkhani et al. (2023) showed that Polyak's momentum attains this lower bound.

These works are confined to *single-level* optimization. The existing studies on Byzantine-robust bilevel problems (Abbas et al., 2024; Li et al.) address the *federated* setting with a central server. Byzantine-resilient *decentralized* bilevel optimization without a server remains unexplored.

C NOTATIONS AND TECHNICAL LEMMAS

Notation convention. Since the convergence analysis is carried out entirely over the honest subnetwork $\mathcal{H}$, it is convenient to collect the per-agent iterates into network-level vectors and to write the mixing step in a matrix form that acts on these vectors. To this end, we adopt the following conventions throughout this section: boldface lowercase letters denote vectors stacked across all n_h honest agents, while boldface uppercase letters denote the corresponding block matrices. The precise definitions are as follows.

- *Honest and Byzantine agents.* The set of honest agents is denoted by $\mathcal{H} \subseteq [n]$, with cardinality $n_h := |\mathcal{H}| = (1-\delta)n$, and the remaining indices $\mathcal{B} := [n] \setminus \mathcal{H}$ correspond to the Byzantine agents.
- *Stacked iterates.* For any local variable $x_i^k \in \mathbb{R}^d$ held by an honest agent $i \in \mathcal{H}$, we write $\mathbf{x}^k := \mathrm{col}\{x_i^k\}_{i\in\mathcal{H}} \in \mathbb{R}^{dn_h}$, and similarly for $\mathbf{y}^k$ and $\mathbf{z}^k$.
- *Honest averages and consensus vectors.* The honest-average iterate is $\bar{x}^k := \frac{1}{n_h}\sum_{i\in\mathcal{H}} x_i^k$, and its stacked consensus counterpart is $\bar{\mathbf{x}}^k := \mathbf{1}_{n_h} \otimes \bar{x}^k \in \mathbb{R}^{dn_h}$, where $\mathbf{1}_{n_h}$ is the all-ones vector in $\mathbb{R}^{n_h}$. The quantities $\bar{y}^k, \bar{z}^k$ and $\bar{\mathbf{y}}^k, \bar{\mathbf{z}}^k$ are defined analogously.
- *Stacked gradients.* The stacked upper-level gradient is $\nabla \mathbf{f}_{\mathcal{H}}(\mathbf{x}^k) := \mathrm{col}\{\nabla f_i(x_i^k)\}_{i\in\mathcal{H}} \in \mathbb{R}^{dn_h}$, and the stacked lower-level and Hessian-vector terms $\mathbf{D}_x^k, \mathbf{D}_y^k, \mathbf{D}_z^k$ are defined in the same manner from the per-agent estimates in Algorithm 1.
- *Mixing matrix.* Let $W \in \mathbb{R}^{n_h \times n_h}$ denote the mixing matrix restricted to honest agents. Its block-diagonal lift to the stacked space is $\mathbf{W} := W \otimes I_d \in \mathbb{R}^{dn_h \times dn_h}$, so that $\mathbf{W}\mathbf{x}^k$ implements the mixing step on $\mathbf{x}^k$.
- *Norms.* Unless otherwise specified, $\|\cdot\|$ denotes the Euclidean (ℓ_2) norm for vectors and the spectral norm for matrices.

Auxiliary shorthand. To keep the recursions compact, we adopt the following shorthand that will recur throughout the proof. We define the composite noise and smoothness constants

$$\sigma_1^2 := (1+\rho^2)\sigma^2, \qquad L_1^2 := \tfrac{1}{2}L_f^2 + \rho^2 L_g^2, \qquad L_2^2 := L_f^2 + \rho^2 L_g^2,$$

and the aggregate consensus error

$$\boldsymbol{\Delta}^k := \frac{L_2^2}{L_g^2}\Big(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\Big) + \mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right],$$

which jointly tracks the dispersion of $\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k$ around their honest means with weights chosen to reflect their relative influence on the hyperobjective. Finally, since the optimal lower-level response and Hessian-inverse vector are always evaluated at the honest-average iterate, we write $\bar{y}_\star^k := y^\star(\bar{x}^k)$ and, analogously, $\bar{z}_\star^k := z^\star(\bar{x}^k)$.

Technical lemmas. With the notation in place, we recall two standard results from bilevel optimization that will be used repeatedly in the convergence analysis. The first lemma states the Lipschitz continuity of the hyperobjective gradient $\nabla\Phi_{\mathcal{H}}$ and of the optimal lower-level maps $y^\star(x), z^\star(x)$, which is needed for the descent argument in Lemma 4. It also gives a uniform bound on $z^\star(x)$ used to control the Hessian-inverse error.

Lemma 1 (Chen et al. (2024), Lemma B.2)**.** *Under Assumptions 2, 3, there exist positive constants $L_{\nabla\Phi}, L_{y^\star}, L_{z^\star}$, such that $\nabla\Phi_{\mathcal{H}}(x), y^\star(x), z^\star(x)$ are $L_{\nabla\Phi}, L_{y^\star}, L_{z^\star}$-Lipschitz continuous, respectively. Moreover, we have $\|z^\star(x)\| \le \frac{C_f}{\mu_g}$ for all $x \in \mathbb{R}^{d_x}$.*

The second lemma gives the one-step contraction of the lower-level gradient step under strong convexity. This inequality is the standard tool for quantifying how $\bar{y}^k$ tracks its optimal response $\bar{y}_\star^k$, and will be used in the proof of the y-tracking recursion (Lemma 5).

Lemma 2 (Chen et al. (2024),Lemma B.1)**.** *Under Assumptions 2, 3, if $\beta < \frac{2}{L_g+\mu_g}$, the following inequality holds:*

$$\left\|\mathbf{y}^k - \beta\nabla_{\mathbf{y}} g^k - \mathbf{y}_\star^k\right\| \le (1-\beta\mu_g)\left\|\mathbf{y}^k - \mathbf{y}_\star^k\right\|.$$

D CONVERGENCE ANALYSIS: PROOF OF THEOREM 1

D.1 PROOF OUTLINE

The proof establishes convergence by tracking four interdependent quantities: the descent of the hyperobjective, the lower-level tracking errors, the consensus errors among honest agents, and the aggregation errors induced by Byzantine-robust aggregation. After recalling the algorithmic updates and introducing a few auxiliary variables, we bound each of these quantities through a series of technical lemmas, and finally combine them into a unified recursion that yields the convergence rate. To keep the outline compact, all proofs are deferred to the corresponding appendices.

We begin by recalling the update rule in Algorithm 1. At each iteration k, every honest agent $i \in \mathcal{H}$ computes local stochastic gradients from its current iterates (x_i^k, y_i^k, z_i^k) and sampled data (ϕ_i^k, φ_i^k):

$$\begin{aligned} D_{x,i}^k &= \nabla_x F_i(x_i^k, y_i^k; \phi_i^k) + \nabla_{xy}^2 G_i(x_i^k, y_i^k; \varphi_i^k) z_i^k, \\ D_{y,i}^k &= \nabla_y G_i(x_i^k, y_i^k; \varphi_i^k), \\ D_{z,i}^k &= \nabla_y F_i(x_i^k, y_i^k; \phi_i^k) + \nabla_{yy}^2 G_i(x_i^k, y_i^k; \varphi_i^k) z_i^k. \end{aligned} \tag{20}$$

Each agent then performs a local gradient step to obtain intermediate iterates $x_i^{k+\frac{1}{2}}, y_i^{k+\frac{1}{2}}, z_i^{k+\frac{1}{2}}$, and then aggregates its neighbors' intermediate iterates through a robust aggregation rule $\mathcal{A}(\cdot)$ that is designed to be robust to Byzantine agents:

$$x_i^{k+\frac{1}{2}} = x_i^k - \eta_x D_{x,i}^k,\ x_i^{k+1} = R_{x,i}^{k+\frac{1}{2}} = \mathcal{A}\big(x_i^{k+\frac{1}{2}}, \{x_m^{k+\frac{1}{2}}\}_{m\in\mathcal{N}_i}\big), \tag{21}$$

$$y_i^{k+\frac{1}{2}} = y_i^k - \eta_y D_{y,i}^k,\ y_i^{k+1} = R_{y,i}^{k+\frac{1}{2}} = \mathcal{A}\big(y_i^{k+\frac{1}{2}}, \{y_m^{k+\frac{1}{2}}\}_{m\in\mathcal{N}_i}\big), \tag{22}$$

$$z_i^{k+\frac{1}{2}} = z_i^k - \eta_z D_{z,i}^k,\ z_i^{k+1} = R_{z,i}^{k+\frac{1}{2}} = \mathcal{A}\big(z_i^{k+\frac{1}{2}}, \{z_m^{k+\frac{1}{2}}\}_{m\in\mathcal{N}_i}\big), \tag{23}$$

where $R_{x,i}^{k+\frac{1}{2}}, R_{y,i}^{k+\frac{1}{2}}, R_{z,i}^{k+\frac{1}{2}}$ are the aggregator outputs and $\eta_x, \eta_y, \eta_z > 0$ are the step sizes.

To measure how much the robust aggregator deviates from ideal averaging, we introduce the honest-only reference aggregates, i.e., the outputs an oracle would obtain with exact averaging over the honest neighbors alone. For each $i \in \mathcal{H}$,

$$\bar{x}_i^{k+\frac{1}{2}} := \sum_{j\in\mathcal{H}_i} w_{ij} x_j^{k+\frac{1}{2}},\ \bar{y}_i^{k+\frac{1}{2}} := \sum_{j\in\mathcal{H}_i} w_{ij} y_j^{k+\frac{1}{2}},\ \bar{z}_i^{k+\frac{1}{2}} := \sum_{j\in\mathcal{H}_i} w_{ij} z_j^{k+\frac{1}{2}}, \tag{24}$$

where $\mathcal{H}_i$ denotes the set of honest neighbors of agent i (including i itself) and the weights satisfy $\sum_{j\in\mathcal{H}_i} w_{ij} = 1$.

Aggregation error. We define the *aggregation error* as the discrepancy between the robust aggregator output and the honest-only reference, which measures the error introduced by robust aggregation. Normalizing by the step size, we let

$$\xi_{x,i}^k = \frac{1}{\eta_x}\left(R_{x,i}^{k+\frac{1}{2}} - \bar{x}_i^{k+\frac{1}{2}}\right),\ \boldsymbol{\xi}_x^k = col\{\xi_{x,i}^k\}_{i\in\mathcal{H}}, \tag{25}$$

$$\xi_{y,i}^k = \frac{1}{\eta_y}\left(R_{y,i}^{k+\frac{1}{2}} - \bar{y}_i^{k+\frac{1}{2}}\right),\ \boldsymbol{\xi}_y^k = col\{\xi_{y,i}^k\}_{i\in\mathcal{H}}, \tag{26}$$

$$\xi_{z,i}^k = \frac{1}{\eta_z}\left(R_{z,i}^{k+\frac{1}{2}} - \bar{z}_i^{k+\frac{1}{2}}\right),\ \boldsymbol{\xi}_z^k = col\{\xi_{z,i}^k\}_{i\in\mathcal{H}}. \tag{27}$$

The following lemma bounds this aggregation error in terms of consensus errors and gradient variance.

Lemma 3. *Suppose Assumptions 1 and 2, 4, 5 hold. Then*

$$\begin{aligned} \mathbb{E}\left[\left\|R_{x,i}^{k+\frac{1}{2}} - \bar{x}_i^{k+\frac{1}{2}}\right\|^2\right] &\le 2\kappa \sum_{j\in\mathcal{H}_i} w_{ij}\left(\mathbb{E}\left[\left\|x_j^k - \bar{x}_i^k\right\|^2\right] + L_g^2\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right]\right) \\ &\quad + \kappa\eta_x^2\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right), \\ \mathbb{E}\left[\left\|R_{y,i}^{k+\frac{1}{2}} - \bar{y}_i^{k+\frac{1}{2}}\right\|^2\right] &\le 2\kappa \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|y_j^k - \bar{y}_i^k\right\|^2\right] + \kappa\eta_y^2\left(2\zeta_g^2 + 4\sigma^2\right), \\ \mathbb{E}\left[\left\|R_{z,i}^{k+\frac{1}{2}} - \bar{z}_i^{k+\frac{1}{2}}\right\|^2\right] &\le 2\kappa(1+L_g^2) \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right] + \kappa\eta_z^2\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right). \end{aligned}$$

Proof: see Appendix D.3.1.

Descent lemma. Treating this aggregation error as a perturbation, we next bound the per-iteration decrease of the hyperobjective $\Phi(\bar{x}^k)$. Averaging the x-update across honest agents and applying the $L_{\nabla\Phi}$-smoothness of Φ from Lemma 1, together with a Young-type splitting of the cross term between $\nabla\Phi(\bar{x}^k)$ and the stochastic gradient $\bar{D}_x^k$, produces a one-step descent inequality in which the tracking errors, the consensus error, and the aggregation error all enter as explicit perturbation terms.

Lemma 4. *Suppose that Assumptions 1 and 2, 3, 4 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \le \mathbb{E}\left[\Phi(\bar{x}^k)\right] - \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \left(\frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2}\right)\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k)\right\|^2\right] \\ + \frac{3L_g^2\eta_x}{n_h}\mathbf{\Delta}^k + 3L_2^2\eta_x\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 3L_g^2\eta_x\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] \\ + \left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h}.\end{aligned}$$

Proof: see Appendix D.3.2.

Lower-level convergence. To close the recursion in Lemma 4, we must control the tracking errors $\|\bar{y}^k - \bar{y}_\star^k\|^2$ and $\|\bar{z}^k - \bar{z}_\star^k\|^2$ on its right-hand side. The μ_g-strong convexity of g_i drives both errors to contract at rates $(1-\eta_y\mu_g)$ and $(1-\eta_z\mu_g)$, while coupling to the upper-level update, the consensus error, and the aggregation error act as perturbations. We first state the bound for the y-tracking error.

Lemma 5. *Suppose that Assumptions 2, 3, and 4 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^{k+1}\right\|^2\right] \le (1-\eta_y\mu_g)\,\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right] \\ + 2(1+\frac{3}{\eta_y\mu_g})\frac{\eta_y^2L_g^4}{n_hL_2^2}\mathbf{\Delta}^k + 2\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + 2\eta_y^2\frac{\sigma^2}{n_h}.\end{aligned}$$

Proof: see Appendix D.3.3.

The z-tracking error admits an analogous recursion. Since the definition of $z_\star^k(x)$ involves the Hessian inverse $\nabla_{yy}^2 g_i(\cdot, y_i^k)^{-1}$, the z-update inherits an additional dependence on the y-tracking error through $\nabla_{yy}^2 G_i$, as reflected in the next bound.

Lemma 6. *Suppose that Assumptions 2, 3, and 4 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1} - \bar{z}_\star^{k+1}\right\|^2\right] \le (1-\eta_z\mu_g)\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] + 12\eta_z^2L_2^2\left(1 + \frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\ + \frac{12L_g^2\eta_z^2}{n_h}\left(1 + \frac{3}{\eta_z\mu_g}\right)\mathbf{\Delta}^k + \left(2\eta_z^2 + \frac{6\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] \\ + 2\eta_z^2\frac{\sigma_1^2}{n_h} + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right].\end{aligned}$$

Proof: see Appendix D.3.4.

Consensus error. Both Lemma 4 and Lemmas 5, 6 depend on the aggregate consensus error $\mathbf{\Delta}^k$, so it remains to describe its evolution. We derive the recursion for each variable separately and then combine them into a single bound on $\mathbf{\Delta}^k$.

Writing the stacked x-update in compact form gives

$$\mathbf{x}^{k+1} = \mathbf{W}\left(\mathbf{x}^k - \eta_x\mathbf{D}_x^k\right) + \boldsymbol{\xi}_x^k. \tag{28}$$

Left-multiplying both sides of (28) by $\frac{1}{N}\mathbf{1}\mathbf{1}^\top$ and using the doubly stochastic property of W (Assumption 1) yields

$$\bar{\mathbf{x}}^{k+1} = \frac{1}{N}\mathbf{1}\mathbf{1}^\top\left(\mathbf{x}^k - \eta_x\mathbf{D}_x^k\right) + \bar{\boldsymbol{\xi}}_x^k. \tag{29}$$

Subtracting (29) from (28) yields the consensus error recursion for x:

$$\begin{aligned}\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1} &= \left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k - \eta_x \mathbf{D}_x^k\right) + \boldsymbol{\xi}_x^k - \bar{\boldsymbol{\xi}}_x^k \\ &= \left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k - \bar{\mathbf{x}}^k - \eta_x \mathbf{D}_x^k\right) + \boldsymbol{\xi}_x^k - \bar{\boldsymbol{\xi}}_x^k.\end{aligned} \tag{30}$$

The recursions for y and z follow identically:

$$\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1} = \left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^k - \bar{\mathbf{y}}^k - \eta_y \mathbf{D}_y^k\right) + \boldsymbol{\xi}_y^k - \bar{\boldsymbol{\xi}}_y^k, \tag{31}$$

$$\mathbf{z}^{k+1} - \bar{\mathbf{z}}^{k+1} = \left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^k - \bar{\mathbf{z}}^k - \eta_z \mathbf{D}_z^k\right) + \boldsymbol{\xi}_z^k - \bar{\boldsymbol{\xi}}_z^k. \tag{32}$$

Since $\|\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\|_2 = \lambda < 1$, the mixing matrix contracts the consensus error at each step while the gradient and aggregation-error terms act as perturbations. We now state bounds for each variable in turn.

Consensus error for y. We begin with the y-variable, which is the simplest of the three because the local gradient $D_{y,i}^k = \nabla_y G_i(x_i^k, y_i^k; \varphi_i^k)$ does not involve the Hessian-vector product and therefore introduces no additional coupling to other variables. Bounding $\|\mathbf{y}^k - \bar{\mathbf{y}}^k\|^2$ via the recursion above and using the gradient heterogeneity assumption to control the stochastic term yields the following.

Lemma 7. *Suppose that Assumptions 1 and 2, 4, 5 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1}\right\|^2\right] \le{}& \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right] + \frac{3L_g^4\eta_y^2\lambda^{3/2}}{(1-\lambda)L_2^2}\boldsymbol{\Delta}^k + n_h\lambda^{3/2}\eta_y^2\sigma^2 \\ &+ \frac{3n_hL_g^2\eta_y^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \frac{8\kappa n_h\eta_y^2}{1-\lambda}(\zeta_g^2 + 2\sigma^2) + \frac{3n_h\eta_y^2\lambda^{3/2}\zeta_g^2}{(1-\lambda)}.\end{aligned}$$

Proof: see Appendix D.3.5.

Consensus error for x. The x-recursion is more involved because the upper-level gradient $D_{x,i}^k$ contains the term $\nabla_{xy}^2 G_i z_i^k$, which ties the x-dispersion to the z-dispersion through the bounded norm of $z^\star(\cdot)$ given in Lemma 1. Tracking this coupling term, together with the usual gradient heterogeneity and variance contributions, gives the next bound.

Lemma 8. *Suppose that Assumptions 1 and 2, 4, 5 hold. Then, we have*

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\right\|^2\right] \\ &\le \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \frac{8\kappa L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] + \frac{2\eta_x^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right] + \frac{6\eta_x^2\lambda^{3/2}L_g^2}{1-\lambda}\boldsymbol{\Delta}^k \\ &+ \frac{6\eta_x^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2 + \rho^2\zeta_g^2\right) + n_h\lambda^{3/2}\eta_x^2\sigma_1^2 + \frac{8\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2 + 3\zeta_g^2\rho^2 + 2\sigma_1^2).\end{aligned}$$

Proof: see Appendix D.3.6.

Consensus error for z. Finally, the z-recursion inherits couplings to both the y- and z-tracking errors, since $D_{z,i}^k$ contains the Hessian $\nabla_{yy}^2 G_i$ acting on z_i^k and the definition of $z^\star(x)$ itself involves the Hessian inverse at the optimal $y^\star(x)$. Accounting for these two couplings together with the aggregation and stochastic terms yields the following.

Lemma 9. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^{k+1}\right\|^2\right] \le{}& \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] + \frac{4\eta_z^2\lambda^{3/2}L_g^2}{1-\lambda}\boldsymbol{\Delta}^k \\ &+ \frac{4n_h\eta_z^2\lambda^{3/2}L_2^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \frac{4n_h\eta_z^2\lambda^{3/2}L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] \\ &+ n_h\lambda^{3/2}\eta_z^2\sigma_1^2 + \frac{8\kappa n_h\eta_z^2}{1-\lambda}(3\zeta_f^2 + 3\zeta_g^2\rho^2 + 2\sigma_1^2) + \frac{4\eta_z^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2 + \zeta_g^2\rho^2\right).\end{aligned}$$

Proof: see Appendix D.3.7.

Unified consensus recursion. Since the three consensus errors are coupled through the algorithm dynamics, we merge them into the single aggregate $\mathbf{\Delta}^k$ already introduced in the notation section, where the weights L_2^2/L_g^2 and L_g^2 reflect their relative influence on the convergence rate. Combining Lemmas 7, 8, and 9 yields the following unified recursion.

Lemma 10. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}\mathbf{\Delta}^{k+1} \le \frac{1+\lambda}{2}\mathbf{\Delta}^k + C_y\frac{n_h\eta_x^2L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] \\ + \frac{2\eta_x^2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right] + n_h\eta_x^2C_{\sigma 1} + \frac{\eta_x^2n_h}{1-\lambda}C_{\zeta 1} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\zeta 2} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\sigma 2},\end{aligned}$$

where the problem-dependent constants $C_y, C_z, C_{\sigma 1}, C_{\zeta 1}, C_{\zeta 2}, C_{\sigma 2}$ are defined in the detailed proof of Lemma 10.

Proof: see Appendix D.3.8.

D.2 Proof of the Theorem 1

Theorem 1. (Convergence of SLBRD) *Suppose Assumptions 1 and 2–5 hold, and the robust aggregation rule $\mathcal{A}(\cdot)$ satisfies Definition 2 with parameter κ. Set the step sizes as $\eta_x = \Theta\big(1/\sqrt{n_hK}\big)$, $\eta_y = c_y\eta_x$, $\eta_z = c_z\eta_x$ for problem-dependent constants $c_y, c_z > 0$. Then the iterates of Algorithm 1 satisfy*

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla\Phi_{\mathcal{H}}(\bar{x}^k)\|^2\big] \le \mathcal{O}\left(\frac{1}{\sqrt{n_hK}}\right) + \mathcal{O}\left(\frac{n_h\sigma^2}{K(1-\lambda)} + \frac{n_h\left(\zeta_f^2+\zeta_g^2\right)}{K(1-\lambda)^2}\right) \\ + \mathcal{O}\left(\frac{\kappa\, n_h\sigma^2}{K(1-\lambda)^2} + \frac{\kappa\, n_h\left(\zeta_f^2+\zeta_g^2\right)}{K(1-\lambda)^2}\right) + \mathcal{O}\Big(\kappa\big(\sigma^2+\zeta_f^2+\zeta_g^2\big)\Big),\end{aligned} \tag{33}$$

where $\bar{x}^k := \frac{1}{n_h}\sum_{i\in\mathcal{H}}x_i^k$ is the honest iterate average and λ is the second-largest eigenvalue magnitude of W.

Proof. The proof proceeds via a Lyapunov argument. Define the potential function as follows

$$\mathcal{L}^k := \Phi(\bar{x}^k) + A_1\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2 + A_2\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2 + A_\Delta\frac{\mathbf{\Delta}^k}{n_h}, \tag{34}$$

where the constants are defined as

$$\begin{aligned}&A_1 := \frac{5L_2^2}{c_y\mu_g} + \frac{240L_2^2L_g^2}{c_y\mu_g^3},\ A_2 := \frac{5L_g^2}{c_z\mu_g},\ A_\Delta := A_3\eta_x + \frac{\kappa A_4}{\eta_x}, \\ &A_3 := \frac{(8+4A_1C_{y^\star}+4A_2C_{z^\star})L_g^2}{1-\lambda},\ A_4 := \frac{2}{1-\lambda}\left((1+L_g^2) + \frac{16A_1}{c_y\mu_g} + \frac{16A_2(1+L_g^2)}{c_z\mu_g}\right).\end{aligned} \tag{35}$$

The one-step change in $\mathcal{L}^k$ combines the descent of the hyperobjective (Lemma 4), the evolution of the lower-level and auxiliary variables (Lemmas 5 and 6), and the consensus dynamics (Lemma 10).

Substituting these bounds yields

$$
\begin{aligned}
&\mathcal{L}^{k+1} - \mathcal{L}^k \\
&\le -\frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \\
&\quad + \underbrace{\left(3L_2^2\eta_x - A_1\eta_y\mu_g + A_2\cdot 12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right) + A_\Delta C_y\frac{\eta_x^2L_2^2}{1-\lambda}\right)}_{T_1}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\
&\quad + \underbrace{\left(3L_g^2\eta_x - A_2\eta_z\mu_g + A_\Delta C_z\frac{\eta_x^2L_g^2}{1-\lambda}\right)}_{T_2}\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] \\
&\quad + \underbrace{\left(\frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2} + A_1\left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2 + A_2\left(\eta_x^2 + \frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2 + \frac{2A_\Delta\eta_x^2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\right)}_{T_3}\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k)\right\|^2\right] \\
&\quad + \underbrace{\left(3L_g^2\eta_x + A_1C_{y^\star}L_g^2\eta_x + A_2C_{z^\star}L_g^2\eta_x - \frac{(1-\lambda)A_\Delta}{2}\right)}_{T_4}\frac{\mathbf{\Delta}^k}{n_h} + R_k,
\end{aligned}
\tag{36}
$$

where we define the residual part R_k as follow

$$
\begin{aligned}
R_k := &\left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + 2A_1\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + A_2\left(2\eta_z^2 + \frac{6\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] \\
&+ \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h} + \frac{2A_1\eta_y^2\sigma^2}{n_h} + \frac{2A_2\eta_z^2\sigma_1^2}{n_h} + \eta_x^2A_\Delta C_{\sigma 1} + \frac{\eta_x^2}{1-\lambda}A_\Delta C_{\zeta 1} \\
&+ \frac{8\kappa\eta_x^2}{1-\lambda}A_\Delta C_{\zeta 2} + \frac{8\kappa\eta_x^2}{1-\lambda}A_\Delta C_{\sigma 2}.
\end{aligned}
\tag{37}
$$

To obtain a monotone descent in $\mathcal{L}^k$, we choose the step sizes so that each bracketed coefficient above is non-positive.

