FULLY FIRST-ORDER LAGRANGIAN METHODS FOR BYZANTINE-ROBUST FEDERATED BILEVEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated bilevel optimization (FBO) plays a crucial role in solving hierarchical learning problems over distributed data. However, its distributed nature makes it highly susceptible to faulty or Byzantine clients. Existing Byzantine-robust FBO methods either rely on Hessian or Jacobian vector products or perform repeated robust updates within an inner loop, incurring additional differentiation or synchronization costs. To address these limitations, we propose `FOBFBL`, the first Byzantine-robust FBO method that is both fully first-order and single-loop. With a finite Lagrangian multiplier λ, it forms three coupled blocks of stochastic gradients that the server robustly aggregates at their natural scales in one round. For nonconvex upper-level and strongly convex lower-level objectives, we establish a nonasymptotic stationarity bound that separates the decaying optimization error, the Byzantine-induced floor, and the $\mathcal{O}(1/\lambda^2)$ Lagrangian bias, yielding $\mathcal{O}(\varepsilon^{-2})$ iterations above the residual floor. Under an additional upper-level PL condition, we obtain a last-iterate function-gap bound with $\widetilde{\mathcal{O}}((n_h K)^{-1})$ finite-horizon dependence. We further develop the Polyak-momentum variant `FOBFBL-M`, which removes stochastic variance from the persistent Byzantine error and leaves a limiting neighborhood governed by honest heterogeneity and the Lagrangian bias. Experiments on federated hyperparameter optimization and hyper-representation learning validate both methods.[1]

1 INTRODUCTION

Bilevel optimization has become an important framework for learning problems in which an upper decision depends on the solution of a lower problem. Representative applications include 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). As these applications move to distributed data that cannot be centralized, they give rise to *federated bilevel optimization* (FBO) (Tarzanagh et al., 2022; Xiao & Ji, 2023; Yang et al., 2023b; 2025). In the standard benign setting, n clients collaboratively solve

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

where f_i and g_i are the upper- and lower-level objectives of client i. Each local lower-level objective is strongly convex in y, and each client queries its local objectives through stochastic first-order oracles,

$$f_i(x,y) = \mathbb{E}_{\omega_i^f\sim\mathcal{D}_{f_i}}[F_i(x,y;\omega_i^f)], \qquad g_i(x,y) = \mathbb{E}_{\omega_i^g\sim\mathcal{D}_{g_i}}[G_i(x,y;\omega_i^g)]. \tag{2}$$

Federated learning is inherently vulnerable to faulty or malicious clients. Hardware failures or deliberate attacks can cause a client to depart from the prescribed protocol and transmit arbitrary updates (Armstrong, 2003; Xie et al., 2020; Baruch et al., 2019). Such clients are commonly modeled as Byzantine. Even when only a small fraction of clients are Byzantine, standard averaging may fail

[1] Code is available at `https://anonymous.4open.science/r/iclr27-123E`.

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 representative single-level Byzantine-robust and federated bilevel methods in terms of structure, smoothness, and sample complexity.

Single-Level Methods	Byz	SL	First-Order	(UL) f	(LL) g	Sample Complexity
Bucketing (Karimireddy et al., 2022)	✓	-	✓	$\mathcal{C}_L^1$	N. A.	$\tilde{\mathcal{O}}(\varepsilon^{-2})$
CC (Karimireddy et al., 2021)	✓	-	✓	$\mathcal{C}_L^1$	N. A.	$\tilde{\mathcal{O}}(\varepsilon^{-2})$
Bilevel Methods	**Byz**	**SL**	**First-Order**	**(UL) f**	**(LL) g**	**Sample Complexity**
FedNest (Tarzanagh et al., 2022)	✗	✗	✗	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\tilde{\mathcal{O}}(\varepsilon^{-2})$
SimFBO (Yang et al., 2023b)	✗	✓	✗	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	$\mathcal{O}(\varepsilon^{-2})$
MemFBO (Yang et al., 2025)	✗	✓	✓	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\varepsilon^{-3.5})$
BILANTINE (Abbas et al., 2024)	✓	✗	✗	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\tilde{\mathcal{O}}(\varepsilon^{-2})$
BR-FedBi (Li et al., 2026)	✓	✓	✗	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{3,3}$	$\mathcal{O}(\varepsilon^{-2})$
FOBFBL (ours)	✓	✓	✓	$\mathcal{C}_L^{2,2}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\varepsilon^{-2})$
FOBFBL-M (ours)	✓	✓	✓	$\mathcal{C}_L^{1,1}$	SC and $\mathcal{C}_L^{2,2}$	$\mathcal{O}(\varepsilon^{-2})$

"Byz" indicates Byzantine robustness, "SL" the single-loop structure, "First-Order" computation without Hessian or Jacobian vector products, and "SC" strong convexity. $\mathcal{C}_L^r$ (single-level) and $\mathcal{C}_L^{r,r}$ (bilevel) denote Lipschitz derivatives through order r. For the last two rows, ε is measured above the residual floor.

and prevent the honest population from converging. This risk is particularly serious in safety-critical applications such as healthcare, fraud detection, and autonomous driving (Han et al., 2023; Kasyap et al., 2024; Chen et al., 2021).

Although Byzantine robustness has been extensively studied for single-level federated optimization, extending it to FBO entails three fundamental challenges. *First*, the global hypergradient depends nonlinearly on population derivatives and an inverse lower-level Hessian, so it is generally not the average of clientwise hypergradients (Tarzanagh et al., 2022; Yang et al., 2023b). Perturbations of either lower-level tracker therefore propagate into the upper-level direction. Controlling this propagation requires a joint analysis of robust aggregation and lower-level tracking. *Second*, existing Byzantine-robust FBO methods retain Hessian or Jacobian vector products or perform multiple robust inner updates. These operations increase computation and communication while exposing more messages to adversarial corruption. *Third*, stochastic noise and data heterogeneity also disperse honest updates. A useful convergence guarantee must identify which components decay, which persist in the Byzantine floor, and how the finite multiplier affects the resulting neighborhood.

We develop the first fully first-order, single-loop framework for Byzantine-robust FBO. Table 1 provides a structural comparison, and our contributions are as follows:

- We introduce `FOBFBL`, which derives three stochastic first-order directions from a finite-multiplier Lagrangian and robustly aggregates each direction in one server exchange per iteration. It requires no Hessian or Jacobian vector products and no repeated robust inner updates, thereby avoiding the associated computation and communication.
- Under nonconvex upper-level and strongly convex lower-level objectives, we establish $\mathcal{O}(\varepsilon^{-2})$ iteration complexity for `FOBFBL` above the residual floor. We further obtain a $\widetilde{\mathcal{O}}((n_h K)^{-1})$ last-iterate function-gap rate under an upper-level PL condition. Both guarantees characterize the limiting accuracy in terms of Byzantine-induced error and finite-multiplier bias.
- We develop `FOBFBL-M`, a Polyak-momentum variant of `FOBFBL`, and prove that momentum eliminates stochastic variance from the asymptotic error, yielding the residual bound $\mathcal{O}(\kappa(\zeta_f^2 + \lambda^2\zeta_g^2) + \lambda^{-2})$. Our analysis relaxes the upper-level smoothness requirement from $\mathcal{C}_L^{2,2}$ to $\mathcal{C}_L^{1,1}$. For fixed λ and $\kappa = \mathcal{O}(\delta)$, the Byzantine-induced component matches the lower bound of Li et al. (2026) in its dependence on the Byzantine fraction and honest-client heterogeneity.
- Experiments on federated hyperparameter optimization and hyper-representation learning demonstrate the effectiveness of our methods on real-world datasets. Efficiency analysis further highlights the computational and communication benefits of our fully first-order, single-loop design.

1.1 Related Work

Federated and fully first-order bilevel optimization. The hypergradient couples upper-level derivatives with an inverse lower-level Hessian, so existing FBO methods differ primarily in how

they approximate it. FedNest and AggITD distribute nested hypergradient computations, whereas SimFBO uses a loopless recursion. All three retain Hessian or Jacobian vector products (Tarzanagh et al., 2022; Xiao & Ji, 2023; Yang et al., 2023b). Fully first-order alternatives avoid these products through value-function, Lagrangian, minimax, penalty, or finite-difference reformulations (Liu et al., 2022; Kwon et al., 2023; Wang et al., 2023; Shen & Chen, 2023; Yang et al., 2023a). Recent analyses also improve complexity under relaxed smoothness conditions (Chen et al., 2024). In distributed settings, related work studies loopless tracking under client heterogeneity and federated automatic differentiation (Niu et al., 2025; Rush et al., 2024). MemFBO is closest to our architecture because its finite-multiplier Lagrangian supports simultaneous updates using ordinary gradients (Yang et al., 2025). These methods assume protocol-compliant clients and do not analyze arbitrary corruption of the coupled upper- and lower-level message streams.

Byzantine robustness in distributed and federated bilevel optimization. In single-level learning, robust aggregators such as Krum, coordinatewise median, and geometric median bound the influence of Byzantine reports (Blanchard et al., 2017; Yin et al., 2018; Acharya et al., 2022). Bucketing, history, and mixing-based methods account for honest-client dispersion (Karimireddy et al., 2022; 2021; Allouah et al., 2023; Wu et al., 2023). These analyses control a single aggregated gradient relative to its honest target. In FBO, corruption in either lower-level tracker also propagates into the upper-level direction. Among Byzantine-robust FBO methods, BILANTINE uses an approximate implicit hypergradient. BR-FedBi uses an auxiliary single-loop recursion, and BR-FedBiM adds Polyak momentum that removes stochastic variance from its asymptotic Byzantine error. These methods retain Hessian or Jacobian vector products (Abbas et al., 2024; Li et al., 2026). Rob-FedBOB is fully first-order but performs multiple robust inner updates per outer iteration (Maralappanavar & Bharath, 2025). Prior Byzantine-robust FBO methods therefore do not combine a single-loop architecture with fully first-order computation. Our approach instead robustly aggregates three first-order Lagrangian blocks in one server exchange and analyzes their coupled perturbations jointly.

2 PRELIMINARIES

Problem formulation. Consider a federated system with a central server and n clients indexed by $[n] := \{1, \dots, n\}$. Let $\mathcal{B} \subseteq [n]$ denote the Byzantine set and $\mathcal{H} := [n] \setminus \mathcal{B}$ the honest set. Let $n_b := |\mathcal{B}|$ and $n_h := |\mathcal{H}| = n - n_b$, and assume $n_b \le \delta n$ for some $\delta < 1/2$. The server does not know either set. Byzantine clients may deviate arbitrarily from the prescribed protocol and transmit arbitrary messages to the server (Baruch et al., 2019; Xie et al., 2020). Accordingly, the target bilevel problem is defined by the honest clients' objectives:

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

Here f_i and g_i denote the local upper- and lower-level objectives of client i, respectively, as introduced in (2). For notational convenience, let $f_{\mathcal{H}} := \frac{1}{n_h} \sum_{i \in \mathcal{H}} f_i$ and $g_{\mathcal{H}} := \frac{1}{n_h} \sum_{i \in \mathcal{H}} g_i$ denote the corresponding averages over the honest clients. Assumption 2 implies that $g_{\mathcal{H}}(x, \cdot)$ is strongly convex. It follows that $y^\star_{\mathcal{H}}(x)$ is unique. We omit the subscript $\mathcal{H}$ when no ambiguity arises. Since $\mathcal{H}$ is unknown to the server, it cannot evaluate these honest averages directly.

Byzantine robustness. Since $\Phi_{\mathcal{H}}$ may be nonconvex, we assess optimality in terms of first-order stationarity. We use the following notion of Byzantine robustness (Liu et al., 2021).

Definition 1 ((δ, ε)-Byzantine robustness)**.** *A federated bilevel algorithm is said to be (δ, ε)-Byzantine-robust if, in the presence of at most a δ-fraction of Byzantine clients, it outputs a point $\hat{x}$ satisfying*

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

where the expectation is taken over the randomness of the algorithm and stochastic oracles.

Robust aggregation. To limit the influence of arbitrary Byzantine messages, the server replaces arithmetic averaging with a robust aggregation rule. We characterize its error in estimating the mean of the honest inputs using the (δ, κ)-robustness criterion (Allouah et al., 2023).

Definition 2 ((δ, κ)-robustness)**.** *Let $\delta < 1/2$ and $\kappa \ge 0$. An aggregation rule $\mathcal{A} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is (δ, κ)-robust if, for any input vectors $u_1, \dots, u_n \in \mathbb{R}^d$ and any index set $\mathcal{H} \subseteq [n]$ satisfying*

$|\mathcal{H}| = n_h \geq (1-\delta)n$, *the following inequality holds:*

$$\|\mathcal{A}(u_1, \ldots, u_n) - \bar{u}_{\mathcal{H}}\|^2 \leq \frac{\kappa}{n_h} \sum_{i \in \mathcal{H}} \|u_i - \bar{u}_{\mathcal{H}}\|^2, \qquad \bar{u}_{\mathcal{H}} := \frac{1}{n_h} \sum_{i \in \mathcal{H}} u_i. \tag{5}$$

The parameter κ, called the robustness coefficient, bounds the squared aggregation error relative to the empirical variance of the honest inputs. In particular, the error vanishes when all honest inputs coincide. Several standard aggregation rules satisfy this criterion with rule-dependent values of κ (Blanchard et al., 2017; Yin et al., 2018; Acharya et al., 2022; Allouah et al., 2023).

3 Methodology

We derive the finite-multiplier first-order reformulation behind `FOBFBL` and its Byzantine-robust, single-loop federated implementation. Following the value function perspective of fully first-order bilevel methods (Kwon et al., 2023; Yang et al., 2025), the protocol protects each direction estimated from the unknown honest population before the server update.

3.1 First-Order Reformulation

A central challenge in solving the federated bilevel problem in (3) is that the lower-level solution $y^\star(x)$ is defined implicitly by the aggregate objective of the unknown honest population. It is therefore not directly available to either an individual client or the server, and the honest hypergradient $\nabla\Phi_{\mathcal{H}}(x)$ cannot be evaluated in a straightforward manner. Let $f := f_{\mathcal{H}}$ and $g := g_{\mathcal{H}}$. Under Assumption 2, $y^\star(x)$ is uniquely defined. The implicit function theorem then gives

$$\nabla\Phi_{\mathcal{H}}(x) = \nabla_x f\big(x, y^\star(x)\big) - \nabla^2_{xy} g\big(x, y^\star(x)\big)\big[\nabla^2_{yy} g\big(x, y^\star(x)\big)\big]^{-1} \nabla_y f\big(x, y^\star(x)\big). \tag{6}$$

A direct evaluation of (6) requires both $y^\star(x)$ and the inverse of the aggregate lower-level Hessian. In particular, the inverse of an average Hessian is not the average of the clientwise inverses, so the global hypergradient cannot be recovered by simply averaging local hypergradients. Computing it instead requires higher-order products or an auxiliary variable, which is especially undesirable when client reports may be Byzantine.

To obtain directions that clients can evaluate using only first-order oracles, we adopt the value-function reformulation of Kwon et al. (2023) and Yang et al. (2025). Define $g^\star(x) := \min_z g(x,z) = g(x, y^\star(x))$. Strong convexity makes the bilevel formulation in (3) equivalent to

$$\min_{x,y} \; f(x,y) \qquad \text{s.t.} \qquad g(x,y) - g^\star(x) \leq 0.$$

For a fixed finite multiplier $\lambda > 0$, the Lagrangian of this constraint yields the surrogate

$$\min_{x,y} \max_{z} \; \mathcal{L}_\lambda(x,y,z), \qquad \mathcal{L}_\lambda(x,y,z) := f(x,y) + \lambda\big[g(x,y) - g(x,z)\big]. \tag{7}$$

For fixed x, the maximizer over z is uniquely given by $y^\star(x)$. When $\lambda \geq 2l_{f,1}/\mu_g$, the function $f(x,\cdot) + \lambda g(x,\cdot)$ is also strongly convex. Hence, the finite-multiplier minimizer $y^\star_\lambda(x) := \arg\min_y \mathcal{L}_\lambda(x,y,y^\star(x))$ is unique. Defining $\mathcal{L}^\star_\lambda(x) := \min_y \max_z \mathcal{L}_\lambda(x,y,z)$, the envelope theorem gives

$$\nabla\mathcal{L}^\star_\lambda(x) = \nabla_x f\big(x, y^\star_\lambda(x)\big) + \lambda\big[\nabla_x g\big(x, y^\star_\lambda(x)\big) - \nabla_x g\big(x, y^\star(x)\big)\big]. \tag{8}$$

Unlike (6), every term in (8) is an ordinary first-order partial gradient. The two lower-level variables play distinct roles: $y^\star_\lambda(x)$ minimizes the finite-multiplier objective, whereas $y^\star(x)$ minimizes the original lower-level objective. A finite λ therefore introduces a controlled approximation rather than an exact hypergradient. For the problem-dependent constant $C_{\text{lag}} > 0$, Lemma 1 shows that $\|\nabla\mathcal{L}^\star_\lambda(x) - \nabla\Phi_{\mathcal{H}}(x)\| \leq C_{\text{lag}}/\lambda$, yielding the Lagrangian bias $\mathcal{O}(C^2_{\text{lag}}/\lambda^2)$ in the final convergence bound.

For any shared triple (x,y,z), the surrogate is an average of local functions:

$$\mathcal{L}_\lambda(x,y,z) = \frac{1}{n_h} \sum_{i \in \mathcal{H}} \mathcal{L}_{\lambda,i}(x,y,z), \qquad \mathcal{L}_{\lambda,i}(x,y,z) := f_i(x,y) + \lambda\big[g_i(x,y) - g_i(x,z)\big]. \tag{9}$$

Each honest client can therefore estimate the three partial gradients of $\mathcal{L}_{\lambda,i}$ using only stochastic gradients of F_i and G_i. Since $\nabla_z \mathcal{L}_{\lambda,i} = -\lambda\nabla_y g_i(x,z)$, ascent in z amounts to descent on $g_i(x,z)$, with λ absorbed into the stepsize. The server estimates the honest mean of each direction by robustly aggregating the client reports, since the honest set $\mathcal{H}$ is unknown.

Algorithm 1 FOBFBL and FOBFBL−M

Require: (x^0, y^0, z^0), $\lambda > 0$, $\{\eta_v > 0,\ \beta_v \in [0,1)\}_{v\in\{x,y,z\}}$, K, and robust aggregator $\mathcal{A}$

Each honest client initializes $m_{v,i}^{-1} \leftarrow 0$ for $v \in \{x, y, z\}$.
for $k = 0, 1, \ldots, K-1$ **do**
  Server broadcasts (x^k, y^k, z^k) to all clients.
  for each honest $i \in \mathcal{H}$ **in parallel do**
    Draw independent $\omega_{f,x,i}^k, \omega_{f,y,i}^k \sim \mathcal{D}_{f_i}$ and $\omega_{g,xy,i}^k, \omega_{g,xz,i}^k, \omega_{g,y,i}^k, \omega_{g,z,i}^k \sim \mathcal{D}_{g_i}$.
    Form $(h_{x,i}^k, h_{y,i}^k, h_{z,i}^k)$ by (10).
    $m_{v,i}^k \leftarrow \beta_v m_{v,i}^{k-1} + (1-\beta_v) h_{v,i}^k$ for $v \in \{x, y, z\}$ ▷ $\beta_v = 0$ for FOBFBL
    Send $(m_{x,i}^k, m_{y,i}^k, m_{z,i}^k)$ to the server.
  end for
  Byzantine clients may transmit arbitrary blockwise messages.
  $\widehat{m}_v^k \leftarrow \mathcal{A}(\{m_{v,i}^k\}_{i\in[n]})$ for $v \in \{x, y, z\}$.
  $v^{k+1} \leftarrow v^k - \eta_v \widehat{m}_v^k$ for $v \in \{x, y, z\}$.
end for
Draw $R \sim \mathrm{Unif}\{0, \ldots, K-1\}$ independently of all preceding algorithmic randomness.
return x^R for the nonconvex guarantee, or x^K under the PL condition.