First, for the coefficient T_1, take $\eta_x \le \sqrt{\frac{1-\lambda}{A_3C_y}}$, $\kappa \le \frac{1-\lambda}{A_4C_y}$, we obtain

$$
\begin{aligned}
T_1 &= 3L_2^2\eta_x - A_1\eta_y\mu_g + A_2 12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right) + A_\Delta C_y\frac{\eta_x^2L_2^2}{1-\lambda} \\
&\le 4L_2^2\eta_x - A_1\eta_y\mu_g + A_2 12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right) + \kappa A_4C_y\frac{\eta_xL_2^2}{1-\lambda} \\
&\le 4L_2^2\eta_x - A_1\eta_y\mu_g + A_2\frac{48\eta_zL_2^2}{\mu_g} + \kappa A_4C_y\frac{\eta_xL_2^2}{1-\lambda} = 0.
\end{aligned}
\tag{38}
$$

For the coefficient T_2, take $\eta_x \le \sqrt{\frac{1-\lambda}{A_3C_z}}$ and $\kappa \le \frac{1-\lambda}{A_4C_z}$; then

$$
T_2 = 3L_g^2\eta_x - A_2\eta_z\mu_g + A_\Delta C_z\frac{\eta_x^2L_g^2}{1-\lambda} \le 4L_g^2\eta_x - A_2\eta_z\mu_g + \kappa A_4C_z\frac{\eta_xL_g^2}{1-\lambda} = 0. \tag{39}
$$

The coefficient T_3 is non-positive when $\eta_x \le \frac{\eta_y \mu_g}{72 A_1 L_{y^\star}^2}$, $\eta_x \le \frac{\eta_z \mu_g}{72 A_1 L_{z^\star}^2}$, $\eta_x \le \frac{1}{A_3}$, $\kappa \le \frac{(1-\lambda) L_g^2}{24 A_4 L_2^2 \lambda^{3/2}}$, and $\eta_x \le \frac{1}{4 D_0}$; in that case

$$\begin{aligned} T_3 &= \frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2} + A_1\left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right) L_{y^\star}^2 + A_2\left(\eta_x^2 + \frac{3\eta_x^2}{\eta_z\mu_g}\right) L_{z^\star}^2 \\ &\quad + \frac{2A_3\eta_x^3 L_2^2 \lambda^{3/2}}{(1-\lambda)L_g^2} + \frac{2\kappa A_4 \eta_x L_2^2 \lambda^{3/2}}{(1-\lambda)L_g^2} \\ &\le \frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{4} + A_1 L_{y^\star}^2 \eta_x^2 + A_2 L_{z^\star}^2 \eta_x^2 + \frac{2A_3\eta_x^3 L_2^2 \lambda^{3/2}}{(1-\lambda)L_g^2} \\ &\le \frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{4} + \eta_x^2 \underbrace{\left(A_1 L_{y^\star}^2 + A_2 L_{z^\star}^2 + \frac{2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\right)}_{D_0} \le -\frac{\eta_x}{4} + \eta_x^2 D_0 \le 0. \end{aligned} \tag{40}$$

Thus the stochastic-direction term contributes at most zero. Hence the one-step bound (36) reduces to

$$\begin{aligned} &\mathcal{L}^{k+1} - \mathcal{L}^k \\ &\le -\frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \underbrace{\left(3L_g^2\eta_x + A_1 C_{y^\star} L_g^2 \eta_x + A_2 C_{z^\star} L_g^2 \eta_x - \frac{(1-\lambda)A_\Delta}{2}\right)}_{T_4} \frac{\boldsymbol{\Delta}^k}{n_h} + R_k. \end{aligned} \tag{41}$$

The only remaining undesired term is the coefficient of $\boldsymbol{\Delta}^k/n_h$ (consensus error). To control it, we bound the aggregation-error terms in R_k, using Lemma 3:

$$\begin{aligned} &2\eta_x\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{8A_1\eta_y}{\mu_g}\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + \frac{8A_2\eta_z}{\mu_g}\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] \\ &\le \frac{4\kappa}{\eta_x n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + L_g^2\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right]\right) + \frac{16A_1\kappa}{\eta_y\mu_g n_h}\mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right] + \frac{16A_2\kappa(1+L_g^2)}{\eta_z\mu_g n_h}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] \\ &\quad + \frac{8A_1\kappa\eta_y}{\mu_g}\left(2\zeta_g^2 + 4\sigma^2\right) + 2\kappa\eta_x\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right) + \frac{8A_2\kappa\eta_z}{\mu_g}\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right) \end{aligned} \tag{42}$$

Using the definition of $\boldsymbol{\Delta}^k$, we have

$$\begin{aligned} &2\eta_x\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{8A_1\eta_y}{\mu_g}\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + \frac{8A_2\eta_z}{\mu_g}\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] \\ &\le \frac{\kappa}{\eta_x}\left((1+L_g^2) + \frac{16A_1}{c_y\mu_g} + \frac{16A_2(1+L_g^2)}{c_z\mu_g}\right)\frac{\boldsymbol{\Delta}^k}{n_h} \\ &\quad + 2\kappa\eta_x\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right) + \frac{8A_1\kappa\eta_y}{\mu_g}\left(2\zeta_g^2 + 4\sigma^2\right) + \frac{8A_2\kappa\eta_z}{\mu_g}\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right). \end{aligned} \tag{43}$$

Substituting this into the recursion, the combined coefficient in front of $\boldsymbol{\Delta}^k/n_h$ is non-positive because the κ/η_x contribution cancels with $\frac{\kappa(1-\lambda)A_4}{2\eta_x}$ inside $\frac{(1-\lambda)A_\Delta}{2}$; indeed, using $A_\Delta = A_3\eta_x + \kappa A_4/\eta_x$ and $A_4 = \frac{2}{1-\lambda}\left((1+L_g^2) + \frac{16A_1}{c_y\mu_g} + \frac{16A_2(1+L_g^2)}{c_z\mu_g}\right)$,

$$\begin{aligned} T_4' &= 3L_g^2\eta_x + \left(A_1 C_{y^\star} + A_2 C_{z^\star}\right) L_g^2\eta_x - \frac{(1-\lambda)A_\Delta}{2} + \frac{\kappa}{\eta_x}\left((1+L_g^2) + \frac{16A_1}{c_y\mu_g} + \frac{16A_2(1+L_g^2)}{c_z\mu_g}\right) \\ &= \left(3 + A_1 C_{y^\star} + A_2 C_{z^\star}\right) L_g^2\eta_x - \frac{(1-\lambda)A_3\eta_x}{2} = -\left(1 + A_1 C_{y^\star} + A_2 C_{z^\star}\right) L_g^2\eta_x \le 0, \end{aligned} \tag{44}$$

where the last equality substitutes $A_3 = \frac{(8 + 4A_1 C_{y^\star} + 4A_2 C_{z^\star}) L_g^2}{1-\lambda}$.

Collecting the remaining terms, we obtain the simplified Lyapunov recursion

$$\begin{aligned}\mathcal{L}^{k+1} - \mathcal{L}^k \le -\frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h} + \frac{2A_1\eta_y^2\sigma^2}{n_h} + \frac{2A_2\eta_z^2\sigma_1^2}{n_h} \\ + \eta_x^2 A_\Delta C_{\sigma 1} + \frac{\eta_x^2}{1-\lambda}A_\Delta C_{\zeta 1} + \frac{8\kappa\eta_x^2}{1-\lambda}\left(A_\Delta C_{\zeta 2} + A_\Delta C_{\sigma 2}\right) \\ + \left(2\kappa\eta_x + \frac{8A_2\kappa\eta_z}{\mu_g}\right)\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right) + \frac{8A_1\kappa\eta_y}{\mu_g}\left(2\zeta_g^2 + 4\sigma^2\right).\end{aligned} \tag{45}$$

Let R_k' denote the sum over $k = 0, \dots, K-1$ of the constant (non-gradient) terms on the right-hand side of (45). Summing (45) over $k = 0, \dots, K-1$ yields

$$\mathcal{L}^K - \mathcal{L}^0 \le -\frac{\eta_x}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + {R_k}'. \tag{46}$$

Because $\mathcal{L}^K \ge 0$ (each term in the potential is non-negative), rearranging gives

$$\frac{\eta_x}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \mathcal{L}^0 + {R_k}'. \tag{47}$$

The initial potential satisfies

$$\mathcal{L}^0 = \Phi(\bar{x}^0) + A_1\left\|\bar{y}^0 - \bar{y}_\star^0\right\|^2 + A_2\left\|\bar{z}^0 - \bar{z}_\star^0\right\|^2 + A_\Delta\frac{\mathbf{\Delta}^0}{n_h}. \tag{48}$$

Expressing the remainder R_k in terms of the problem parameters yields an explicit bound. Introduce non-negative constants $a_1, \dots, a_7$ (depending on $L_{\nabla\Phi}$, ρ, A_1, A_2, A_3, A_4, κ, μ_g, c_y, c_z, ζ_f, ζ_g, σ) so that the right-hand side of (47) is bounded above by the expression below; one may take

$$\begin{aligned}&a_1 := \left(L_{\nabla\Phi} + 2A_1c_y^2 + 2A_2c_z^2\right)(1+\rho^2),\ a_2 := \frac{A_3C_{\zeta 1}}{1-\lambda},\ a_3 := A_3C_{\sigma 1},\ a_4 := \frac{A_3C_{\zeta 2}}{1-\lambda}, \\ &a_5 := \frac{A_3C_{\sigma 2}}{1-\lambda},\ a_6 := \left(12 + \frac{48A_2c_z}{\mu_g}\right)\left(\zeta_f^2 + \rho^2\zeta_g^2\right) + \frac{16A_1c_y}{\mu_g}\zeta_g^2 + \frac{A_4}{1-\lambda}\left(C_{\zeta 1} + \kappa C_{\zeta 2}\right), \\ &a_7 := \left(8 + \frac{32A_2c_z}{\mu_g}\right)(1+\rho^2) + \frac{32A_1c_y}{\mu_g} + A_4C_{\sigma 1} + \frac{A_4C_{\sigma 2}}{1-\lambda}.\end{aligned} \tag{49}$$

Then, we have

$$\begin{aligned}\frac{\eta_x}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \mathcal{L}^0 + a_1\frac{\eta_x^2\sigma^2}{n_h}K + a_2\eta_x^3K + a_3\eta_x^3\sigma^2K + \kappa a_4\eta_x^3K + \kappa a_5\eta_x^3\sigma^2K \\ + a_6\kappa\eta_xK + a_7\kappa\eta_x\sigma^2K.\end{aligned} \tag{50}$$

Dividing both sides by $\eta_x K/4$ gives the time-average gradient norm bound

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \frac{4\mathcal{L}^0}{\eta_xK} + a_1\frac{4\eta_x\sigma^2}{n_h} + 4a_2\eta_x^2 + 4a_3\eta_x^2\sigma^2 + 4\kappa a_4\eta_x^2 + 4\kappa a_5\eta_x^2\sigma^2 \\ + 4a_6\kappa + 4a_7\kappa\sigma^2.\end{aligned} \tag{51}$$

Choosing the step size $\eta_x = \mathcal{O}(1/\sqrt{n_hK})$ and absorbing problem-dependent constants into the $\mathcal{O}$-notation yields the convergence rate stated in the theorem:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \mathcal{O}\left(\frac{1}{\sqrt{n_hK}} + \frac{n_h\sigma^2}{K(1-\lambda)} + \frac{n_h(\zeta_f^2+\zeta_g^2)}{K(1-\lambda)^2}\right) \\ + \mathcal{O}\left(\frac{\kappa n_h\sigma^2}{K(1-\lambda)^2} + \frac{\kappa n_h(\zeta_f^2+\zeta_g^2)}{K(1-\lambda)^2}\right) + \mathcal{O}\left(\kappa\left(\sigma^2 + \zeta_f^2 + \zeta_g^2\right)\right).\end{aligned} \tag{52}$$

This completes the proof. □

Corollary 1. *Under the conditions of Theorem 1, suppose $\mathcal{A}(\cdot)$ satisfies the (δ,κ)-robustness criterion with $\kappa=\mathcal{O}(\delta)$. The transient time of* `SLBRD` *in Algorithm 1 is*

$$\tau=\mathcal{O}\left(\max\left\{\frac{n_h^3}{(1-\lambda)^2},\frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4},\frac{\delta^2 n_h^3}{(1-\lambda)^4}\right\}\right). \tag{53}$$

Proof. The algorithm achieves linear speedup when the higher-order terms in (52) are dominated by the $\mathcal{O}(1/\sqrt{n_hK})$ term. This requires each of the following conditions to hold:

$$\frac{n_h\sigma^2}{(1-\lambda)K}\lesssim\frac{1}{\sqrt{n_hK}},\quad \frac{n_h\left(\zeta_f^2+\zeta_g^2\right)}{(1-\lambda)^2K}\lesssim\frac{1}{\sqrt{n_hK}}, \tag{54}$$

$$\frac{\kappa n_h\sigma^2}{(1-\lambda)^2K}\lesssim\frac{1}{\sqrt{n_hK}},\quad \frac{\kappa n_h\left(\zeta_f^2+\zeta_g^2\right)}{(1-\lambda)^2K}\lesssim\frac{1}{\sqrt{n_hK}}. \tag{55}$$

Rearranging each condition in terms of K yields the respective requirements:

$$K\gtrsim\frac{n_h^3}{(1-\lambda)^2},\quad K\gtrsim\frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4},\quad K\gtrsim\frac{\kappa^2 n_h^3}{(1-\lambda)^4}. \tag{56}$$

Taking the maximum over all conditions. Suppose $\mathcal{A}(\cdot)$ satisfies the (δ,κ)-robustness criterion with $\kappa=\mathcal{O}(\delta)$. Then the transient time satisfies

$$K\gtrsim\mathcal{O}\left(\max\left\{\frac{n_h^3}{(1-\lambda)^2},\frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4},\frac{\delta^2 n_h^3}{(1-\lambda)^4}\right\}\right), \tag{57}$$

which completes the proof. □

D.3 Proof of Supporting Lemmas

D.3.1 Proof of Lemma 3

Lemma 3. *Suppose Assumptions 1 and 2, 4, 5 hold. Then*

$$\begin{aligned}
\mathbb{E}\left[\left\|R_{x,i}^{k+\frac{1}{2}}-\bar{x}_i^{k+\frac{1}{2}}\right\|^2\right]&\le 2\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\left(\mathbb{E}\left[\left\|x_j^k-\bar{x}_i^k\right\|^2\right]+L_g^2\mathbb{E}\left[\left\|z_j^k-\bar{z}_i^k\right\|^2\right]\right)\\
&\quad+\kappa\eta_x^2\left(6\zeta_f^2+6\rho^2\zeta_g^2+4\sigma_1^2\right),\\
\mathbb{E}\left[\left\|R_{y,i}^{k+\frac{1}{2}}-\bar{y}_i^{k+\frac{1}{2}}\right\|^2\right]&\le 2\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\mathbb{E}\left[\left\|y_j^k-\bar{y}_i^k\right\|^2\right]+\kappa\eta_y^2\left(2\zeta_g^2+4\sigma^2\right),\\
\mathbb{E}\left[\left\|R_{z,i}^{k+\frac{1}{2}}-\bar{z}_i^{k+\frac{1}{2}}\right\|^2\right]&\le 2\kappa(1+L_g^2)\sum_{j\in\mathcal{H}_i}w_{ij}\mathbb{E}\left[\left\|z_j^k-\bar{z}_i^k\right\|^2\right]+\kappa\eta_z^2\left(6\zeta_f^2+6\rho^2\zeta_g^2+4\sigma_1^2\right).
\end{aligned}$$

Proof. We prove each bound in turn.

Bound for x. Fix $k\in\{0,\cdots,K-1\}$. By the (δ,κ)-robustness property (Definition 2), we have

$$\mathbb{E}\left[\left\|R_{x,i}^{k+\frac{1}{2}}-\bar{x}_i^{k+\frac{1}{2}}\right\|^2\right]\le\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\mathbb{E}\left[\left\|x_j^{k+\frac{1}{2}}-\bar{x}_i^{k+\frac{1}{2}}\right\|^2\right], \tag{58}$$

where $R_{x,i}^{k+\frac{1}{2}}:=\mathcal{A}_i\left(x_i^{k+\frac{1}{2}},\{x_m^{k+\frac{1}{2}}\}_{m\in\mathcal{N}_i}\right)$ and $\bar{x}_i^{k+\frac{1}{2}}:=\sum_{j\in\mathcal{H}_i}w_{ij}x_j^{k+\frac{1}{2}}$.

Substituting the update rule into (58) and applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|R_{x,i}^{k+\frac{1}{2}}-\bar{x}_i^{k+\frac{1}{2}}\right\|^2\right]&\le\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\mathbb{E}\left[\left\|x_j^k-\bar{x}_i^k-\eta_x\left(D_{x,j}^k-\bar{D}_{x,i}^k\right)\right\|^2\right]\\
&\le 2\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\left(\mathbb{E}\left[\left\|x_j^k-\bar{x}_i^k\right\|^2\right]+\eta_x^2\mathbb{E}\left[\left\|D_{x,j}^k-\bar{D}_{x,i}^k\right\|^2\right]\right).
\end{aligned} \tag{59}$$

To bound the second term on the right-hand side of (59), we apply the bias-variance decomposition under Assumption 4:

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|D_{x,j}^k - \bar{D}_{x,i}^k\right\|^2\right] \\
&= \sum_{j\in\mathcal{H}_i} w_{ij}\Big(\mathbb{E}\left[\left\|\mathbb{E}(D_{x,j}^k) - \mathbb{E}(\bar{D}_{x,i}^k)\right\|^2\right] + \mathbb{E}\left[\left\|D_{x,j}^k - \mathbb{E}(D_{x,j}^k)\right\|^2\right] + \mathbb{E}\left[\left\|\bar{D}_{x,i}^k - \mathbb{E}(\bar{D}_{x,i}^k)\right\|^2\right]\Big) \\
&\le \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|\mathbb{E}(D_{x,j}^k) - \mathbb{E}(\bar{D}_{x,i}^k)\right\|^2\right] + 2\sigma_1^2.
\end{aligned} \tag{60}$$

To bound the bias term, we expand the gradient and Hessian differences using the smoothness and heterogeneity assumptions. Applying Assumptions 2 and 5 together with Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|\mathbb{E}(D_{x,j}^k) - \mathbb{E}(\bar{D}_{x,i}^k)\right\|^2\right] \\
&\le \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\Bigg[\Bigg\|\nabla_x f_j(x_j^k, y_j^k) - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_x f_c(x_c^k, y_c^k) \\
&\qquad + \nabla_{xy}^2 g_j(x_j^k, y_j^k)z_j^k - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_{xy}^2 g_c(x_c^k, y_c^k)z_c^k\Bigg\|^2\Bigg].
\end{aligned} \tag{61}$$

To handle the cross-Hessian terms, we use the decomposition

$$\begin{aligned}
&\nabla_{xy}^2 g_j(x_j^k, y_j^k)z_j^k - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_{xy}^2 g_c(x_c^k, y_c^k)z_c^k \\
&= \left(\nabla_{xy}^2 g_j(x_j^k, y_j^k) - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_{xy}^2 g_c(x_c^k, y_c^k)\right)z_j^k - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_{xy}^2 g_c(x_c^k, y_c^k)(z_c^k - z_j^k).
\end{aligned} \tag{62}$$

Collecting like terms and applying Assumptions 2 and 5 together with Cauchy-Schwarz inequalityyields

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|\mathbb{E}(D_{x,j}^k) - \mathbb{E}(\bar{D}_{x,i}^k)\right\|^2\right] \\
&\le \sum_{j\in\mathcal{H}_i} w_{ij}\Bigg(3\mathbb{E}\left[\left\|\nabla_x f_j(x_j^k, y_j^k) - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_x f_c(x_c^k, y_c^k)\right\|^2\right] + 3L_g^2\mathbb{E}\left[\left\|z_j^k - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i} z_c^k\right\|^2\right] \\
&+ 3\rho^2\mathbb{E}\left[\left\|\nabla_{xy}^2 g_j(x_j^k, y_j^k) - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_{xy}^2 g_c(x_c^k, y_c^k)\right\|^2\right]\Bigg) \\
&\le 3\zeta_f^2 + 3\rho^2\zeta_g^2 + 3L_g^2\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right].
\end{aligned} \tag{63}$$

Substituting (63) and (60) into (59), we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|R_{x,i}^{k+\frac{1}{2}} - \bar{x}_i^{k+\frac{1}{2}}\right\|^2\right] &\le 2\kappa\sum_{j\in\mathcal{H}_i} w_{ij}\left(\mathbb{E}\left[\left\|x_j^k - \bar{x}_i^k\right\|^2\right] + 3\eta_x^2 L_g^2\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right]\right) \\
&\quad + \kappa\eta_x^2\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right).
\end{aligned} \tag{64}$$

Under the step-size condition $\eta_x \le 1/\sqrt{3}$ adopted in the convergence analysis, we have $3\eta_x^2 \le 1$, hence $3\eta_x^2 L_g^2 \le L_g^2$, which yields the bound stated in Lemma 3. *Bound for y.* The argument parallels that for x. Applying Definition 2 and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|R_{y,i}^{k+\frac{1}{2}} - \bar{y}_i^{k+\frac{1}{2}}\right\|^2\right] &\le \kappa\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|y_j^k - \bar{y}_i^k - \eta_y\left(D_{y,j}^k - \bar{D}_{y,i}^k\right)\right\|^2\right] \\
&\le 2\kappa\sum_{j\in\mathcal{H}_i} w_{ij}\left(\mathbb{E}\left[\left\|y_j^k - \bar{y}_i^k\right\|^2\right] + \eta_y^2\mathbb{E}\left[\left\|D_{y,j}^k - \bar{D}_{y,i}^k\right\|^2\right]\right).
\end{aligned} \tag{65}$$

For the second term on the right-hand side of (65), applying Assumption 4 gives

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|D_{y,j}^k-\bar{D}_{y,i}^k\right\|^2\right]\\
&\le \sum_{j\in\mathcal{H}_i} w_{ij}\Big(\mathbb{E}\left[\left\|\mathbb{E}(D_{y,j}^k)-\mathbb{E}(\bar{D}_{y,i}^k)\right\|^2\right]+\mathbb{E}\left[\left\|D_{y,j}^k-\mathbb{E}(D_{y,j}^k)\right\|^2\right]+\mathbb{E}\left[\left\|\bar{D}_{y,i}^k-\mathbb{E}(\bar{D}_{y,i}^k)\right\|^2\right]\Big)\\
&\le \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|\mathbb{E}(D_{y,j}^k)-\mathbb{E}(\bar{D}_{y,i}^k)\right\|^2\right]+2\sigma^2.
\end{aligned} \tag{66}$$

Bounding the gradient dissimilarity by Assumption 5 yields

$$\begin{aligned}
\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|D_{y,j}^k-\bar{D}_{y,i}^k\right\|^2\right] &\le \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|\nabla_y g_j(x_j^k,y_j^k)-\frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_y g_c(x_c^k,y_c^k)\right\|^2\right]+2\sigma^2\\
&\le \zeta_g^2+2\sigma^2.
\end{aligned} \tag{67}$$

Plugging (67) into (65) gives

$$\mathbb{E}\left[\left\|R_{y,i}^{k+\frac{1}{2}}-\bar{y}_i^{k+\frac{1}{2}}\right\|^2\right]\le 2\kappa\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|y_j^k-\bar{y}_i^k\right\|^2\right]+2\kappa\eta_y^2\left(\zeta_g^2+2\sigma^2\right). \tag{68}$$

Bound for z. The argument parallels that for x. By Definition 2, we have

$$\mathbb{E}\left[\left\|R_{z,i}^{k+\frac{1}{2}}-\bar{z}_i^{k+\frac{1}{2}}\right\|^2\right]\le\kappa\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|z_j^{k+\frac{1}{2}}-\bar{z}_i^{k+\frac{1}{2}}\right\|^2\right], \tag{69}$$

where $R_{z,i}^{k+\frac{1}{2}}:=\mathcal{A}_i\left(z_i^{k+\frac{1}{2}},\{z_m^{k+\frac{1}{2}}\}_{m\in\mathcal{N}_i}\right)$ and $\bar{z}_i^{k+\frac{1}{2}}:=\sum_{j\in\mathcal{H}_i} w_{ij}z_j^{k+\frac{1}{2}}$.

Substituting the update rule into (69) and applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|R_{z,i}^{k+\frac{1}{2}}-\bar{z}_i^{k+\frac{1}{2}}\right\|^2\right] &\le \kappa\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|z_j^k-\bar{z}_i^k-\eta_z\left(D_{z,j}^k-\bar{D}_{z,i}^k\right)\right\|^2\right]\\
&\le 2\kappa\sum_{j\in\mathcal{H}_i} w_{ij}\left(\mathbb{E}\left[\left\|z_j^k-\bar{z}_i^k\right\|^2\right]+\eta_z^2\mathbb{E}\left[\left\|D_{z,j}^k-\bar{D}_{z,i}^k\right\|^2\right]\right).
\end{aligned} \tag{70}$$

For the second term on the right-hand side of (70), applying Assumption 4 gives

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|D_{z,j}^k-\bar{D}_{z,i}^k\right\|^2\right]\\
&= \sum_{j\in\mathcal{H}_i} w_{ij}\Big(\mathbb{E}\left[\left\|\mathbb{E}(D_{z,j}^k)-\mathbb{E}(\bar{D}_{z,i}^k)\right\|^2\right]+\mathbb{E}\left[\left\|D_{z,j}^k-\mathbb{E}(D_{z,j}^k)\right\|^2\right]+\mathbb{E}\left[\left\|\bar{D}_{z,i}^k-\mathbb{E}(\bar{D}_{z,i}^k)\right\|^2\right]\Big)\\
&\le \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|\mathbb{E}(D_{z,j}^k)-\mathbb{E}(\bar{D}_{z,i}^k)\right\|^2\right]+2\sigma_1^2.
\end{aligned} \tag{71}$$

Applying Assumptions 2 and 5 together with Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|D_{z,j}^k - \bar{D}_{z,i}^k\right\|^2\right] \\
&\le \sum_{j\in\mathcal{H}_i} w_{ij}\Big(3\mathbb{E}\left[\left\|\nabla_y f_j(x_j^k,y_j^k) - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_y f_c(x_c^k,y_c^k)\right\|^2\right] + 3L_g^2\mathbb{E}\left[\left\|z_j^k - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i} z_c^k\right\|^2\right] \\
&+ 3\rho^2\mathbb{E}\left[\left\|\nabla_{yy}^2 g_j(x_j^k,y_j^k) - \frac{1}{|\mathcal{H}_i|}\sum_{c\in\mathcal{H}_i}\nabla_{yy}^2 g_c(x_c^k,y_c^k)\right\|^2\right]\Big) + 2\sigma_1^2 \\
&\le 3\zeta_f^2 + 3\rho^2\zeta_g^2 + 3L_g^2\sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right] + 2\sigma_1^2.
\end{aligned} \tag{72}$$

Substituting (72) into (70) yields

$$\mathbb{E}\left[\left\|R_{z,i}^{k+\frac{1}{2}} - \bar{z}_i^{k+\frac{1}{2}}\right\|^2\right] \le 2\kappa\sum_{j\in\mathcal{H}_i} w_{ij}\left(1+3\eta_z^2L_g^2\right)\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right] + \kappa\eta_z^2\left(6\zeta_f^2 + 6\rho^2\zeta_g^2 + 4\sigma_1^2\right). \tag{73}$$

Under the step-size condition $\eta_z \le 1/\sqrt{3}$ adopted in the convergence analysis, we have $3\eta_z^2 \le 1$, hence $1 + 3\eta_z^2L_g^2 \le 1 + L_g^2$, which gives the bound stated in Lemma 3. This completes the proof.

□

D.3.2 Proof of Lemma 4

Lemma 4. *Suppose that Assumptions 1 and 2, 3, 4 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \le{}& \mathbb{E}\left[\Phi(\bar{x}^k)\right] - \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \left(\frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2}\right)\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k)\right\|^2\right] \\
&+ \frac{3L_g^2\eta_x}{n_h}\boldsymbol{\Delta}^k + 3L_2^2\eta_x\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 3L_g^2\eta_x\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] \\
&+ \left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h}.
\end{aligned}$$

Proof. By the $L_{\nabla\Phi}$-smoothness of Φ (Lemma 1), we have

$$\Phi(\bar{x}^{k+1}) \le \Phi(\bar{x}^k) + \langle\nabla\Phi(\bar{x}^k), \bar{x}^{k+1} - \bar{x}^k\rangle + \frac{L_{\nabla\Phi}}{2}\left\|\bar{x}^{k+1} - \bar{x}^k\right\|^2. \tag{74}$$

Taking expectations on both sides of (74) gives

$$\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \le \mathbb{E}\left[\Phi(\bar{x}^k)\right] + \mathbb{E}\left[\langle\nabla\Phi(\bar{x}^k), \bar{x}^{k+1} - \bar{x}^k\rangle\right] + \frac{L_{\nabla\Phi}}{2}\mathbb{E}\left[\left\|\bar{x}^{k+1} - \bar{x}^k\right\|^2\right]. \tag{75}$$

Substituting the averaged update $\bar{x}^{k+1} = \bar{x}^k - \eta_x\bar{D}_x^k + \eta_x\bar{\xi}_x^k$ into the right-hand side of (75) yields

$$\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \le \mathbb{E}\left[\Phi(\bar{x}^k)\right] + \eta_x\mathbb{E}\left[\langle\nabla\Phi(\bar{x}^k), \bar{\xi}_x^k - \bar{D}_x^k\rangle\right] + L_{\nabla\Phi}\eta_x^2\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right] + L_{\nabla\Phi}\eta_x^2\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right].$$

Separating the stochastic noise via Assumption 4, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \le{}& \mathbb{E}\left[\Phi(\bar{x}^k)\right] + \eta_x\mathbb{E}\left[\langle\nabla\Phi(\bar{x}^k), \bar{\xi}_x^k - \mathbb{E}(\bar{D}_x^k)\rangle\right] + L_{\nabla\Phi}\eta_x^2\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k)\right\|^2\right] \\
&+ L_{\nabla\Phi}\eta_x^2\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h}.
\end{aligned} \tag{76}$$

Applying the polarization identity $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$, we get

$$-\eta_x\mathbb{E}\left[\langle\nabla\Phi(\bar{x}^k), \mathbb{E}(\bar{D}_x^k)\rangle\right] = \frac{\eta_x}{2}\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k) - \nabla\Phi(\bar{x}^k)\right\|^2\right] - \frac{\eta_x}{2}\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k)\right\|^2\right] - \frac{\eta_x}{2}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right]. \tag{77}$$

By Young's inequality $2\langle a, b\rangle \leq c\|a\|^2 + \frac{1}{c}\|b\|^2$, we have

$$\eta_x \mathbb{E}\left[\langle \nabla\Phi(\bar{x}^k), \bar{\xi}_x^k\rangle\right] \leq \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \eta_x \mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right]. \tag{78}$$

Plugging (77) into (76) and rearranging terms, we obtain

$$\begin{aligned}\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \leq \mathbb{E}\left[\Phi(\bar{x}^k)\right] + \frac{\eta_x}{2}\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k) - \nabla\Phi(\bar{x}^k)\right\|^2\right] - \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \\ + \left(L_{\nabla\Phi}\eta_x^2 - \frac{\eta_x}{2}\right)\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k)\right\|^2\right] + \left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h}.\end{aligned} \tag{79}$$

Finally, applying Lemma 11 to bound the bias term $\|\mathbb{E}(\bar{D}_x^k) - \nabla\Phi(\bar{x}^k)\|^2$ yields

$$\begin{aligned}\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \leq \mathbb{E}\left[\Phi(\bar{x}^k)\right] - \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \left(\frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2}\right)\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_x^k)\right\|^2\right] + \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h} \\ + \left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{3L_g^2\eta_x}{n_h}\boldsymbol{\Delta}^k \\ + 3L_2^2\eta_x\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 3L_g^2\eta_x\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right].\end{aligned} \tag{80}$$

This completes the proof.

□

Lemma 11. *Suppose that Assumptions 1 and 2 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbb{E}\left(\bar{D}_x^k\right) - \nabla\Phi(\bar{\mathbf{x}}^k)\right\|^2\right] \leq \frac{6L_2^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right) + \frac{6L_g^2}{n_h}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] \\ + 6L_2^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 6L_g^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right].\end{aligned}$$

Proof. By the Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}\left[\left\|\mathbb{E}\left(\bar{D}_x^k\right) - \nabla\Phi(\bar{\mathbf{x}}^k)\right\|^2\right] \leq 2\mathbb{E}\left[\left\|\mathbb{E}\left(D_{\bar{x}}^k\right) - \mathbb{E}\left(\bar{D}_x^k\right)\right\|^2\right] + 2\mathbb{E}\left[\left\|\mathbb{E}\left(D_{\bar{x}}^k\right) - \nabla\Phi(\bar{\mathbf{x}}^k)\right\|^2\right], \tag{81}$$

where $\mathbb{E}\left(D_{\bar{x}}^k\right) = \nabla_x f(\bar{x}^k, \bar{y}^k) + \nabla_{xy}^2 g(\bar{x}^k, \bar{y}^k)\bar{z}^k$. We bound each term on the right-hand side of (81) separately. For the first term, by the definition of $\bar{D}_x^k$ and $D_{\bar{x}}^k$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbb{E}\left(D_{\bar{x}}^k\right) - \mathbb{E}\left(\bar{D}_x^k\right)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla_x f_i(x_i^k, y_i^k) - \nabla_x f_i(\bar{x}^k, \bar{y}^k) + \nabla_{xy}^2 g_i(x_i^k, y_i^k)z_i^k - \nabla_{xy}^2 g_i(\bar{x}^k, \bar{y}^k)\bar{z}^k\right)\right\|^2\right] \\ &\leq \Big\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\Big(\nabla_x f_i(x_i^k, y_i^k) - \nabla_x f_i(\bar{x}^k, \bar{y}^k) + \nabla_{xy}^2 g_i(x_i^k, y_i^k)z_i^k - \nabla_{xy}^2 g_i(x_i^k, y_i^k)\bar{z}^k \\ &\quad + \nabla_{xy}^2 g_i(x_i^k, y_i^k)\bar{z}^k - \nabla_{xy}^2 g_i(\bar{x}^k, \bar{y}^k)\bar{z}^k\Big)\Big\|^2 \\ &\leq \frac{1}{n_h}((3L_f^2 + 3L_g^2\rho^2)\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right) + 3L_g^2\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right]).\end{aligned} \tag{82}$$

For the second term, we use the optimality conditions of the lower-level problem:

$$\nabla_y g(\bar{x}^k, \bar{y}_\star^k) = 0,\ \nabla_{yy}^2 g(\bar{x}^k, \bar{y}_\star^k)\bar{z}_\star^k + \nabla_y f(\bar{x}^k, \bar{y}_\star^k) = 0. \tag{83}$$

Since $\nabla\Phi(\bar{x}^k) = \nabla_x f(\bar{x}^k, \bar{y}_\star^k) + \nabla^2_{xy} g(\bar{x}^k, \bar{y}_\star^k)\bar{z}_\star^k$, applying Assumption 2 gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}\left(\bar{D}_x^k\right) - \nabla\Phi(\bar{x}^k)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla_x f(\bar{x}^k, \bar{y}^k) - \nabla_x f(\bar{x}^k, \bar{y}_\star^k) + \left(\nabla^2_{xy} g(\bar{x}^k, \bar{y}^k) - \nabla^2_{xy} g(\bar{x}^k, \bar{y}_\star^k)\right)\bar{z}_\star^k + \nabla^2_{xy} g(\bar{x}^k, \bar{y}^k)\left(\bar{z}^k - \bar{z}_\star^k\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla_x f(\bar{x}^k, \bar{y}^k) - \nabla_x f(\bar{x}^k, \bar{y}_\star^k)\right\|^2\right] + \mathbb{E}\left[\left\|\left(\nabla^2_{xy} g(\bar{x}^k, \bar{y}^k) - \nabla^2_{xy} g(\bar{x}^k, \bar{y}_\star^k)\right)\bar{z}_\star^k\right\|^2\right] \\
&+ \mathbb{E}\left[\left\|\nabla^2_{xy} g(\bar{x}^k, \bar{y}^k)\left(\bar{z}^k - \bar{z}_\star^k\right)\right\|^2\right] \\
&\leq \left(3L_f^2 + 3L_g^2\rho^2\right)\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 3L_g^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right].
\end{aligned} \tag{84}$$

Substituting (82) and (84) into (81) yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbb{E}\left(\bar{D}_x^k\right) - \nabla\Phi(\bar{\mathbf{x}}^k)\right\|^2\right] \leq \frac{6L_2^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right) + \frac{6L_g^2}{n_h}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] \\
+ 6L_2^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 6L_g^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right].
\end{aligned} \tag{85}$$

This completes the proof.

□

D.3.3 Proof of Lemma 5

Lemma 5. *Suppose that Assumptions 2, 3, and 4 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^{k+1}\right\|^2\right] \leq (1 - \eta_y\mu_g)\,\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2\,\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right] \\
+ 2(1 + \frac{3}{\eta_y\mu_g})\frac{\eta_y^2 L_g^4}{n_h L_2^2}\mathbf{\Delta}^k + 2\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + 2\eta_y^2\frac{\sigma^2}{n_h}.
\end{aligned}$$

Proof. By Jensen's inequality and the Lipschitz continuity of $y^\star(\cdot)$ (Lemma 1), we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^{k+1}\right\|^2\right] &\leq \left(1 + \frac{\eta_y\mu_g}{3}\right)\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^k\right\|^2\right] + \left(1 + \frac{3}{\eta_y\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}_\star^{k+1} - \bar{y}_\star^k\right\|^2\right] \\
&\leq \left(1 + \frac{\eta_y\mu_g}{3}\right)\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^k\right\|^2\right] + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2\,\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right].
\end{aligned} \tag{86}$$

To bound the first term on the right-hand side of (86), we expand the averaged update:

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^k\right\|^2\right] = \mathbb{E}\left[\left\|\bar{y}^k - \eta_y\mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k\right\|^2\right] + \eta_y^2\mathbb{E}\left[\left\|\bar{D}_y^k - \mathbb{E}(\bar{D}_y^k) - \bar{\xi}_y^k\right\|^2\right] \\
+ 2\eta_y\mathbb{E}\left[\left\langle\bar{y}^k - \eta_y\mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k, \bar{\xi}_y^k\right\rangle\right].
\end{aligned} \tag{87}$$

By Young's inequality, we obtain

$$2\mathbb{E}\left[\left\langle\bar{y}^k - \eta_y\mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k, \eta_y\bar{\xi}_y^k\right\rangle\right] \leq v\mathbb{E}\left[\left\|\bar{y}^k - \eta_y\mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k\right\|^2\right] + \frac{\eta_y^2}{v}\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right]. \tag{88}$$

Substituting (88) into (87) yields

$$\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^k\right\|^2\right] \leq (1 + v)\mathbb{E}\left[\left\|\bar{y}^k - \eta_y\mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k\right\|^2\right] + \left(\eta_y^2 + \frac{\eta_y^2}{v}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + \eta_y^2\frac{\sigma^2}{n_h}. \tag{89}$$

For the first term on the right-hand side of (89), applying Cauchy-Schwarz inequalitygives

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{y}^k - \eta_y\mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k\right\|^2\right] \leq (1 + u)\mathbb{E}\left[\left\|\bar{y}^k - \eta_y\mathbb{E}(D_{\bar{y}}^k) - \bar{y}_\star^k\right\|^2\right] \\
+ (1 + \frac{1}{u})\eta_y^2\mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_y^k) - \mathbb{E}(D_{\bar{y}}^k)\right\|^2\right].
\end{aligned} \tag{90}$$

Since g is μ_g-strongly convex in y (Assumption 3), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{y}^k - \eta_y \mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k\right\|^2\right] &\leq (1+u)\left(1-\eta_y\mu_g\right)^2 \mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\ &\quad + (1+\frac{1}{u})\eta_y^2 \mathbb{E}\left[\left\|\mathbb{E}(\bar{D}_y^k) - \mathbb{E}(D_{\bar{y}}^k)\right\|^2\right].\end{aligned} \tag{91}$$

Bounding the bias term $\|\mathbb{E}(\bar{D}_y^k) - \mathbb{E}(D_{\bar{y}}^k)\|^2$ using Assumption 2 gives

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{y}^k - \eta_y \mathbb{E}(\bar{D}_y^k) - \bar{y}_\star^k\right\|^2\right] &\leq (1+u)\left(1-\eta_y\mu_g\right)^2 \mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\ &\quad + (1+\frac{1}{u})\frac{\eta_y^2 L_g^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right).\end{aligned} \tag{92}$$

Substituting (92) into (89) yields

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^k\right\|^2\right] &\leq (1+u)(1+v)\left(1-\eta_y\mu_g\right)^2 \mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\ &\quad + (1+\frac{1}{u})(1+v)\frac{\eta_y^2 L_g^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right) \\ &\quad + \left(\eta_y^2 + \frac{\eta_y^2}{v}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + \eta_y^2\frac{\sigma^2}{n_h}.\end{aligned} \tag{93}$$

Setting $u = v = \frac{\eta_y\mu_g}{3}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^k\right\|^2\right] &\leq (1+\frac{\eta_y\mu_g}{3})(1+\frac{\eta_y\mu_g}{3})\left(1-\eta_y\mu_g\right)^2 \mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\ &\quad + (1+\frac{3}{\eta_y\mu_g})(1+\frac{\eta_y\mu_g}{3})\frac{\eta_y^2 L_g^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right) \\ &\quad + \left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + \eta_y^2\frac{\sigma^2}{n_h}.\end{aligned} \tag{94}$$

Plugging (94) into (86) gives

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^{k+1}\right\|^2\right] \\ &\leq \left(1+\frac{\eta_y\mu_g}{3}\right)^3\left(1-\eta_y\mu_g\right)^2 \mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2 \mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right] \\ &+ \left(1+\frac{\eta_y\mu_g}{3}\right)(1+\frac{3}{\eta_y\mu_g})(1+\frac{\eta_y\mu_g}{2})\frac{\eta_y^2 L_g^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right) \\ &+ \left(1+\frac{\eta_y\mu_g}{3}\right)\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + \left(1+\frac{\eta_y\mu_g}{3}\right)\eta_y^2\frac{\sigma^2}{n_h}.\end{aligned} \tag{95}$$

Since $\eta_y\mu_g \leq 1$, simplifying the constants yields

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^{k+1}\right\|^2\right] &\leq \left(1-\eta_y\mu_g\right)\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2 \mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right] \\ &\quad + 2(1+\frac{3}{\eta_y\mu_g})\frac{\eta_y^2 L_g^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\right) \\ &\quad + 2\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + 2\eta_y^2\frac{\sigma^2}{n_h}.\end{aligned} \tag{96}$$

Hence, substituting the definition of $\mathbf{\Delta}^k$, we arrive at

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{y}^{k+1} - \bar{y}_\star^{k+1}\right\|^2\right] &\leq \left(1-\eta_y\mu_g\right)\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2 \mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right] \\ &\quad + 2(1+\frac{3}{\eta_y\mu_g})\frac{\eta_y^2 L_g^4}{n_h L_2^2}\mathbf{\Delta}^k + 2\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] + 2\eta_y^2\frac{\sigma^2}{n_h}.\end{aligned} \tag{97}$$

This completes the proof.

□

D.3.4 Proof of Lemma 6

Lemma 6. *Suppose that Assumptions 2, 3, and 4 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right] \le (1-\eta_z\mu_g)\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] + 12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] \\ + \frac{12L_g^2\eta_z^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\boldsymbol{\Delta}^k + \left(2\eta_z^2+\frac{6\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] \\ + 2\eta_z^2\frac{\sigma_1^2}{n_h} + \left(\eta_x^2+\frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right].\end{aligned}$$

Proof. Firstly, from Jensen's Inequality and the Lipschitz continuity of $\bar{z}_\star$, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right] &\le \left(1+\frac{\eta_z\mu_g}{3}\right)\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^k\right\|^2\right] + \left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{z}_\star^{k+1}-\bar{z}_\star^k\right\|^2\right] \\ &\le \left(1+\frac{\eta_z\mu_g}{3}\right)\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^k\right\|^2\right] + \left(\eta_x^2+\frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right].\end{aligned} \tag{98}$$

To bound the first term on the right-hand side of (98), we expand the averaged update:

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^k\right\|^2\right] = \mathbb{E}\left[\left\|\bar{z}^k-\eta_z\mathbb{E}(\bar{D}_z^k)-\bar{z}_\star^k\right\|^2\right] + \eta_z^2\mathbb{E}\left[\left\|\bar{D}_z^k-\mathbb{E}(\bar{D}_z^k)-\bar{\xi}_z^k\right\|^2\right] \\ + 2\eta_z\mathbb{E}\left[\left\langle\bar{z}^k-\eta_z\mathbb{E}(\bar{D}_z^k)-\bar{z}_\star^k,\bar{\xi}_z^k\right\rangle\right].\end{aligned} \tag{99}$$

By Young's inequality, we obtain

$$2\mathbb{E}\left[\left\langle\bar{z}^k-\eta_z\mathbb{E}(\bar{D}_z^k)-\bar{z}_\star^k,\eta_z\bar{\xi}_z^k\right\rangle\right] \le \frac{\eta_z\mu_g}{3}\mathbb{E}\left[\left\|\bar{z}^k-\eta_z\mathbb{E}(\bar{D}_z^k)-\bar{z}_\star^k\right\|^2\right] + \frac{3\eta_z}{\mu_g}\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right]. \tag{100}$$

Substituting (100) into (99) yields

$$\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^k\right\|^2\right] \le (1+\frac{\eta_z\mu_g}{3})\mathbb{E}\left[\left\|\bar{z}^k-\eta_z\mathbb{E}(\bar{D}_z^k)-\bar{z}_\star^k\right\|^2\right] + \left(\eta_z^2+\frac{3\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] + \eta_z^2\frac{\sigma_1^2}{n_h}. \tag{101}$$

By the definition of $\bar{z}_\star^k$, we can decompose the update as

$$\begin{aligned}&\bar{z}^k-\eta_z\mathbb{E}(\bar{D}_z^k)-\bar{z}_\star^k \\ &= \frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(I-\eta_z\nabla_{yy}^2g_i(x_i^k,y_i^k)\right)\left(\bar{z}^k-\bar{z}_\star^k\right) - \frac{\eta_z}{n_h}\sum_{i\in\mathcal{H}}\nabla_{yy}^2g_i(x_i^k,y_i^k)\left(z_i^k-\bar{z}^k\right) \\ &- \frac{\eta_z}{n_h}\sum_{i\in\mathcal{H}}\left[\left(\nabla_{yy}^2g_i(x_i^k,y_i^k)-\nabla_{yy}^2g(\bar{x}^k,\bar{y}_\star^k)\right)\bar{z}_\star^k + \left(\nabla_yf_i(x_i^k,y_i^k)-\nabla_yf(\bar{x}^k,\bar{y}_\star^k)\right)\right]\end{aligned} \tag{102}$$

where the equality uses $\nabla_{yy}^2g(\bar{x}^k,\bar{y}_\star^k)\bar{z}_\star^k+\nabla_yf(\bar{x}^k,\bar{y}_\star^k)=0$. Applying Jensen's inequality to (102), we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{z}^k-\eta_z\mathbb{E}(D_z^k)-\bar{z}_\star^k\right\|^2\right] \\ &\le \left(1+\frac{\eta_z\mu_g}{3}\right)\frac{1}{n_h}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\left(I-\eta_z\nabla_{yy}^2g_i(x_i^k,y_i^k)\right)\left(\bar{z}^k-\bar{z}_\star^k\right)\right\|^2\right] \\ &+ 3\eta_z^2L_g^2\left(1+\frac{3}{\eta_z\mu_g}\right)\frac{1}{n_h}\mathbb{E}\left[\left\|\mathbf{z}^k-\bar{\mathbf{z}}^k\right\|^2\right] \\ &+ 3\eta_z^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla_{yy}^2g_i(x_i^k,y_i^k)-\nabla_{yy}^2g(\bar{x}^k,\bar{y}^k)\right)\right\|^2\right]\mathbb{E}\left[\left\|\bar{z}_\star^k\right\|^2\right] \\ &+ 3\eta_z^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla_yf_i(x_i^k,y_i^k)-\nabla_yf(\bar{x}^k,\bar{y}^k)\right)\right\|^2\right].\end{aligned} \tag{103}$$

Since g_i is μ_g-strongly convex in y (Assumption 3) and $\eta_z\mu_g \leq 1$, we have

$$\mathbb{E}\left[\left\|\left(I-\eta_z\nabla^2_{yy}g_i(x_i^k,y_i^k)\right)\left(\bar{z}^k-\bar{z}_\star^k\right)\right\|^2\right] \leq (1-\eta_z\mu_g)^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]. \tag{104}$$

By the Lipschitz continuity of $\nabla^2_{yy}g_i$ and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla^2_{yy}g_i(x_i^k,y_i^k)-\nabla^2_{yy}g(\bar{x}^k,\bar{y}^k)\right)\right\|^2\right]\mathbb{E}\left[\left\|\bar{z}_\star^k\right\|^2\right]\\
&\leq 2\rho^2\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla^2_{yy}g_i(x_i^k,y_i^k)-\nabla^2_{yy}g_i(\bar{x}^k,\bar{y}^k)\right)\right\|^2\right]\\
&+2\rho^2\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla^2_{yy}g_i(\bar{x}^k,\bar{y}^k)-\nabla^2_{yy}g_i(\bar{x}^k,\bar{y}_\star^k)\right)\right\|^2\right]\\
&\leq \frac{2\rho^2L_g^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right)+2\rho^2L_g^2\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right].
\end{aligned} \tag{105}$$

Similarly, by the Lipschitz continuity of $\nabla_y f_i$ and Cauchy-Schwarz inequality, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla_y f_i(x_i^k,y_i^k)-\nabla_y f(\bar{x}^k,\bar{y}^k)\right)\right\|^2\right]\\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla_y f_i(x_i^k,y_i^k)-\nabla_y f_i(\bar{x}^k,\bar{y}^k)\right)\right\|^2\right]\\
&\quad+2\mathbb{E}\left[\left\|\frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\nabla_y f_i(\bar{x}^k,\bar{y}^k)-\nabla_y f_i(\bar{x}^k,\bar{y}_\star^k)\right)\right\|^2\right]\\
&\leq \frac{2L_f^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right)+2L_f^2\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right].
\end{aligned} \tag{106}$$