3.2 The Proposed Algorithms

Based on the first-order reformulation, we propose FOBFBL and its momentum variant FOBFBL−M, summarized in Algorithm 1. The global iterates x^k, y^k, and z^k represent the upper-level variable and approximations to $y_\lambda^\star(x^k)$ and $y^\star(x^k)$, respectively. The two lower-level sequences track distinct targets while sharing the same upper iterate.

At round k, the server broadcasts (x^k, y^k, z^k) to the clients. Each honest client $i \in \mathcal{H}$ draws independent local samples and computes the stochastic directions

$$\begin{aligned} h_{x,i}^k &= \nabla_x F_i(x^k, y^k) + \lambda\big[\nabla_x G_i(x^k, y^k) - \nabla_x G_i(x^k, z^k)\big], \\ h_{y,i}^k &= \nabla_y F_i(x^k, y^k) + \lambda \nabla_y G_i(x^k, y^k), \qquad h_{z,i}^k = \nabla_y G_i(x^k, z^k), \end{aligned} \tag{10}$$

where $h_{x,i}^k$ estimates $\nabla_x \mathcal{L}_{\lambda,i}$, while $h_{y,i}^k$ and $h_{z,i}^k$ are stochastic gradients of the local penalized and original lower-level objectives, respectively. The x correction uses both lower iterates, whose estimates are maintained by the y and z directions. The sample arguments are suppressed; Algorithm 1 specifies the independent oracle samples. All directions are computed before the server update.

In FOBFBL−M, each honest client smooths the three directions using Polyak's heavy-ball momentum:

$$m_{v,i}^k = \beta_v m_{v,i}^{k-1} + (1-\beta_v) h_{v,i}^k, \qquad m_{v,i}^{-1} = 0, \qquad v \in \{x, y, z\}, \tag{11}$$

where $\beta_x, \beta_y, \beta_z \in [0, 1)$ are the momentum parameters. Setting all three coefficients to zero gives $m_{v,i}^k = h_{v,i}^k$ and recovers FOBFBL. Momentum remains local and adds no transmitted vectors.

Each honest client sends $(m_{x,i}^k, m_{y,i}^k, m_{z,i}^k)$ to the server; Byzantine clients may send arbitrary vectors. All clients submit their messages before observing the current round's honest reports. The server then robustly aggregates each block and updates the global iterates:

$$\widehat{m}_v^k = \mathcal{A}\big(\{m_{v,i}^k\}_{i\in[n]}\big), \qquad v^{k+1} = v^k - \eta_v \widehat{m}_v^k, \qquad v \in \{x, y, z\}, \tag{12}$$

where $\eta_x, \eta_y, \eta_z > 0$ are the stepsizes and $\mathcal{A}$ satisfies the robustness criterion in Definition 2. The updated triple is broadcast at the beginning of the next round. This blockwise aggregation protects both lower-level tracking sequences as well as the upper update. Both algorithms update all three variables once per round in a single loop, using only the stochastic first-order directions in (10).

4 Convergence Analysis

4.1 Assumptions

We now state the regularity and stochastic assumptions used throughout the analysis.

Assumption 1 (Smoothness). *For every $i \in \mathcal{H}$, ∇f_i and ∇g_i are respectively $l_{f,1}$- and $l_{g,1}$-Lipschitz in (x, y). Moreover, $\|\nabla_y f_i(x, y)\| \le l_{f,0}$ for all (x, y).*

Assumption 2 (μ_g-strong convexity). *For every $i \in \mathcal{H}$ and $x \in \mathbb{R}^{d_x}$, $g_i(x, \cdot)$ is μ_g-strongly convex.*

Assumption 3 (Bounded stochastic noise). *For every $i \in \mathcal{H}$, the six stochastic block oracles in (10) are conditionally unbiased and independent across blocks, honest clients, and rounds, with conditional mean-square errors bounded by σ_f^2 for the F_i-oracles and σ_g^2 for the G_i-oracles at their respective query points.*

Assumption 4 (Bounded gradient dissimilarity). *For some $\zeta_f, \zeta_g \ge 0$, every $i \in \mathcal{H}$ and (x, y) satisfy*

$$\|\nabla f_i(x, y) - \nabla f(x, y)\|^2 \le \zeta_f^2, \qquad \|\nabla g_i(x, y) - \nabla g(x, y)\|^2 \le \zeta_g^2, \tag{13}$$

where $f = n_h^{-1} \sum_{i\in\mathcal{H}} f_i$ and $g = n_h^{-1} \sum_{i\in\mathcal{H}} g_i$.

Assumption 5 (Second-order regularity). *The twice continuously differentiable honest averages f and g have respectively $l_{f,2}$- and $l_{g,2}$-Lipschitz Hessians in (x, y), and $\Phi_* := \inf_x \Phi(x) > -\infty$.*

Assumption 6 (Upper-level PL condition). *There exists a constant $\mu_{\mathrm{PL}} > 0$ such that the honest hyperobjective Φ satisfies the following Polyak–Lojasiewicz (PL) inequality:*

$$\frac{1}{2}\|\nabla\Phi(x)\|^2 \ge \mu_{\mathrm{PL}}\big(\Phi(x) - \Phi_*\big), \qquad x \in \mathbb{R}^{d_x}. \tag{14}$$

Following Kwon et al. (2023), Assumptions 1, 2, and 5 ensure regularity of the finite-multiplier formulation and its optimizer maps. These assumptions are needed only for analysis and require no second-order oracle evaluations. Assumption 6 is used only for the last-iterate guarantees in Theorems 2 and 4.

4.2 Convergence Results

We next present nonasymptotic convergence guarantees for `FOBFBL` and its momentum variant, `FOBFBL-M`. Complete proofs are deferred to Appendices D and E. We use $\Sigma_\lambda := \sigma_f^2 + \lambda^2\sigma_g^2$ for the aggregate stochastic-noise scale.

Theorem 1 (Convergence of `FOBFBL` for nonconvex upper-level objectives). *Suppose Assumptions 1–5 hold and $\mathcal{A}$ satisfies Definition 2 almost surely. Fix $\lambda \ge 2l_{f,1}/\mu_g$, assume $\Sigma_\lambda > 0$, and let $\Delta_0 > 0$ denote the initial Lyapunov gap defined in Appendix D. For sufficiently large K, choose $\eta = \sqrt{n_h\Delta_0/(K\Sigma_\lambda)}$, set $\eta_x = \eta$, $\eta_y = c_y^\circ\eta_x$, and $\eta_z = c_z^\circ\eta_x$ for the positive problem-dependent constants c_y°, c_z°, and take $\beta_x = \beta_y = \beta_z = 0$. With deterministic initialization, x^R satisfies*

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 = \mathcal{O}\Bigg(\underbrace{\sqrt{\frac{\Delta_0\Sigma_\lambda}{n_hK}}}_{\text{finite-horizon error}} + \underbrace{\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\mathrm{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}} \Bigg). \tag{15}$$

Theorem 2 (Convergence of `FOBFBL` under the PL condition). *Under the same assumptions on the problem, aggregation rule, and multiplier as in Theorem 1, suppose Assumption 6 also holds. There exists $\gamma_{\mathrm{F}} > 0$ such that, for sufficiently large K, choosing $\eta = \log(n_hK)/(\gamma_{\mathrm{F}}K)$ with the stepsize ratios and zero momentum of Theorem 1 gives*

$$\mathbb{E}[\Phi(x^K) - \Phi_*] = \mathcal{O}\Bigg(\underbrace{\frac{\Delta_0}{n_hK} + \frac{\Sigma_\lambda\log(n_hK)}{n_hK}}_{\text{finite-horizon error}} + \underbrace{\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\mathrm{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}} \Bigg). \tag{16}$$

Remark 1 (Finite-horizon and residual errors). The bound in (15) separates terms with different asymptotic behavior. The stochastic finite-horizon term scales as $\mathcal{O}((n_hK)^{-1/2})$, yielding linear speedup in the number of honest clients. At fixed λ, increasing K drives this contribution to zero but does not reduce the Byzantine-induced or finite-multiplier terms. Under the PL condition, Theorem 2 replaces random-iterate stationarity with last-iterate function-gap convergence at $\widetilde{\mathcal{O}}((n_hK)^{-1})$ finite-horizon dependence, again yielding linear speedup in the number of honest clients without changing the residual structure. When $\kappa = \mathcal{O}(\delta)$, the Byzantine-induced error contains the persistent component $\mathcal{O}(\delta\Sigma_\lambda)$. This component arises because robust aggregation is governed by dispersion across individual client messages rather than the variance of their average. Theorem 3 shows that momentum removes this variance contribution from the nonvanishing Byzantine error.

Remark 2 (Finite-multiplier tradeoff)**.** Theorems 1 and 3 keep λ fixed. Increasing λ reduces the approximation bias $C_{\mathrm{lag}}^2/\lambda^2$, but increases the lower-level noise and heterogeneity contributions through $\lambda^2\sigma_g^2$ and $\lambda^2\zeta_g^2$ and tightens the admissible stepsize conditions. The multiplier therefore balances approximation accuracy against sensitivity to lower-level stochasticity and heterogeneity. In particular, increasing λ need not decrease the total residual. A time-varying multiplier would introduce an additional drift term that is not controlled by the present Lyapunov analysis.

Theorem 3 (Convergence of FOBFBL-M for nonconvex upper-level objectives)**.** *Suppose Assumptions 1–4 hold, g has an $l_{g,2}$-Lipschitz Hessian, $\Phi_* > -\infty$, and $\mathcal{A}$ satisfies Definition 2 almost surely. Fix $\lambda \geq 2l_{f,1}/\mu_g$ and assume $\Sigma_\lambda > 0$. Define the effective aggregation factor $\chi_\kappa := 1 + n_h\kappa$ and let $\overline{\Delta}_0^{\mathrm{m}} > 0$ denote the uniform initial-gap bound in Appendix E. For sufficiently large K, choose $\eta = \sqrt{n_h\overline{\Delta}_0^{\mathrm{m}}/(K\chi_\kappa\Sigma_\lambda)}$, $\eta_v = c_v^\circ\eta$, and $\beta_v = \sqrt{1-\rho_v\eta}$ for $v \in \{x, y, z\}$, where $c_x^\circ = 1$ and $c_v^\circ, \rho_v > 0$ are problem-dependent constants. With deterministic initialization, x^R satisfies*

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 = \mathcal{O}\Bigg(\underbrace{\sqrt{\frac{\overline{\Delta}_0^{\mathrm{m}}\chi_\kappa\Sigma_\lambda}{n_hK}}}_{\text{finite-horizon error}} + \underbrace{\kappa(\zeta_f^2 + \lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\mathrm{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}}\Bigg). \tag{17}$$

Theorem 4 (Convergence of FOBFBL-M under the PL condition)**.** *Under the same assumptions on the problem, aggregation rule, and multiplier as in Theorem 3, suppose Assumption 6 also holds. There exists $\gamma_{\mathrm{M}} > 0$ such that, for sufficiently large K, choosing $\eta = \log(n_hK)/(\gamma_{\mathrm{M}}K)$ with the stepsize ratios and momentum parameters of Theorem 3 gives*

$$\mathbb{E}[\Phi(x^K) - \Phi_*] = \mathcal{O}\Bigg(\underbrace{\frac{\overline{\Delta}_0^{\mathrm{m}}}{n_hK} + \frac{\chi_\kappa\Sigma_\lambda\log(n_hK)}{n_hK}}_{\text{finite-horizon error}} + \underbrace{\kappa(\zeta_f^2 + \lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\mathrm{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}}\Bigg). \tag{18}$$

Remark 3 (Momentum and the Byzantine-induced error)**.** Theorems 1 and 3 both yield finite-horizon stationarity errors of order $\mathcal{O}(K^{-1/2})$. Their distinction lies in the nonvanishing Byzantine term. For FOBFBL-M, stochastic noise enters only through $\chi_\kappa\Sigma_\lambda/n_h$ in the finite-horizon term, so its contribution vanishes as K grows. The definition $\chi_\kappa = 1+n_h\kappa$ shows that linear speedup in n_h holds when $\kappa = \mathcal{O}(1/n_h)$, including the Byzantine-free case. The remaining Byzantine-induced error, $\kappa(\zeta_f^2 + \lambda^2\zeta_g^2)$, depends on honest-client heterogeneity but not on stochastic-gradient variance. This recovers, for a fully first-order finite-multiplier method, the variance-removal property established for BR-FedBiM by Li et al. (2026). When $\kappa = \mathcal{O}(\delta)$ and λ is fixed, this term has the same dependence on the Byzantine fraction and honest-client heterogeneity as their lower bound $\Omega(\delta(\zeta_f^2 + \zeta_g^2))$. This comparison concerns only Byzantine-induced error. The finite-multiplier bias $C_{\mathrm{lag}}^2/\lambda^2$ remains.

5 Experiments

We evaluate FOBFBL and FOBFBL-M on two federated bilevel tasks: hyperparameter optimization and hyper-representation learning. The main accuracy results use MNIST, while Appendix B.2.2 provides complementary hyper-representation results on CIFAR-10. We compare against FedNest (Tarzanagh et al., 2022), AggITD (Xiao & Ji, 2023), BR-FedBi (Li et al., 2026), and Rob-FedBOB (Maralappanavar & Bharath, 2025) under ALIE (Baruch et al., 2019), Sign-Flipping (SF), IPM (Xie et al., 2020), and Random Noise (RN) attacks. We use Krum (Blanchard et al., 2017), geometric median (Median), Trimmed Mean (Yin et al., 2018), and coordinatewise median (CWMed). Main comparisons report mean test accuracy with one standard deviation over three seeds; ablations use seed 42. Appendix B gives the supporting accuracy results and efficiency measurement settings.

5.1 Main Results

Federated hyperparameter optimization. We formulate this task as the following bilevel problem:

$$\min_w \frac{1}{n_h}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{D}_{\mathrm{val}}^i|}\sum_{\phi\in\mathcal{D}_{\mathrm{val}}^i} L\big(\theta^\star(w);\phi\big),\ \text{s.t.}\ \theta^\star(w) = \operatorname*{argmin}_\theta \frac{1}{n_h}\sum_{i\in\mathcal{H}}\frac{1}{|\mathcal{D}_{\mathrm{tr}}^i|}\sum_{\varphi\in\mathcal{D}_{\mathrm{tr}}^i} L(\theta;\varphi) + \mathcal{R}(w,\theta).$$

Here w and θ correspond to the upper- and lower-level variables x and y, respectively, while L and $\mathcal{R}$ denote the sample loss and ℓ_2 regularizer. We evaluate multinomial logistic regression on

Table 2: **Hyperparameter optimization on MNIST.** Test accuracy (%, mean ± SD over three seeds). "Worst" is the minimum mean across attacks. Best and second-best results within each aggregator are bold and underlined. Shaded rows denote the proposed methods.

Krum

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$67.40_{\pm 2.53}$	$62.54_{\pm 2.42}$	$59.82_{\pm 1.59}$	$65.96_{\pm 1.65}$	59.82
AggITD	$67.63_{\pm 0.26}$	$66.14_{\pm 0.22}$	$61.78_{\pm 2.05}$	$65.99_{\pm 1.52}$	61.78
BR-FedBi	$66.84_{\pm 4.78}$	$62.61_{\pm 6.29}$	$58.66_{\pm 8.59}$	$63.04_{\pm 1.37}$	58.66
Rob-FedBOB	$\mathbf{69.78_{\pm 2.06}}$	$62.67_{\pm 3.22}$	$54.65_{\pm 7.38}$	$\mathbf{67.99_{\pm 0.51}}$	54.65
FOBFBL	$69.26_{\pm 3.51}$	$\mathbf{66.19_{\pm 1.48}}$	$\mathbf{66.26_{\pm 2.31}}$	$66.50_{\pm 1.56}$	**66.19**
FOBFBL−M	$67.29_{\pm 3.47}$	$65.99_{\pm 1.55}$	$65.94_{\pm 2.93}$	$66.71_{\pm 0.97}$	65.94

Median

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$54.74_{\pm 5.88}$	$62.58_{\pm 3.16}$	$59.42_{\pm 3.13}$	$67.66_{\pm 2.87}$	54.74
AggITD	$59.41_{\pm 5.53}$	$64.39_{\pm 2.25}$	$62.79_{\pm 3.17}$	$66.33_{\pm 1.59}$	59.41
BR-FedBi	$62.47_{\pm 9.84}$	$62.96_{\pm 6.06}$	$56.03_{\pm 5.06}$	$66.98_{\pm 4.59}$	56.03
Rob-FedBOB	$70.55_{\pm 2.83}$	$62.12_{\pm 4.42}$	$54.25_{\pm 8.83}$	$\mathbf{68.98_{\pm 1.54}}$	54.25
FOBFBL	$\mathbf{72.79_{\pm 1.07}}$	$67.72_{\pm 2.13}$	$65.09_{\pm 1.88}$	$68.77_{\pm 3.30}$	65.09
FOBFBL−M	$71.57_{\pm 1.92}$	$\mathbf{67.95_{\pm 2.82}}$	$\mathbf{66.20_{\pm 2.66}}$	$68.48_{\pm 4.52}$	**66.20**

Trimmed Mean

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$64.23_{\pm 0.40}$	$67.35_{\pm 0.87}$	$63.43_{\pm 2.86}$	$71.41_{\pm 0.45}$	63.43
AggITD	$65.99_{\pm 3.56}$	$\mathbf{69.49_{\pm 0.56}}$	$\mathbf{69.68_{\pm 0.33}}$	$71.64_{\pm 0.57}$	**65.99**
BR-FedBi	$65.01_{\pm 1.49}$	$56.30_{\pm 3.30}$	$52.62_{\pm 0.41}$	$65.50_{\pm 1.74}$	52.62
Rob-FedBOB	$71.11_{\pm 0.72}$	$56.29_{\pm 2.24}$	$46.29_{\pm 2.05}$	$69.45_{\pm 0.17}$	46.29
FOBFBL	$\mathbf{73.05_{\pm 1.04}}$	$64.52_{\pm 2.81}$	$61.58_{\pm 1.85}$	$71.73_{\pm 0.69}$	61.58
FOBFBL−M	$72.71_{\pm 1.20}$	$64.65_{\pm 3.00}$	$62.75_{\pm 2.51}$	$\mathbf{71.89_{\pm 0.50}}$	62.75

CWMed

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$64.03_{\pm 0.17}$	$56.12_{\pm 1.33}$	$55.35_{\pm 1.18}$	$70.85_{\pm 0.42}$	55.35
AggITD	$69.45_{\pm 1.29}$	$58.02_{\pm 1.12}$	$58.37_{\pm 0.77}$	$71.27_{\pm 0.65}$	58.02
BR-FedBi	$33.51_{\pm 6.61}$	$57.04_{\pm 1.03}$	$56.64_{\pm 1.12}$	$66.18_{\pm 0.79}$	33.51
Rob-FedBOB	$67.92_{\pm 0.63}$	$50.13_{\pm 1.10}$	$49.80_{\pm 1.70}$	$67.81_{\pm 0.79}$	49.80
FOBFBL	$76.16_{\pm 0.97}$	$\mathbf{64.01_{\pm 1.21}}$	$62.97_{\pm 1.31}$	$71.81_{\pm 0.38}$	62.97
FOBFBL−M	$\mathbf{76.55_{\pm 0.96}}$	$63.75_{\pm 1.34}$	$\mathbf{64.58_{\pm 2.34}}$	$\mathbf{71.88_{\pm 0.66}}$	**63.75**

Table 3: **Hyper-representation learning on MNIST.** Test accuracy (%, mean ± SD over three seeds). "Worst" is the minimum mean across attacks. Best and second-best results within each aggregator are bold and underlined. Shaded rows denote the proposed methods.