Substituting (103)–(106) into (101) and collecting terms yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^k\right\|^2\right] \leq{}& \left(1+\frac{\eta_z\mu_g}{3}\right)^2(1-\eta_z\mu_g)^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\
&+\frac{3\eta_z^2L_g^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\mathbf{z}^k-\bar{\mathbf{z}}^k\right\|^2\right]\\
&+\frac{6\eta_z^2L_2^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right)\\
&+6\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+\left(\eta_z^2+\frac{3\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right]+\eta_z^2\frac{\sigma_1^2}{n_h}.
\end{aligned} \tag{107}$$

Plugging (107) into (98) gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right] \leq{}& \left(1+\frac{\eta_z\mu_g}{3}\right)^3(1-\eta_z\mu_g)^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\
&+\frac{3\eta_z^2L_g^2}{n_h}\left(1+\frac{\eta_z\mu_g}{3}\right)\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\mathbf{z}^k-\bar{\mathbf{z}}^k\right\|^2\right]\\
&+\frac{6\eta_z^2L_2^2}{n_h}\left(1+\frac{\eta_z\mu_g}{3}\right)\left(1+\frac{3}{\eta_z\mu_g}\right)\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right)\\
&+6\eta_z^2L_2^2\left(1+\frac{\eta_z\mu_g}{3}\right)\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]\\
&+\left(1+\frac{\eta_z\mu_g}{3}\right)\left(\eta_z^2+\frac{3\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right]+\left(1+\frac{\eta_z\mu_g}{3}\right)\eta_z^2\frac{\sigma_1^2}{n_h}.
\end{aligned} \tag{108}$$

Since $\eta_z\mu_g \leq 1$, simplifying the constants yields

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right] \leq (1-\eta_z\mu_g)\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] + \frac{6\eta_z^2L_g^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\mathbf{z}^k-\bar{\mathbf{z}}^k\right\|^2\right] \\ + \frac{12\eta_z^2L_2^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right) \\ + 12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] \\ + \left(2\eta_z^2+\frac{6\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] + 2\eta_z^2\frac{\sigma_1^2}{n_h} + \left(\eta_x^2+\frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right].\end{aligned} \tag{109}$$

Hence, substituting the definition of $\mathbf{\Delta}^k$, we arrive at

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right] \leq (1-\eta_z\mu_g)\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] + 12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] \\ + \frac{12L_g^2\eta_z^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbf{\Delta}^k + \left(2\eta_z^2+\frac{6\eta_z}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] \\ + 2\eta_z^2\frac{\sigma_1^2}{n_h} + \left(\eta_x^2+\frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{D}_x^k\right\|^2\right].\end{aligned} \tag{110}$$

This completes the proof. □

D.3.5 Proof of Lemma 7

Lemma 7. *Suppose that Assumptions 1 and 2, 4, 5 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\right\|^2\right] \leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right] + \frac{3L_g^4\eta_y^2\lambda^{3/2}}{(1-\lambda)L_2^2}\mathbf{\Delta}^k + n_h\lambda^{3/2}\eta_y^2\sigma^2 \\ + \frac{3n_hL_g^2\eta_y^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + \frac{8\kappa n_h\eta_y^2}{1-\lambda}(\zeta_g^2+2\sigma^2) + \frac{3n_h\eta_y^2\lambda^{3/2}\zeta_g^2}{(1-\lambda)}.\end{aligned}$$

Proof. By Young's inequality $\mathbb{E}\left[\|a+b\|^2\right] \leq \frac{1}{c}\mathbb{E}\left[\|a\|^2\right] + \frac{1}{1-c}\mathbb{E}\left[\|b\|^2\right]$ for $c \in (0,1)$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\right\|^2\right] \\ &= \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^k-\bar{\mathbf{y}}^k-\eta_y\mathbf{D}_y^k\right)\right\|^2\right] + \frac{1}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_y^k-\bar{\boldsymbol{\xi}}_y^k\right\|^2\right] \\ &= \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^k-\bar{\mathbf{y}}^k-\eta_y\mathbf{D}_y^k\right)\right\|^2\right] + \frac{4}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_y^k\right\|^2\right].\end{aligned} \tag{111}$$

For the first term on the right-hand side of (111), applying Cauchy-Schwarz inequalityand Assumption 4 gives

$$\begin{aligned}&\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^k-\bar{\mathbf{y}}^k-\eta_y\mathbf{D}_y^k\right)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^k-\bar{\mathbf{y}}^k-\eta_y\mathbb{E}(\mathbf{D}_y^k)\right)\right\|^2\right] + \lambda^2\eta_y^2\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)-\mathbf{D}_y^k\right\|^2\right] \\ &\leq \frac{1}{u}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^k-\bar{\mathbf{y}}^k\right)\right\|^2\right] + \frac{\eta_y^2}{1-u}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + n_h\lambda^2\eta_y^2\sigma^2 \\ &\leq \frac{\lambda^2}{u}\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right] + \frac{\eta_y^2\lambda^2}{1-u}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + n_h\lambda^2\eta_y^2\sigma^2.\end{aligned} \tag{112}$$

Substituting (112) into (111) and applying Lemma 3 yields

$$\mathbb{E}\left[\left\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\right\|^2\right] \leq \left(\frac{\lambda^2}{uv}+\frac{8\kappa}{1-v}\right)\mathbb{E}\left[\left\|\mathbf{y}^{k}-\bar{\mathbf{y}}^{k}\right\|^2\right]+\frac{\eta_y^2\lambda^2}{(1-u)v}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)\right\|^2\right]+\frac{n_h\lambda^2\eta_y^2\sigma^2}{v} + \frac{8\kappa n_h\eta_y^2}{1-v}(\zeta_g^2+2\sigma^2). \tag{113}$$

Setting $v=u=\sqrt{\lambda}$, we obtain

$$\mathbb{E}\left[\left\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\right\|^2\right] \leq \left(\lambda+\frac{8\kappa}{1-\sqrt{\lambda}}\right)\mathbb{E}\left[\left\|\mathbf{y}^{k}-\bar{\mathbf{y}}^{k}\right\|^2\right]+\frac{\eta_y^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)\right\|^2\right]+n_h\lambda^{3/2}\eta_y^2\sigma^2 + \frac{8\kappa n_h\eta_y^2}{1-\lambda}(\zeta_g^2+2\sigma^2). \tag{114}$$

To bound the second term on the right-hand side of (114), we use the lower-level optimality condition $\nabla_y g(\bar{x}^k,\bar{y}_\star^k)=0$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)\right\|^2\right]\\
&\leq 3\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y g_i(x_i^k,y_i^k)-\nabla_y g_i(\bar{x}^k,\bar{y}^k)\right\|^2\right]+3n_h\mathbb{E}\left[\left\|\nabla_y g_i(\bar{x}^k,\bar{y}^k)-\nabla_y g(\bar{x}^k,\bar{y}_\star^k)\right\|^2\right]\\
&+3\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y g_i(\bar{x}^k,\bar{y}^k)-\nabla_y g(\bar{x}^k,\bar{y}^k)\right\|^2\right]\\
&\leq 3L_g^2\sum_{i\in\mathcal{H}}\left(\mathbb{E}\left[\left\|x_i^k-\bar{x}^k\right\|^2\right]+\mathbb{E}\left[\left\|y_i^k-\bar{y}^k\right\|^2\right]\right)+3n_h\zeta_g^2+3n_hL_g^2\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right].
\end{aligned} \tag{115}$$

Plugging (115) into (114) yields

$$\mathbb{E}\left[\left\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\right\|^2\right] \leq \left(\lambda+\frac{8\kappa}{1-\sqrt{\lambda}}\right)\mathbb{E}\left[\left\|\mathbf{y}^{k}-\bar{\mathbf{y}}^{k}\right\|^2\right]+\frac{3L_g^2\eta_y^2\lambda^{3/2}}{(1-\lambda)}\left(\mathbb{E}\left[\left\|\mathbf{x}^{k}-\bar{\mathbf{x}}^{k}\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^{k}-\bar{\mathbf{y}}^{k}\right\|^2\right]\right) + n_h\lambda^{3/2}\eta_y^2\sigma^2+\frac{3n_hL_g^2\eta_y^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+\frac{8\kappa n_h\eta_y^2}{1-\lambda}(\zeta_g^2+2\sigma^2)+\frac{3n_h\eta_y^2\lambda^{3/2}\zeta_g^2}{(1-\lambda)}. \tag{116}$$

Under the condition $\kappa\leq(1-\lambda)(1-\sqrt{\lambda})/2$ and substituting the definition of $\mathbf{\Delta}^k$, we obtain

$$\mathbb{E}\left[\left\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\right\|^2\right] \leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{y}^{k}-\bar{\mathbf{y}}^{k}\right\|^2\right]+\frac{3L_g^4\eta_y^2\lambda^{3/2}}{(1-\lambda)L_2^2}\mathbf{\Delta}^k+n_h\lambda^{3/2}\eta_y^2\sigma^2 + \frac{3n_hL_g^2\eta_y^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+\frac{8\kappa n_h\eta_y^2}{1-\lambda}(\zeta_g^2+2\sigma^2)+\frac{3n_h\eta_y^2\lambda^{3/2}\zeta_g^2}{(1-\lambda)}. \tag{117}$$

□

D.3.6 Proof of Lemma 8

Lemma 8. *Suppose that Assumptions 1 and 2, 4, 5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}\right\|^2\right]\\
&\leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k}-\bar{\mathbf{x}}^{k}\right\|^2\right]+\frac{8\kappa L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{2\eta_x^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right]+\frac{6\eta_x^2\lambda^{3/2}L_g^2}{1-\lambda}\mathbf{\Delta}^k\\
&+\frac{6\eta_x^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\rho^2\zeta_g^2\right)+n_h\lambda^{3/2}\eta_x^2\sigma_1^2+\frac{8\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2).
\end{aligned}$$

Proof. By Young's inequality $\mathbb{E}\left[\|a+b\|^2\right] \leq \frac{1}{c}\mathbb{E}\left[\|a\|^2\right] + \frac{1}{1-c}\mathbb{E}\left[\|b\|^2\right]$ for $c \in (0,1)$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\right\|^2\right] &= \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k - \bar{\mathbf{x}}^k - \eta_x \mathbf{D}_x^k\right)\right\|^2\right] + \frac{1}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_x^k - \bar{\boldsymbol{\xi}}_x^k\right\|^2\right] \\ &= \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k - \bar{\mathbf{x}}^k - \eta_x \mathbf{D}_x^k\right)\right\|^2\right] + \frac{4}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_x^k\right\|^2\right].\end{aligned} \tag{118}$$

For the first term on the right-hand side of (118), applying Cauchy-Schwarz inequalityand Assumption 4 gives

$$\begin{aligned}&\mathbb{E}\left[\left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k - \bar{\mathbf{x}}^k - \eta_x \mathbf{D}_x^k\right)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k - \bar{\mathbf{x}}^k - \eta_x \mathbb{E}(\mathbf{D}_x^k)\right)\right\|^2\right] + \lambda^2\eta_x^2\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_x^k) - \mathbf{D}_x^k\right\|^2\right] \\ &\leq \frac{1}{u}\mathbb{E}\left[\left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k - \bar{\mathbf{x}}^k\right)\right\|^2\right] + \frac{\eta_x^2}{1-u}\mathbb{E}\left[\left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] + n_h\lambda^2\eta_x^2\sigma_1^2 \\ &\leq \frac{\lambda^2}{u}\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \frac{\eta_x^2\lambda^2}{1-u}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] + n_h\lambda^2\eta_x^2\sigma_1^2.\end{aligned} \tag{119}$$

Substituting (119) into (118) and applying Lemma 3 yields

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\right\|^2\right] \\ &\leq \left(\frac{\lambda^2}{uv} + \frac{8\kappa}{1-v}\right)\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \frac{8\kappa L_g^2}{(1-v)}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] + \frac{\eta_x^2\lambda^2}{(1-u)v}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] \\ &+ \frac{n_h\lambda^2\eta_x^2\sigma_1^2}{v} + \frac{8\kappa n_h\eta_x^2}{1-v}(3\zeta_f^2 + 3\zeta_g^2\rho^2 + 2\sigma_1^2).\end{aligned} \tag{120}$$

Setting $v = u = \sqrt{\lambda}$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\right\|^2\right] \\ &\leq \left(\lambda + \frac{8\kappa}{1-\sqrt{\lambda}}\right)\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \frac{8\kappa L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] + \frac{\eta_x^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] \\ &+ n_h\lambda^{3/2}\eta_x^2\sigma_1^2 + \frac{8\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2 + 3\zeta_g^2\rho^2 + 2\sigma_1^2).\end{aligned} \tag{121}$$

Applying Cauchy-Schwarz inequalityto separate the bias and variance, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\right\|^2\right] \leq \left(\lambda + \frac{8\kappa}{1-\sqrt{\lambda}}\right)\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \frac{8\kappa L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right] + \frac{2\eta_x^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right] \\ + \frac{2\eta_x^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}\left(\bar{\mathbf{D}}_x^k\right) - \mathbb{E}\left(\mathbf{D}_x^k\right)\right\|^2\right] + n_h\lambda^{3/2}\eta_x^2\sigma_1^2 + \frac{8\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2 + 3\zeta_g^2\rho^2 + 2\sigma_1^2).\end{aligned} \tag{122}$$

For the bias term on the right-hand side of (122), by Assumption 5 we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbb{E}\left(\bar{\mathbf{D}}_x^k\right) - \mathbb{E}\left(\mathbf{D}_x^k\right)\right\|^2\right] \\ &\leq 3\zeta_f^2 n_h + 3\rho^2\zeta_g^2 n_h + 3L_g^2\mathbb{E}\left[\left\|\mathbf{z}^k - \bar{\mathbf{z}}^k\right\|^2\right].\end{aligned} \tag{123}$$

Substituting (123) into (120) and collecting terms yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}\right\|^2\right] \\
&\leq \left(\lambda+\frac{8\kappa}{1-\sqrt{\lambda}}\right)\mathbb{E}\left[\left\|\mathbf{x}^{k}-\bar{\mathbf{x}}^{k}\right\|^2\right]+\frac{8\kappa L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{2\eta_x^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right] \\
&+\frac{6\eta_x^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\rho^2\zeta_g^2\right)+\frac{6\eta_x^2\lambda^{3/2}L_g^2}{1-\lambda}\mathbf{\Delta}^k \\
&+n_h\lambda^{3/2}\eta_x^2\sigma_1^2+\frac{8\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2).
\end{aligned} \tag{124}$$

Under the condition $\kappa \leq \frac{(1-\lambda)(1-\sqrt{\lambda})}{16}$ and substituting the definition of $\mathbf{\Delta}^k$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}\right\|^2\right] \\
&\leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k}-\bar{\mathbf{x}}^{k}\right\|^2\right]+\frac{8\kappa L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{2\eta_x^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right]+\frac{6\eta_x^2\lambda^{3/2}L_g^2}{1-\lambda}\mathbf{\Delta}^k \\
&+\frac{6\eta_x^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\rho^2\zeta_g^2\right)+n_h\lambda^{3/2}\eta_x^2\sigma_1^2+\frac{8\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2).
\end{aligned} \tag{125}$$

This completes the proof. □

D.3.7 Proof of Lemma 9

Lemma 9. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] \leq{}& \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{4\eta_z^2\lambda^{3/2}L_g^2}{1-\lambda}\mathbf{\Delta}^k \\
&+\frac{4n_h\eta_z^2\lambda^{3/2}L_2^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right]+\frac{4n_h\eta_z^2\lambda^{3/2}L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{z}^{k}-\bar{z}_\star^{k}\right\|^2\right] \\
&+n_h\lambda^{3/2}\eta_z^2\sigma_1^2+\frac{8\kappa n_h\eta_z^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2)+\frac{4\eta_z^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\zeta_g^2\rho^2\right).
\end{aligned}$$

Proof. By Young's inequality $\mathbb{E}\left[\|a+b\|^2\right] \leq \frac{1}{c}\mathbb{E}\left[\|a\|^2\right]+\frac{1}{1-c}\mathbb{E}\left[\|b\|^2\right]$ for $c\in(0,1)$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] &= \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^k-\bar{\mathbf{z}}^k-\eta_z\mathbf{D}_z^k\right)\right\|^2\right]+\frac{1}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_z^k-\bar{\boldsymbol{\xi}}_z^k\right\|^2\right] \\
&= \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^k-\bar{\mathbf{z}}^k-\eta_z\mathbf{D}_z^k\right)\right\|^2\right]+\frac{4}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_z^k\right\|^2\right].
\end{aligned} \tag{126}$$

For the first term on the right-hand side of (126), applying Cauchy-Schwarz inequalityand Assumption 4 gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^k-\bar{\mathbf{z}}^k-\eta_z\mathbf{D}_z^k\right)\right\|^2\right] \\
&=\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^k-\bar{\mathbf{z}}^k-\eta_z\mathbb{E}(\mathbf{D}_z^k)\right)\right\|^2\right]+\lambda^2\eta_z^2\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)-\mathbf{D}_z^k\right\|^2\right] \\
&\leq \frac{1}{u}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^k-\bar{\mathbf{z}}^k\right)\right\|^2\right]+\frac{\eta_z^2}{1-u}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]+n_h\lambda^2\eta_z^2\sigma_1^2 \\
&\leq \frac{\lambda^2}{u}\mathbb{E}\left[\left\|\mathbf{z}^k-\bar{\mathbf{z}}^k\right\|^2\right]+\frac{\eta_z^2\lambda^2}{1-u}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]+n_h\lambda^2\eta_z^2\sigma_1^2.
\end{aligned} \tag{127}$$

Substituting (127) into (126) and applying Lemma 3 yields

$$\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] \leq \left(\frac{\lambda^2}{uv}+\frac{8\kappa(1+L_g^2)}{1-v}\right)\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{\eta_z^2\lambda^2}{(1-u)v}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] + \frac{n_h\lambda^2\eta_z^2\sigma_1^2}{v}+\frac{8\kappa n_h\eta_z^2}{1-v}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2). \tag{128}$$

Setting $v=u=\sqrt{\lambda}$, we obtain

$$\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] \leq \left(\lambda+\frac{8\kappa(1+L_g^2)}{1-\sqrt{\lambda}}\right)\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{\eta_z^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] + n_h\lambda^{3/2}\eta_z^2\sigma_1^2+\frac{8\kappa n_h\eta_z^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2). \tag{129}$$

To bound the second term on the right-hand side of (128), we use the lower-level optimality condition $\nabla_{yy}^2 g(\bar{x}^k,\bar{y}_\star^k)\bar{z}_\star^k+\nabla_y f(\bar{x}^k,\bar{y}_\star^k)=0$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] &\leq 4\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_{yy}^2 g_i(x_i^k,y_i^k)z_i^k-\nabla_{yy}^2 g(\bar{x}^k,\bar{y}^k)\bar{z}_\star^k\right\|^2\right]\\
&\quad+4\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y f_i(x_i^k,y_i^k)-\nabla_y f(\bar{x}^k,\bar{y}^k)\right\|^2\right]+4n_h\mathbb{E}\left[\left\|\left(\nabla_{yy}^2 g(\bar{x}^k,\bar{y}^k)-\nabla_{yy}^2 g(\bar{x}^k,\bar{y}_\star^k)\right)\bar{z}_\star^k\right\|^2\right]\\
&\quad+4n_h\mathbb{E}\left[\left\|\nabla_y f(\bar{x}^k,\bar{y}^k)-\nabla_y f(\bar{x}^k,\bar{y}_\star^k)\right\|^2\right]\\
&\leq 4\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_{yy}^2 g_i(x_i^k,y_i^k)z_i^k-\nabla_{yy}^2 g(\bar{x}^k,\bar{y}^k)\bar{z}_\star^k\right\|^2\right]\\
&\quad+4\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y f_i(x_i^k,y_i^k)-\nabla_y f(\bar{x}^k,\bar{y}^k)\right\|^2\right]+4n_h\left(L_f^2+L_g^2\rho^2\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right].
\end{aligned} \tag{130}$$

To handle the first term on the right-hand side of (130), we use the decomposition

$$\begin{aligned}
&\nabla_{yy}^2 g_i(x_i^k,y_i^k)z_i^k-\nabla_{yy}^2 g(\bar{x}^k,\bar{y}^k)\bar{z}_\star^k\\
&=\nabla_{yy}^2 g_i(x_i^k,y_i^k)(z_i^k-\bar{z}^k)+\nabla_{yy}^2 g_i(x_i^k,y_i^k)(\bar{z}^k-\bar{z}_\star^k)\\
&+\left(\nabla_{yy}^2 g_i(x_i^k,y_i^k)-\nabla_{yy}^2 g_i(\bar{x}^k,\bar{y}^k)\right)\bar{z}_\star^k+\left(\nabla_{yy}^2 g_i(\bar{x}^k,\bar{y}^k)-\nabla_{yy}^2 g(\bar{x}^k,\bar{y}^k)\right)\bar{z}_\star^k.
\end{aligned} \tag{131}$$

Taking the norm on both sides and applying Assumption 5 yields

$$\begin{aligned}
&\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_{yy}^2 g_i(x_i^k,y_i^k)z_i^k-\nabla_{yy}^2 g(\bar{x}^k,\bar{y}^k)\bar{z}_\star^k\right\|^2\right]\\
&\leq 4L_g^2\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|z_i^k-\bar{z}^k\right\|^2\right]+4n_hL_g^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\
&+4L_g^2\rho^2\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right)+4n_h\rho^2\zeta_g^2.
\end{aligned} \tag{132}$$

For the second term on the right-hand side of (130), by Assumption 2 we have

$$\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y f_i(x_i^k,y_i^k)-\nabla_y f(\bar{x}^k,\bar{y}^k)\right\|^2\right]\leq 4L_f^2\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right)+4n_h\zeta_f^2. \tag{133}$$

Substituting (130)–(133) into (129) and collecting terms yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] &\leq \left(\lambda+\frac{8\kappa(1+L_g^2)}{1-\sqrt{\lambda}}\right)\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{4\eta_z^2\lambda^{3/2}L_2^2}{1-\lambda}\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right)\\
&+\frac{4n_h\eta_z^2\lambda^{3/2}L_2^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+\frac{4n_h\eta_z^2\lambda^{3/2}L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\
&+n_h\lambda^{3/2}\eta_z^2\sigma_1^2+\frac{8\kappa n_h\eta_z^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2)+\frac{4\eta_z^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\zeta_g^2\rho^2\right).
\end{aligned} \tag{134}$$

Under the condition $\kappa \leq \frac{(1-\lambda)(1-\sqrt{\lambda})}{16(1+L_g^2)}$ and substituting the definition of $\mathbf{\Delta}^k$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] \leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right] + \frac{4\eta_z^2\lambda^{3/2}L_g^2}{1-\lambda}\mathbf{\Delta}^k \\ + \frac{4n_h\eta_z^2\lambda^{3/2}L_2^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + \frac{4n_h\eta_z^2\lambda^{3/2}L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] \\ + n_h\lambda^{3/2}\eta_z^2\sigma_1^2 + \frac{8\kappa n_h\eta_z^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2) + \frac{4\eta_z^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\zeta_g^2\rho^2\right).\end{aligned} \tag{135}$$

This completes the proof. □

D.3.8 Proof of Lemma 10

Lemma 10. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}\mathbf{\Delta}^{k+1} \leq \frac{1+\lambda}{2}\mathbf{\Delta}^k + C_y\frac{n_h\eta_x^2L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] \\ + \frac{2\eta_x^2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right] + n_h\eta_x^2C_{\sigma 1} + \frac{\eta_x^2n_h}{1-\lambda}C_{\zeta 1} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\zeta 2} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\sigma 2},\end{aligned}$$

where the problem-dependent constants $C_y, C_z, C_{\sigma 1}, C_{\zeta 1}, C_{\zeta 2}, C_{\sigma 2}$ are defined in the detailed proof of Lemma 10.

Proof. Combining Lemmas 7, 8, and 9 with the definition of $\mathbf{\Delta}^k$ yields

$$\begin{aligned}\mathbf{\Delta}^{k+1} \leq \left(\lambda + \frac{8\kappa(1+L_g^2)}{1-\sqrt{\lambda}}\right)\mathbf{\Delta}^k + C_\Delta\frac{L_g^2\eta_x^2}{1-\lambda}\mathbf{\Delta}^k + C_y\frac{n_h\eta_x^2L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] \\ + \frac{2\eta_x^2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right] + n_h\eta_x^2C_{\sigma 1} + \frac{\eta_x^2n_h}{1-\lambda}C_{\zeta 1} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\zeta 2} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\sigma 2},\end{aligned} \tag{136}$$

where the constants are defined as

$$\begin{aligned}C_\Delta &:= 3c_y^2\lambda^{3/2} + \frac{6L_2^2\lambda^{3/2}}{L_g^2} + 4c_z^2\lambda^{3/2},\ C_y := \left(3c_y^2\lambda^{3/2}+4c_z^2\lambda^{3/2}\right),\ C_z := 4c_z^2\lambda^{3/2}, \\ C_{\zeta 1} &:= \frac{6L_2^2\lambda^{3/2}}{L_g^2}\left(\zeta_f^2+\zeta_g^2\rho^2\right) + 4c_z^2\lambda^{3/2}\left(\zeta_f^2+\zeta_g^2\rho^2\right) + \frac{3c_y^2L_2^2\lambda^{3/2}\zeta_g^2}{L_g^2}, \\ C_{\sigma 1} &:= \lambda^{3/2}\left(\frac{L_2^2}{L_g^2}\left(\sigma_1^2+c_y^2\sigma^2\right)+c_z^2\sigma_1^2\right), \\ C_{\zeta 2} &:= \left(\frac{L_2^2}{L_g^2}(3\zeta_f^2+3\zeta_g^2\rho^2) + \frac{c_y^2L_2^2}{L_g^2}\zeta_g^2 + c_z^2(3\zeta_f^2+3\zeta_g^2\rho^2)\right), \\ C_{\sigma 2} &:= \left(\frac{2L_2^2}{L_g^2}\sigma_1^2 + \frac{c_y^2L_2^2}{L_g^2}\sigma^2 + 2c_z^2\sigma_1^2\right).\end{aligned} \tag{137}$$

Under the conditions $\kappa \leq \frac{(1-\lambda)(1-\sqrt{\lambda})}{32(1+L_g^2)}$ and $\eta_x \leq \frac{1-\lambda}{2L_g\sqrt{C_\Delta}}$, simplifying gives

$$\begin{aligned}\mathbf{\Delta}^{k+1} \leq \frac{1+\lambda}{2}\mathbf{\Delta}^k + C_y\frac{n_h\eta_x^2L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right] \\ + \frac{2\eta_x^2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\mathbb{E}(\bar{\mathbf{D}}_x^k)\right\|^2\right] + n_h\eta_x^2C_{\sigma 1} + \frac{\eta_x^2n_h}{1-\lambda}C_{\zeta 1} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\zeta 2} + \frac{8\kappa\eta_x^2n_h}{1-\lambda}C_{\sigma 2}.\end{aligned} \tag{138}$$

This finishes the proof. □

E Convergence Analysis: Proof of Theorem 2

E.1 Proof Outline

The proof follows the same structure as that of Theorem 1, with one additional ingredient: each local update replaces the stochastic gradient $D_{u,i}^k$ with a momentum buffer $m_{u,i}^k$ for $u \in \{x, y, z\}$, defined via the recursion

$$m_{u,i}^k = \beta_u m_{u,i}^{k-1} + (1-\beta_u) D_{u,i}^k, \qquad u \in \{x, y, z\}, \tag{139}$$

and the local steps $x_i^{k+\frac{1}{2}} = x_i^k - \eta_x m_{x,i}^k$, $y_i^{k+\frac{1}{2}} = y_i^k - \eta_y m_{y,i}^k$, $z_i^{k+\frac{1}{2}} = z_i^k - \eta_z m_{z,i}^k$ are then passed through the robust aggregator $\mathcal{A}(\cdot)$, exactly as in Algorithm 1. The honest-only reference aggregates $\bar{x}_i^{k+\frac{1}{2}}, \bar{y}_i^{k+\frac{1}{2}}, \bar{z}_i^{k+\frac{1}{2}}$ and the normalized aggregation errors $\boldsymbol{\xi}_x^k, \boldsymbol{\xi}_y^k, \boldsymbol{\xi}_z^k$ are defined exactly as in the proof of Theorem 1, and will be reused without restating.

Because the momentum buffer tracks a weighted average of past stochastic directions rather than the current gradient, the Lyapunov argument must additionally control how much $\mathbf{m}_u^k$ deviates from $\mathbb{E}(\mathbf{D}_u^k)$. We therefore extend the four quantities bounded in the proof of Theorem 1 (aggregation error, descent, lower-level tracking, consensus error) with a fifth quantity, the *momentum drift* $\|\mathbf{m}_u^k - \mathbb{E}(\mathbf{D}_u^k)\|^2$. To keep the outline compact, all proofs are deferred to the corresponding appendices.

Aggregation error. Following the same argument as in Lemma 3, the aggregation error now inherits the momentum averaging, which damps the variance contribution by a factor $(1-\beta_u)/(1+\beta_u)$ for each $u \in \{x, y, z\}$.

Lemma 12. *Suppose that Assumptions 1 and 2, 4, 5 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|R_{x,i}^{k+\frac{1}{2}} - \bar{x}_i^{k+\frac{1}{2}}\right\|^2\right] &\le 2\kappa \sum_{j\in\mathcal{H}_i} w_{ij}\left(\mathbb{E}\left[\left\|x_j^k - \bar{x}_i^k\right\|^2\right] + L_g^2 \mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right]\right) \\
&\quad + \kappa\eta_x^2\left(18\zeta_f^2 + 18\rho^2\zeta_g^2 + 12\frac{1-\beta_x}{1+\beta_x}\sigma_1^2\right), \\
\mathbb{E}\left[\left\|R_{y,i}^{k+\frac{1}{2}} - \bar{y}_i^{k+\frac{1}{2}}\right\|^2\right] &\le 2\kappa \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|y_j^k - \bar{y}_i^k\right\|^2\right] + \kappa\eta_y^2\left(6\zeta_g^2 + 12\frac{1-\beta_y}{1+\beta_y}\sigma^2\right), \\
\mathbb{E}\left[\left\|R_{z,i}^{k+\frac{1}{2}} - \bar{z}_i^{k+\frac{1}{2}}\right\|^2\right] &\le 2(1+L_g^2)\kappa \sum_{j\in\mathcal{H}_i} w_{ij}\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right] \\
&\quad + \kappa\eta_z^2\left(18\zeta_f^2 + 18\rho^2\zeta_g^2 + 12\frac{1-\beta_z}{1+\beta_z}\sigma_1^2\right).
\end{aligned}$$

Proof: see Appendix E.3.1.

Descent lemma. Averaging the x-update across honest agents and invoking $L_{\nabla\Phi}$-smoothness as in Lemma 4, the one-step descent inequality now contains one additional term involving the momentum drift $\|\bar{m}_x^k - \mathbb{E}(\bar{D}_x^k)\|^2$; the other terms are of the same form as in Lemma 4.

Lemma 13. *Suppose that Assumptions 1 and 2, 3, 4 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] &\le \mathbb{E}\left[\Phi(\bar{x}^k)\right] - \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \left(\frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2}\right)\mathbb{E}\left[\left\|\mathbb{E}(\bar{m}_x^k)\right\|^2\right] \\
&\quad + \left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{6L_g^2\eta_x}{n_h}\boldsymbol{\Delta}^k + \eta_x\mathbb{E}\left[\left\|\bar{m}_x^k - \mathbb{E}(\bar{D}_x^k)\right\|^2\right] \\
&\quad + 6L_2^2\eta_x\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 6L_g^2\eta_x\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right].
\end{aligned}$$

Proof: see Appendix E.3.2.

Lower-level convergence. The y- and z-tracking recursions retain the same linear contraction $(1-\eta_y\mu_g)$, $(1-\eta_z\mu_g)$ as in Lemmas 5 and 6; each of them now carries an additional momentum-drift perturbation in the appropriate variable.

Lemma 14. *Suppose that Assumptions 2–5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{y}^{k+1}-\bar{y}_\star^{k+1}\right\|^2\right]\\
&\le (1-\eta_y\mu_g)\,\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right]+\left(\eta_x^2+\frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2\mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right]+4(1+\frac{3}{\eta_y\mu_g})\frac{\eta_y^2L_g^4}{n_hL_2^2}\mathbf{\Delta}^k\\
&+4(1+\frac{3}{\eta_y\mu_g})\eta_y^2\mathbb{E}\left[\left\|\bar{m}_y^k-\mathbb{E}(\bar{D}_y^k)\right\|^2\right]+2\left(\eta_y^2+\frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right].
\end{aligned}$$

Proof: see Appendix E.3.3.

Lemma 15. *Suppose that Assumptions 2–5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right]\\
&\le(1-\eta_z\mu_g)\mathbb{E}\left[\left\|\bar{z}^{k}-\bar{z}_\star^{k}\right\|^2\right]+\frac{12\eta_z^2L_g^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbf{\Delta}^k+12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right]\\
&+\left(2\eta_z^2+\frac{6\eta_z}{\mu_g}\right)\left(\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right]+\mathbb{E}\left[\left\|\bar{m}_z^k-\mathbb{E}(\bar{D}_z^k)\right\|^2\right]\right)+\left(\eta_x^2+\frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right].
\end{aligned}$$

Proof: see Appendix E.3.4.

Consensus error. The consensus error recursions for x, y, and z are obtained exactly as in the proof of Theorem 1, with the stochastic gradient $\mathbf{D}_u^k$ replaced by the momentum buffer $\mathbf{m}_u^k$. The resulting per-variable bounds differ from Lemmas 7, 8, and 9 only through this substitution and an extra momentum-drift term.

Lemma 16. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{y}^{k+1}-\bar{\mathbf{y}}^{k+1}\right\|^2\right]&\le\frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{y}^{k}-\bar{\mathbf{y}}^{k}\right\|^2\right]+\frac{3L_g^4\eta_y^2\lambda^{3/2}}{(1-\lambda)L_2^2}\mathbf{\Delta}^k+\frac{4\eta_y^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)-\mathbf{m}_y^k\right\|^2\right]\\
&+\frac{3n_hL_g^2\eta_y^2\lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right]+\frac{48\kappa n_h\eta_y^2}{1-\lambda}(\zeta_g^2+2\sigma^2\frac{1-\beta_y}{1+\beta_y})+\frac{3n_h\eta_y^2\lambda^{3/2}\zeta_g^2}{(1-\lambda)}.
\end{aligned}$$

Proof: see Appendix E.4.

Lemma 17. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right]&\le\frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{4\eta_z^2\lambda^{3/2}L_g^2}{1-\lambda}\mathbf{\Delta}^k+\frac{4\eta_z^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{m}_z^k-\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]\\
&+\frac{4n_h\eta_z^2\lambda^{3/2}L_2^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right]+\frac{4n_h\eta_z^2\lambda^{3/2}L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{z}^{k}-\bar{z}_\star^{k}\right\|^2\right]\\
&+\frac{48\kappa n_h\eta_z^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2\frac{1-\beta_z}{1+\beta_z})+\frac{4\eta_z^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\zeta_g^2\rho^2\right).
\end{aligned}$$

Proof: see Appendix E.5.

Lemma 18. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}\right\|^2\right]\\
&\le\frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k}-\bar{\mathbf{x}}^{k}\right\|^2\right]+\frac{16\kappa L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{\eta_x^2\lambda^{3/2}}{1-\lambda}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^k\right\|^2\right]+\frac{4\eta_x^2L_g^2}{1-\lambda}\mathbf{\Delta}^k\\
&+\frac{24\eta_x^2n_h\sigma_1^2}{1-\lambda}\cdot\frac{1-\beta_x}{1+\beta_x}+\frac{36\eta_x^2n_h}{1-\lambda}\left(\zeta_f^2+\rho^2\zeta_g^2\right)+\frac{48\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2\frac{1-\beta_x}{1+\beta_x}).
\end{aligned}$$

Proof: see Appendix E.6.

Combining the three bounds above with the aggregate $\mathbf{\Delta}^k$ introduced in the notation section yields the unified consensus recursion.

Lemma 19. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}\boldsymbol{\Delta}^{k+1} \le{}& \frac{1+\lambda}{2}\boldsymbol{\Delta}^k + C_y\frac{n_h\eta_x^2L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\ &+ \frac{2\eta_x^2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^k\right\|^2\right] + \frac{24n_h\eta_x^2\sigma_1^2}{1-\lambda}\cdot\frac{1-\beta_x}{1+\beta_x}\\ &+ \frac{\eta_x^2n_h}{1-\lambda}C_{\zeta 1} + \frac{48\kappa\eta_x^2n_h}{1-\lambda}C_{\zeta 2} + \frac{48\kappa\eta_x^2n_h}{1-\lambda}\frac{1-\beta_x}{1+\beta_x}C_{\sigma 2},\end{aligned}$$

where the problem-dependent constants $C_y, C_z, C_{\zeta 1}, C_{\zeta 2}, C_{\sigma 2}$ are the same as in Lemma 10.

Proof: see Appendix E.7.

Momentum drift. This is the only ingredient not present in the proof of Theorem 1, and it is needed to close the Lyapunov recursion. Since the momentum buffer averages past stochastic directions, its deviation $\|\mathbf{m}_u^k - \mathbb{E}(\mathbf{D}_u^k)\|^2$ contracts at the rate β_u^2 up to variance, aggregation, and coupling terms. We state the bounds for y, x, and z below.

Lemma 20. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{m}_y^{k+1} - \mathbb{E}(\mathbf{D}_y^{k+1})\|^2\right]\\ &\le \beta_y^2(1+L_g\eta_y)(1+3\eta_yL_g)\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + (1-\beta_y)^2n_h\sigma^2\\ &\quad+ \frac{2\beta_y^2}{c_y^2}\eta_y(1+L_g\eta_y)L_g\left(\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right] + \mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]\right)\\ &\quad+ \beta_y^2\eta_y(1+L_g\eta_y)L_g\left(9n_hL_g^2\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + \frac{9L_g^4}{L_2^2}\boldsymbol{\Delta}^k + 9n_h\zeta_g^2 + 3\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right]\right).\end{aligned}$$

Proof: see Appendix E.8.

Lemma 21. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{m}_x^{k+1} - \mathbb{E}(\mathbf{D}_x^{k+1})\right\|^2\right]\\ &\le \beta_x^2\left(1+L_g\eta_z\right)\mathbb{E}\left[\left\|\mathbf{m}_x^k - \mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] + (1-\beta_x)^2n_h\sigma_1^2 + 16L_1^2\eta_x^2\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right]\\ &\quad+ 24L_1^2\eta_y^2\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + 24L_g^2\eta_z^2\mathbb{E}\left[\left\|\mathbf{m}_z^k - \mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]\\ &\quad+ 24n_h\left(3L_1^2L_g^2\eta_y^2 + 16L_g^2L_2^2\eta_z^2\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right] + 384n_hL_g^4\eta_z^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\ &\quad+ 24L_1^2\eta_y^2\left(\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right] + \frac{3L_g^4}{L_2^2}\boldsymbol{\Delta}^k + 3n_h\zeta_g^2\right)\\ &\quad+ 24L_g^2\eta_z^2\left(\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right] + 16L_g^2\boldsymbol{\Delta}^k + 16n_h(\zeta_f^2+\rho^2\zeta_g^2)\right)\\ &\quad+ 32L_1^2\eta_x^2\left(\mathbb{E}\left[\left\|\xi_x^k\right\|^2\right] + L_g^2\boldsymbol{\Delta}^k + 6\frac{1-\beta_x}{1+\beta_x}n_h\sigma_1^2 + 9n_h(\zeta_f^2+\rho^2\zeta_g^2)\right).\end{aligned}$$

Proof: see Appendix E.9.

Lemma 22. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{m}_z^{k+1}-\mathbb{E}(\mathbf{D}_z^{k+1})\right\|^2\right]\\
&\le \beta_z^2\left(1+L_g\eta_z\right)\left(1+12L_g\eta_z\right)\mathbb{E}\left[\left\|\mathbf{m}_z^{k}-\mathbb{E}(\mathbf{D}_z^{k})\right\|^2\right]+(1-\beta_z)^2 n_h\sigma_1^2+16L_1^2\eta_x^2\mathbb{E}\left[\left\|\mathbf{m}_x^{k}\right\|^2\right]\\
&+24L_1^2\eta_y^2\mathbb{E}\left[\left\|\mathbf{m}_y^{k}-\mathbb{E}(\mathbf{D}_y^{k})\right\|^2\right]\\
&+24n_h\left(3L_1^2L_g^2\eta_y^2+16L_g^2L_2^2\eta_z^2\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+384n_hL_g^4\eta_z^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\
&+24L_1^2\eta_y^2\left(\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right]+\frac{3L_g^4}{L_2^2}\mathbf{\Delta}^k+3n_h\zeta_g^2\right)\\
&+24L_g^2\eta_z^2\left(\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right]+16L_g^2\mathbf{\Delta}^k+16n_h(\zeta_f^2+\rho^2\zeta_g^2)\right)\\
&+32L_1^2\eta_x^2\left(\mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]+L_g^2\mathbf{\Delta}^k+6\frac{1-\beta_x}{1+\beta_x}n_h\sigma_1^2+9n_h(\zeta_f^2+\rho^2\zeta_g^2)\right).
\end{aligned}
$$

Proof: see Appendix E.10.

E.2 Proof of Theorem 2

Theorem 3 (Restatement of Theorem 2)**.** *Under the same conditions as Theorem 1, by choosing the momentum coefficients such that $1-\beta_v=\Theta(\eta_v L_{\beta_v})$ for all $v\in\{x,y,z\}$, where $L_{\beta_x},L_{\beta_y},L_{\beta_z}>0$ are problem-dependent constants, the iterates generated by Algorithm 1 satisfy:*

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla\Phi_{\mathcal{H}}(\bar{x}^k)\|^2\right]&\le\mathcal{O}\left(\frac{1}{\sqrt{n_hK}}+\frac{n_h^{3/2}\sigma^2}{K^{3/2}(1-\lambda)}+\frac{n_h\left(\zeta_f^2+\zeta_g^2\right)}{K(1-\lambda)^2}\right)\\
&\quad+\mathcal{O}\left(\frac{\kappa\, n_h^{3/2}\sigma^2}{K^{3/2}(1-\lambda)^2}+\frac{\kappa\, n_h\left(\zeta_f^2+\zeta_g^2\right)}{K(1-\lambda)^2}\right)+\mathcal{O}\left(\kappa\left(\zeta_f^2+\zeta_g^2\right)\right).
\end{aligned}
\tag{140}
$$

Proof. We define the Lyapunov function $\mathcal{L}^k$ as follows

$$
\begin{aligned}
\mathcal{L}^k:=\Phi(\bar{x}^k)&+B_1\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2+B_2\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2+B_\Delta\frac{\mathbf{\Delta}^k}{n_h}\\
&+\frac{B_3}{n_h}\left\|\mathbf{m}_x^k-\mathbb{E}(\mathbf{D}_x^k)\right\|^2+\frac{B_4}{n_h}\left\|\mathbf{m}_y^k-\mathbb{E}(\mathbf{D}_y^k)\right\|^2+\frac{B_5}{n_h}\left\|\mathbf{m}_z^k-\mathbb{E}(\mathbf{D}_z^k)\right\|^2,
\end{aligned}
\tag{141}
$$

where the Lyapunov weights are chosen as

$$
\begin{aligned}
&B_1:=\frac{5L_2^2}{c_y\mu_g}+\frac{240L_2^2L_g^2}{c_y\mu_g^3},\quad B_2:=\frac{5L_g^2}{c_z\mu_g},\quad B_\Delta:=B_3^\circ\eta_x+\frac{\kappa B_4^\circ}{\eta_x},\\
&B_3^\circ:=\frac{(8+4B_1C_{y^\star}+4B_2C_{z^\star})L_g^2}{1-\lambda},\quad B_4^\circ:=\frac{2}{1-\lambda}\left((1+L_g^2)+\frac{16B_1}{c_y\mu_g}+\frac{16B_2(1+L_g^2)}{c_z\mu_g}\right),\\
&B_3:=\frac{1}{L_g},\qquad B_4:=\frac{1}{L_g},\qquad B_5:=\frac{1}{L_g}.
\end{aligned}
\tag{142}
$$

The weights B_1,B_2,B_Δ coincide with A_1,A_2,A_Δ in (35), and the momentum weights $B_3,B_4,B_5=\Theta(1/L_g)$ are chosen so that the coefficients T_{mx},T_{my},T_{mz} defined below admit an annihilation by β_x,β_y,β_z.

Combining the descent inequality (Lemma 13), the lower-level tracking bounds (Lemmas 14 and 15), the unified consensus recursion (Lemma 19), the aggregation-error bound (Lemma 12), and the three momentum-drift recursions (Lemmas 21, 20, and 22), we obtain the recursion below. Whenever

an averaged drift $\left\|\bar{m}_u^k - \mathbb{E}(\bar{D}_u^k)\right\|^2$ appears in the descent or tracking lemmas, we upper-bound it by $\frac{1}{n_h}\mathbb{E}\left[\left\|\mathbf{m}_u^k - \mathbb{E}(\mathbf{D}_u^k)\right\|^2\right]$ via Jensen's inequality so that all drift contributions match the scaling of

the Lyapunov function.

$$
\begin{aligned}
&\mathcal{L}^{k+1} - \mathcal{L}^k \\
&\leq -\frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \\
&+ \underbrace{\left(6L_2^2\eta_x - B_1\eta_y\mu_g + 12B_2\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right) + B_\Delta C_y\frac{\eta_x^2L_2^2}{1-\lambda}\right.}_{T_1(1)} \\
&\quad \underbrace{\left. +9B_4L_g^3\eta_y + (B_3+B_5)\big(72L_1^2L_g^2\eta_y^2 + 384L_g^2L_2^2\eta_z^2\big)\right)}_{T_1(2)}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\
&+ \underbrace{\left(6L_g^2\eta_x - B_2\eta_z\mu_g + B_\Delta C_z\frac{\eta_x^2L_g^2}{1-\lambda} + 384(B_3+B_5)L_g^4\eta_z^2\right)}_{T_2}\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] \\
&+ \underbrace{\left(\frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2} + B_1\left(\eta_x^2 + \frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2 + B_2\left(\eta_x^2 + \frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\right.}_{T_3(1)} \\
&\quad \underbrace{\left. +\frac{2B_\Delta\eta_x^2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\right)}_{T_3(2)}\mathbb{E}\left[\left\|\mathbb{E}(\bar{m}_x^k)\right\|^2\right] \\
&+ \underbrace{\left(\frac{2B_4\beta_y^2L_g\eta_y(1+L_g\eta_y)}{c_y^2} + 16(B_3+B_5)L_1^2\eta_x^2\right)}_{T_3'}\frac{\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right]}{n_h} \\
&+ \underbrace{\left(6L_g^2\eta_x + 4B_1\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right)\frac{L_g^4}{L_2^2} + 12B_2\left(\eta_z^2 + \frac{3\eta_z}{\mu_g}\right)L_g^2 - \frac{(1-\lambda)B_\Delta}{2} + \frac{9B_4L_g^5\eta_y}{L_2^2}\right.}_{T_4(1)} \\
&\quad \underbrace{\left. +(B_3+B_5)\left(\frac{72L_1^2L_g^4\eta_y^2}{L_2^2} + 384L_g^4\eta_z^2 + 288L_1^2L_g^2\eta_x^2\right)\right)}_{T_4(2)}\frac{\boldsymbol{\Delta}^k}{n_h} \\
&+ \underbrace{\big(\beta_x^2(1+L_g\eta_z) - 1 + \eta_x\big)}_{T_{mx}}\frac{B_3\mathbb{E}\left[\left\|\mathbf{m}_x^k - \mathbb{E}(\mathbf{D}_x^k)\right\|^2\right]}{n_h} \\
&+ \underbrace{\left(\begin{array}{l}\beta_y^2(1+L_g\eta_y)(1+3L_g\eta_y) - 1 \\ \quad + \dfrac{4B_1}{B_4}\left(\eta_y^2 + \dfrac{3\eta_y}{\mu_g}\right) + \dfrac{24(B_3+B_5)L_1^2\eta_y^2}{B_4}\end{array}\right)}_{T_{my}}\frac{B_4\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right]}{n_h} \\
&+ \underbrace{\left(\beta_z^2(1+L_g\eta_z)(1+12L_g\eta_z) - 1 + \frac{2B_2}{B_5}\left(\eta_z^2 + \frac{3\eta_z}{\mu_g}\right) + \frac{24B_3L_g^2\eta_z^2}{B_5}\right)}_{T_{mz}}\frac{B_5\mathbb{E}\left[\left\|\mathbf{m}_z^k - \mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]}{n_h} \\
&+ R_k,
\end{aligned}
$$

where the residual R_k collects all aggregation-error, gradient-heterogeneity, and stochastic-variance contributions:

$$\begin{aligned}
R_k := &\left(L_{\nabla\Phi}\eta_x^2 + \eta_x + \frac{32(B_3+B_5)L_1^2\eta_x^2}{n_h} + \frac{2B_4\beta_y^2 L_g\eta_y(1+L_g\eta_y)}{c_y^2 n_h}\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] \\
&+ \left(2B_1\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right) + \frac{3B_4\beta_y^2 L_g\eta_y(1+L_g\eta_y)}{n_h} + \frac{24(B_3+B_5)L_1^2\eta_y^2}{n_h}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right] \\
&+ \left(2B_2\left(\eta_z^2 + \frac{3\eta_z}{\mu_g}\right) + \frac{24(B_3+B_5)L_g^2\eta_z^2}{n_h}\right)\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] \\
&+ B_\Delta\left(\frac{\eta_x^2}{1-\lambda}\left(C_{\zeta 1} + C_{\sigma 1}\frac{1-\beta_x}{1+\beta_x}\right) + \frac{48\kappa\eta_x^2}{1-\lambda}\left(C_{\zeta 2} + C_{\sigma 2}\frac{1-\beta_x}{1+\beta_x}\right)\right) \\
&+ B_4(1-\beta_y)^2\sigma^2 + (B_3+B_5)(1-\beta_x)^2\sigma_1^2 \\
&+ \left(9B_4\beta_y^2 L_g\eta_y(1+L_g\eta_y) + 24(B_3+B_5)L_1^2\eta_y^2\right)\zeta_g^2 \\
&+ 16(B_3+B_5)\left(L_g^2\eta_z^2 + 2L_1^2\eta_x^2\right)\left(\zeta_f^2 + \rho^2\zeta_g^2\right) \\
&+ 192(B_3+B_5)L_1^2\eta_x^2\frac{1-\beta_x}{1+\beta_x}\sigma_1^2.
\end{aligned} \tag{143}$$

To obtain a monotone descent in $\mathcal{L}^k$, we choose the step sizes and momentum coefficients so that every bracketed coefficient in the Lyapunov recursion is non-positive. We first dispose of the three momentum-correction terms by selecting $\beta_x, \beta_y, \beta_z$ so that their coefficients vanish, after which the remaining coefficients T_1, T_2, T_3, T_4 can be handled in the same way as in the proof of Theorem 1.