Krum

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$58.40_{\pm 4.16}$	$55.80_{\pm 10.00}$	$67.94_{\pm 14.43}$	$58.73_{\pm 7.52}$	55.80
AggITD	$82.68_{\pm 3.65}$	$\mathbf{87.14_{\pm 4.12}}$	$80.99_{\pm 4.39}$	$66.97_{\pm 4.15}$	66.97
BR-FedBi	$\mathbf{95.93_{\pm 0.36}}$	$80.32_{\pm 6.83}$	$\mathbf{86.59_{\pm 4.89}}$	$\mathbf{96.46_{\pm 0.20}}$	**80.32**
Rob-FedBOB	$53.57_{\pm 29.94}$	$64.92_{\pm 25.99}$	$59.43_{\pm 22.92}$	$94.79_{\pm 0.33}$	53.57
FOBFBL	$82.29_{\pm 3.58}$	$74.76_{\pm 3.16}$	$65.92_{\pm 1.62}$	$78.06_{\pm 5.81}$	65.92
FOBFBL−M	$82.06_{\pm 1.48}$	$72.95_{\pm 3.58}$	$74.12_{\pm 3.45}$	$81.19_{\pm 1.67}$	72.95

Median

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$52.88_{\pm 10.98}$	$72.39_{\pm 13.43}$	$69.05_{\pm 14.05}$	$58.18_{\pm 6.07}$	52.88
AggITD	$78.49_{\pm 3.40}$	$73.39_{\pm 18.59}$	$77.43_{\pm 4.34}$	$75.12_{\pm 12.10}$	73.39
BR-FedBi	$\mathbf{96.65_{\pm 0.14}}$	$\mathbf{90.96_{\pm 1.32}}$	$\mathbf{82.07_{\pm 4.91}}$	$\mathbf{96.17_{\pm 0.49}}$	**82.07**
Rob-FedBOB	$62.54_{\pm 22.64}$	$49.65_{\pm 32.21}$	$58.20_{\pm 22.85}$	$94.21_{\pm 0.17}$	49.65
FOBFBL	$87.09_{\pm 1.56}$	$78.47_{\pm 3.10}$	$76.31_{\pm 2.13}$	$83.73_{\pm 3.44}$	76.31
FOBFBL−M	$84.39_{\pm 3.81}$	$74.42_{\pm 2.19}$	$68.88_{\pm 5.70}$	$77.79_{\pm 0.28}$	68.88

Trimmed Mean

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$71.15_{\pm 2.14}$	$76.85_{\pm 4.24}$	$60.91_{\pm 10.84}$	$75.75_{\pm 1.13}$	60.91
AggITD	$86.16_{\pm 4.12}$	$\mathbf{92.04_{\pm 1.13}}$	$\mathbf{90.66_{\pm 1.13}}$	$81.34_{\pm 4.04}$	81.34
BR-FedBi	$\mathbf{98.12_{\pm 0.14}}$	$91.57_{\pm 4.93}$	$71.48_{\pm 14.08}$	$\mathbf{98.05_{\pm 0.16}}$	71.48
Rob-FedBOB	$85.06_{\pm 15.57}$	$63.55_{\pm 21.19}$	$49.25_{\pm 14.76}$	$95.73_{\pm 0.53}$	49.25
FOBFBL	$84.81_{\pm 6.93}$	$85.68_{\pm 2.33}$	$84.47_{\pm 2.85}$	$89.62_{\pm 1.77}$	**84.47**
FOBFBL−M	$81.91_{\pm 6.24}$	$83.66_{\pm 4.86}$	$76.43_{\pm 7.23}$	$87.66_{\pm 1.55}$	76.43

CWMed

Algorithm	ALIE	SF	IPM	RN	Worst
FedNest	$64.90_{\pm 2.31}$	$19.45_{\pm 3.65}$	$24.34_{\pm 6.69}$	$71.26_{\pm 3.74}$	19.45
AggITD	$65.80_{\pm 5.20}$	$13.99_{\pm 0.73}$	$14.43_{\pm 1.68}$	$88.09_{\pm 1.23}$	13.99
BR-FedBi	$69.41_{\pm 1.89}$	$\mathbf{93.81_{\pm 0.95}}$	$\mathbf{92.53_{\pm 0.88}}$	$\mathbf{98.07_{\pm 0.03}}$	69.41
Rob-FedBOB	$\mathbf{89.44_{\pm 7.80}}$	$58.90_{\pm 4.62}$	$75.00_{\pm 8.02}$	$96.67_{\pm 0.19}$	58.90
FOBFBL	$89.22_{\pm 1.44}$	$84.61_{\pm 2.49}$	$85.68_{\pm 3.51}$	$92.89_{\pm 1.60}$	84.61
FOBFBL−M	$89.10_{\pm 1.98}$	$85.33_{\pm 1.24}$	$85.08_{\pm 3.36}$	$93.15_{\pm 0.81}$	**85.08**

long-tailed MNIST (LeCun et al., 1998) with $n = 25$ clients and $n_b = 7$ Byzantine clients ($\delta = 0.28$). Appendix B.1 gives the training curves.

Federated hyper-representation learning. Following Franceschi et al. (2018), we solve

$$\min_{w} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}^i_{\text{val}}|} \sum_{\phi \in \mathcal{D}^i_{\text{val}}} F_i(w, \theta^\star(w); \phi), \quad \text{s.t.} \quad \theta^\star(w) = \operatorname*{argmin}_{\theta} \frac{1}{n_h} \sum_{i \in \mathcal{H}} \frac{1}{|\mathcal{D}^i_{\text{tr}}|} \sum_{\varphi \in \mathcal{D}^i_{\text{tr}}} G_i(w, \theta; \varphi),$$

where $\mathcal{D}^i_{\text{tr}}$ and $\mathcal{D}^i_{\text{val}}$ are client i's training and validation sets. The final layer is θ, and the remaining parameters form w. We use a two-block CNN with 32 and 64 channels on MNIST (LeCun et al., 1998) and a Dirichlet split with $\alpha_{\text{Dir}} = 0.5$. All $n = 25$ clients participate, including $n_b = 7$ Byzantine clients ($\delta = 0.28$). Training uses a 1000-communication-round budget, batch size 128, and five local stochastic steps per update (one for BR-FedBi). Appendices B.2.1 and B.2.2 give the MNIST training curves and CIFAR-10 results.

Results. Tables 2 and 3 show that our fully first-order methods maintain strong predictive performance under Byzantine attacks on both MNIST tasks. In hyperparameter optimization, FOBFBL and FOBFBL−M achieve the two highest worst-case accuracies with Krum, Median, and CWMed. Both variants outperform all baselines under every attack with CWMed and show clear gains under SF and IPM with Median. In hyper-representation learning, FOBFBL and FOBFBL−M attain the best worst-case accuracy with Trimmed Mean and CWMed, respectively. With CWMed, both variants outperform FedNest and AggITD under SF and IPM, and BR-FedBi under ALIE. Across both tasks and all aggregation rules, our methods outperform the fully first-order Rob-FedBOB baseline under SF and IPM. On CIFAR-10, FOBFBL achieves the highest worst-case accuracy with Krum and Median, while FOBFBL−M does so with CWMed (Appendix B.2.2). Both variants outperform Rob-FedBOB in every tested configuration. These results demonstrate that our methods combine fully first-order updates with Byzantine robustness across datasets, attacks, and aggregation rules.

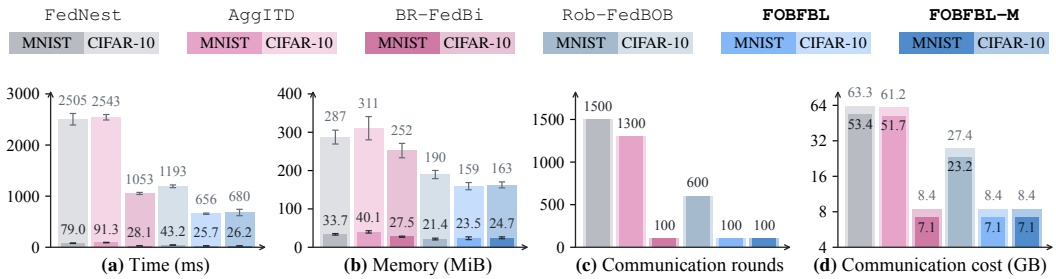

Figure 1: Computation and communication costs on MNIST and CIFAR-10. **(a)** Client time per outer update; **(b)** additional client memory; **(c)** communication rounds; **(d)** total upload and download volume (log scale). Panels (c,d) show expected totals over 100 outer updates with 25 clients.

5.2 Efficiency Analysis

We evaluate whether our fully first-order, single-loop design reduces the cost of each outer update. We examine local computation and client–server communication separately to assess the benefits of the two design choices. Figure 1 summarizes the results, with measurement details in Appendix B.3.

(1) Computation Efficiency. The fully first-order design uses only ordinary gradients and avoids the Hessian-vector and mixed-derivative products required by FedNest, AggITD, and BR-FedBi, simplifying derivative evaluation and its intermediate storage. Panels (a,b) show that `FOBFBL` and `FOBFBL-M` have the lowest measured update times on both datasets. On CIFAR-10, they take 655.8 and 679.7 ms, compared with 2505.0 and 2543.3 ms for FedNest and AggITD, and use less additional peak memory than all baselines (159.3–162.7 MiB versus 189.7–310.5 MiB). Memory use also depends on the model and implementation: Rob-FedBOB uses the least on MNIST. Results average three client/seed pairs per method and dataset; error bars show one standard deviation.

(2) Communication Efficiency. Our single-loop design transmits all three update blocks in one communication round per outer update, comprising client uploads and a server broadcast. In panel (c), `FOBFBL` and `FOBFBL-M` require only 100 communication rounds over 100 outer updates, compared with 600 for Rob-FedBOB, 1300 for AggITD, and an expected 1500 for FedNest (see Appendix B.3.3 for the round-count derivation). In panel (d), `FOBFBL` and `FOBFBL-M` transfer 7.12 GB on MNIST and 8.44 GB on CIFAR-10 over the same updates, counting uploads and downloads across all 25 clients. Both costs match BR-FedBi and are lower than those of the other baselines. The local momentum in `FOBFBL-M` adds no communication cost.

5.3 Ablation Study

We conduct ablation studies on the attack fraction, the number of participating clients, and the heterogeneity level (Appendix B.4). First, as the attack fraction increases, our methods show smaller average accuracy drops than the baselines, with the momentum variant exhibiting the smallest decline. Second, as the number of participating clients increases from 10 to 40, both methods achieve higher accuracy across the tested attacks and aggregation rules. Third, as the heterogeneity level increases through greater class imbalance, both methods maintain competitive accuracy, including under the most severely imbalanced setting. Overall, these results demonstrate the robustness of our methods in heterogeneous federated environments.

6 Conclusion

We proposed `FOBFBL` and its momentum variant `FOBFBL-M`, fully first-order Lagrangian methods for Byzantine-robust federated bilevel learning. The Lagrangian reformulation uses only ordinary gradients. Our single-loop design robustly aggregates three gradient blocks in one client–server exchange per iteration, reducing computational and synchronization overhead. We established convergence guarantees under stochastic noise and client heterogeneity, and showed that momentum removes the stochastic-variance contribution to the persistent Byzantine error without additional communication. Our bounds separate optimization, Byzantine, and finite-multiplier errors. Experiments on hyperparameter optimization and hyper-representation learning demonstrate the effectiveness of both methods across diverse attacks and aggregation rules. Future work includes extending the framework beyond strongly convex lower-level objectives while preserving its fully first-order structure, computational efficiency, and robustness to Byzantine attacks.

AI USE STATEMENT

We used generative AI tools only to edit the manuscript to improve readability, and not for any task with required disclosure. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Momin Abbas, Yi Zhou, Nathalie Baracaldo, Horst Samulowitz, Parikshit Ram, and Theodoros Salonidis. Byzantine-resilient bilevel federated learning. In *2024 IEEE 13th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 1–5. IEEE, 2024. doi: 10.1109/SAM60225.2024.10636694.

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.

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, Royal Institute of Technology, Stockholm, Sweden, 2003.

Moran Baruch, Gilad 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. doi: 10.1162/089976600300015187.

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.

Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109):1–56, 2025. URL https://jmlr.org/papers/v26/23-1104.html.

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.

Tianshu Chu, Dachuan Xu, Wei Yao, and Jin Zhang. SPABA: A single-loop and probabilistic stochastic bilevel algorithm achieving optimal sample complexity. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 8848–8903. PMLR, 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.

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.

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.

Eduard Gorbunov, Samuel Horváth, Peter Richtárik, and Gauthier Gidel. Variance reduction is an antidote to Byzantine workers: Better rates, weaker assumptions and communication compression as a cherry on the top. In *The Eleventh International Conference on Learning Representations*, 2023.

Baoru Han, Rutvij H Jhaveri, Han Wang, Dawei Qiao, and Jinglong Du. Application of robust zero-watermarking scheme based on federated learning for securing the healthcare data. *IEEE Journal of Biomedical and Health Informatics*, 27(2):804–813, 2023. doi: 10.1109/JBHI.2021.3123936.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pp. 4882–4892. PMLR, 2021.

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.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=jXKKDEi5vJt.

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.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34:30271–30283, 2021.

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

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits, 1998. URL http://yann.lecun.com/exdb/mnist/.

Junyi Li, Feihu Huang, and Heng Huang. Communication-efficient federated bilevel optimization with global and local lower level problems. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1326–1338, 2023. doi: 10.52202/075280-0065.

Yangnan Li, Shenghui Song, and Xuanyu Cao. Single-loop Byzantine-resilient federated bilevel optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. BOME! Bilevel Optimization Made Easy: A Simple First-Order Approach. *Advances in Neural Information Processing Systems*, 35: 17248–17262, 2022.

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, 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, 2022.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024. doi: 10.1137/23M1566753.

Shruti P Maralappanavar and BN Bharath. Byzantine-robust and Hessian-free federated bilevel optimization. *Transactions on Machine Learning Research*, 2025.

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.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

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.

Keith Rush, Zachary Charles, and Zachary Garrett. Federated automatic differentiation. *Journal of Machine Learning Research*, 25(357):1–39, 2024. URL https://jmlr.org/papers/v25/23-0223.html.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pp. 30992–31015. PMLR, 2023.

Qiankun Shi, Jie Peng, Kun Yuan, Xiao Wang, and Qing Ling. Optimal complexity in Byzantine-robust distributed stochastic optimization with data heterogeneity. *Journal of Machine Learning Research*, 26(268):1–58, 2025. URL https://jmlr.org/papers/v26/25-0613.html.

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

Xiaoyu Wang, Rui Pan, Renjie Pi, and Jipeng Zhang. Effective bilevel optimization via minimax reformulation. *arXiv preprint arXiv:2305.13153*, 2023.

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.

Peiyao Xiao and Kaiyi Ji. Communication-efficient federated hypergradient computation via aggregated iterative differentiation. In *International Conference on Machine Learning*, pp. 38059–38086. PMLR, 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.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in Hessian/Jacobian-free stochastic bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. doi: 10.52202/075280-1715.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. SimFBO: Towards simple, flexible and communication-efficient federated bilevel learning. *Advances in Neural Information Processing Systems*, 36, 2023b. doi: 10.52202/075280-1433.

Yifan Yang, Peiyao Xiao, Shiqian Ma, and Kaiyi Ji. First-Order Federated Bilevel Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22029–22037, 2025. doi: 10.1609/aaai.v39i21.34355.

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

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 "Fully First-Order Lagrangian Methods for Byzantine-Robust Federated Bilevel Learning"

CONTENTS

1 Introduction **1**
1.1 Related Work . . . 2

2 Preliminaries **3**

3 Methodology **4**
3.1 First-Order Reformulation . . . 4
3.2 The Proposed Algorithms . . . 5

4 Convergence Analysis **5**
4.1 Assumptions . . . 5
4.2 Convergence Results . . . 6

5 Experiments **7**
5.1 Main Results . . . 7
5.2 Efficiency Analysis . . . 9
5.3 Ablation Study . . . 9

6 Conclusion **9**

Appendix

A Additional Related Work **15**

B Experiments **16**
B.1 Federated Hyperparameter Optimization . . . 16
B.2 Federated Hyper-Representation Learning . . . 17
B.3 Efficiency Analysis . . . 19
B.4 Ablation Study . . . 23

C Technical Lemmas **26**

D Convergence Analysis of `FOBFBL` **33**
D.1 Auxiliary Lemmas . . . 33
D.2 Proof of Theorem 1: Nonconvex Convergence . . . 35
D.3 Proof of Theorem 2: Convergence under the PL Condition . . . 38
D.4 Proofs of the Auxiliary Lemmas . . . 40

E Convergence Analysis of `FOBFBL-M` **46**
E.1 Auxiliary Lemmas . . . 46
E.2 Proof of Theorem 3: Nonconvex Convergence . . . 48
E.3 Proof of Theorem 4: Convergence under the PL Condition . . . 52
E.4 Proofs of the Auxiliary Lemmas . . . 55

A Additional Related Work

Bilevel optimization. Computing an upper-level update requires accounting for the dependence of the lower-level solution on the upper variable. Hypergradient methods do so by differentiating through inner iterations or approximating implicit differentiation (Ji et al., 2021). One line of work reduces nested computation by updating the upper variable, lower solution, and an auxiliary variable jointly, as in SOBA and SABA (Dagréou et al., 2022). Momentum and variance reduction further improve stochastic sample complexity, as illustrated by SUSTAIN and SPABA (Khanduri et al., 2021; Chu et al., 2024); related analyses extend complexity guarantees to relaxed smoothness conditions (Chen et al., 2024). These methods retain Hessian or Jacobian vector products. Fully first-order methods address this derivative cost through value-function, Lagrangian, penalty, and finite-difference formulations (Liu et al., 2022; Kwon et al., 2023; Wang et al., 2023; Shen & Chen, 2023; Yang et al., 2023a). This line includes near-optimal deterministic oracle complexity via two-timescale updates (Chen et al., 2025) and penalty methods based on minimax subproblems for convex lower-level objectives (Lu & Mei, 2024). Thus, simplifying the update structure and eliminating higher-order derivative products are distinct aspects of bilevel algorithm design.

Federated bilevel optimization. Distributing a bilevel problem introduces communication costs and client heterogeneity into hypergradient estimation. Federated automatic differentiation provides a general framework for computing derivatives through federated programs (Rush et al., 2024). FedNest and AggITD use nested updates, whereas SimFBO uses a single-loop recursion; all three require Hessian or Jacobian vector products (Tarzanagh et al., 2022; Xiao & Ji, 2023; Yang et al., 2023b). FedBiOAcc and its local variant improve communication efficiency through momentum-based variance reduction for global and client-specific lower-level problems, respectively (Li et al., 2023). Related decentralized methods address coordination across nodes through loopless tracking (Niu et al., 2025), augmented-Lagrangian consensus in SLAM (Lu et al., 2022), and primal-dual heterogeneity correction in SPARKLE (Zhu et al., 2024). For fully first-order FBO, MemFBO uses a finite-multiplier Lagrangian with simultaneous ordinary gradient updates (Yang et al., 2025). It provides the formulation closest to ours, but assumes honest communication. Protecting the coupled updates against arbitrary client reports requires additional aggregation mechanisms and error analysis.

Byzantine-robust optimization. In single-level learning, robust aggregation limits the influence of arbitrary client reports. Representative rules include Krum, coordinatewise median, and geometric median (Blanchard et al., 2017; Yin et al., 2018; Acharya et al., 2022), with RFA applying an approximate geometric median to federated model updates (Pillutla et al., 2022). Honest-client heterogeneity motivates preprocessing, mixing, history-based aggregation, and the gradient splitting used by GAS (Karimireddy et al., 2022; Allouah et al., 2023; Karimireddy et al., 2021; Wu et al., 2023; Liu et al., 2023). Stochastic noise motivates a complementary set of techniques: RESAM combines worker-side momentum with robust averaging (Farhadkhani et al., 2022), while Byz-VR-MARINA uses variance reduction and compressed communication for nonconvex and PL objectives (Gorbunov et al., 2023). Minimax analyses characterize the attainable robustness under uniform gradient heterogeneity (Murata et al., 2024) and the fundamental limits on Byzantine error and stochastic oracle complexity (Shi et al., 2025).

In bilevel learning, robustness additionally requires controlling how errors in the lower-level updates affect the upper direction. BILANTINE robustly aggregates an approximate implicit hypergradient (Abbas et al., 2024), while BR-FedBi uses an auxiliary recursion and BR-FedBiM adds Polyak momentum to remove stochastic variance from the asymptotic Byzantine error (Li et al., 2026). These methods retain Hessian or Jacobian vector products. Rob-FedBOB avoids these products and permits PL lower-level objectives, but uses multiple robust inner updates per outer round (Maralappanavar & Bharath, 2025). Its lower-level PL assumption differs from the upper-level PL condition used for our last-iterate guarantees. `FOBFBL` combines fully first-order computation with one client–server exchange by robustly aggregating three ordinary gradient blocks. Our analysis therefore controls the upper-direction error and both lower-level tracking errors jointly. `FOBFBL-M` extends the variance-removal effect to this finite-multiplier formulation, with a limiting neighborhood governed by honest heterogeneity and finite-multiplier bias.

B EXPERIMENTS

B.1 FEDERATED HYPERPARAMETER OPTIMIZATION

Figures 2–5 show long-tailed MNIST HPO accuracy. These curves supplement Section 5.1; the separate efficiency measurements are detailed in Appendix B.3. Each figure fixes one aggregator, with panels for ALIE, SF, IPM, and RN. Every point shows the mean over all three seeds; shading is population SD (divisor 3). Curves end at each method's last feasible communication round. The MLP lower problem is nonconvex, so these HPO experiments extend beyond the strongly convex lower-level setting of our theory.

Relative performance and the benefit of momentum depend on the aggregator and attack. For example, under ALIE with CWMed, `FOBFBL` and `FOBFBL-M` reach 76.16% and 76.55%, respectively, matching Table 2.

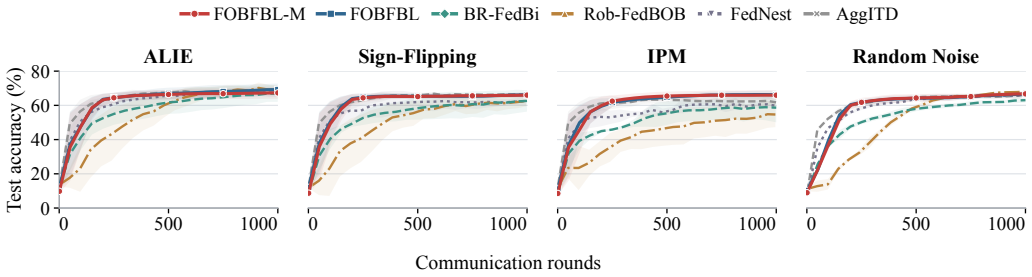

Figure 2: MNIST HPO with **Krum** (mean ±1 SD over three seeds).

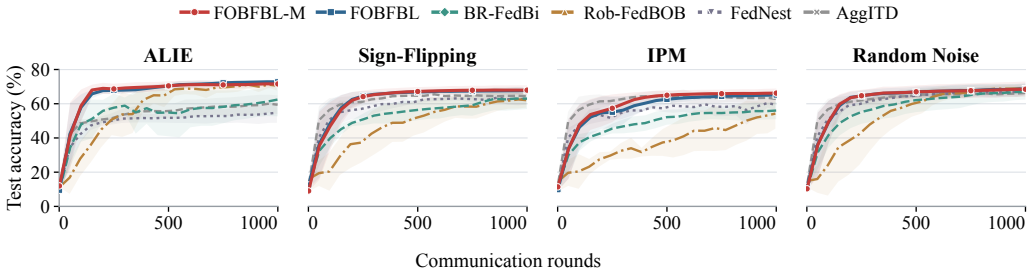

Figure 3: MNIST HPO with **Geometric Median** (mean ±1 SD over three seeds).

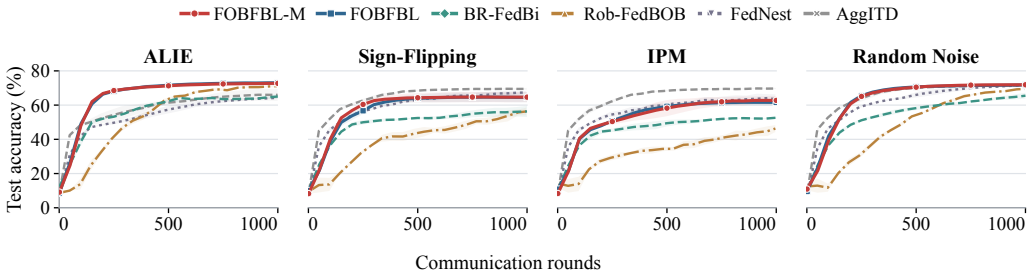

Figure 4: MNIST HPO with **Trimmed Mean** (mean ±1 SD over three seeds).

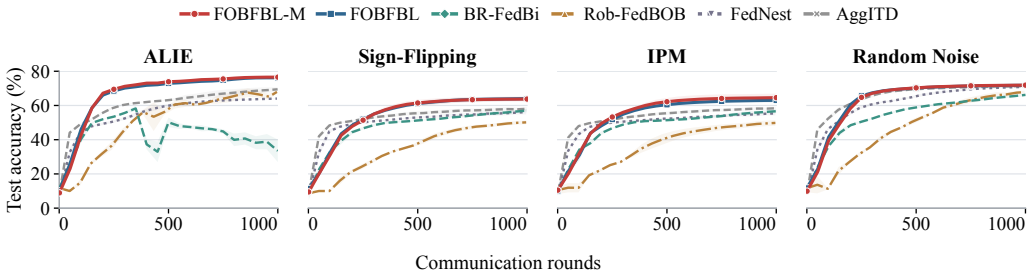

Figure 5: MNIST HPO with **CWMed** (mean ±1 SD over three seeds).

B.2 Federated Hyper-Representation Learning

B.2.1 MNIST

Figures 6–9 give the training curves corresponding to the MNIST results in Table 3. The layout is the same as in Appendix B.1: one figure per aggregator and one panel per attack. Curves and shaded regions are the mean and one standard deviation over three seeds. The BR-FedBi trajectories use one local stochastic step per round, matching Table 3. The remaining methods use five local stochastic steps. All methods use population SD (divisor 3).

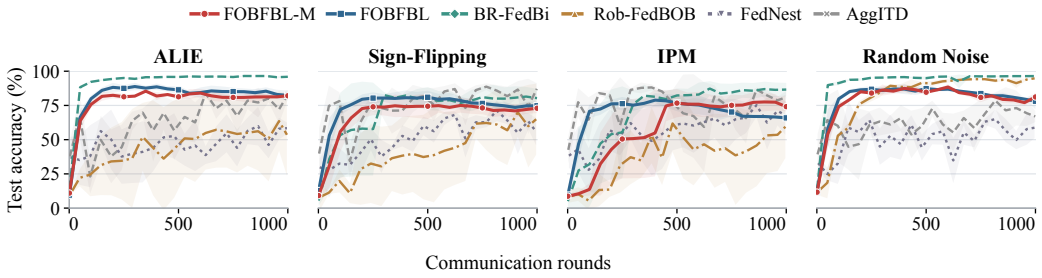

Figure 6: MNIST hyper-representation learning with **Krum** (mean ± 1 SD).

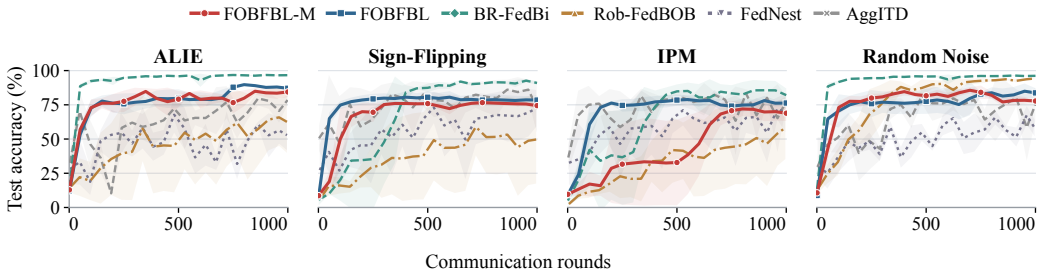

Figure 7: MNIST hyper-representation learning with **Geometric Median** (mean ± 1 SD).

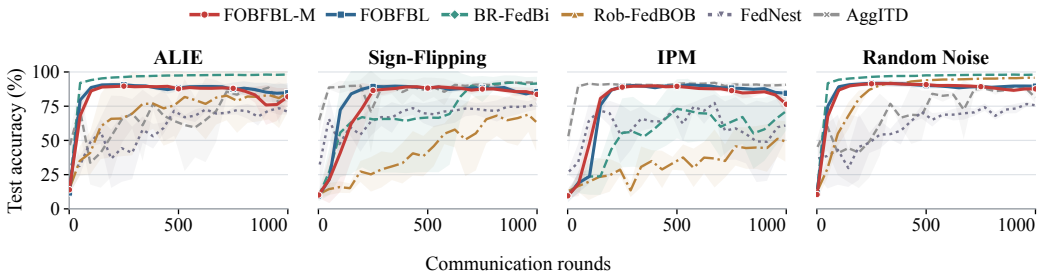

Figure 8: MNIST hyper-representation learning with **Trimmed Mean** (mean ± 1 SD).

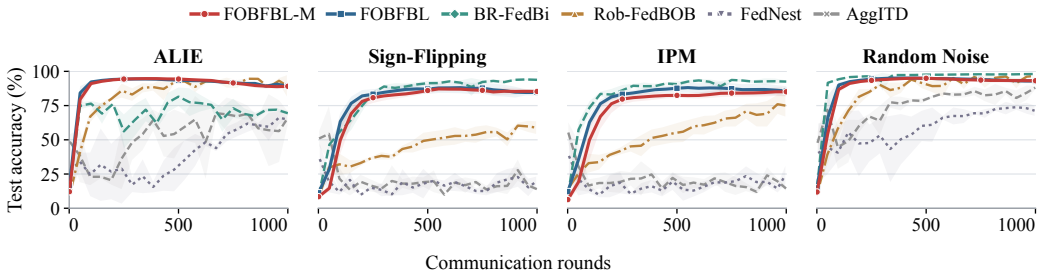

Figure 9: MNIST hyper-representation learning with **CWMed** (mean ± 1 SD).

B.2.2 CIFAR-10

Table 4 gives the complementary CIFAR-10 hyper-representation results omitted from the main text. We use a CNN with three convolutional blocks, $n = 25$, $n_b = 7$, a 1000-communication-round budget, five local stochastic steps, and three seeds. Among fully first-order single-loop methods, FOBFBL and FOBFBL-M have the strongest worst-case performance. They lead with Krum (34.80%) and Median (32.04%), and take the top two positions with CWMed (29.50% and 27.99%). FedNest stays near chance in most settings; AggITD reaches 26.70% under IPM with Trimmed Mean. Rob-FedBOB is weaker under adaptive attacks. BR-FedBi leads with Trimmed Mean (30.24%), but requires products of a Hessian or Jacobian with a vector.

Table 4: **Hyper-representation learning on CIFAR-10.** Test accuracy (%, mean ± SD over three seeds). "Worst" is the minimum mean across attacks. Best and second-best results within each aggregator are bold and underlined. Shaded rows denote the proposed methods.

Krum					
Algorithm	ALIE	SF	IPM	RN	**Worst**
FedNest	$12.80_{\pm 0.55}$	$14.78_{\pm 0.67}$	$14.26_{\pm 5.62}$	$11.22_{\pm 0.96}$	11.22
AggITD	$18.81_{\pm 4.95}$	$14.98_{\pm 4.30}$	$16.92_{\pm 5.27}$	$16.83_{\pm 5.58}$	14.98
BR-FedBi	$\mathbf{43.37_{\pm 2.10}}$	$\mathbf{37.78_{\pm 1.87}}$	$26.02_{\pm 11.33}$	$\mathbf{40.59_{\pm 2.49}}$	<u>26.02</u>
Rob-FedBOB	$29.13_{\pm 3.62}$	$17.55_{\pm 1.34}$	$17.80_{\pm 3.08}$	$25.97_{\pm 2.01}$	17.55
FOBFBL	<u>$38.99_{\pm 1.84}$</u>	<u>$35.77_{\pm 3.24}$</u>	$\mathbf{34.80_{\pm 1.56}}$	$35.72_{\pm 1.06}$	**34.80**
FOBFBL-M	$38.57_{\pm 2.11}$	$19.65_{\pm 11.97}$	<u>$34.54_{\pm 4.64}$</u>	<u>$35.94_{\pm 3.15}$</u>	19.65
Median					
Algorithm	ALIE	SF	IPM	RN	**Worst**
FedNest	$13.04_{\pm 1.90}$	$10.00_{\pm 0.00}$	$11.35_{\pm 1.91}$	$11.28_{\pm 0.51}$	10.00
AggITD	$11.83_{\pm 1.10}$	$10.21_{\pm 0.29}$	$10.27_{\pm 0.38}$	$16.95_{\pm 6.66}$	10.21
BR-FedBi	$\mathbf{50.86_{\pm 1.33}}$	$22.66_{\pm 9.87}$	$23.08_{\pm 9.29}$	$\mathbf{40.20_{\pm 3.00}}$	22.66
Rob-FedBOB	$25.03_{\pm 3.19}$	$21.70_{\pm 5.66}$	$16.43_{\pm 6.83}$	$26.47_{\pm 2.37}$	16.43
FOBFBL	<u>$40.61_{\pm 1.13}$</u>	$\mathbf{36.10_{\pm 2.74}}$	$\mathbf{32.04_{\pm 0.61}}$	<u>$38.13_{\pm 1.82}$</u>	**32.04**
FOBFBL-M	$39.28_{\pm 1.14}$	<u>$35.85_{\pm 2.65}$</u>	<u>$24.82_{\pm 10.48}$</u>	$36.78_{\pm 2.37}$	<u>24.82</u>
Trimmed Mean					
Algorithm	ALIE	SF	IPM	RN	**Worst**
FedNest	$13.62_{\pm 0.39}$	$12.40_{\pm 1.70}$	$11.97_{\pm 2.37}$	$14.15_{\pm 0.61}$	11.97
AggITD	$15.70_{\pm 0.72}$	$11.20_{\pm 1.00}$	$26.70_{\pm 0.83}$	$14.30_{\pm 0.68}$	11.20
BR-FedBi	$\mathbf{53.07_{\pm 0.92}}$	$\mathbf{30.24_{\pm 6.37}}$	$\mathbf{33.83_{\pm 1.94}}$	$\mathbf{50.83_{\pm 3.80}}$	**30.24**
Rob-FedBOB	$32.03_{\pm 5.74}$	$21.79_{\pm 1.69}$	$17.85_{\pm 0.98}$	$35.81_{\pm 1.80}$	17.85
FOBFBL	$45.27_{\pm 1.92}$	<u>$30.15_{\pm 6.60}$</u>	$26.43_{\pm 5.40}$	$43.89_{\pm 1.34}$	26.43
FOBFBL-M	<u>$45.49_{\pm 2.06}$</u>	$30.14_{\pm 2.69}$	<u>$28.75_{\pm 2.56}$</u>	<u>$45.43_{\pm 2.91}$</u>	<u>28.75</u>
CWMed					
Algorithm	ALIE	SF	IPM	RN	**Worst**
FedNest	$13.24_{\pm 1.08}$	$12.56_{\pm 4.31}$	$10.12_{\pm 0.17}$	$13.32_{\pm 1.35}$	10.12
AggITD	$11.05_{\pm 1.04}$	$10.50_{\pm 2.67}$	$10.19_{\pm 0.27}$	$13.32_{\pm 2.49}$	10.19
BR-FedBi	$38.13_{\pm 1.59}$	$25.96_{\pm 1.41}$	$27.12_{\pm 2.14}$	$\mathbf{49.55_{\pm 3.43}}$	25.96
Rob-FedBOB	$16.79_{\pm 5.42}$	$16.77_{\pm 0.37}$	$17.95_{\pm 0.80}$	$31.75_{\pm 4.86}$	16.77
FOBFBL	<u>$41.97_{\pm 2.44}$</u>	<u>$27.99_{\pm 3.26}$</u>	<u>$28.84_{\pm 1.32}$</u>	$44.25_{\pm 3.59}$	<u>27.99</u>
FOBFBL-M	$\mathbf{42.13_{\pm 2.41}}$	$\mathbf{29.50_{\pm 2.11}}$	$\mathbf{29.98_{\pm 2.70}}$	<u>$44.44_{\pm 4.47}$</u>	**29.50**

B.3 Efficiency Analysis

This section specifies the HPO experiment in Figure 1. The six methods use the same model, losses, numerical precision, and local-update budgets on each dataset. Panels (a,b) measure the time and additional memory of a single client on real data. Panels (c,d) count communication across all 25 clients. These measurements concern the cost of an update; they do not measure the time or communication needed to reach a target accuracy.