Eliminating the x-momentum correction. Choosing β_x so that $1-\beta_x^2 = L_{\beta_x}\eta_x$ with $L_{\beta_x} := 1 + L_g c_z$ yields

$$T_{mx} = \beta_x^2(1+L_g\eta_z) - 1 + \eta_x = \beta_x^2 - 1 + \eta_x(1+L_g c_z) = 0. \tag{144}$$

Eliminating the y-momentum correction. Under the step-size condition $\eta_y \le \frac{1}{8L_g}$, we absorb the quadratic terms via $\eta_y^2 \le \frac{\eta_y}{8L_g}$, so that

$$(1+L_g\eta_y)(1+3L_g\eta_y) \le 1+4L_g\eta_y+\tfrac{3}{8}\eta_y \le 1+5L_g\eta_y, \quad \frac{4B_1}{B_4}\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right) \le \frac{4B_1\eta_y}{B_4}\left(\frac{1}{8L_g} + \frac{3}{\mu_g}\right),$$

and $\frac{24(B_3+B_5)L_1^2\eta_y^2}{B_4} \le \frac{3(B_3+B_5)L_1^2}{B_4 L_g}\eta_y$. Choosing β_y so that $1-\beta_y^2 = L_{\beta_y}\eta_y$ with $L_{\beta_y} := 5L_g + \frac{4B_1}{B_4}\left(\frac{1}{8L_g} + \frac{3}{\mu_g}\right) + \frac{3(B_3+B_5)L_1^2}{B_4 L_g}$ yields

$$\begin{aligned}
T_{my} &= \beta_y^2(1+L_g\eta_y)(1+3L_g\eta_y) - 1 + \frac{4B_1}{B_4}\left(\eta_y^2 + \frac{3\eta_y}{\mu_g}\right) + \frac{24(B_3+B_5)L_1^2\eta_y^2}{B_4} \\
&\le \beta_y^2 - 1 + L_{\beta_y}\eta_y = 0.
\end{aligned} \tag{145}$$

Eliminating the z-momentum correction. Similarly, under $\eta_z \le \frac{1}{12L_g}$ we have $\eta_z^2 \le \frac{\eta_z}{12L_g}$ and $(1+L_g\eta_z)(1+12L_g\eta_z) \le 1+14L_g\eta_z$. Choosing β_z so that $1-\beta_z^2 = L_{\beta_z}\eta_z$ with $L_{\beta_z} := 14L_g + \frac{2B_2}{B_5}\left(\frac{1}{12L_g} + \frac{3}{\mu_g}\right) + \frac{2B_3 L_g}{B_5}$ yields

$$\begin{aligned}
T_{mz} &= \beta_z^2(1+L_g\eta_z)(1+12L_g\eta_z) - 1 + \frac{2B_2}{B_5}\left(\eta_z^2 + \frac{3\eta_z}{\mu_g}\right) + \frac{24B_3L_g^2\eta_z^2}{B_5} \\
&\le \beta_z^2 - 1 + L_{\beta_z}\eta_z = 0.
\end{aligned} \tag{146}$$

With the three momentum-correction coefficients set to zero, we dispose of the cross term $T_3'\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right]/n_h$ via the decomposition

$$\|\mathbf{m}_x^k\|^2 \;\le\; 2\|\mathbf{m}_x^k - \mathbb{E}(\mathbf{D}_x^k)\|^2 + 2\|\mathbb{E}(\mathbf{D}_x^k)\|^2.$$

The first piece is absorbed into the already-eliminated T_{mx} coefficient (after enlarging B_3 by a constant factor). The second piece is bounded by Lemma 25, which expresses $\frac{1}{n_h}\|\mathbb{E}(\mathbf{D}_x^k)\|^2$ as a combination of $\mathbb{E}\left[\left\|\mathbb{E}(\bar{m}_x^k)\right\|^2\right]$ (absorbed into T_3), tracking errors $\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right]$ and $\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right]$ (absorbed into T_1 and T_2), and the heterogeneity term $\mathcal{O}(\zeta_f^2 + \rho^2\zeta_g^2)$ (absorbed into R_k).

We now eliminate T_1, T_2, T_3 following the argument in Theorem 1. The only difference is that each coefficient carries extra higher-order terms from the momentum-drift lemmas, which are absorbed by enlarging the constants in (142) by at most a constant factor.

Eliminating T_1. Under $\eta_x \le \sqrt{\frac{1-\lambda}{B_3^\circ C_y}}$, $\kappa \le \frac{1-\lambda}{B_4^\circ C_y}$ (identical to (38)), together with $\eta_y \le \frac{L_2^2}{72L_1^2L_g}$ and $\eta_z \le \frac{\mu_g}{384L_g^2L_2^2}$, the momentum-induced quadratic drift terms in $T_1(2)$ reduce to linear ones,

$$72L_1^2L_g^2\eta_y^2 \le L_g\eta_y, \qquad 384L_g^2L_2^2\eta_z^2 \le \mu_g\eta_z,$$

and are dominated by $-B_1\eta_y\mu_g$ once B_1 is enlarged by at most a constant factor:

$$T_1 \le \underbrace{4L_2^2\eta_x - B_1\eta_y\mu_g + B_2\frac{48\eta_z L_2^2}{\mu_g} + \kappa B_4^\circ C_y\frac{\eta_x L_2^2}{1-\lambda}}_{\text{identical to (38)}} + \underbrace{9B_4L_g^3\eta_y + (B_3+B_5)(L_g\eta_y + \mu_g\eta_z)}_{\text{momentum-induced, linear in } \eta_y,\eta_z} \le 0. \tag{147}$$

Eliminating T_2. Under $\eta_x \le \sqrt{\frac{1-\lambda}{B_3^\circ C_z}}$, $\kappa \le \frac{1-\lambda}{B_4^\circ C_z}$, together with $\eta_z \le \frac{1}{384L_g^2}$ (so that $384L_g^4\eta_z^2 \le L_g^2\eta_z$), the analogue of (39) gives

$$T_2 \le \underbrace{4L_g^2\eta_x - B_2\eta_z\mu_g + \kappa B_4^\circ C_z\frac{\eta_x L_g^2}{1-\lambda}}_{\text{identical to (39)}} + \underbrace{(B_3+B_5)L_g^2\eta_z}_{\text{momentum-induced, linear in } \eta_z} \le 0. \tag{148}$$

Eliminating T_3. Under the same step-size conditions as (40) ($\eta_x \le \frac{\eta_y\mu_g}{72B_1L_{y^\star}^2}$, $\eta_x \le \frac{\eta_z\mu_g}{72B_2L_{z^\star}^2}$, $\eta_x \le 1/B_3^\circ$, $\kappa \le \frac{(1-\lambda)L_g^2}{24B_4^\circ L_2^2\lambda^{3/2}}$, $\eta_x \le 1/(4D_0)$ with $D_0 := B_1L_{y^\star}^2 + B_2L_{z^\star}^2 + \frac{2L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}$), we have

$$T_3 \le -\tfrac{\eta_x}{4} + \eta_x^2 D_0 \le 0. \tag{149}$$

Momentum introduces no new term into T_3, since the cross term T_3' has already been absorbed above.

Variance-related aggregation error. Before eliminating T_4, we control the stochastic noise terms in R_k induced by the momentum updates. Using the chosen momentum coefficients (144), (145), and (146), together with $\eta_y = c_y\eta_x$ and $\eta_z = c_z\eta_x$, we have the *first-order* bound

$$\frac{1-\beta_u}{1+\beta_u} \le 1-\beta_u \le 1-\beta_u^2 = L_{\beta_u}\eta_u, \qquad u \in \{x,y,z\}, \tag{150}$$

and the *second-order* bound (using $(1-\beta_u)^2 = \frac{(1-\beta_u^2)^2}{(1+\beta_u)^2} \le (1-\beta_u^2)^2$)

$$\begin{aligned}(1-\beta_x)^2 &\le (1-\beta_x^2)^2 = L_{\beta_x}^2\eta_x^2,\\ (1-\beta_y)^2 &\le (1-\beta_y^2)^2 = L_{\beta_y}^2\eta_y^2 = L_{\beta_y}^2c_y^2\eta_x^2,\\ (1-\beta_z)^2 &\le (1-\beta_z^2)^2 = L_{\beta_z}^2\eta_z^2 = L_{\beta_z}^2c_z^2\eta_x^2.\end{aligned} \tag{151}$$

Applying (151) to the momentum-drift noise $B_4(1-\beta_y)^2\sigma^2$ and $(B_3+B_5)(1-\beta_x)^2\sigma_1^2$, and (150) to the factors $\frac{1-\beta_x}{1+\beta_x}$ multiplying $C_{\sigma 1}, C_{\sigma 2}$ and the aggregation noise $192(B_3+B_5)L_1^2\eta_x^2\frac{1-\beta_x}{1+\beta_x}\sigma_1^2$ in R_k, we convert every stochastic term into an $\mathcal{O}(\eta_x^2\sigma^2)$ or $\mathcal{O}(\eta_x^3\sigma^2)$ contribution. The heterogeneity constants $C_{\zeta 1}, C_{\zeta 2}$ and ζ_f^2, ζ_g^2 are not compressed by β_u and retain their non-momentum scaling.

Eliminating the consensus coefficient T_4. With T_1, T_2, T_3 non-positive and the variance bounds (150)–(151) in hand, we bound the aggregation errors $\mathbb{E}\left[\left\|\bar{\xi}_u^k\right\|^2\right]$ in R_k via Lemma 12 and fold

them into T_4, mirroring the argument of (42)–(44). Together with the momentum-specific step-size conditions $\eta_y \le \frac{L_2^2}{72L_1^2L_g}$, $\eta_z \le \frac{1}{384L_g^2}$, $\eta_x \le \frac{1}{288L_1^2}$ (so that $\frac{72L_1^2L_g^4\eta_y^2}{L_2^2} \le L_g^3\eta_y$, $384L_g^4\eta_z^2 \le L_g^2\eta_z$, and $288L_1^2L_g^2\eta_x^2 \le L_1^2\eta_x$), the combined consensus coefficient satisfies

$$\begin{aligned} T_4' &:= T_4 + \frac{\kappa}{\eta_x}\left((1+L_g^2) + \frac{16B_1}{c_y\mu_g} + \frac{16B_2(1+L_g^2)}{c_z\mu_g}\right) \\ &\le \underbrace{-\left(1 + B_1C_{y^\star} + B_2C_{z^\star}\right)L_g^2\eta_x}_{\text{identical to (44)}} \\ &\quad + \underbrace{\frac{9B_4L_g^5\eta_y}{L_2^2} + (B_3+B_5)\big(L_g^3\eta_y + L_g^2\eta_z + L_1^2\eta_x\big)}_{\text{momentum-induced, linear in } \eta_x,\eta_y,\eta_z} \le 0. \end{aligned} \tag{152}$$

As in (147)–(149), the momentum-induced additions are absorbed by enlarging B_Δ in (142) by at most a constant factor, i.e., replacing B_3° by a constant multiple so that the negative drift $-(1-\lambda)B_3^\circ\eta_x/2$ dominates the momentum-induced linear terms.

Collecting the remaining stochastic, heterogeneity, and aggregation contributions, the Lyapunov recursion reduces to

$$\begin{aligned} \mathcal{L}^{k+1} - \mathcal{L}^k \le &-\frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \frac{L_{\nabla\Phi}\eta_x^2\sigma_1^2}{n_h} \\ &+ \frac{\eta_x^2}{1-\lambda}B_\Delta C_{\zeta 1} + \frac{48\kappa\eta_x^2}{1-\lambda}B_\Delta C_{\zeta 2} \\ &+ \frac{L_{\beta_x}\eta_x^3}{1-\lambda}B_\Delta C_{\sigma 1} + \frac{48L_{\beta_x}\kappa\eta_x^3}{1-\lambda}B_\Delta C_{\sigma 2} + 192(B_3+B_5)L_1^2L_{\beta_x}\eta_x^3\sigma_1^2 \\ &+ B_4L_{\beta_y}^2c_y^2\eta_x^2\sigma^2 + (B_3+B_5)L_{\beta_x}^2\eta_x^2\sigma_1^2 \\ &+ \left(2\kappa\eta_x + \frac{8B_2\kappa\eta_z}{\mu_g}\right)\left(6\zeta_f^2 + 6\rho^2\zeta_g^2\right) + \frac{8B_1\kappa\eta_y}{\mu_g}\cdot 2\zeta_g^2 \\ &+ \left(2\kappa\eta_x + \frac{8B_2\kappa\eta_z}{\mu_g}\right)\cdot 4L_{\beta_x}\eta_x\sigma_1^2 + \frac{8B_1\kappa\eta_y}{\mu_g}\cdot 4L_{\beta_y}\eta_y\sigma^2. \end{aligned} \tag{153}$$

Summing (153) over $k = 0, \dots, K-1$ and letting R_k' denote the sum of all non-gradient terms on the right-hand side, we obtain

$$\mathcal{L}^K - \mathcal{L}^0 \le -\frac{\eta_x}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + R_k'. \tag{154}$$

Because $\mathcal{L}^K \ge 0$ (each term in the potential is non-negative), rearranging gives

$$\frac{\eta_x}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \mathcal{L}^0 + R_k'. \tag{155}$$

The initial potential satisfies

$$\begin{aligned} \mathcal{L}^0 = \Phi(\bar{x}^0) &+ B_1\left\|\bar{y}^0 - \bar{y}_\star^0\right\|^2 + B_2\left\|\bar{z}^0 - \bar{z}_\star^0\right\|^2 + B_\Delta\frac{\mathbf{\Delta}^0}{n_h} \\ &+ \frac{B_3}{n_h}\left\|\mathbf{m}_x^0 - \mathbb{E}(\mathbf{D}_x^0)\right\|^2 + \frac{B_4}{n_h}\left\|\mathbf{m}_y^0 - \mathbb{E}(\mathbf{D}_y^0)\right\|^2 + \frac{B_5}{n_h}\left\|\mathbf{m}_z^0 - \mathbb{E}(\mathbf{D}_z^0)\right\|^2. \end{aligned} \tag{156}$$

We now bound the three momentum-drift terms in (156). Using the momentum update (12) and the initialization $\mathbf{m}_u^{-1} = 0$ for $u \in \{x, y, z\}$, we have $\mathbf{m}_u^0 = (1-\beta_u)\mathbf{D}_u^0$. Taking $u = y$ as an example,

$$\mathbf{m}_y^0 - \mathbb{E}(\mathbf{D}_y^0) = (1-\beta_y)\big(\mathbf{D}_y^0 - \mathbb{E}(\mathbf{D}_y^0)\big) - \beta_y\,\mathbb{E}(\mathbf{D}_y^0). \tag{157}$$

By Assumption 4, $\mathbb{E}[\mathbf{D}_y^0 - \mathbb{E}(\mathbf{D}_y^0)] = 0$ and $\mathbb{E}\left[\left\|\mathbf{D}_y^0 - \mathbb{E}(\mathbf{D}_y^0)\right\|^2\right] \le n_h\sigma^2$. Taking the squared norm and expectation on both sides of (157), the cross term vanishes by unbiasedness, and we obtain

$$\mathbb{E}\left[\left\|\mathbf{m}_y^0 - \mathbb{E}(\mathbf{D}_y^0)\right\|^2\right] \le (1-\beta_y)^2 n_h\sigma^2 + \beta_y^2\left\|\mathbb{E}(\mathbf{D}_y^0)\right\|^2 \le L_{\beta_y}^2\eta_x^2\cdot n_h\sigma^2 + \left\|\mathbb{E}(\mathbf{D}_y^0)\right\|^2, \quad (158)$$

where in the last inequality we applied the second-order bound (151) to $(1-\beta_y)^2$ and used $\beta_y^2 < 1$. Analogous arguments for $u = x, z$ yield

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{m}_x^0 - \mathbb{E}(\mathbf{D}_x^0)\right\|^2\right] &\le L_{\beta_x}^2\eta_x^2\cdot n_h\sigma_1^2 + \left\|\mathbb{E}(\mathbf{D}_x^0)\right\|^2,\\ \mathbb{E}\left[\left\|\mathbf{m}_z^0 - \mathbb{E}(\mathbf{D}_z^0)\right\|^2\right] &\le L_{\beta_z}^2\eta_x^2\cdot n_h\sigma_1^2 + \left\|\mathbb{E}(\mathbf{D}_z^0)\right\|^2.\end{aligned} \quad (159)$$

Substituting (158)–(159) into (156) and dividing by n_h, the stochastic contributions are of order $\mathcal{O}(\eta_x^2\sigma^2)$ and can be absorbed into the step-size-dependent remainder below, while $\left\|\mathbb{E}(\mathbf{D}_u^0)\right\|^2/n_h$ is a constant depending only on the initial full gradients. Hence $\mathcal{L}^0 = \mathcal{O}(1)$ with respect to K.

To write the remainder R_k' explicitly, we introduce non-negative constants $b_1, \dots, b_9$ defined as follows:

$$\begin{aligned}
&b_1 := L_{\nabla\Phi}(1+\rho^2), \quad b_2 := \frac{B_3^\circ C_{\zeta 1}}{1-\lambda}, \quad b_3 := \frac{B_3^\circ L_{\beta_x} C_{\sigma 1}}{1-\lambda},\\
&b_4 := \frac{B_3^\circ C_{\zeta 2}}{1-\lambda}, \quad b_5 := \frac{B_3^\circ L_{\beta_x} C_{\sigma 2}}{1-\lambda}, \quad b_6 := B_4 L_{\beta_y}^2 c_y^2 + (B_3+B_5)L_{\beta_x}^2,\\
&b_7 := 192(B_3+B_5)L_1^2 L_{\beta_x},\\
&b_8 := \left(12 + \frac{48B_2c_z}{\mu_g}\right)\left(\zeta_f^2 + \rho^2\zeta_g^2\right) + \frac{16B_1c_y}{\mu_g}\zeta_g^2 + \frac{B_4^\circ}{1-\lambda}(C_{\zeta 1} + \kappa C_{\zeta 2}),\\
&b_9 := \left(8 + \frac{32B_2c_z}{\mu_g}\right)L_{\beta_x}(1+\rho^2) + \frac{32B_1c_y}{\mu_g}L_{\beta_y} + \frac{B_4^\circ L_{\beta_x}}{1-\lambda}(C_{\sigma 1} + \kappa C_{\sigma 2}).
\end{aligned} \quad (160)$$

Compared with (49), b_1, b_2, b_4, b_8 match a_1, a_2, a_4, a_6; the extra L_{β_u} factor in b_3, b_5, b_7, b_9 comes from the linear compression (150), and b_6 is a new contribution from the stacked drift recursions. Summing (153) over $k = 0, \dots, K-1$,

$$\begin{aligned}\frac{\eta_x}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \mathcal{L}^0 + b_1\frac{\eta_x^2\sigma^2}{n_h}K + b_2\eta_x^3K + b_3\eta_x^4\sigma^2K + \kappa b_4\eta_x^3K + \kappa b_5\eta_x^4\sigma^2K\\ + b_6\,\eta_x^2\sigma^2K + b_7\,\eta_x^3\sigma^2K + b_8\,\kappa\eta_xK + b_9\,\kappa\eta_x^2\sigma^2K.\end{aligned} \quad (161)$$

Dividing both sides by $\eta_x K/4$ gives the time-average gradient-norm bound

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \frac{4\mathcal{L}^0}{\eta_xK} + 4b_1\frac{\eta_x\sigma^2}{n_h} + 4b_2\eta_x^2 + 4b_3\eta_x^3\sigma^2 + 4\kappa b_4\eta_x^2 + 4\kappa b_5\eta_x^3\sigma^2\\ + 4b_6\,\eta_x\sigma^2 + 4b_7\,\eta_x^2\sigma^2 + 4b_8\,\kappa + 4b_9\,\kappa\eta_x\sigma^2.\end{aligned} \quad (162)$$

Choosing the step size $\eta_x = \Theta(1/\sqrt{n_hK})$ and absorbing problem-dependent constants into the $\mathcal{O}$-notation yields the convergence rate stated in the theorem. Compared with the non-momentum rate (52), the extra η_x factor provided by (150) on the noise constants $C_{\sigma 1}, C_{\sigma 2}$ upgrades the non-momentum $\eta_x^2\sigma^2 \sim n_h\sigma^2/K$ term into $\eta_x^3\sigma^2 \sim n_h^{3/2}\sigma^2/K^{3/2}$, while the heterogeneity terms are unchanged because $\zeta_f^2 + \zeta_g^2$ is not compressed by momentum.

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \le \mathcal{O}\left(\frac{1}{\sqrt{n_hK}} + \frac{n_h^{3/2}\sigma^2}{K^{3/2}(1-\lambda)} + \frac{n_h(\zeta_f^2+\zeta_g^2)}{K(1-\lambda)^2}\right)\\ + \mathcal{O}\left(\frac{\kappa n_h^{3/2}\sigma^2}{K^{3/2}(1-\lambda)^2} + \frac{\kappa n_h(\zeta_f^2+\zeta_g^2)}{K(1-\lambda)^2}\right) + \mathcal{O}\left(\kappa\left(\zeta_f^2+\zeta_g^2\right)\right).\end{aligned} \quad (163)$$

This completes the proof. □

Corollary 2. *Under the conditions of Theorem 2, suppose $\mathcal{A}(\cdot)$ satisfies the (δ,κ)-robustness criterion with $\kappa=\mathcal{O}(\delta)$. The transient time of* `SLBRD-M` *in Algorithm 1 is*

$$\tau=\mathcal{O}\left(\max\left\{\frac{n_h^2}{(1-\lambda)},\frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4},\frac{\delta^2 n_h^3}{(1-\lambda)^4}\right\}\right). \tag{164}$$

Proof. The algorithm achieves linear speedup when the higher-order terms in (163) are dominated by the $\mathcal{O}(1/\sqrt{n_hK})$ term. This requires each of the following conditions to hold:

$$\frac{{n_h}^{3/2}\sigma^2}{K^{3/2}(1-\lambda)}\lesssim\frac{1}{\sqrt{n_hK}},\qquad\frac{n_h\left(\zeta_f^2+\zeta_g^2\right)}{(1-\lambda)^2K}\lesssim\frac{1}{\sqrt{n_hK}}, \tag{165}$$

$$\frac{\kappa{n_h}^{3/2}\sigma^2}{(1-\lambda)^2K^{3/2}}\lesssim\frac{1}{\sqrt{n_hK}},\qquad\frac{\kappa n_h\left(\zeta_f^2+\zeta_g^2\right)}{(1-\lambda)^2K}\lesssim\frac{1}{\sqrt{n_hK}}. \tag{166}$$

Rearranging each condition in terms of K yields the respective requirements:

$$K\gtrsim\frac{n_h^2}{(1-\lambda)},\quad K\gtrsim\frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4},\quad K\gtrsim\frac{\kappa^2n_h^3}{(1-\lambda)^4}. \tag{167}$$

Taking the maximum over all conditions. Suppose $\mathcal{A}(\cdot)$ satisfies the (δ,κ)-robustness criterion with $\kappa=\mathcal{O}(\delta)$. Then the transient time satisfies

$$K\gtrsim\mathcal{O}\left(\max\left\{\frac{n_h^2}{(1-\lambda)},\frac{n_h^3(\zeta_f^2+\zeta_g^2)^2}{(1-\lambda)^4},\frac{\delta^2n_h^3}{(1-\lambda)^4}\right\}\right), \tag{168}$$

which completes the proof. □

E.3 Proof of Supporting Lemmas

E.3.1 Proof of Lemma 12

Lemma 12. *Suppose that Assumptions 1 and 2, 4, 5 hold. Then, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|R_{x,i}^{k+\frac{1}{2}}-\bar{x}_i^{k+\frac{1}{2}}\right\|^2\right]&\le 2\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\left(\mathbb{E}\left[\left\|x_j^k-\bar{x}_i^k\right\|^2\right]+L_g^2\mathbb{E}\left[\left\|z_j^k-\bar{z}_i^k\right\|^2\right]\right)\\
&\quad+\kappa\eta_x^2\left(18\zeta_f^2+18\rho^2\zeta_g^2+12\frac{1-\beta_x}{1+\beta_x}\sigma_1^2\right),\\
\mathbb{E}\left[\left\|R_{y,i}^{k+\frac{1}{2}}-\bar{y}_i^{k+\frac{1}{2}}\right\|^2\right]&\le 2\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\mathbb{E}\left[\left\|y_j^k-\bar{y}_i^k\right\|^2\right]+\kappa\eta_y^2\left(6\zeta_g^2+12\frac{1-\beta_y}{1+\beta_y}\sigma^2\right),\\
\mathbb{E}\left[\left\|R_{z,i}^{k+\frac{1}{2}}-\bar{z}_i^{k+\frac{1}{2}}\right\|^2\right]&\le 2(1+L_g^2)\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\mathbb{E}\left[\left\|z_j^k-\bar{z}_i^k\right\|^2\right]\\
&\quad+\kappa\eta_z^2\left(18\zeta_f^2+18\rho^2\zeta_g^2+12\frac{1-\beta_z}{1+\beta_z}\sigma_1^2\right).
\end{aligned}$$

Proof. We follow the proof of Lemma 3 with $D_{u,i}^k$ replaced by $m_{u,i}^k$ for $u\in\{x,y,z\}$. By the (δ,κ)-robustness property (Definition 2), the Cauchy–Schwarz inequality, and the local update $u_j^{k+\frac{1}{2}}=u_j^k-\eta_u m_{u,j}^k$,

$$\mathbb{E}\left[\left\|R_{u,i}^{k+\frac{1}{2}}-\bar{u}_i^{k+\frac{1}{2}}\right\|^2\right]\le 2\kappa\sum_{j\in\mathcal{H}_i}w_{ij}\left(\mathbb{E}\left[\left\|u_j^k-\bar{u}_i^k\right\|^2\right]+\eta_u^2\mathbb{E}\left[\left\|m_{u,j}^k-\bar{m}_{u,i}^k\right\|^2\right]\right). \tag{169}$$