B.3.1 Models, Data, and Update Settings

Both datasets use hyperparameter optimization with 10 classes. The upper variable $x = (\boldsymbol{\Delta}, \boldsymbol{\iota})$ contains 20 classwise VS-loss parameters, initialized to $\boldsymbol{\Delta} = \mathbf{1}$ and $\boldsymbol{\iota} = \mathbf{0}$. The lower variable y contains the entire classifier, including all biases. For classifier logits s and label t, the lower loss averages

$$\ell_{\mathrm{VS}}(s, t; x) = -\log \frac{\exp(\Delta_t s_t - \iota_t)}{\sum_{c=1}^{10} \exp(\Delta_c s_c - \iota_c)} \tag{19}$$

over the training minibatch and adds $0.01\|y\|_2^2/2$. The upper loss averages ordinary cross-entropy within each class present in the validation minibatch, then averages over those classes. Table 5 summarizes the two model and data configurations.

Table 5: HPO models and data settings used for Figure 1.

Setting	MNIST	CIFAR-10
Classifier	MLP: $784 \to 200 \to 100 \to 10$	CNN: channels $3 \to 16 \to 32$; FC $2048 \to 100 \to 10$
Network parameters d_y	178,110	210,998
Upper parameters d_x	20	20
Training/test images downloaded	60,000/10,000	50,000/10,000
Client partition	Global long tail, $\rho_{\mathrm{imb}} = 0.01$; shuffled equal shards	Dirichlet label split, $\alpha_{\mathrm{Dir}} = 1$
Clients / validation fraction	25 / 0.2	25 / 0.2
Partition seed	42	42
Configured batch size	128	128
CPU lower/upper batch sizes	128/117	128/128

The MNIST MLP applies ReLU after each hidden layer. The CIFAR-10 CNN uses 3×3 convolutions with stride 1 and padding 1, each followed by ReLU and 2×2 max pooling; its hidden fully connected layer also uses ReLU. For MNIST, images are converted to tensors without further normalization. The long-tailed subset retains $\max\{1, \lfloor n_0 \rho_{\mathrm{imb}}^{c/9} \rfloor\}$ images of class $c \in \{0, \dots, 9\}$, where n_0 is the original class-0 count. The selected indices are shuffled before splitting into 25 shards. CIFAR-10 uses channel means $(0.4914, 0.4822, 0.4465)$ and standard deviations $(0.2470, 0.2435, 0.2616)$, without random cropping or flipping. Each client reserves 20% of its shard for the upper loss.

All six methods use float32 and five local stochastic steps wherever their local-update routine is called. These are five minibatch steps, not five full passes over a shard. FedNest uses $T = 5$ lower stages and Neumann depth $N = 5$; AggITD uses $N = 5$; Rob-FedBOB uses $T = 5$ inner stages. The optimization stepsizes are 10^{-4}, including the auxiliary update of BR-FedBi and the inner update of Rob-FedBOB. The gradient-clipping threshold is 5. The finite multiplier is 1 for `FOBFBL`, `FOBFBL-M`, and Rob-FedBOB. `FOBFBL-M` uses $(\beta_x, \beta_y, \beta_z) = (0.5, 0.5, 0.5)$; `FOBFBL` uses no momentum. Learning-rate warmup and decay are disabled, and the server stepsize ratio is 1. FedNest's inverse-curvature scale is 0.1; AggITD's Hessian-recursion scale is 10^{-4}; BR-FedBi projects its auxiliary vector onto a ball of radius 10.

B.3.2 Time and Memory

CPU and client selection. The measurements run on an Apple M1 Pro with 32 GiB RAM, using PyTorch 2.7.1 and torchvision 0.22.1 on the CPU, with two intra-op threads and one inter-op thread. The three repeats use client indices 7, 8, and 9 (zero based), with model and minibatch-sampling

seeds 42, 43, and 44, respectively. Each MNIST client has 471 lower-training and 117 upper-validation examples. The corresponding CIFAR-10 sizes are $2508/627$, $1004/250$, and $1972/493$. The implementation samples with replacement and caps each batch at the size of its local split, giving the effective batch sizes in Table 5. All data preprocessing and client-cache loading finish before measurement.

Each process runs one selected client's complete update through the core algorithm, with a single-client coordinator and mean aggregation. The measured work includes gradients, parameter copies, algorithm state, and local coordination. It excludes network transport, attack generation, data loading, and accuracy evaluation. Methods run sequentially, in a fixed randomized order for each repeat.

Time. Each repeat discards two warmup outer updates and times the next 18 updates using a monotonic wall clock. Algorithm states, including momentum, persist across updates. FedNest samples a recursion length from $\{0, \dots, 4\}$; AggITD samples an initialization index from $\{0, \dots, 5\}$. Their random seeds are selected deterministically so that every branch occurs at least twice during timing. We first average within each branch and then weight the branch means uniformly. Figure 1(a) reports the mean and sample standard deviation of the three repeat estimates, in milliseconds. Timing uses separate processes from memory sampling, so the memory monitor does not affect the timed updates.

Memory. Each repeat starts a fresh process and records its memory use after loading the model, client cache, and reset copies of the initial parameters. It then executes one update for every possible random branch, restoring the initial parameters between branches. We sample the process's resident memory at a requested interval of 1 ms and report the largest observed value minus the initial value. Thus, the plotted quantity is the *additional memory used while running the algorithm*, including algorithm buffers, allocator overhead, and library workspaces. This process-memory measure is also called resident set size (RSS). The largest observed sampling interval is 26.2 ms, so very short-lived peaks may not be captured. Figure 1(b) reports the mean and sample standard deviation over the three fresh processes, in MiB ($1\ \mathrm{MiB} = 2^{20}$ bytes).

Table 6: Single-client CPU measurements for Figure 1(a,b). Values are mean $\pm$ sample standard deviation over three client/initialization combinations.

	MNIST		CIFAR-10	
Method	Time (ms)	Memory (MiB)	Time (ms)	Memory (MiB)
FedNest	79.0 ± 5.1	33.7 ± 2.0	2505.0 ± 112.5	287.4 ± 18.1
AggITD	91.3 ± 4.8	40.1 ± 3.1	2543.3 ± 53.4	310.5 ± 30.4
BR-FedBi	28.1 ± 1.7	27.5 ± 1.5	1052.6 ± 20.3	252.2 ± 18.8
Rob-FedBOB	43.2 ± 1.6	21.4 ± 2.3	1192.6 ± 26.8	189.7 ± 10.6
FOBFBL	25.7 ± 3.0	23.5 ± 3.6	655.8 ± 13.4	159.3 ± 9.4
FOBFBL-M	26.2 ± 1.3	24.7 ± 2.4	679.7 ± 62.4	162.7 ± 7.8

B.3.3 Communication Rounds and Volume

Panel (c), labeled *Communication rounds*, counts communication synchronizations. One event consists of the recorded client reports and their corresponding server broadcast; message blocks sent together count as one event. Panel (d), labeled *Communication cost*, sums the bytes uploaded and downloaded for model parameters, gradients, and auxiliary vectors; it does not count dataset images. Both panels cover 25 fully participating clients and report expected totals per 100 complete outer updates, rather than per 100 communication rounds.

Rounds per outer update. For method m, let r_m denote the number of client–server synchronizations needed to complete one outer update. The five local minibatch steps within a client routine add no communication rounds. Likewise, reports from 25 clients sent in parallel count as one round; their sizes are summed when computing communication volume. The following counts use the update settings in Appendix B.3.1.

`FOBFBL`, `FOBFBL-M`, and BR-FedBi. Each client transmits its three update blocks together, and the server aggregates the blocks and broadcasts the updated variables. No additional synchronization is

needed between the block updates. Thus $r_m = 1$ for each of these methods. The momentum buffers of FOBFBL−M remain on the clients and add no exchanges.

Rob-FedBOB. Each of the T inner stages exchanges a lower-level update and the corresponding server state. After these stages, the two outer update blocks are transmitted together in one further synchronization. Hence $r_{\text{Rob-FedBOB}} = T + 1 = 6$ for $T = 5$.

AggITD. At stages $t = 0, \dots, N$, the lower-gradient and auxiliary reports, when both are present, are transmitted together. These stages contribute $N + 1$ rounds. The first N stages also exchange the locally updated lower models, contributing another N rounds. Finally, the global hypergradient and the locally updated outer models require one round each. Therefore,

$$r_{\text{AggITD}} = (N + 1) + N + 1 + 1 = 2N + 3 = 13 \qquad (N = 5),$$

consistent with Table 1 of Xiao & Ji (2023). The sampled initialization index $Q \in \{0, \dots, N\}$ changes which auxiliary reports are present, but the number of synchronization stages remains 13 for every branch.

FedNest. Each of the T lower stages has two synchronizations: one for the lower gradients and one for the locally updated lower models, giving $2T$ rounds. Initializing the inverse-Hessian-gradient estimator adds one round. The estimator then performs N' Hessian-vector exchanges, where N' is uniform on $\{0, \dots, N - 1\}$ (Algorithm 3 of Tarzanagh et al. (2022)). The outer hypergradient and outer model exchanges add one round each. Consequently,

$$r_{\text{FedNest}} = 2T + 1 + N' + 1 + 1, \qquad \mathbb{E}[r_{\text{FedNest}}] = 2T + \frac{N - 1}{2} + 3 = 15 \quad (T = N = 5).$$

The five possible branch counts are $13, 14, 15, 16, 17$, whose uniform average is 15. We report this expectation of the sampled recursion length. The expression $2T + N + 3$ in Section 2.6 of Tarzanagh et al. (2022) instead uses the configured depth N; substituting $T = N = 5$ into that budget gives 18 rather than the branch average.

Totals over 100 outer updates. With the same update settings at every iteration, the expected total for method m over K outer updates is $K\mathbb{E}[r_m]$. For $K = 100$, this gives $100 \times 1 = 100$ rounds for FOBFBL, FOBFBL−M, and BR-FedBi; $100 \times 6 = 600$ for Rob-FedBOB; $100 \times 13 = 1300$ for AggITD; and $100 \times 15 = 1500$ in expectation for FedNest. A particular sequence of FedNest updates can have a different total because its recursion lengths are random. These are the counts plotted in Figure 1(c) and reported in Table 7.

Communication-volume measurement. The communication probe uses the same HPO models, losses, batch size 128, and update settings described above. It executes each uniformly sampled branch once on synthetic images of the appropriate shape, recording actual tensor dimensions and dtypes. There are 15 complete branch probes and 162 recorded communication events per dataset. The exact branch average is multiplied by 100. Every float32 entry contributes four bytes. Downloads count one copy per recipient; protocol headers, transport overhead, and additional initialization traffic are excluded. The probe uses coordinatewise median aggregation and seven same-size sign-flipped client reports. These choices preserve message dimensions. The real-data memory probes independently record the same message sizes and reproduce every total below.

Table 7: Expected communication per 100 complete outer updates for Figure 1(c,d). Volume includes uploads and one download copy per client.

Method	Rounds (both datasets)	MNIST (GB)	CIFAR-10 (GB)
FedNest	1500	53.43540	63.30180
AggITD	1300	51.65390	61.19142
BR-FedBi	100	7.12480	8.44032
Rob-FedBOB	600	23.15470	27.43014
FOBFBL	100	7.12480	8.44032
FOBFBL-M	100	7.12480	8.44032

Both proposed methods upload and download all three blocks, totaling $d_x + d_y + d_z = d_x + 2d_y$ float32 entries in each direction per client and outer update. Their total communication volume is

therefore

$$V = 8nK(d_x + 2d_y) \text{ bytes}, \tag{20}$$

which gives 7.1248 GB for MNIST and 8.44032 GB for CIFAR-10 when $n = 25$ and $K = 100$ (1 GB $= 10^9$ bytes). BR-FedBi sends the same number of entries. Client momentum adds no messages. Panel (d) uses a logarithmic axis; numerical labels show the original GB values. These volumes describe the implemented protocol, rather than physical network measurements.

Implementation and reproducibility. FedNest includes lower- and upper-level SVRG updates and the N/L initialization of its inverse-Hessian estimator. AggITD includes the endpoint $Q = N$, the $\lambda_{\mathrm{ITD}}(N + 1)$ scaling, both SVRG stages, and the HessIV product in Eq. (4) of Xiao & Ji (2023). Their use with sign-flipped reports and robust aggregation is an experimental adaptation of the original benign algorithms. The probes use neural-network lower objectives and practical multi-step updates; they do not establish the global strong-convexity assumptions or the theoretical stepsize–momentum schedule. The resource probe does not evaluate accuracy.

The accompanying measurement project provides the dataset checksums, client indices, source-code snapshots, and all per-update records. It contains 72 fresh worker records: 36 for time and 36 for memory, covering 648 timed updates after 72 warmup updates and 90 memory-branch updates. Starting in `communication_volume_audit`, the real-data CPU measurements are reproduced by

```
python3 download_hpo_data.py
python3 benchmark_cpu_hpo.py prepare
python3 benchmark_cpu_hpo.py run
python3 report_cpu_hpo.py
```

The communication probes are reproduced by

```
python3 measure_cnn.py --scenario mnist_hpo_mlp \
    --batch-size 128 --local-steps 5
python3 measure_cnn.py --scenario cifar_hpo_cnn \
    --batch-size 128 --local-steps 5
python3 prepare_figure1d.py --task hpo
```

The measurement scripts check the raw-record and source hashes, recompute the means and standard deviations, and verify agreement with the communication counts.

B.4 Ablation Study

We study the attack fraction, number of participating clients, and heterogeneity level in MNIST HPO. All ablations use seed 42 and report class-balanced accuracy on the official test set; attack panels have independent y-axis ranges.

B.4.1 Attack Fraction

We fix $n = 25$ and vary the attack fraction n_b/n over 20%, 28%, and 36%, corresponding to $n_b \in \{5, 7, 9\}$. Figures 10–13 report the final test accuracy for each aggregator and attack.

Across the 16 aggregator–attack pairs, increasing n_b from 5 to 9 reduces final accuracy by 4.14 and 5.72 points on average for `FOBFBL-M` and `FOBFBL`, respectively. Every baseline has a larger average reduction (at least 6.78 points). The advantage depends on the aggregator: at $n_b = 9$, the better proposed variant exceeds the strongest baseline in all four CWMed panels by 0.73–10.93 points, whereas with the geometric median a clear advantage appears only under ALIE. BR-FedBi reaches chance accuracy under ALIE with Krum and CWMed, and Rob-FedBOB falls below 30% under IPM with Trimmed Mean.

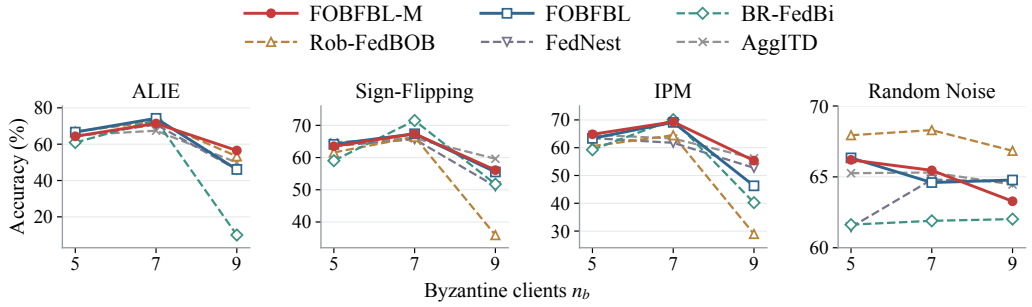

Figure 10: MNIST HPO under varying attack fractions with **Krum** ($n = 25$).

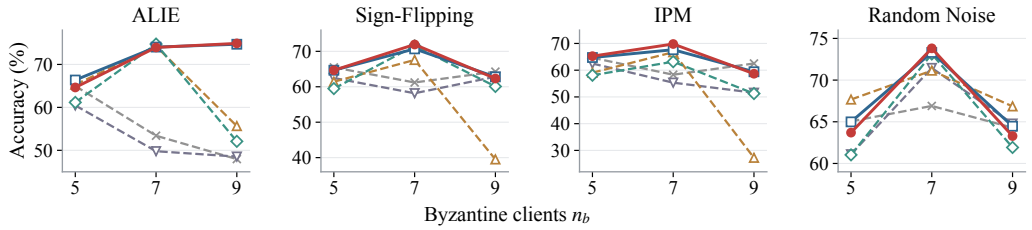

Figure 11: MNIST HPO under varying attack fractions with **Geometric Median** ($n = 25$).

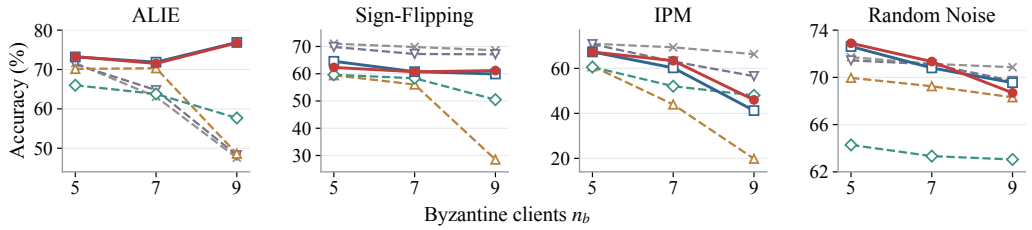

Figure 12: MNIST HPO under varying attack fractions with **Trimmed Mean** ($n = 25$).

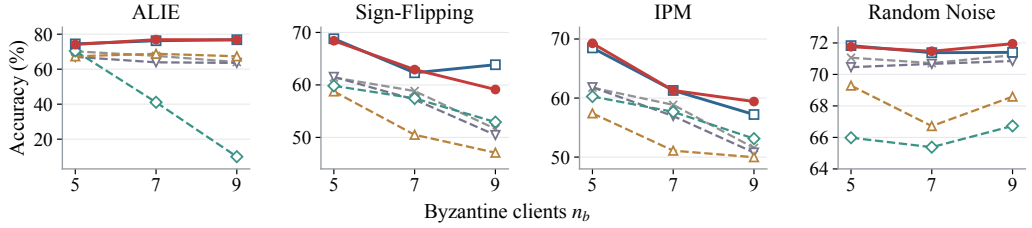

Figure 13: MNIST HPO under varying attack fractions with **CWMed** ($n = 25$).

B.4.2 Number of Participating Clients

We vary the number of participating clients over $n \in \{10, 25, 40\}$ while holding the attack fraction near 30%. The cases $n = 10$ and $n = 40$ use $\delta = 30\%$, while $n = 25$ uses $\delta = 28\%$. Final test accuracy is shown in Figures 14–17.

From $n = 10$ to $n = 40$, FOBFBL−M and FOBFBL improve in all 16 aggregator–attack pairs, by 8.72 and 8.75 points on average. At $n = 40$, the better proposed variant outperforms the strongest baseline in 14 of the 16 pairs, including all four CWMed panels by 7.36–10.34 points. This finite-budget trend is compatible with the honest-client averaging terms in Theorems 1 and 3, but it does not by itself verify linear speedup. For FOBFBL−M, such a guarantee additionally requires $\kappa = \mathcal{O}(1/n_h)$.