Bounding $\sum_j w_{ij}\mathbb{E}\left[\left\|m_{u,j}^k-\bar{m}_{u,i}^k\right\|^2\right]$ by the bias-variance decomposition in Lemma 3, and using the momentum variance bound

$$\mathbb{E}\left[\left\|m_{u,j}^k-\mathbb{E}(m_{u,j}^k)\right\|^2\right]\le\frac{1-\beta_u}{1+\beta_u}\sigma_u^2 \tag{170}$$

in place of $\mathbb{E}\left[\left\|D_{u,j}^k - \mathbb{E}(D_{u,j}^k)\right\|^2\right] \le \sigma_u^2$, we obtain the three stated bounds. □

Lemma 23. *Suppose that Assumptions 1 and 2, 4, 5 hold. Then, for every honest $i \in \mathcal{H}$,*

$$\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|m_{j,y}^k - \bar{m}_{i,y}^k\right\|^2\right] \le 6\frac{1-\beta_y}{1+\beta_y}\sigma^2 + 3\zeta_g^2, \tag{171}$$

$$\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|m_{j,x}^k - \bar{m}_{i,x}^k\right\|^2\right] \le 6\frac{1-\beta_x}{1+\beta_x}\sigma_1^2 + 9\zeta_f^2 + 9\rho^2\zeta_g^2 + L_g^2\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right], \tag{172}$$

$$\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|m_{j,z}^k - \bar{m}_{i,z}^k\right\|^2\right] \le 6\frac{1-\beta_z}{1+\beta_z}\sigma_1^2 + 9\zeta_f^2 + 9\rho^2\zeta_g^2 + L_g^2\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right]. \tag{173}$$

Proof. We prove the bound for $u = y$; the bounds for $u \in \{x, z\}$ follow by the same argument with β_y replaced by β_u. Recall the momentum update and its neighborhood average

$$m_{i,y}^k = \beta_y\, m_{i,y}^{k-1} + (1-\beta_y)D_{i,y}^k, \qquad \bar{m}_{i,y}^k = \sum_{j\in\mathcal{H}_i} w_{ij}\, m_{j,y}^k, \tag{174}$$

where $D_{i,y}^k := \nabla G_i(x_i^k;\varphi_i^k)$ and $\bar{D}_{i,y}^k := \sum_{j\in\mathcal{H}_i} w_{ij}\, D_{j,y}^k$. Unrolling (174) gives $m_{i,y}^k = (1-\beta_y)\sum_{t=0}^k \beta_y^{k-t} D_{i,y}^t$. Applying Cauchy-Schwarz inequalityyields

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|m_{j,y}^k - \bar{m}_{i,y}^k\right\|^2\right] \\
&\le 3(1-\beta_y)^2\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|\sum_{t=0}^k \beta_y^{k-t}\left(D_{j,y}^t - \nabla g_j(x_j^t, y_j^t)\right)\right\|^2\right] \\
&\quad + 3(1-\beta_y)^2\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|\sum_{t=0}^k \beta_y^{k-t}\left(\bar{D}_{i,y}^t - \sum_{c\in\mathcal{H}_i} w_{ic}\nabla g_c(x_c^t, y_c^t)\right)\right\|^2\right] \\
&\quad + 3(1-\beta_y)^2\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|\sum_{t=0}^k \beta_y^{k-t}\left(\nabla g_j(x_j^t, y_j^t) - \sum_{c\in\mathcal{H}_i} w_{ic}\nabla g_c(x_c^t, y_c^t)\right)\right\|^2\right].
\end{aligned} \tag{175}$$

Stochastic-noise terms. Define $A^k := \mathbb{E}\left[\left\|\sum_{t=0}^k \beta_y^{k-t}(D_{j,y}^t - \nabla g_j(x_j^t, y_j^t))\right\|^2\right]$. Unbiasedness of $D_{j,y}^t$ (Assumption 4) kills the cross terms when expanding the square, so

$$A^k \le \beta_y^2 A^{k-1} + \sigma^2, \tag{176}$$

and, with $A^0 \le \sigma^2$, recursion gives $A^k \le \sigma^2/(1-\beta_y^2)$. An identical argument for $B^k := \mathbb{E}\left[\left\|\sum_{t=0}^k \beta_y^{k-t}(\bar{D}_{i,y}^t - \sum_{c\in\mathcal{H}_i} w_{ic}\nabla g_c(x_c^t, y_c^t))\right\|^2\right]$ yields $B^k \le \sigma^2/(1-\beta_y^2)$.

Heterogeneity term. By Assumption 5 and Cauchy-Schwarz inequality,

$$\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|\sum_{t=0}^k \beta_y^{k-t}\left(\nabla g_j(x_j^t, y_j^t) - \sum_{c\in\mathcal{H}_i} w_{ic}\nabla g_c(x_c^t, y_c^t)\right)\right\|^2\right] \le \left(\sum_{t=0}^k \beta_y^{k-t}\right)^2 \zeta_g^2 \le \frac{\zeta_g^2}{(1-\beta_y)^2}. \tag{177}$$

Substituting the three bounds into (175) gives

$$\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|m_{j,y}^k - \bar{m}_{i,y}^k\right\|^2\right] \le 3(1-\beta_y)^2\left(\frac{2\sigma^2}{1-\beta_y^2} + \frac{\zeta_g^2}{(1-\beta_y)^2}\right) = 6\frac{1-\beta_y}{1+\beta_y}\sigma^2 + 3\zeta_g^2, \tag{178}$$

which establishes (171).

Bounds for x and z. Repeating the derivation with β_y replaced by β_u, $D_{j,y}^t$ by $D_{j,u}^t$, and σ^2 by σ_1^2 for $u \in \{x, z\}$, the only modification is that the heterogeneity term (177) is no longer given directly by Assumption 5, but must be decomposed through the expected directions. Using Assumption 2 together with Assumption 5 and Cauchy-Schwarz inequality, we obtain

$$\sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|\mathbb{E}(D_{j,u}^k) - \mathbb{E}(\bar{D}_{i,u}^k)\right\|^2\right] \le 3\zeta_f^2 + 3\rho^2\zeta_g^2 + 3L_g^2 \sum_{j\in\mathcal{H}_i} w_{ij}\,\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right], \qquad u \in \{x, z\}, \tag{179}$$

in parallel to the bound (72) established in the proof of Lemma 9. Combining (179) with the stochastic-noise bounds and the identity $3(1-\beta_u)^2\sum_{t=0}^k(\beta_u^{k-t})^2 \le 3(1-\beta_u)/(1+\beta_u)$, and using that the per-iterate z-consensus is non-decreasing in k up to the step-size factors absorbed in the Lyapunov analysis, so that $\sum_{t=0}^k(\beta_u^{k-t})^2\,\mathbb{E}\left[\left\|z_j^t - \bar{z}_i^t\right\|^2\right] \le \frac{1}{1-\beta_u^2}\mathbb{E}\left[\left\|z_j^k - \bar{z}_i^k\right\|^2\right]$ to leading order, yields (172) and (173) after collecting the resulting factor 9 on the heterogeneity constants $\zeta_f^2, \rho^2\zeta_g^2$. This completes the proof. □

E.3.2 Proof of Lemma 13

Lemma 13. *Suppose that Assumptions 1 and 2, 3, 4 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \le \mathbb{E}\left[\Phi(\bar{x}^k)\right] &- \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] + \left(\frac{L_{\nabla\Phi}\eta_x^2}{2} - \frac{\eta_x}{2}\right)\mathbb{E}\left[\left\|\mathbb{E}(\bar{m}_x^k)\right\|^2\right] \\ &+ \left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right] + \frac{6L_g^2\eta_x}{n_h}\boldsymbol{\Delta}^k + \eta_x\mathbb{E}\left[\left\|\bar{m}_x^k - \mathbb{E}(\bar{D}_x^k)\right\|^2\right] \\ &+ 6L_2^2\eta_x\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 6L_g^2\eta_x\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right].\end{aligned}$$

Proof. We follow the proof of Lemma 4 with $\bar{D}_x^k$ replaced by $\bar{m}_x^k$. By the $L_{\nabla\Phi}$-smoothness of Φ (Lemma 1), the momentum update $\bar{x}^{k+1} = \bar{x}^k - \eta_x\bar{m}_x^k + \eta_x\bar{\xi}_x^k$, the polarization identity, and Young's inequality,

$$\begin{aligned}\mathbb{E}\left[\Phi(\bar{x}^{k+1})\right] \le \mathbb{E}\left[\Phi(\bar{x}^k)\right] &+ \frac{\eta_x}{2}\mathbb{E}\left[\left\|\bar{m}_x^k - \nabla\Phi(\bar{x}^k)\right\|^2\right] - \frac{\eta_x}{4}\mathbb{E}\left[\left\|\nabla\Phi(\bar{x}^k)\right\|^2\right] \\ &+ \left(L_{\nabla\Phi}\eta_x^2 - \frac{\eta_x}{2}\right)\mathbb{E}\left[\left\|\mathbb{E}(\bar{m}_x^k)\right\|^2\right] + \left(L_{\nabla\Phi}\eta_x^2 + \eta_x\right)\mathbb{E}\left[\left\|\bar{\xi}_x^k\right\|^2\right].\end{aligned} \tag{180}$$

Bounding the residual $\mathbb{E}\left[\left\|\bar{m}_x^k - \nabla\Phi(\bar{x}^k)\right\|^2\right]$ by Lemma 24 in place of Lemma 11 yields the stated recursion. □

Lemma 24. *Suppose that Assumptions 1 and 2 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{m}_x^k - \nabla\Phi(\bar{\mathbf{x}}^k)\right\|^2\right] \le 2\mathbb{E}\left[\left\|\bar{m}_x^k - \mathbb{E}\left(\bar{D}_x^k\right)\right\|^2\right] &+ \frac{12L_g^2}{n_h}\boldsymbol{\Delta}^k \\ &+ 12L_2^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 12L_g^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right].\end{aligned} \tag{181}$$

Proof. By Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left\|\bar{m}_x^k - \nabla\Phi(\bar{\mathbf{x}}^k)\right\|^2\right] \le 2\mathbb{E}\left[\left\|\bar{m}_x^k - \mathbb{E}\left(\bar{D}_x^k\right)\right\|^2\right] + 2\mathbb{E}\left[\left\|\mathbb{E}\left(\bar{D}_x^k\right) - \nabla\Phi(\bar{\mathbf{x}}^k)\right\|^2\right]. \tag{182}$$

Bounding the second term on the right-hand side of (182) by Lemma 11 yields the stated bound. □

E.3.3 Proof of Lemma 14

Lemma 14. *Suppose that Assumptions 2–5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{y}^{k+1}-\bar{y}_\star^{k+1}\right\|^2\right] \\
&\leq (1-\eta_y\mu_g)\,\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right]+\left(\eta_x^2+\frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2\mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right]+4(1+\frac{3}{\eta_y\mu_g})\frac{\eta_y^2L_g^4}{n_hL_2^2}\boldsymbol{\Delta}^k \\
&+4(1+\frac{3}{\eta_y\mu_g})\eta_y^2\mathbb{E}\left[\left\|\bar{m}_y^k-\mathbb{E}(\bar{D}_y^k)\right\|^2\right]+2\left(\eta_y^2+\frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right].
\end{aligned}$$

Proof. We follow the proof of Lemma 5 with $\mathbb{E}(\bar{D}_y^k)$ replaced by $\bar{m}_y^k$. By Jensen's inequality and the Lipschitz continuity of $\bar{y}_\star(\cdot)$ (Lemma 1),

$$\mathbb{E}\left[\left\|\bar{y}^{k+1}-\bar{y}_\star^{k+1}\right\|^2\right]\leq\left(1+\frac{\eta_y\mu_g}{3}\right)\mathbb{E}\left[\left\|\bar{y}^{k+1}-\bar{y}_\star^{k}\right\|^2\right]+\left(\eta_x^2+\frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2\mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right]. \tag{183}$$

Substituting the momentum update $\bar{y}^{k+1}=\bar{y}^k-\eta_y\bar{m}_y^k+\eta_y\bar{\xi}_y^k$ into (183) and applying Young's inequality with parameter $v\in(0,1)$ gives

$$\mathbb{E}\left[\left\|\bar{y}^{k+1}-\bar{y}_\star^{k}\right\|^2\right]\leq(1+v)\mathbb{E}\left[\left\|\bar{y}^{k}-\eta_y\bar{m}_y^k-\bar{y}_\star^{k}\right\|^2\right]+\left(\eta_y^2+\frac{\eta_y^2}{v}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right]. \tag{184}$$

Expanding $\mathbb{E}\left[\left\|\bar{y}^{k}-\eta_y\bar{m}_y^k-\bar{y}_\star^{k}\right\|^2\right]$ as in Lemma 5, the only modification is the bound

$$\mathbb{E}\left[\left\|\bar{m}_y^k-\mathbb{E}(D_{\bar{y}}^k)\right\|^2\right]\leq 2\mathbb{E}\left[\left\|\bar{m}_y^k-\mathbb{E}(\bar{D}_y^k)\right\|^2\right]+\frac{2L_g^2}{n_h}\left(\mathbb{E}\left[\left\|\mathbf{x}^k-\bar{\mathbf{x}}^k\right\|^2\right]+\mathbb{E}\left[\left\|\mathbf{y}^k-\bar{\mathbf{y}}^k\right\|^2\right]\right), \tag{185}$$

which introduces the momentum drift $\mathbb{E}\left[\left\|\bar{m}_y^k-\mathbb{E}(\bar{D}_y^k)\right\|^2\right]$. Setting $u=v=\eta_y\mu_g/3$ and proceeding as in Lemma 5,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{y}^{k+1}-\bar{y}_\star^{k+1}\right\|^2\right] \\
&\leq (1-\eta_y\mu_g)\,\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right]+\left(\eta_x^2+\frac{3\eta_x^2}{\eta_y\mu_g}\right)L_{y^\star}^2\mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right]+4\left(1+\frac{3}{\eta_y\mu_g}\right)\frac{\eta_y^2L_g^4}{n_hL_2^2}\boldsymbol{\Delta}^k \\
&+4\left(1+\frac{3}{\eta_y\mu_g}\right)\eta_y^2\mathbb{E}\left[\left\|\bar{m}_y^k-\mathbb{E}(\bar{D}_y^k)\right\|^2\right]+2\left(\eta_y^2+\frac{3\eta_y}{\mu_g}\right)\mathbb{E}\left[\left\|\bar{\xi}_y^k\right\|^2\right].
\end{aligned} \tag{186}$$

This completes the proof.

□

E.3.4 Proof of Lemma 15

Lemma 15. *Suppose that Assumptions 2–5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right] \\
&\leq(1-\eta_z\mu_g)\mathbb{E}\left[\left\|\bar{z}^{k}-\bar{z}_\star^{k}\right\|^2\right]+\frac{12\eta_z^2L_g^2}{n_h}\left(1+\frac{3}{\eta_z\mu_g}\right)\boldsymbol{\Delta}^k+12\eta_z^2L_2^2\left(1+\frac{3}{\eta_z\mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^{k}-\bar{y}_\star^{k}\right\|^2\right] \\
&+\left(2\eta_z^2+\frac{6\eta_z}{\mu_g}\right)\left(\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right]+\mathbb{E}\left[\left\|\bar{m}_z^k-\mathbb{E}(\bar{D}_z^k)\right\|^2\right]\right)+\left(\eta_x^2+\frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right].
\end{aligned}$$

Proof. We follow the proof of Lemma 6 with $\mathbb{E}(\bar{D}_z^k)$ replaced by $\bar{m}_z^k$. By Jensen's inequality and the Lipschitz continuity of $\bar{z}_\star(\cdot)$ (Lemma 1),

$$\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k+1}\right\|^2\right]\leq\left(1+\frac{\eta_z\mu_g}{3}\right)\mathbb{E}\left[\left\|\bar{z}^{k+1}-\bar{z}_\star^{k}\right\|^2\right]+\left(\eta_x^2+\frac{3\eta_x^2}{\eta_z\mu_g}\right)L_{z^\star}^2\mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right]. \tag{187}$$

Substituting the momentum update $\bar{z}^{k+1} = \bar{z}^k - \eta_z \bar{m}_z^k + \eta_z \bar{\xi}_z^k$ into (187) and applying Young's inequality gives

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1} - \bar{z}_\star^k\right\|^2\right] \leq \left(1 + \frac{\eta_z \mu_g}{3}\right) \mathbb{E}\left[\left\|\bar{z}^k - \eta_z \mathbb{E}(\bar{D}_z^k) - \bar{z}_\star^k\right\|^2\right] \\ + 2\left(1 + \frac{3}{\eta_z \mu_g}\right) \eta_z^2 \left(\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] + \mathbb{E}\left[\left\|\bar{m}_z^k - \mathbb{E}(\bar{D}_z^k)\right\|^2\right]\right).\end{aligned} \tag{188}$$

Decomposing $\bar{z}^k - \eta_z \mathbb{E}(\bar{D}_z^k) - \bar{z}_\star^k$ via the optimality condition $\nabla_{yy}^2 g(\bar{x}^k, \bar{y}_\star^k)\bar{z}_\star^k + \nabla_y f(\bar{x}^k, \bar{y}_\star^k) = 0$ and proceeding as in Lemma 6 yields

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{z}^{k+1} - \bar{z}_\star^{k+1}\right\|^2\right] \leq (1 - \eta_z \mu_g)\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] + \frac{12\eta_z^2 L_g^2}{n_h}\left(1 + \frac{3}{\eta_z \mu_g}\right)\boldsymbol{\Delta}^k \\ + 12\eta_z^2 L_2^2 \left(1 + \frac{3}{\eta_z \mu_g}\right)\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \left(\eta_x^2 + \frac{3\eta_x^2}{\eta_z \mu_g}\right) L_{z^\star}^2 \mathbb{E}\left[\left\|\bar{m}_x^k\right\|^2\right] \\ + \left(2\eta_z^2 + \frac{6\eta_z}{\mu_g}\right)\left(\mathbb{E}\left[\left\|\bar{\xi}_z^k\right\|^2\right] + \mathbb{E}\left[\left\|\bar{m}_z^k - \mathbb{E}(\bar{D}_z^k)\right\|^2\right]\right).\end{aligned} \tag{189}$$

This completes the proof. □

E.4 Proof of Lemma 16

Lemma 16. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1}\right\|^2\right] \leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right] + \frac{3L_g^4 \eta_y^2 \lambda^{3/2}}{(1-\lambda)L_2^2}\boldsymbol{\Delta}^k + \frac{4\eta_y^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k) - \mathbf{m}_y^k\right\|^2\right] \\ + \frac{3n_h L_g^2 \eta_y^2 \lambda^{3/2}}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \frac{48\kappa n_h \eta_y^2}{1-\lambda}(\zeta_g^2 + 2\sigma^2\frac{1-\beta_y}{1+\beta_y}) + \frac{3n_h\eta_y^2\lambda^{3/2}\zeta_g^2}{(1-\lambda)}.\end{aligned}$$

Proof. We follow the proof of Lemma 7 with $\mathbf{D}_y^k$ replaced by $\mathbf{m}_y^k$. By Young's inequality with parameter $v \in (0,1)$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1}\right\|^2\right] = \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{y}^k - \bar{\mathbf{y}}^k - \eta_y \mathbb{E}(\mathbf{D}_y^k)\right)\right\|^2\right] \\ + \frac{8}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_y^k\right\|^2\right] + \frac{4\eta_y^2}{1-v}\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right].\end{aligned} \tag{190}$$

The only modification with respect to (111) is the additional momentum drift $\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right]$. Bounding the first term on the right-hand side of (190) as in (112), applying Lemma 12 in place of Lemma 3, setting $u = v = \sqrt{\lambda}$, and invoking the lower-level optimality condition $\nabla_y g(\bar{x}^k, \bar{y}_\star^k) = 0$ gives, under $\kappa \leq (1-\lambda)(1-\sqrt{\lambda})/4$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \bar{\mathbf{y}}^{k+1}\right\|^2\right] \leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right] + \frac{3L_g^4 \eta_y^2 \lambda^{3/2}}{(1-\lambda)L_2^2}\boldsymbol{\Delta}^k + \frac{4\eta_y^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] \\ + \frac{3n_h L_g^2 \eta_y^2 \lambda^{3/2}}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \frac{3n_h\eta_y^2\lambda^{3/2}\zeta_g^2}{1-\lambda} \\ + \frac{48\kappa n_h \eta_y^2}{1-\lambda}\left(\zeta_g^2 + 2\sigma^2\frac{1-\beta_y}{1+\beta_y}\right).\end{aligned} \tag{191}$$

This completes the proof. □

E.5 Proof of Lemma 17

Lemma 17. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] \le{}& \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right] + \frac{4\eta_z^2\lambda^{3/2}L_g^2}{1-\lambda}\boldsymbol{\Delta}^k + \frac{4\eta_z^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{m}_z^k-\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]\\ &+\frac{4n_h\eta_z^2\lambda^{3/2}L_2^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+\frac{4n_h\eta_z^2\lambda^{3/2}L_g^2}{(1-\lambda)}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\ &+\frac{48\kappa n_h\eta_z^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2\frac{1-\beta_z}{1+\beta_z})+\frac{4\eta_z^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\zeta_g^2\rho^2\right).\end{aligned}$$

Proof. We follow the proof of Lemma 9 with $\mathbf{D}_z^k$ replaced by $\mathbf{m}_z^k$. By Young's inequality with parameter $v\in(0,1)$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] ={}& \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{z}^k-\bar{\mathbf{z}}^k-\eta_z\mathbb{E}(\mathbf{D}_z^k)\right)\right\|^2\right]\\ &+\frac{8}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_z^k\right\|^2\right]+\frac{4\eta_z^2}{1-v}\mathbb{E}\left[\left\|\mathbf{m}_z^k-\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right].\end{aligned}\tag{192}$$

Bounding the first term on the right-hand side of (192) as in (127), applying Lemma 12 in place of Lemma 3, setting $u=v=\sqrt{\lambda}$, and invoking the lower-level optimality condition $\nabla_{yy}^2 g(\bar{x}^k,\bar{y}_\star^k)\bar{z}_\star^k+\nabla_y f(\bar{x}^k,\bar{y}_\star^k)=0$ gives, under $\kappa\le(1-\lambda)(1-\sqrt{\lambda})/(32(1+L_g^2))$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k+1}\right\|^2\right] \le{}& \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right] + \frac{4\eta_z^2\lambda^{3/2}L_g^2}{1-\lambda}\boldsymbol{\Delta}^k + \frac{4\eta_z^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{m}_z^k-\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]\\ &+\frac{4n_h\eta_z^2\lambda^{3/2}L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+\frac{4n_h\eta_z^2\lambda^{3/2}L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\ &+\frac{4\eta_z^2\lambda^{3/2}n_h}{1-\lambda}\left(\zeta_f^2+\zeta_g^2\rho^2\right)+\frac{48\kappa n_h\eta_z^2}{1-\lambda}\left(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2\frac{1-\beta_z}{1+\beta_z}\right).\end{aligned}\tag{193}$$

This completes the proof. □

E.6 Proof of Lemma 18

Lemma 18. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}\right\|^2\right]\\ &\le\frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k}-\bar{\mathbf{x}}^{k}\right\|^2\right]+\frac{16\kappa L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\bar{\mathbf{z}}^{k}\right\|^2\right]+\frac{\eta_x^2\lambda^{3/2}}{1-\lambda}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^k\right\|^2\right]+\frac{4\eta_x^2L_g^2}{1-\lambda}\boldsymbol{\Delta}^k\\ &+\frac{24\eta_x^2n_h\sigma_1^2}{1-\lambda}\cdot\frac{1-\beta_x}{1+\beta_x}+\frac{36\eta_x^2n_h}{1-\lambda}\left(\zeta_f^2+\rho^2\zeta_g^2\right)+\frac{48\kappa n_h\eta_x^2}{1-\lambda}(3\zeta_f^2+3\zeta_g^2\rho^2+2\sigma_1^2\frac{1-\beta_x}{1+\beta_x}).\end{aligned}$$

Proof. We follow the proof of Lemma 8 with $\mathbf{D}_x^k$ replaced by $\mathbf{m}_x^k$. By Young's inequality with parameter $v\in(0,1)$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}\right\|^2\right] ={}& \frac{1}{v}\mathbb{E}\left[\left\|\left(\mathbf{W}-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\left(\mathbf{x}^k-\bar{\mathbf{x}}^k-\eta_x\bar{\mathbf{m}}_x^k\right)\right\|^2\right]\\ &+\frac{8}{1-v}\mathbb{E}\left[\left\|\boldsymbol{\xi}_x^k\right\|^2\right]+\frac{4\eta_x^2}{1-v}\mathbb{E}\left[\left\|\mathbf{m}_x^k-\bar{\mathbf{m}}_x^k\right\|^2\right].\end{aligned}\tag{194}$$

Bounding the first term on the right-hand side of (194) as in (119), applying Lemma 12 in place of Lemma 3, and setting $u = v = \sqrt{\lambda}$ gives, under $\kappa \leq (1-\lambda)(1-\sqrt{\lambda})/32$,

$$\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\right\|^2\right] \leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k} - \bar{\mathbf{x}}^{k}\right\|^2\right] + \frac{16\kappa L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{z}^{k} - \bar{\mathbf{z}}^{k}\right\|^2\right] + \frac{\eta_x^2\lambda^{3/2}}{1-\lambda}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^{k}\right\|^2\right] + \frac{4\eta_x^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{m}_x^{k} - \bar{\mathbf{m}}_x^{k}\right\|^2\right] + \frac{48\kappa n_h \eta_x^2}{1-\lambda}\left(3\zeta_f^2 + 3\zeta_g^2\rho^2 + 2\sigma_1^2\frac{1-\beta_x}{1+\beta_x}\right). \tag{195}$$

By Assumption 5, the momentum consensus gap is bounded as

$$\mathbb{E}\left[\left\|\mathbf{m}_x^{k} - \bar{\mathbf{m}}_x^{k}\right\|^2\right] \leq 6n_h\sigma_1^2\frac{1-\beta_x}{1+\beta_x} + 9n_h\left(\zeta_f^2 + \zeta_g^2\rho^2\right) + L_g^2\mathbb{E}\left[\left\|\mathbf{z}^{k} - \bar{\mathbf{z}}^{k}\right\|^2\right]. \tag{196}$$

Substituting (196) into (195) yields

$$\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\right\|^2\right] \leq \frac{1+\lambda}{2}\mathbb{E}\left[\left\|\mathbf{x}^{k} - \bar{\mathbf{x}}^{k}\right\|^2\right] + \frac{16\kappa L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\mathbf{z}^{k} - \bar{\mathbf{z}}^{k}\right\|^2\right] + \frac{\eta_x^2\lambda^{3/2}}{1-\lambda}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^{k}\right\|^2\right] + \frac{4\eta_x^2 L_g^2}{1-\lambda}\boldsymbol{\Delta}^k + \frac{24\eta_x^2 n_h\sigma_1^2}{1-\lambda}\cdot\frac{1-\beta_x}{1+\beta_x} + \frac{36\eta_x^2 n_h}{1-\lambda}\left(\zeta_f^2 + \rho^2\zeta_g^2\right) + \frac{48\kappa n_h\eta_x^2}{1-\lambda}\left(3\zeta_f^2 + 3\zeta_g^2\rho^2 + 2\sigma_1^2\frac{1-\beta_x}{1+\beta_x}\right). \tag{197}$$

This completes the proof. □

E.7 Proof of Lemma 19

Lemma 19. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\boldsymbol{\Delta}^{k+1} \leq \frac{1+\lambda}{2}\boldsymbol{\Delta}^k + C_y\frac{n_h\eta_x^2 L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2 L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] + \frac{2\eta_x^2 L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^{k}\right\|^2\right] + \frac{24n_h\eta_x^2\sigma_1^2}{1-\lambda}\cdot\frac{1-\beta_x}{1+\beta_x} + \frac{\eta_x^2 n_h}{1-\lambda}C_{\zeta 1} + \frac{48\kappa\eta_x^2 n_h}{1-\lambda}C_{\zeta 2} + \frac{48\kappa\eta_x^2 n_h}{1-\lambda}\frac{1-\beta_x}{1+\beta_x}C_{\sigma 2},$$

where the problem-dependent constants $C_y, C_z, C_{\zeta 1}, C_{\zeta 2}, C_{\sigma 2}$ are the same as in Lemma 10.