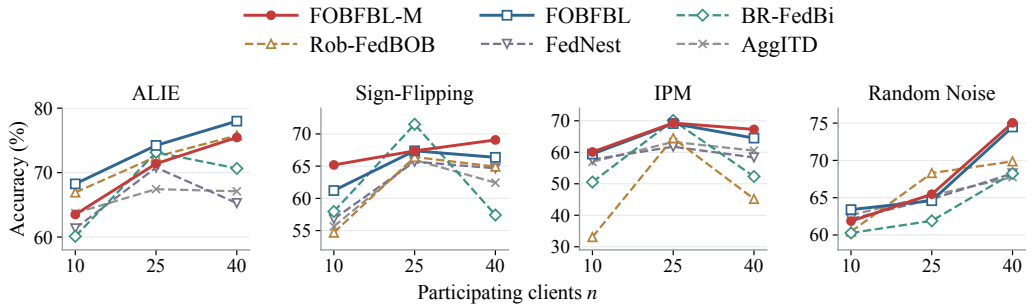

Figure 14: MNIST HPO with different numbers of participating clients using **Krum** ($\delta \approx 30\%$).

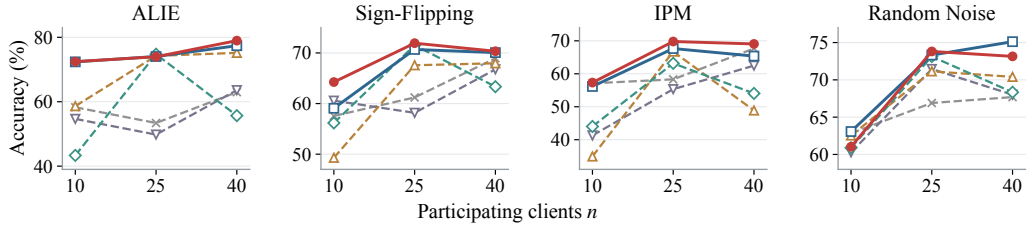

Figure 15: MNIST HPO with different numbers of participating clients using **Geometric Median** ($\delta \approx 30\%$).

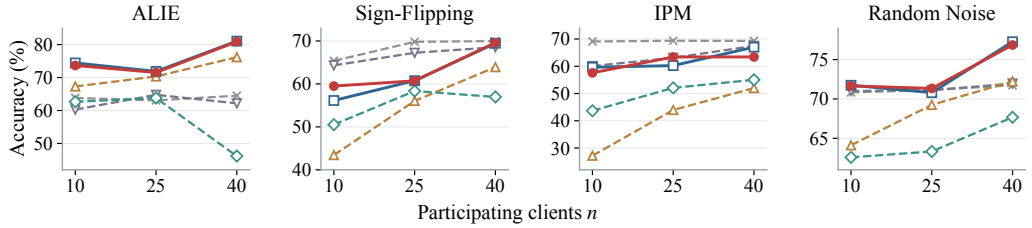

Figure 16: MNIST HPO with different numbers of participating clients using **Trimmed Mean** ($\delta \approx 30\%$).

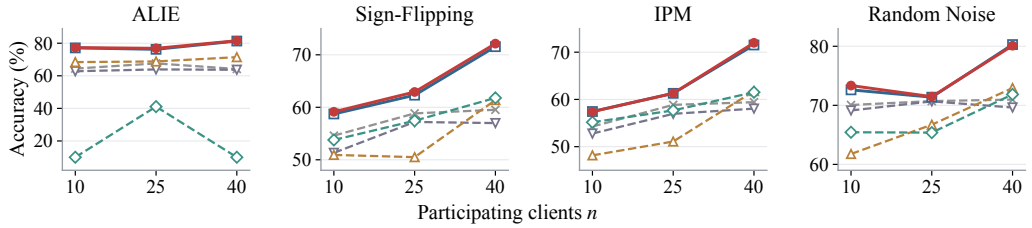

Figure 17: MNIST HPO with different numbers of participating clients using **CWMed** ($\delta \approx 30\%$).

B.4.3 HETEROGENEITY LEVEL

We control the heterogeneity level through class imbalance in long-tailed MNIST HPO, setting $\rho_{\text{imb}} \in \{0.1, 0.05, 0.01\}$. These values give approximate tail ratios of 10:1, 20:1, and 100:1. We keep $n = 25$, $n_b = 7$, and the communication budget fixed. Figures 18–21 report final accuracy.

Both proposed methods lose accuracy as ρ_{imb} decreases, confirming that heavier class imbalance makes the task harder. At $\rho_{\text{imb}} = 0.01$, the better proposed variant outperforms the strongest baseline in 10 of the 16 aggregator–attack pairs, including all four CWMed panels by 0.75–8.08 points. With the geometric median it leads in three of four panels, whereas Krum and Trimmed Mean give mixed rankings. The result therefore supports robustness primarily under CWMed rather than uniform dominance across every aggregator.

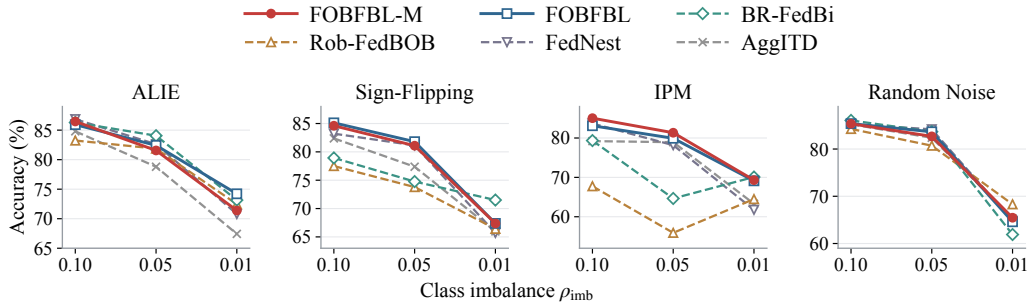

Figure 18: MNIST HPO at different heterogeneity levels (class imbalance) with **Krum** ($n = 25$, $n_b = 7$).

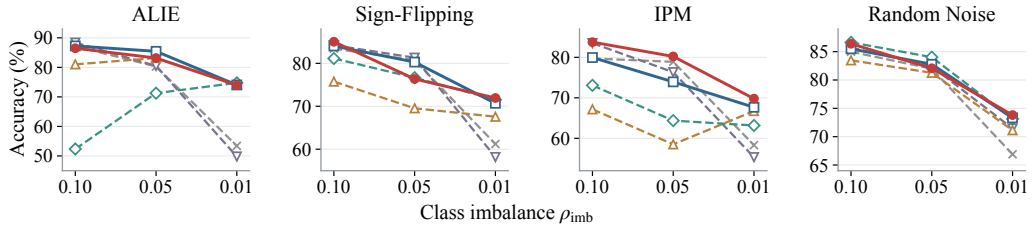

Figure 19: MNIST HPO at different heterogeneity levels (class imbalance) with **Geometric Median** ($n = 25$, $n_b = 7$).

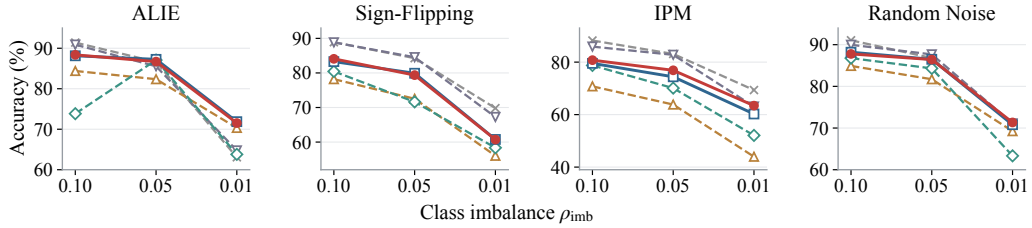

Figure 20: MNIST HPO at different heterogeneity levels (class imbalance) with **Trimmed Mean** ($n = 25$, $n_b = 7$).

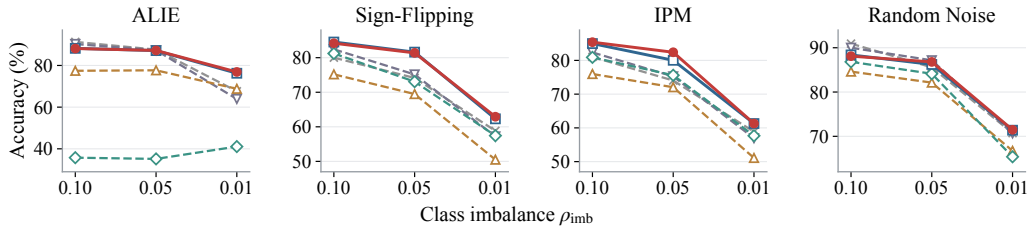

Figure 21: MNIST HPO at different heterogeneity levels (class imbalance) with **CWMed** ($n = 25$, $n_b = 7$).

C TECHNICAL LEMMAS

This section collects the technical lemmas used in the convergence proofs in Appendices D and E. Lemmas 1 and 2 bound the finite-multiplier gradient bias and establish regularity of the optimizer maps and the hyperobjective. Lemma 3 bounds the block noise, client heterogeneity, and upper-level direction bias. Finally, Lemma 4 bounds the optimizer drift caused by an upper-level update, which is needed to control the lower-level tracking errors.

Under Assumptions 1 and 2, fix $\lambda > 0$ with $\lambda \geq 2l_{f,1}/\mu_g$ and recall

$$\mu_\lambda := \lambda\mu_g - l_{f,1} \geq \frac{\lambda\mu_g}{2}, \qquad L_\lambda := l_{f,1} + \lambda l_{g,1}. \tag{21}$$

For fixed x, $f(x,\cdot) + \lambda g(x,\cdot)$ is μ_λ-strongly convex and L_λ-smooth.

Lemma 1 (Finite-multiplier gradient)**.** *Suppose Assumptions 1 and 2 hold, and g is twice continuously differentiable with an $l_{g,2}$-Lipschitz Hessian. For $\lambda > 0$ satisfying $\lambda \geq 2l_{f,1}/\mu_g$, define $\mathcal{L}_\lambda^\star(x) := \mathcal{L}_\lambda(x, y_\lambda^\star(x), y^\star(x))$. Then*

$$\nabla\mathcal{L}_\lambda^\star(x) = \nabla_x f(x, y_\lambda^\star(x)) + \lambda\big[\nabla_x g(x, y_\lambda^\star(x)) - \nabla_x g(x, y^\star(x))\big], \tag{22}$$

and

$$\|\nabla\mathcal{L}_\lambda^\star(x) - \nabla\Phi(x)\| \leq \frac{C_{\mathrm{lag}}}{\lambda}, \qquad C_{\mathrm{lag}} := \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right). \tag{23}$$

Proof. Fix x and write $y_\lambda^\star = y_\lambda^\star(x)$ and $y^\star = y^\star(x)$. Since $f(x,\cdot)$ is $l_{f,1}$-weakly convex, $\psi_{\lambda,x}(y) := f(x,y) + \lambda g(x,y)$ is μ_λ-strongly convex and has a unique minimizer. The optimality conditions give

$$\nabla_y f(x, y_\lambda^\star) + \lambda\nabla_y g(x, y_\lambda^\star) = 0, \qquad \nabla_y g(x, y^\star) = 0. \tag{24}$$

Let $\delta_\lambda := y_\lambda^\star - y^\star$ and $d_\lambda := \|\delta_\lambda\|$. Strong monotonicity of $\nabla_y g(x,\cdot)$ and (24) give

$$\begin{aligned}\mu_g d_\lambda^2 &\leq \langle\nabla_y g(x, y_\lambda^\star) - \nabla_y g(x, y^\star), \delta_\lambda\rangle \\ &= -\lambda^{-1}\langle\nabla_y f(x, y_\lambda^\star), \delta_\lambda\rangle \leq \frac{l_{f,0}}{\lambda} d_\lambda.\end{aligned} \tag{25}$$

If $d_\lambda > 0$, divide by $\mu_g d_\lambda$; if $d_\lambda = 0$, the bound is immediate. In particular, we will use the relaxed estimate

$$\|y_\lambda^\star(x) - y^\star(x)\| \leq \frac{2l_{f,0}}{\lambda\mu_g}. \tag{26}$$

We first prove the gradient identity. Recall that $\mathcal{L}_\lambda^\star(x) = \min_y \psi_{\lambda,x}(y) - \lambda g^\star(x)$, where $g^\star(x) := \min_z g(x,z)$. By Danskin's theorem, we have

$$\nabla\mathcal{L}_\lambda^\star(x) = \nabla_x f(x, y_\lambda^\star) + \lambda\nabla_x g(x, y_\lambda^\star) - \lambda\nabla g^\star(x), \tag{27}$$

$$\nabla g^\star(x) = \nabla_x g(x, y^\star). \tag{28}$$

Substituting the second identity into the first proves (22).

We now bound the difference between this gradient and $\nabla\Phi(x)$. Define

$$H_\star := \nabla_{yy}^2 g(x, y^\star(x)), \qquad B_\star := \nabla_{yx}^2 g(x, y^\star(x)). \tag{29}$$

The implicit function theorem gives $\nabla y^\star(x) = -H_\star^{-1}B_\star$ with $\|H_\star^{-1}\| \leq 1/\mu_g$ and $\|B_\star\| \leq l_{g,1}$. To make the perturbation argument explicit (Kwon et al., 2023, Lemma A.2), expand both partial gradients of g at $y^\star$:

$$\nabla_y g(x, y_\lambda^\star) = H_\star\delta_\lambda + R_y, \tag{30}$$

$$\nabla_x g(x, y_\lambda^\star) - \nabla_x g(x, y^\star) = B_\star^\top\delta_\lambda + R_x. \tag{31}$$

The remainder in each line is the integral of the corresponding Hessian difference along $y^\star + t\delta_\lambda$, $0 \leq t \leq 1$. Hessian Lipschitz continuity bounds its norm by $l_{g,2}d_\lambda^2/2$, while the bound $l_{g,1}$ on each Hessian block gives $2l_{g,1}d_\lambda$. Thus, for $v \in \{x, y\}$,

$$\|R_v\| \leq d_\lambda \min\{2l_{g,1}, l_{g,2}d_\lambda\}. \tag{32}$$

The optimality condition now reads $\lambda H_\star \delta_\lambda = -\nabla_y f(x, y_\lambda^\star) - \lambda R_y$. Substitute this identity into (22) and subtract $\nabla\Phi(x) = \nabla_x f(x, y^\star) - B_\star^\top H_\star^{-1} \nabla_y f(x, y^\star)$ to obtain

$$\begin{aligned}\nabla\mathcal{L}_\lambda^\star(x) - \nabla\Phi(x) = \nabla_x f(x, y_\lambda^\star) - \nabla_x f(x, y^\star) \\ - B_\star^\top H_\star^{-1}[\nabla_y f(x, y_\lambda^\star) - \nabla_y f(x, y^\star)] \\ + \lambda(R_x - B_\star^\top H_\star^{-1} R_y).\end{aligned} \tag{33}$$

The first two terms are bounded by $(1 + l_{g,1}/\mu_g) l_{f,1} d_\lambda$. Since $l_{g,1} \geq \mu_g$, we have $1 + l_{g,1}/\mu_g \leq 2l_{g,1}/\mu_g$. Applying the remainder bounds therefore gives

$$\begin{aligned}\|\nabla\Phi(x) - \nabla\mathcal{L}_\lambda^\star(x)\| &\leq 2\frac{l_{g,1}}{\mu_g} l_{f,1} d_\lambda \\ &\quad + 2\frac{l_{g,1}}{\mu_g}\lambda d_\lambda \min\{2l_{g,1}, l_{g,2} d_\lambda\} \\ &= 2\frac{l_{g,1}}{\mu_g} d_\lambda \left(l_{f,1} + \lambda \min\{2l_{g,1}, l_{g,2} d_\lambda\}\right).\end{aligned} \tag{34}$$

Substituting $\min\{2l_{g,1}, l_{g,2} d_\lambda\} \leq l_{g,2} d_\lambda$ and the bound on d_λ in (26), we obtain

$$\begin{aligned}\|\nabla\mathcal{L}_\lambda^\star(x) - \nabla\Phi(x)\| &\leq \frac{4 l_{f,0} l_{g,1} l_{f,1}}{\lambda\mu_g^2} + \frac{8 l_{f,0}^2 l_{g,1} l_{g,2}}{\lambda\mu_g^3} \\ &= \frac{4 l_{f,0} l_{g,1}}{\lambda\mu_g^2}\left(l_{f,1} + \frac{2 l_{f,0} l_{g,2}}{\mu_g}\right) = \frac{C_{\mathrm{lag}}}{\lambda}.\end{aligned} \tag{35}$$

This proves (23) and completes the proof. □

We next establish the regularity of the optimizer maps and the hyperobjective.

Lemma 2 (Optimizer and value regularity)**.** *Under Assumptions 1 and 2, the optimizer maps are Lipschitz: there are finite $L_*, L_{\lambda,*}$ such that*

$$\|y^\star(x') - y^\star(x)\| \leq L_* \|x' - x\|, \tag{36}$$

$$\|y_\lambda^\star(x') - y_\lambda^\star(x)\| \leq L_{\lambda,*} \|x' - x\|. \tag{37}$$

One may take $L_* \leq l_{g,1}/\mu_g$ and $L_{\lambda,*} \leq L_\lambda/\mu_\lambda$. If g is twice continuously differentiable with an $l_{g,2}$-Lipschitz Hessian, then there is a finite S_* such that

$$\|\nabla y^\star(x') - \nabla y^\star(x)\| \leq S_* \|x' - x\|, \tag{38}$$

and Φ is smooth: for some finite L_Φ,

$$\|\nabla\Phi(x') - \nabla\Phi(x)\| \leq L_\Phi \|x' - x\|. \tag{39}$$

If, in addition, f has an $l_{f,2}$-Lipschitz Hessian, then there is a finite $S_{\lambda,*}$ such that

$$\|\nabla y_\lambda^\star(x') - \nabla y_\lambda^\star(x)\| \leq S_{\lambda,*} \|x' - x\|. \tag{40}$$

Proof. Let x, x' be arbitrary, and write $y = y^\star(x)$ and $y' = y^\star(x')$. Recall that $\nabla_y g(x, y) = \nabla_y g(x', y') = 0$. By strong monotonicity, we obtain

$$\begin{aligned}\mu_g \|y' - y\|^2 &\leq \langle \nabla_y g(x, y') - \nabla_y g(x, y), y' - y\rangle \\ &= \langle \nabla_y g(x, y') - \nabla_y g(x', y'), y' - y\rangle \\ &\leq l_{g,1} \|x' - x\| \|y' - y\|,\end{aligned} \tag{41}$$

where the last inequality uses gradient Lipschitz continuity and Cauchy–Schwarz. If $y' = y$, the claim is immediate; otherwise, dividing by $\mu_g \|y' - y\|$ proves (36) with $L_* := l_{g,1}/\mu_g$. For $f(x, \cdot) + \lambda g(x, \cdot)$, the same argument proves (37) with $L_{\lambda,*} := L_\lambda/\mu_\lambda$, since

$$\|y_\lambda^\star(x') - y_\lambda^\star(x)\| \leq \frac{L_\lambda}{\mu_\lambda} \|x' - x\|. \tag{42}$$

Next, consider the Jacobians of the optimizer maps. Under the stated second-order regularity conditions, differentiating the optimality conditions gives

$$\nabla y^\star(x) = -[\nabla^2_{yy} g(x, y^\star(x))]^{-1} \nabla^2_{yx} g(x, y^\star(x)), \tag{43}$$

$$\nabla y^\star_\lambda(x) = -[\nabla^2_{yy} (f + \lambda g)(x, y^\star_\lambda(x))]^{-1} \nabla^2_{yx} (f + \lambda g)(x, y^\star_\lambda(x)). \tag{44}$$

For (43), set $H(x) := \nabla^2_{yy} g(x, y^\star(x))$ and $B(x) := \nabla^2_{yx} g(x, y^\star(x))$. The inverse matrices satisfy $\|H(x)^{-1}\| \le 1/\mu_g$, while $\|B(x)\| \le l_{g,1}$. Moreover, Hessian Lipschitzness and (41) give

$$\|H(x') - H(x)\| \le l_{g,2}(1 + L_*)\|x' - x\|, \tag{45}$$

$$\|B(x') - B(x)\| \le l_{g,2}(1 + L_*)\|x' - x\|. \tag{46}$$

Recall the identity $H(x')^{-1} - H(x)^{-1} = H(x')^{-1}[H(x) - H(x')]H(x)^{-1}$. Adding and subtracting $H(x')^{-1}B(x)$, we can write

$$\begin{aligned} \nabla y^\star(x') - \nabla y^\star(x) = &-H(x')^{-1}[B(x') - B(x)] \\ &- [H(x')^{-1} - H(x)^{-1}]B(x). \end{aligned} \tag{47}$$

Taking norms and applying the bounds above yields

$$\begin{aligned} \|\nabla y^\star(x') - \nabla y^\star(x)\| &\le \frac{1}{\mu_g}\|B(x') - B(x)\| + \frac{l_{g,1}}{\mu_g^2}\|H(x') - H(x)\| \\ &\le l_{g,2}(1 + L_*)\left(\frac{1}{\mu_g} + \frac{l_{g,1}}{\mu_g^2}\right)\|x' - x\|. \end{aligned} \tag{48}$$

This proves (38), with the coefficient on the right as S_*. For $y^\star_\lambda$, set $L_{\lambda,2} := l_{f,2} + \lambda l_{g,2}$ and apply the same decomposition to (44). We obtain

$$\|\nabla y^\star_\lambda(x') - \nabla y^\star_\lambda(x)\| \le L_{\lambda,2}(1 + L_{\lambda,*})\left(\frac{1}{\mu_\lambda} + \frac{L_\lambda}{\mu_\lambda^2}\right)\|x' - x\|. \tag{49}$$

Taking the coefficient on the right as $S_{\lambda,*}$ proves (40).

Finally, recall that $\Phi(x) = f(x, y^\star(x))$. Applying the chain rule, we have

$$\nabla\Phi(x) = \nabla_x f(x, y^\star(x)) + \nabla y^\star(x)^\top \nabla_y f(x, y^\star(x)). \tag{50}$$

Evaluate (50) at x' and x, add and subtract $\nabla y^\star(x')^\top \nabla_y f(x, y^\star(x))$, and use $\|\nabla y^\star\| \le L_*$ and $\|\nabla_y f\| \le l_{f,0}$. Writing $y' = y^\star(x')$ and $y = y^\star(x)$, we obtain

$$\begin{aligned} \|\nabla\Phi(x') - \nabla\Phi(x)\| \le &\|\nabla_x f(x', y') - \nabla_x f(x, y)\| \\ &+ L_*\|\nabla_y f(x', y') - \nabla_y f(x, y)\| \\ &+ l_{f,0}\|\nabla y^\star(x') - \nabla y^\star(x)\|. \end{aligned} \tag{51}$$

Since $\|(x', y') - (x, y)\| \le (1 + L_*)\|x' - x\|$, gradient Lipschitz continuity and (38) imply

$$\begin{aligned} \|\nabla\Phi(x') - \nabla\Phi(x)\| \le &\, l_{f,1}(1 + L_*)\|x' - x\| + L_* l_{f,1}(1 + L_*)\|x' - x\| \\ &+ l_{f,0} S_*\|x' - x\|. \end{aligned} \tag{52}$$

Thus (39) holds with $L_\Phi := l_{f,1}(1 + L_*)^2 + l_{f,0}S_*$. Only the Hessian regularity of g is needed for this last conclusion; see also Kwon et al. (2023, Lemma A.1). □

Assumption 1 bounds only $\nabla_y f_i$ and allows nonconstant, globally PL hyperobjectives. For example, let every honest client have scalar objectives $f_i(x, y) = ax^2/2 + by$ and $g_i(x, y) = \mu_g(y - cx)^2/2$, where $a > 0$ and $b, c \neq 0$. Then $y^\star(x) = cx$, $\Phi(x) = ax^2/2 + bcx$, and $\|\nabla\Phi(x)\|^2/2 = a(\Phi(x) - \Phi_*)$ with $\Phi_* = -b^2c^2/(2a)$. Both Hessians are constant and $\|\nabla_y f_i\| = |b|$, so the regularity assumptions and the global PL condition hold simultaneously.

We now bound the noise and bias in the three update blocks. Let $\mathcal{F}_k$ be the history before the oracle draws in round k, and write $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$. For $v \in \{x, y, z\}$, define $\bar{h}^k_{v,i} := \mathbb{E}_k[h^k_{v,i}]$ and $\bar{h}^k_v := n_h^{-1} \sum_{i \in \mathcal{H}} \bar{h}^k_{v,i}$. Define the honest stochastic mean and its sampling error by

$$h^k_v := \frac{1}{n_h} \sum_{i \in \mathcal{H}} h^k_{v,i}, \qquad \epsilon^k_v := h^k_v - \bar{h}^k_v. \tag{53}$$

The deterministic block directions are

$$\begin{aligned}\bar{h}_x^k &= \nabla_x f(x^k, y^k) + \lambda[\nabla_x g(x^k, y^k) - \nabla_x g(x^k, z^k)],\\ \bar{h}_y^k &= \nabla_y f(x^k, y^k) + \lambda \nabla_y g(x^k, y^k),\\ \bar{h}_z^k &= \nabla_y g(x^k, z^k).\end{aligned} \tag{54}$$

Lemma 3 (Block noise and direction bias)**.** *Suppose Assumptions 1–4 hold, together with the Hessian regularity of g and the multiplier condition in Lemma 1. Define*

$$\Sigma_x := \sigma_f^2 + 2\lambda^2\sigma_g^2, \qquad \Sigma_y := \sigma_f^2 + \lambda^2\sigma_g^2, \qquad \Sigma_z := \sigma_g^2, \tag{55}$$

$$Z_x := 3\zeta_f^2 + 6\lambda^2\zeta_g^2, \qquad Z_y := 2\zeta_f^2 + 2\lambda^2\zeta_g^2, \qquad Z_z := \zeta_g^2. \tag{56}$$

Then $\mathbb{E}_k[\epsilon_v^k] = 0$ and

$$\mathbb{E}_k\|\epsilon_v^k\|^2 \le \frac{\Sigma_v}{n_h}, \qquad v \in \{x, y, z\}. \tag{57}$$

The individual noise and client heterogeneity also satisfy

$$\frac{1}{n_h}\sum_{i\in\mathcal{H}} \mathbb{E}_k\|h_{v,i}^k - \bar{h}_{v,i}^k\|^2 \le \Sigma_v, \qquad \frac{1}{n_h}\sum_{i\in\mathcal{H}} \|\bar{h}_{v,i}^k - \bar{h}_v^k\|^2 \le Z_v. \tag{58}$$

Furthermore, for constants $D_y, D_z > 0$ depending only on the smoothness parameters and the fixed multiplier λ,

$$\|\bar{h}_x^k - \nabla\Phi(x^k)\|^2 \le D_y\|y^k - y_\lambda^\star(x^k)\|^2 + D_z\|z^k - y^\star(x^k)\|^2 + 3\frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{59}$$

Proof. Fix a round k. We separate the oracle noise from the variation across honest clients by defining

$$\xi_{v,i}^k := h_{v,i}^k - \bar{h}_{v,i}^k, \qquad d_{v,i}^k := \bar{h}_{v,i}^k - \bar{h}_v^k. \tag{60}$$

Here $\mathbb{E}_k[\xi_{v,i}^k] = 0$, while $d_{v,i}^k$ is $\mathcal{F}_k$-measurable and $\sum_{i\in\mathcal{H}} d_{v,i}^k = 0$.

We first bound the oracle noise in each block. Denote the errors of the six oracle calls in (10) by $\xi_{f,x,i}^k$, $\xi_{g,xy,i}^k$, $\xi_{g,xz,i}^k$, $\xi_{f,y,i}^k$, $\xi_{g,y,i}^k$, and $\xi_{g,z,i}^k$. The block errors are

$$\xi_{x,i}^k = \xi_{f,x,i}^k + \lambda\xi_{g,xy,i}^k - \lambda\xi_{g,xz,i}^k, \tag{61}$$

$$\xi_{y,i}^k = \xi_{f,y,i}^k + \lambda\xi_{g,y,i}^k, \tag{62}$$

$$\xi_{z,i}^k = \xi_{g,z,i}^k. \tag{63}$$

For two distinct oracle errors ξ and ξ' in the same round, Assumption 3 gives $\mathbb{E}_k\langle\xi, \xi'\rangle = \langle\mathbb{E}_k\xi, \mathbb{E}_k\xi'\rangle = 0$. In particular, expanding the squared norm of the x-block error yields

$$\begin{aligned}\mathbb{E}_k\|\xi_{x,i}^k\|^2 &= \mathbb{E}_k\|\xi_{f,x,i}^k\|^2 + \lambda^2\mathbb{E}_k\|\xi_{g,xy,i}^k\|^2 + \lambda^2\mathbb{E}_k\|\xi_{g,xz,i}^k\|^2\\ &\le \sigma_f^2 + 2\lambda^2\sigma_g^2 = \Sigma_x.\end{aligned} \tag{64}$$

The same calculation for the other two blocks gives

$$\mathbb{E}_k\|\xi_{y,i}^k\|^2 \le \sigma_f^2 + \lambda^2\sigma_g^2 = \Sigma_y, \tag{65}$$

$$\mathbb{E}_k\|\xi_{z,i}^k\|^2 \le \sigma_g^2 = \Sigma_z. \tag{66}$$

For any $v \in \{x, y, z\}$, recall that $\epsilon_v^k = n_h^{-1}\sum_{i\in\mathcal{H}} \xi_{v,i}^k$. Thus $\mathbb{E}_k[\epsilon_v^k] = 0$. Conditional independence across honest clients also makes $\mathbb{E}_k\langle\xi_{v,i}^k, \xi_{v,j}^k\rangle = 0$ whenever $i \ne j$, so

$$\mathbb{E}_k\|\epsilon_v^k\|^2 = \frac{1}{n_h^2}\sum_{i\in\mathcal{H}} \mathbb{E}_k\|\xi_{v,i}^k\|^2 \le \frac{\Sigma_v}{n_h}. \tag{67}$$

For the client deviations, apply Assumption 4 to the three directions in (54). Using $\|\sum_{j=1}^m a_j\|^2 \le m\sum_{j=1}^m \|a_j\|^2$ with $m = 3, 2, 1$, respectively, yields

$$\frac{1}{n_h}\sum_{i\in\mathcal{H}} \|d_{x,i}^k\|^2 \le 3\zeta_f^2 + 6\lambda^2\zeta_g^2 = Z_x, \tag{68}$$

$$\frac{1}{n_h}\sum_{i\in\mathcal{H}} \|d_{y,i}^k\|^2 \le 2\zeta_f^2 + 2\lambda^2\zeta_g^2 = Z_y, \tag{69}$$

$$\frac{1}{n_h}\sum_{i\in\mathcal{H}} \|d_{z,i}^k\|^2 \le \zeta_g^2 = Z_z. \tag{70}$$

These estimates prove (57) and (58). The client-moment bounds will be used to control aggregation errors in Lemmas 5 and 9.

Finally, we establish (59). Write $y_\lambda^\star = y_\lambda^\star(x^k)$ and $y^\star = y^\star(x^k)$. Adding and subtracting $\nabla\mathcal{L}_\lambda^\star(x^k)$ gives

$$\nabla\Phi(x^k) - \bar{h}_x^k = b_{\lambda,k} + b_{y,k} + b_{z,k}, \tag{71}$$

where, by (22),

$$b_{\lambda,k} := \nabla\Phi(x^k) - \nabla\mathcal{L}_\lambda^\star(x^k), \tag{72}$$

$$b_{y,k} := \nabla_x f(x^k, y_\lambda^\star) - \nabla_x f(x^k, y^k) + \lambda[\nabla_x g(x^k, y_\lambda^\star) - \nabla_x g(x^k, y^k)], \tag{73}$$

$$b_{z,k} := \lambda[\nabla_x g(x^k, z^k) - \nabla_x g(x^k, y^\star)]. \tag{74}$$

Applying Lemma 1 to $b_{\lambda,k}$ and gradient Lipschitz continuity to the other two terms yields

$$\begin{aligned} \|b_{\lambda,k}\|^2 &\le \frac{C_{\mathrm{lag}}^2}{\lambda^2}, \\ \|b_{y,k}\|^2 &\le L_\lambda^2\|y^k - y_\lambda^\star(x^k)\|^2, \\ \|b_{z,k}\|^2 &\le \lambda^2 l_{g,1}^2\|z^k - y^\star(x^k)\|^2. \end{aligned} \tag{75}$$

Apply $\|a+b+c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$ to (71). This proves (59) with

$$D_y := 3L_\lambda^2, \qquad D_z := 3\lambda^2 l_{g,1}^2. \tag{76}$$

□

The conditional estimates below are integrable at every finite iteration. Indeed, globally Lipschitz gradients have at most linear growth, and Assumption 3 bounds the conditional second moments of their noise. Definition 2 bounds the squared server direction by $2(1+\kappa)n_h^{-1}\sum_{i\in\mathcal{H}}\|m_{v,i}^k\|^2$. Starting from deterministic iterates and zero momentum buffers, induction therefore gives finite second moments of all iterates and honest buffers. Lemma 2 then gives integrable tracking errors and function values, without a bounded-iterate assumption.

The following lemma controls the change in an optimizer map after an upper-level update. We separate the deterministic honest direction from its sampling noise, expand the optimizer map around the deterministic update, and then bound the displacement due to robust aggregation. This refinement of the drift argument in Kwon et al. (2023, Lemmas A.5–A.6) uses only conditional second moments of the noise and requires no uniform bound on $\bar{h}_x$.

Lemma 4 (Optimizer drift)**.** *Under the corresponding assumptions of Lemma 2, let q denote either $y^\star$ or $y_\lambda^\star$, with Lipschitz constants L_q for q and S_q for its Jacobian. In particular, $q = y_\lambda^\star$ requires the additional Hessian regularity of f stated there. Suppose*

$$x^{\mathrm{av}} = x - \eta_x(\bar{h}_x + \epsilon_x), \qquad x^+ = x^{\mathrm{av}} - \eta_x r_x. \tag{77}$$

Let u and r_x be square-integrable, with u determined by the history and the other-block reports, independently of the current honest x-block draws conditional on the history. Define the honest-average noise scale

$$\Gamma_{x,\lambda} := \frac{\Sigma_x}{n_h} = \frac{\sigma_f^2 + 2\lambda^2\sigma_g^2}{n_h}. \tag{78}$$

For any $\chi \in (0,1]$, if $\eta_x^2 \Gamma_{x,\lambda} \le c_q \chi$ for a sufficiently small constant $c_q > 0$ depending only on the regularity of the optimizer maps, then

$$\begin{aligned} &2\mathbb{E}_k\langle u, q(x) - q(x^+)\rangle + \mathbb{E}_k\|q(x^+) - q(x)\|^2 \\ &\quad\le \chi\,\mathbb{E}_k\|u\|^2 + \frac{C_q\eta_x^2}{\chi}\big(\|\bar h_x\|^2 + \mathbb{E}_k\|r_x\|^2\big) \\ &\quad\quad + C_q\eta_x^2\left(\frac{\Sigma_x}{n_h} + \mathbb{E}_k\|r_x\|^2\right), \end{aligned} \tag{79}$$

where C_q depends only on the regularity constants, not on the oracle variances.

Proof. Fix a round k and suppress the iteration superscripts. Throughout the proof, C_q denotes a positive constant depending only on L_q, S_q and may increase from line to line.

We first specify the conditioning needed for the proof. Let $\mathcal{K}$ be the sigma-field generated by $\mathcal{F}_k$ and the information from the other blocks that determines u, including their reports and any randomness used in their aggregation. Thus u is $\mathcal{K}$-measurable, while x and $\bar h_x$ are already $\mathcal{F}_k$-measurable. Assumption 3 makes the honest oracle draws conditionally independent across blocks. Moreover, the submission protocol in Section 3 prevents the other-block reports from depending on the current honest x-block draws. We may therefore condition on $\mathcal{K}$ while retaining the centering, variance bounds, and independence across honest clients for the x-block errors. In the notation of Lemma 3, this means

$$\mathbb{E}[\xi_{x,i} \mid \mathcal{K}] = 0, \qquad \mathbb{E}[\|\xi_{x,i}\|^2 \mid \mathcal{K}] \le \Sigma_x, \qquad i \in \mathcal{H}. \tag{80}$$

Since $\epsilon_x = n_h^{-1}\sum_{i\in\mathcal{H}} \xi_{x,i}$, the cross terms vanish on expanding its squared norm. Hence

$$\begin{aligned} \mathbb{E}[\epsilon_x \mid \mathcal{K}] &= 0, \\ \mathbb{E}[\|\epsilon_x\|^2 \mid \mathcal{K}] &= \frac{1}{n_h^2}\sum_{i\in\mathcal{H}} \mathbb{E}[\|\xi_{x,i}\|^2 \mid \mathcal{K}] \le \frac{\Sigma_x}{n_h} = \Gamma_{x,\lambda}, \\ \mathbb{E}[\|\bar h_x + \epsilon_x\|^2 \mid \mathcal{K}] &= \|\bar h_x\|^2 + \mathbb{E}[\|\epsilon_x\|^2 \mid \mathcal{K}] \\ &\le \|\bar h_x\|^2 + \Gamma_{x,\lambda}. \end{aligned} \tag{81}$$

The deterministic direction remains explicit in the last line; it need not be uniformly bounded.