Proof. We follow the proof of Lemma 10 with $\mathbf{D}_u^k$ replaced by $\mathbf{m}_u^k$ for $u \in \{x, y, z\}$. Combining Lemmas 16, 17, and 18 with the definition of $\boldsymbol{\Delta}^k$ yields

$$\boldsymbol{\Delta}^{k+1} \leq \left(\lambda + \frac{16\kappa(1+L_g^2)}{1-\sqrt{\lambda}}\right)\boldsymbol{\Delta}^k + C_\Delta\frac{L_g^2\eta_x^2}{1-\lambda}\boldsymbol{\Delta}^k + C_y\frac{n_h\eta_x^2 L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2 L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] + \frac{2\eta_x^2 L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^{k}\right\|^2\right] + \frac{24n_h\eta_x^2\sigma_1^2}{1-\lambda}\cdot\frac{1-\beta_x}{1+\beta_x} + \frac{\eta_x^2 n_h}{1-\lambda}C_{\zeta 1} + \frac{48\kappa\eta_x^2 n_h}{1-\lambda}C_{\zeta 2} + \frac{48\kappa\eta_x^2 n_h}{1-\lambda}\frac{1-\beta_x}{1+\beta_x}C_{\sigma 2}, \tag{198}$$

where the constants $C_\Delta, C_y, C_z, C_{\zeta 1}, C_{\zeta 2}, C_{\sigma 2}$ are defined as in Lemma 10. Under $\kappa \leq \frac{(1-\lambda)(1-\sqrt{\lambda})}{32(1+L_g^2)}$ and $\eta_x \leq \frac{1-\lambda}{2L_g\sqrt{C_\Delta}}$, (198) simplifies to

$$\boldsymbol{\Delta}^{k+1} \leq \frac{1+\lambda}{2}\boldsymbol{\Delta}^k + C_y\frac{n_h\eta_x^2 L_2^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + C_z\frac{n_h\eta_x^2 L_g^2}{1-\lambda}\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] + \frac{2\eta_x^2 L_2^2\lambda^{3/2}}{(1-\lambda)L_g^2}\mathbb{E}\left[\left\|\bar{\mathbf{m}}_x^{k}\right\|^2\right] + \frac{24n_h\eta_x^2\sigma_1^2}{1-\lambda}\cdot\frac{1-\beta_x}{1+\beta_x} + \frac{\eta_x^2 n_h}{1-\lambda}C_{\zeta 1} + \frac{48\kappa\eta_x^2 n_h}{1-\lambda}C_{\zeta 2} + \frac{48\kappa\eta_x^2 n_h}{1-\lambda}\frac{1-\beta_x}{1+\beta_x}C_{\sigma 2}. \tag{199}$$

This completes the proof. □

E.8 Proof of Lemma 20

Lemma 20. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\big[\|\mathbf{m}_y^{k+1} - \mathbb{E}(\mathbf{D}_y^{k+1})\|^2\big] \\
&\le \beta_y^2(1+L_g\eta_y)(1+3\eta_y L_g)\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + (1-\beta_y)^2 n_h\sigma^2 \\
&\quad + \frac{2\beta_y^2}{c_y^2}\eta_y(1+L_g\eta_y)L_g\big(\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right] + \mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]\big) \\
&\quad + \beta_y^2\eta_y(1+L_g\eta_y)L_g\left(9n_hL_g^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \frac{9L_g^4}{L_2^2}\mathbf{\Delta}^k + 9n_h\zeta_g^2 + 3\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right]\right).
\end{aligned}$$

Proof. Stacking the per-agent momentum updates $m_{y,i}^{k+1} = \beta_y m_{y,i}^k + (1-\beta_y)D_{y,i}^{k+1}$ and summing over $i \in \mathcal{H}$, adding and subtracting $\beta_y\mathbb{E}(\mathbf{D}_y^k) + (1-\beta_y)\mathbb{E}(\mathbf{D}_y^{k+1})$, and using the unbiasedness of $\mathbf{D}_y^{k+1}$ gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{m}_y^{k+1} - \mathbb{E}(\mathbf{D}_y^{k+1})\right\|^2\right] &\le \beta_y^2\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + (1-\beta_y)^2 n_h\sigma^2 \\
&\quad + \beta_y^2\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^{k+1}) - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] \\
&\quad - 2\beta_y^2\left\langle \mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k),\, \mathbb{E}(\mathbf{D}_y^{k+1}) - \mathbb{E}(\mathbf{D}_y^k)\right\rangle,
\end{aligned} \tag{200}$$

where we used $\mathbb{E}\left[\left\|\mathbf{D}_y^{k+1} - \mathbb{E}(\mathbf{D}_y^{k+1})\right\|^2\right] = \sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|D_{y,i}^{k+1} - \mathbb{E}(D_{y,i}^{k+1})\right\|^2\right] \le n_h\sigma^2$ by Assumption 4. By the Lipschitz continuity of $\nabla_y g$ and Young's inequality, the cross term is controlled by

$$\begin{aligned}
&- 2\beta_y^2\left\langle \mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k),\, \mathbb{E}(\mathbf{D}_y^{k+1}) - \mathbb{E}(\mathbf{D}_y^k)\right\rangle \\
&\le \beta_y^2 L_g\eta_y\left(\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + \mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right] + \frac{1}{c_y^2}\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right]\right),
\end{aligned} \tag{201}$$

where we used $\|\mathbb{E}(\mathbf{D}_y^{k+1}) - \mathbb{E}(\mathbf{D}_y^k)\|^2 \le L_g^2\eta_y^2(\|\hat{\mathbf{m}}_y^k\|^2 + \frac{1}{c_y^2}\|\hat{\mathbf{m}}_x^k\|^2)$ summed over agents. Substituting (201) into (200) yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{m}_y^{k+1} - \mathbb{E}(\mathbf{D}_y^{k+1})\right\|^2\right] &\le \beta_y^2\left(1+L_g\eta_y\right)\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + (1-\beta_y)^2 n_h\sigma^2 \\
&\quad + \beta_y^2 L_g\eta_y\left(1+L_g\eta_y\right)\left(\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right] + \frac{1}{c_y^2}\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right]\right).
\end{aligned} \tag{202}$$

Invoking Lemma 25 (stacked form) to bound $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right]$ and $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right]$ gives the stated recursion

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{m}_y^{k+1} - \mathbb{E}(\mathbf{D}_y^{k+1})\right\|^2\right] \\
&\le \beta_y^2\left(1+L_g\eta_y\right)\left(1+3L_g\eta_y\right)\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + (1-\beta_y)^2 n_h\sigma^2 \\
&\quad + \beta_y^2 L_g\eta_y\left(1+L_g\eta_y\right)\left(9n_hL_g^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + \frac{9L_g^4}{L_2^2}\mathbf{\Delta}^k + 9n_h\zeta_g^2 + 3\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right]\right) \\
&\quad + \frac{2\beta_y^2 L_g\eta_y(1+L_g\eta_y)}{c_y^2}\left(\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right] + \mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]\right).
\end{aligned} \tag{203}$$

This completes the proof. □

E.9 Proof of Lemma 21

Lemma 21. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{m}_x^{k+1} - \mathbb{E}(\mathbf{D}_x^{k+1})\right\|^2\right] \\
&\le \beta_x^2\,(1 + L_g\eta_z)\,\mathbb{E}\left[\left\|\mathbf{m}_x^k - \mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] + (1-\beta_x)^2 n_h\sigma_1^2 + 16L_1^2\eta_x^2\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right] \\
&+ 24L_1^2\eta_y^2\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + 24L_g^2\eta_z^2\mathbb{E}\left[\left\|\mathbf{m}_z^k - \mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] \\
&+ 24n_h\left(3L_1^2L_g^2\eta_y^2 + 16L_g^2L_2^2\eta_z^2\right)\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 384n_hL_g^4\eta_z^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] \\
&+ 24L_1^2\eta_y^2\left(\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right] + \frac{3L_g^4}{L_2^2}\boldsymbol{\Delta}^k + 3n_h\zeta_g^2\right) \\
&+ 24L_g^2\eta_z^2\left(\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right] + 16L_g^2\boldsymbol{\Delta}^k + 16n_h(\zeta_f^2 + \rho^2\zeta_g^2)\right) \\
&+ 32L_1^2\eta_x^2\left(\mathbb{E}\left[\left\|\xi_x^k\right\|^2\right] + L_g^2\boldsymbol{\Delta}^k + 6\frac{1-\beta_x}{1+\beta_x}n_h\sigma_1^2 + 9n_h(\zeta_f^2 + \rho^2\zeta_g^2)\right).
\end{aligned}$$

Proof. Stacking the per-agent momentum updates $m_{x,i}^{k+1} = \beta_x m_{x,i}^k + (1-\beta_x)D_{x,i}^{k+1}$, summing over $i \in \mathcal{H}$, and proceeding as in the proof of Lemma 20 gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{m}_x^{k+1} - \mathbb{E}(\mathbf{D}_x^{k+1})\right\|^2\right] &\le \beta_x^2\mathbb{E}\left[\left\|\mathbf{m}_x^k - \mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] + (1-\beta_x)^2 n_h\sigma_1^2 \\
&+ \beta_x^2\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_x^{k+1}) - \mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] \\
&- 2\beta_x^2\left\langle \mathbf{m}_x^k - \mathbb{E}(\mathbf{D}_x^k),\, \mathbb{E}(\mathbf{D}_x^{k+1}) - \mathbb{E}(\mathbf{D}_x^k)\right\rangle.
\end{aligned} \tag{204}$$

Applying Young's inequality to the cross term and the smoothness of the hypergradient surrogate (summed over agents) gives

$$\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_x^{k+1}) - \mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] \le 4L_1^2\eta_x^2\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right] + 4L_1^2\eta_y^2\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right] + 4L_g^2\eta_z^2\mathbb{E}\left[\left\|\hat{\mathbf{m}}_z^k\right\|^2\right]. \tag{205}$$

For brevity, set

$$C_1 := \beta_x^2\left(1 + \frac{1}{\eta_x L_1}\right), \qquad C_2 := \beta_x^2\left(1 + \frac{1}{\eta_x L_g}\right).$$

Substituting (205) into (204) and invoking Lemma 25 (stacked form) to bound $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right]$, $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right]$ and $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_z^k\right\|^2\right]$ yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{m}_x^{k+1} - \mathbb{E}(\mathbf{D}_x^{k+1})\right\|^2\right] \\
&\le \beta_x^2(1 + L_g\eta_z)\mathbb{E}\left[\left\|\mathbf{m}_x^k - \mathbb{E}(\mathbf{D}_x^k)\right\|^2\right] + (1-\beta_x)^2 n_h\sigma_1^2 \\
&\quad + 16C_1L_1^2\eta_x^2\big(\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right] + \mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]\big) + 12C_1L_1^2\eta_y^2\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] \\
&\quad + 12C_2L_g^2\eta_z^2\mathbb{E}\left[\left\|\mathbf{m}_z^k - \mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] + 12C_2\big(2n_hL_1^2L_g^2\eta_y^2 + 12n_hL_g^2L_2^2\eta_z^2\big)\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\
&\quad + 192C_2n_hL_g^4\eta_z^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] + 12C_2L_1^2\eta_y^2\left(\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right] + \frac{3L_g^4}{L_2^2}\boldsymbol{\Delta}^k + 3n_h\zeta_g^2\right) \\
&\quad + 12C_2L_g^2\eta_z^2\big(\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right] + 16L_g^2\boldsymbol{\Delta}^k + 16n_h(\zeta_f^2 + \rho^2\zeta_g^2)\big).
\end{aligned} \tag{206}$$

This completes the proof. □

E.10 Proof of Lemma 22

Lemma 22. *Suppose that Assumptions 1 and 2–5 hold. Then, we have*

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{m}_z^{k+1}-\mathbb{E}(\mathbf{D}_z^{k+1})\right\|^2\right]\\
&\le \beta_z^2\left(1+L_g\eta_z\right)\left(1+12L_g\eta_z\right)\mathbb{E}\left[\left\|\mathbf{m}_z^{k}-\mathbb{E}(\mathbf{D}_z^{k})\right\|^2\right]+(1-\beta_z)^2n_h\sigma_1^2+16L_1^2\eta_x^2\mathbb{E}\left[\left\|\mathbf{m}_x^{k}\right\|^2\right]\\
&+24L_1^2\eta_y^2\mathbb{E}\left[\left\|\mathbf{m}_y^{k}-\mathbb{E}(\mathbf{D}_y^{k})\right\|^2\right]\\
&+24n_h\left(3L_1^2L_g^2\eta_y^2+16L_g^2L_2^2\eta_z^2\right)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+384n_hL_g^4\eta_z^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\
&+24L_1^2\eta_y^2\left(\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right]+\frac{3L_g^4}{L_2^2}\mathbf{\Delta}^k+3n_h\zeta_g^2\right)\\
&+24L_g^2\eta_z^2\left(\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right]+16L_g^2\mathbf{\Delta}^k+16n_h(\zeta_f^2+\rho^2\zeta_g^2)\right)\\
&+32L_1^2\eta_x^2\left(\mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]+L_g^2\mathbf{\Delta}^k+6\frac{1-\beta_x}{1+\beta_x}n_h\sigma_1^2+9n_h(\zeta_f^2+\rho^2\zeta_g^2)\right).
\end{aligned}
$$

Proof. Stacking the per-agent momentum updates $m_{z,i}^{k+1}=\beta_z m_{z,i}^k+(1-\beta_z)D_{z,i}^{k+1}$, summing over $i\in\mathcal{H}$, and proceeding as in the proof of Lemma 21 gives

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{m}_z^{k+1}-\mathbb{E}(\mathbf{D}_z^{k+1})\right\|^2\right] &\le \beta_z^2\mathbb{E}\left[\left\|\mathbf{m}_z^{k}-\mathbb{E}(\mathbf{D}_z^{k})\right\|^2\right]+(1-\beta_z)^2n_h\sigma_1^2\\
&\quad+\beta_z^2\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^{k+1})-\mathbb{E}(\mathbf{D}_z^{k})\right\|^2\right]\\
&\quad-2\beta_z^2\left\langle\mathbf{m}_z^k-\mathbb{E}(\mathbf{D}_z^k),\,\mathbb{E}(\mathbf{D}_z^{k+1})-\mathbb{E}(\mathbf{D}_z^k)\right\rangle.
\end{aligned}
\tag{207}
$$

Applying Young's inequality to the cross term, together with the Lipschitz continuity of $\nabla_y f$ and $\nabla_{yy}^2 g$ summed over agents, gives

$$
\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^{k+1})-\mathbb{E}(\mathbf{D}_z^{k})\right\|^2\right]\le 3L_1^2\eta_x^2\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right]+3L_1^2\eta_y^2\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right]+3L_g^2\eta_z^2\mathbb{E}\left[\left\|\hat{\mathbf{m}}_z^k\right\|^2\right].
\tag{208}
$$

Substituting (208) into (207) and invoking Lemma 25 (stacked form) to bound $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right]$, $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right]$ and $\mathbb{E}\left[\left\|\hat{\mathbf{m}}_z^k\right\|^2\right]$ yields

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{m}_z^{k+1}-\mathbb{E}(\mathbf{D}_z^{k+1})\right\|^2\right]\\
&\le \beta_z^2\left(1+L_g\eta_z\right)\left(1+12L_g\eta_z\right)\mathbb{E}\left[\left\|\mathbf{m}_z^{k}-\mathbb{E}(\mathbf{D}_z^{k})\right\|^2\right]+(1-\beta_z)^2n_h\sigma_1^2\\
&\quad+16L_1^2\eta_x^2\mathbb{E}\left[\left\|\mathbf{m}_x^{k}\right\|^2\right]+24L_1^2\eta_y^2\mathbb{E}\left[\left\|\mathbf{m}_y^{k}-\mathbb{E}(\mathbf{D}_y^{k})\right\|^2\right]\\
&\quad+24\big(3n_hL_1^2L_g^2\eta_y^2+16n_hL_g^2L_2^2\eta_z^2\big)\mathbb{E}\left[\left\|\bar{y}^k-\bar{y}_\star^k\right\|^2\right]+384n_hL_g^4\eta_z^2\mathbb{E}\left[\left\|\bar{z}^k-\bar{z}_\star^k\right\|^2\right]\\
&\quad+24L_1^2\eta_y^2\left(\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right]+\frac{3L_g^4}{L_2^2}\mathbf{\Delta}^k+3n_h\zeta_g^2\right)\\
&\quad+24L_g^2\eta_z^2\left(\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right]+16L_g^2\mathbf{\Delta}^k+16n_h(\zeta_f^2+\rho^2\zeta_g^2)\right)\\
&\quad+32L_1^2\eta_x^2\left(\mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]+L_g^2\mathbf{\Delta}^k+6\frac{1-\beta_x}{1+\beta_x}n_h\sigma_1^2+9n_h(\zeta_f^2+\rho^2\zeta_g^2)\right),
\end{aligned}
\tag{209}
$$

where the term $9\beta_z^2(1+1/(\eta_zL_g))L_g^2\eta_z^2\mathbb{E}\left[\left\|\mathbf{m}_z^k-\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]$ generated by the $\hat{\mathbf{m}}_z^k$ expansion has been absorbed into the leading coefficient via the bound $9\beta_z^2(1+1/(\eta_zL_g))L_g^2\eta_z^2\le 12\beta_z^2L_g\eta_z$, valid for $\eta_zL_g\le 1$. This completes the proof. □

Lemma 25 (Bounded second moment for momentum)**.** *Suppose Assumptions 1 and 2–5 hold. Then, for every $k \in [K]$,*

$$\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right] \le 2\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right] + 2\mathbb{E}\left[\left\|\xi_x^k\right\|^2\right], \tag{210}$$

$$\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right] \le 3\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right] + 3\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + \frac{9L_g^4}{L_2^2}\boldsymbol{\Delta}^k + 9n_hL_g^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 9n_h\zeta_g^2, \tag{211}$$

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\mathbf{m}}_z^k\right\|^2\right] \le{}& 3\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right] + 3\mathbb{E}\left[\left\|\mathbf{m}_z^k - \mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] + 48L_g^2\boldsymbol{\Delta}^k + 48n_hL_g^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] \\ &+ 48n_hL_2^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 48n_h(\zeta_f^2 + \rho^2\zeta_g^2).\end{aligned} \tag{212}$$

Proof. Bound for $\mathbf{m}_x$. By Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k\right\|^2\right] = \mathbb{E}\left[\left\|\hat{\mathbf{m}}_x^k - \mathbf{m}_x^k + \mathbf{m}_x^k\right\|^2\right] \le 2\mathbb{E}\left[\left\|\mathbf{m}_x^k\right\|^2\right] + 2\mathbb{E}\left[\left\|\xi_x^k\right\|^2\right]. \tag{213}$$

Bound for $\mathbf{m}_y$. Decomposing $\hat{\mathbf{m}}_y^k = (\hat{\mathbf{m}}_y^k - \mathbf{m}_y^k) + (\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)) + \mathbb{E}(\mathbf{D}_y^k)$ and applying Cauchy-Schwarz inequalitygives

$$\mathbb{E}\left[\left\|\hat{\mathbf{m}}_y^k\right\|^2\right] \le 3\mathbb{E}\left[\left\|\xi_y^k\right\|^2\right] + 3\mathbb{E}\left[\left\|\mathbf{m}_y^k - \mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] + 3\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)\right\|^2\right]. \tag{214}$$

Using the optimality identity $\nabla_y g(\bar{x}^k, \bar{y}_\star^k) = 0$ and Cauchy-Schwarz inequalitywith a three-way split,

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_y^k)\right\|^2\right] \qquad (215)\\ &= \sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y g_i(x_i^k, y_i^k)\right\|^2\right] \\ &\le 3\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y g_i(x_i^k, y_i^k) - \nabla_y g_i(\bar{x}^k, \bar{y}^k)\right\|^2\right] + 3\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y g_i(\bar{x}^k, \bar{y}^k) - \nabla_y g(\bar{x}^k, \bar{y}^k)\right\|^2\right] \\ &\quad + 3n_h\mathbb{E}\left[\left\|\nabla_y g(\bar{x}^k, \bar{y}^k) - \nabla_y g(\bar{x}^k, \bar{y}_\star^k)\right\|^2\right] \qquad (216)\\ &\le 3L_g^2\big(\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right]\big) + 3n_h\zeta_g^2 + 3n_hL_g^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] \\ &\le \frac{3L_g^4}{L_2^2}\boldsymbol{\Delta}^k + 3n_hL_g^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 3n_h\zeta_g^2,\end{aligned}$$

where the last inequality uses the aggregate-consensus bound $\mathbb{E}\left[\left\|\mathbf{x}^k - \bar{\mathbf{x}}^k\right\|^2\right] + \mathbb{E}\left[\left\|\mathbf{y}^k - \bar{\mathbf{y}}^k\right\|^2\right] \le \frac{L_g^2}{L_2^2}\boldsymbol{\Delta}^k$. Combining (214) and (215) yields (211).

Bound for $\mathbf{m}_z$. Decomposing $\hat{\mathbf{m}}_z^k = (\hat{\mathbf{m}}_z^k - \mathbf{m}_z^k) + (\mathbf{m}_z^k - \mathbb{E}(\mathbf{D}_z^k)) + \mathbb{E}(\mathbf{D}_z^k)$ and applying Cauchy-Schwarz inequalitygives

$$\mathbb{E}\left[\left\|\hat{\mathbf{m}}_z^k\right\|^2\right] \le 3\mathbb{E}\left[\left\|\xi_z^k\right\|^2\right] + 3\mathbb{E}\left[\left\|\mathbf{m}_z^k - \mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] + 3\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right]. \tag{217}$$

Using the optimality identity $\nabla_{yy}^2 g(\bar{x}^k, \bar{y}_\star^k)\bar{z}_\star^k + \nabla_y f(\bar{x}^k, \bar{y}_\star^k) = 0$ and a four-way split via Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] \le{}& 4\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_{yy}^2 g_i(x_i^k, y_i^k)z_i^k - \nabla_{yy}^2 g(\bar{x}^k, \bar{y}^k)\bar{z}_\star^k\right\|^2\right] \\ &+ 4\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla_y f_i(x_i^k, y_i^k) - \nabla_y f(\bar{x}^k, \bar{y}^k)\right\|^2\right] \\ &+ 4n_h\mathbb{E}\left[\left\|\left(\nabla_{yy}^2 g(\bar{x}^k, \bar{y}^k) - \nabla_{yy}^2 g(\bar{x}^k, \bar{y}_\star^k)\right)\bar{z}_\star^k\right\|^2\right] \\ &+ 4n_h\mathbb{E}\left[\left\|\nabla_y f(\bar{x}^k, \bar{y}^k) - \nabla_y f(\bar{x}^k, \bar{y}_\star^k)\right\|^2\right].\end{aligned} \tag{218}$$

Applying Lipschitz smoothness and heterogeneity to each term bounds (218) by

$$\mathbb{E}\left[\left\|\mathbb{E}(\mathbf{D}_z^k)\right\|^2\right] \leq 16L_g^2\boldsymbol{\Delta}^k + 16n_hL_g^2\mathbb{E}\left[\left\|\bar{z}^k - \bar{z}_\star^k\right\|^2\right] + 16n_hL_2^2\mathbb{E}\left[\left\|\bar{y}^k - \bar{y}_\star^k\right\|^2\right] + 16n_h(\zeta_f^2 + \rho^2\zeta_g^2). \tag{219}$$

Combining (217) and (219) yields (212). This completes the proof. □