Set $x^{\mathrm{d}} := x - \eta_x \bar h_x$, which is $\mathcal{F}_k$-measurable, and $\Delta_\epsilon := x^{\mathrm{av}} - x^{\mathrm{d}} = -\eta_x\epsilon_x$. Expand only the noise displacement around x^{d}. The integral form of Taylor's theorem gives

$$q(x) - q(x^{\mathrm{av}}) = q(x) - q(x^{\mathrm{d}}) + \eta_x \nabla q(x^{\mathrm{d}})\epsilon_x - \mathrm{Rem}_x, \tag{82}$$

$$\mathrm{Rem}_x := \int_0^1 [\nabla q(x^{\mathrm{d}} + t\Delta_\epsilon) - \nabla q(x^{\mathrm{d}})]\Delta_\epsilon\, dt. \tag{83}$$

By the S_q-Lipschitz continuity of ∇q, $\|\,\mathrm{Rem}_x\,\| \le S_q\eta_x^2\|\epsilon_x\|^2/2$. Since u and $\nabla q(x^{\mathrm{d}})$ are $\mathcal{K}$-measurable, the tower property and (81) give

$$\mathbb{E}_k\langle u, \nabla q(x^{\mathrm{d}})\epsilon_x\rangle = \mathbb{E}\big[\langle u, \nabla q(x^{\mathrm{d}})\mathbb{E}[\epsilon_x \mid \mathcal{K}]\rangle \,\big|\, \mathcal{F}_k\big] = 0. \tag{84}$$

For the deterministic displacement, Young's inequality and Lipschitz continuity of q yield

$$2\mathbb{E}_k\langle u, q(x) - q(x^{\mathrm{d}})\rangle \le \frac{\chi}{4}\mathbb{E}_k\|u\|^2 + \frac{4L_q^2\eta_x^2}{\chi}\|\bar h_x\|^2. \tag{85}$$

For the remainder, Cauchy–Schwarz and $2\|u\| \le \|u\|^2 + 1$ imply the pointwise bound

$$2|\langle u, \mathrm{Rem}_x\rangle| \le S_q\eta_x^2\|u\|\|\epsilon_x\|^2 \le \frac{S_q\eta_x^2}{2}(\|u\|^2 + 1)\|\epsilon_x\|^2. \tag{86}$$

To control the product on the right, condition on $\mathcal{K}$ before averaging over the other blocks. Since u is $\mathcal{K}$-measurable and $\mathcal{F}_k \subseteq \mathcal{K}$,

$$\begin{aligned} \mathbb{E}_k[\|u\|^2\|\epsilon_x\|^2] &= \mathbb{E}\big[\|u\|^2\mathbb{E}[\|\epsilon_x\|^2 \mid \mathcal{K}] \,\big|\, \mathcal{F}_k\big] \\ &\le \Gamma_{x,\lambda}\mathbb{E}_k\|u\|^2. \end{aligned} \tag{87}$$

The same conditioning without the factor $\|u\|^2$ gives $\mathbb{E}_k\|\epsilon_x\|^2 \le \Gamma_{x,\lambda}$. Consequently,

$$2\mathbb{E}_k|\langle u, \mathrm{Rem}_x\rangle| \le \frac{S_q}{2}\Gamma_{x,\lambda}\eta_x^2\mathbb{E}_k\|u\|^2 + \frac{S_q\eta_x^2\Sigma_x}{2n_h}. \tag{88}$$

This argument uses only conditional second moments of the oracle noise. Lipschitz continuity of q also gives

$$\mathbb{E}_k\|q(x^{\mathrm{av}}) - q(x)\|^2 \le L_q^2\eta_x^2\mathbb{E}_k\|\bar{h}_x + \epsilon_x\|^2 \le L_q^2\eta_x^2\left(\|\bar{h}_x\|^2 + \frac{\Sigma_x}{n_h}\right). \tag{89}$$

Choose $c_q := 1/[2(1+S_q)]$, so that $(S_q/2)c_q \le 1/4$. The stepsize condition then makes the coefficient of $\mathbb{E}_k\|u\|^2$ in the remainder bound at most $\chi/4$. Together with the $\chi/4$ from the deterministic displacement, this gives $\chi/2$. Substituting the preceding estimates into the Taylor expansion, using $\chi \le 1$, and absorbing constants depending on L_q, S_q into C_q, we obtain

$$\begin{aligned} &2\mathbb{E}_k\langle u, q(x) - q(x^{\mathrm{av}})\rangle + 2\mathbb{E}_k\|q(x^{\mathrm{av}}) - q(x)\|^2 \\ &\quad\le \frac{\chi}{2}\mathbb{E}_k\|u\|^2 + \frac{C_q\eta_x^2}{\chi}\|\bar{h}_x\|^2 + C_q\eta_x^2\frac{\Sigma_x}{n_h}. \end{aligned} \tag{90}$$

For example, $C_q := 1 + 6L_q^2 + S_q$ suffices for this bound and all remaining bounds below.

It remains to account for the displacement from x^{av} to x^+. Since $x^+ - x^{\mathrm{av}} = -\eta_x r_x$, Lipschitz continuity of q gives

$$\|q(x^+) - q(x^{\mathrm{av}})\|^2 \le L_q^2\eta_x^2\|r_x\|^2. \tag{91}$$

Set $a := q(x) - q(x^{\mathrm{av}})$ and $b := q(x^{\mathrm{av}}) - q(x^+)$. Young's inequality and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ give

$$\begin{aligned} 2\langle u, a+b\rangle + \|a+b\|^2 \le{}& 2\langle u, a\rangle + 2\|a\|^2 + \frac{\chi}{2}\|u\|^2 \\ &+ \left(2 + \frac{2}{\chi}\right)\|b\|^2. \end{aligned} \tag{92}$$

The first two terms are bounded by (90). For the last term, (91) and $\chi \le 1$ imply

$$\left(2 + \frac{2}{\chi}\right)\mathbb{E}_k\|b\|^2 \le \frac{C_q\eta_x^2}{\chi}\mathbb{E}_k\|r_x\|^2. \tag{93}$$

Taking conditional expectations in (92), the honest-average bound contributes $(\chi/2)\mathbb{E}_k\|u\|^2$ and the final Young bound contributes the remaining $(\chi/2)\mathbb{E}_k\|u\|^2$. Substituting the bounds for the other terms yields

$$\begin{aligned} &2\mathbb{E}_k\langle u, q(x) - q(x^+)\rangle + \mathbb{E}_k\|q(x^+) - q(x)\|^2 \\ &\quad\le \chi\mathbb{E}_k\|u\|^2 + \frac{C_q\eta_x^2}{\chi}\left(\|\bar{h}_x\|^2 + \mathbb{E}_k\|r_x\|^2\right) + C_q\eta_x^2\frac{\Sigma_x}{n_h}. \end{aligned} \tag{94}$$

Adding the nonnegative term $C_q\eta_x^2\mathbb{E}_k\|r_x\|^2$ gives (79). The argument does not require independence between r_x and the honest x-block noise. □

D CONVERGENCE ANALYSIS OF FOBFBL

We prove Theorems 1 and 2 using the aggregation, descent, and tracking bounds below. Throughout this section, the assumptions of Theorem 1 hold, with fixed $\lambda \geq 2l_{f,1}/\mu_g$ and $\beta_x = \beta_y = \beta_z = 0$. The PL condition is imposed in Appendix D.3.

D.1 AUXILIARY LEMMAS

We first decompose the update errors and state the aggregation, descent, and tracking estimates used in both convergence proofs. The proofs of these estimates are given in Appendix D.4.

D.1.1 UPDATE AND ERROR DECOMPOSITION

Let $\mathcal{F}_k$ denote the history before the honest samples in round k are drawn, and write $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$ and $\mathbb{E}[\cdot]$ for conditional and total expectations, respectively. Positive constants denoted by c and C may change from line to line; they depend on the fixed multiplier and problem parameters, but not on k, K, or the stepsizes.

With zero momentum, $m_{v,i}^k = h_{v,i}^k$ and $\widehat{m}_v^k = \widehat{h}_v^k$; the server aggregates the current stochastic directions and updates

$$\widehat{h}_v^k = \mathcal{A}\big(\{h_{v,i}^k\}_{i\in[n]}\big), \qquad v^{k+1} = v^k - \eta_v \widehat{h}_v^k, \qquad v \in \{x, y, z\}. \tag{95}$$

For each block, define the honest mean direction, its conditional mean, the sampling noise, and the aggregation error by

$$h_v^k := \frac{1}{n_h}\sum_{i\in\mathcal{H}} h_{v,i}^k, \qquad \bar{h}_v^k := \mathbb{E}_k[h_v^k], \qquad \epsilon_v^k := h_v^k - \bar{h}_v^k, \qquad r_v^k := \widehat{h}_v^k - h_v^k. \tag{96}$$

The update can therefore be written as

$$v^{k+1} = v^k - \eta_v(\bar{h}_v^k + \epsilon_v^k + r_v^k), \qquad v \in \{x, y, z\}. \tag{97}$$

By Lemma 3, ϵ_v^k is conditionally centered with second moment at most Σ_v/n_h. The aggregation error r_v^k may depend on the honest messages and need not be centered. We retain these errors separately to distinguish the $\mathcal{O}(\eta_v^2)$ contribution of sampling variance from the $\mathcal{O}(\eta_v)$ contribution of robust aggregation.

To quantify the accuracy of the two lower-level iterates, define the tracking errors

$$I_k := \mathbb{E}\|y^k - y_\lambda^\star(x^k)\|^2, \qquad J_k := \mathbb{E}\|z^k - y^\star(x^k)\|^2, \tag{98}$$

and the second moments of the true hypergradient, the honest direction, and the aggregation error

$$G_k := \mathbb{E}\|\nabla\Phi(x^k)\|^2, \qquad H_k := \mathbb{E}\|\bar{h}_x^k\|^2, \qquad R_{v,k} := \mathbb{E}\|r_v^k\|^2. \tag{99}$$

D.1.2 AGGREGATION ERROR

The aggregation error r_v^k measures the deviation of the server direction from the honest stochastic average. The following lemma bounds this error separately from the sampling noise ϵ_v^k.

Lemma 5 (Aggregation error)**.** *Suppose Assumptions 3 and 4 hold and $\mathcal{A}$ satisfies Definition 2 almost surely. For* FOBFBL*, every $k \geq 0$ and $v \in \{x, y, z\}$ satisfy*

$$\mathbb{E}_k\|r_v^k\|^2 \leq \kappa(\Sigma_v + Z_v), \qquad R_{v,k} \leq \kappa(\Sigma_v + Z_v), \tag{100}$$

where Σ_v and Z_v are defined in (55)–(56). In particular,

$$R_{x,k} \leq \kappa\big(\sigma_f^2 + 2\lambda^2\sigma_g^2 + 3\zeta_f^2 + 6\lambda^2\zeta_g^2\big), \tag{101}$$

$$R_{y,k} \leq \kappa\big(\sigma_f^2 + \lambda^2\sigma_g^2 + 2\zeta_f^2 + 2\lambda^2\zeta_g^2\big), \tag{102}$$

$$R_{z,k} \leq \kappa\big(\sigma_g^2 + \zeta_g^2\big). \tag{103}$$

Proof. See Appendix D.4.1.

For the corresponding error in the iterate, let $v_{\mathrm{av}}^{k+1} := v^k - \eta_v h_v^k$ be the update using the honest stochastic average. Then

$$\mathbb{E}\|v^{k+1} - v_{\mathrm{av}}^{k+1}\|^2 = \eta_v^2 R_{v,k} \le \eta_v^2 \kappa(\Sigma_v + Z_v). \tag{104}$$

The direction error is controlled by the dispersion of individual honest reports, so its bound has no $1/n_h$ factor. In Appendix D.2, the terms proportional to $\eta R_{v,k}$ therefore contribute both stochastic variance and client heterogeneity to the Byzantine-induced error floor.

D.1.3 Upper-Level Descent

The following lemma bounds the change in the honest upper-level objective. Its proof uses Lemma 1 to separate the finite-multiplier bias from the tracking errors.

Lemma 6 (Upper-level descent). *There is a stepsize ceiling $\eta_{\mathrm{d}} > 0$ such that, whenever $\eta_x \le \eta_{\mathrm{d}}$,*

$$\begin{aligned}\mathbb{E}\Phi(x^{k+1}) \le \mathbb{E}\Phi(x^k) - c\eta_x G_k - c\eta_x H_k + C\eta_x\left(I_k + J_k + \frac{C_{\mathrm{lag}}^2}{\lambda^2}\right) \\ + C\eta_x R_{x,k} + C\eta_x^2\left(\frac{\Sigma_x}{n_h} + R_{x,k}\right).\end{aligned} \tag{105}$$

Proof. See Appendix D.4.2.

In (105), the G_k term yields the stationarity guarantee, while the negative H_k term absorbs the target-drift terms in the tracking recursions below. The terms involving I_k, J_k, and $C_{\mathrm{lag}}^2/\lambda^2$ account for the error between the honest update direction and the true hypergradient.

D.1.4 Lower-Level Tracking

For the penalized lower-level problem, strong convexity gives contraction toward $y_\lambda^\star(x^k)$. Lemma 4 controls the displacement of this solution under the subsequent x-update, yielding the following recursion. Here $\Gamma_{x,\lambda} = \Sigma_x/n_h$ is the honest-average noise bound in (78).

Lemma 7 (Penalized lower-level tracking). *There are positive constants c_y, C_y and a stepsize ceiling $\eta_{\mathrm{t},y}$ such that*

$$\begin{aligned}I_{k+1} \le (1 - c_y\mu_\lambda\eta_y)I_k + \frac{C_y\eta_x^2}{\mu_\lambda\eta_y}(H_k + R_{x,k}) \\ + C_y\eta_x^2\frac{\Sigma_x}{n_h} + C_y\eta_y^2\frac{\Sigma_y}{n_h} + C_y\left(\frac{\eta_y}{\mu_\lambda} + \eta_y^2\right)R_{y,k}.\end{aligned} \tag{106}$$

This holds when $\eta_y L_\lambda$ and η_x are at most $\eta_{\mathrm{t},y}$ and $\Gamma_{x,\lambda}\eta_x^2 \le c\mu_\lambda\eta_y$.

Proof. See Appendix D.4.3.

The terms involving H_k and $R_{x,k}$ in (106) arise from target drift. The remaining error terms distinguish sampling variance in the x- and y-blocks from robust aggregation in the y-block.

The same argument applies to z^k, which tracks the minimizer of $g(x^k, \cdot)$.

Lemma 8 (Original lower-level tracking). *There are positive constants c_z, C_z and a stepsize ceiling $\eta_{\mathrm{t},z}$ such that*

$$\begin{aligned}J_{k+1} \le (1 - c_z\mu_g\eta_z)J_k + \frac{C_z\eta_x^2}{\mu_g\eta_z}(H_k + R_{x,k}) \\ + C_z\eta_x^2\frac{\Sigma_x}{n_h} + C_z\eta_z^2\frac{\Sigma_z}{n_h} + C_z\left(\frac{\eta_z}{\mu_g} + \eta_z^2\right)R_{z,k}.\end{aligned} \tag{107}$$

This holds when $\eta_z l_{g,1}$ and η_x are at most $\eta_{\mathrm{t},z}$ and $\Gamma_{x,\lambda}\eta_x^2 \le c\mu_g\eta_z$.

Proof. See Appendix D.4.4.

Both tracking recursions consist of a contraction term, target drift, and sampling and aggregation errors. We combine them with Lemma 6 using fixed stepsize ratios and positive Lyapunov weights. The resulting recursion proves Theorem 1 by telescoping and Theorem 2 by applying the PL inequality.

D.2 Proof of Theorem 1: Nonconvex Convergence

Theorem 1 (Convergence of FOBFBL for nonconvex upper-level objectives)**.** *Suppose Assumptions 1–5 hold and $\mathcal{A}$ satisfies Definition 2 almost surely. Fix $\lambda \geq 2l_{f,1}/\mu_g$, assume $\Sigma_\lambda > 0$, and let $\Delta_0 > 0$ denote the initial Lyapunov gap defined in Appendix D. For sufficiently large K, choose $\eta = \sqrt{n_h \Delta_0/(K\Sigma_\lambda)}$, set $\eta_x = \eta$, $\eta_y = c_y^\circ \eta_x$, and $\eta_z = c_z^\circ \eta_x$ for the positive problem-dependent constants c_y°, c_z°, and take $\beta_x = \beta_y = \beta_z = 0$. With deterministic initialization, x^R satisfies*

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 = \mathcal{O}\Bigg(\underbrace{\sqrt{\frac{\Delta_0\Sigma_\lambda}{n_h K}}}_{\text{finite-horizon error}} + \underbrace{\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\text{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}} \Bigg). \tag{15}$$

Proof. Let $\eta_x = \eta$, $\eta_y = c_y^\circ \eta$, and $\eta_z = c_z^\circ \eta$, where the positive ratios c_y°, c_z° will be specified below. For now, take $\eta > 0$ sufficiently small for the descent and tracking bounds to hold. Write $\Delta\Phi_k := \mathbb{E}[\Phi(x^{k+1}) - \Phi(x^k)]$, $\Delta I_k := I_{k+1} - I_k$, and $\Delta J_k := J_{k+1} - J_k$. Substituting the stepsize ratios into Lemmas 6, 7, and 8 gives

$$\begin{aligned}
\Delta\Phi_k &\leq -d_G\,\eta\,G_k - d_H\,\eta\,H_k + \eta\left(D_I\,I_k + D_J\,J_k + D_B\frac{C_{\text{lag}}^2}{\lambda^2}\right) \\
&\quad + (D_R\eta + D_2\eta^2)\,R_{x,k} + D_2\eta^2\frac{\Sigma_x}{n_h}, \qquad (108)\\
\Delta I_k &\leq -t_y\mu_\lambda c_y^\circ\eta\,I_k + \frac{Q_y\eta}{\mu_\lambda c_y^\circ}(H_k + R_{x,k}) + Q_y\eta^2\frac{\Sigma_x + (c_y^\circ)^2\Sigma_y}{n_h} \\
&\quad + Q_y\left(\frac{c_y^\circ\eta}{\mu_\lambda} + (c_y^\circ)^2\eta^2\right)R_{y,k}, \qquad (109)\\
\Delta J_k &\leq -t_z\mu_g c_z^\circ\eta\,J_k + \frac{Q_z\eta}{\mu_g c_z^\circ}(H_k + R_{x,k}) + Q_z\eta^2\frac{\Sigma_x + (c_z^\circ)^2\Sigma_z}{n_h} \\
&\quad + Q_z\left(\frac{c_z^\circ\eta}{\mu_g} + (c_z^\circ)^2\eta^2\right)R_{z,k}. \qquad (110)
\end{aligned}$$

The positive constants in these bounds are independent of k, K, and η.

To combine these inequalities, we choose the stepsize ratios and Lyapunov weights so that tracking contraction absorbs the positive I_k and J_k terms, while upper-level descent absorbs the target-drift terms proportional to H_k. Specifically, take

$$\begin{aligned}
(c_y^\circ)^2 &\geq \frac{8D_I Q_y}{t_y\mu_\lambda^2 d_H}, \qquad (c_z^\circ)^2 \geq \frac{8D_J Q_z}{t_z\mu_g^2 d_H}, \\
A_1 &:= \frac{2D_I}{t_y\mu_\lambda c_y^\circ}, \qquad A_2 := \frac{2D_J}{t_z\mu_g c_z^\circ}.
\end{aligned} \tag{111}$$

These ratios and weights depend only on the fixed multiplier and problem parameters. Define the Lyapunov function and its initial value by

$$V_k := \mathbb{E}[\Phi(x^k) - \Phi_*] + A_1\,I_k + A_2\,J_k, \qquad \Delta_0 := V_0. \tag{112}$$

Since each component is nonnegative, $V_k \geq 0$. The weights are constant over the iterations, so the increment of V_k is exactly

$$V_{k+1} - V_k = \Delta\Phi_k + A_1\Delta I_k + A_2\Delta J_k. \tag{113}$$

Substituting (108)–(110) into (113) and collecting the coefficients of each state gives

$$\begin{aligned} V_{k+1} - V_k \le & -d_G\,\eta\,G_k + \eta\underbrace{\left(D_I - A_1 t_y \mu_\lambda c_y^\circ\right)}_{T_1} I_k \\ & + \eta\underbrace{\left(D_J - A_2 t_z \mu_g c_z^\circ\right)}_{T_2} J_k \\ & + \eta\underbrace{\left(-d_H + \frac{A_1 Q_y}{\mu_\lambda c_y^\circ} + \frac{A_2 Q_z}{\mu_g c_z^\circ}\right)}_{T_3} H_k + \mathrm{Res}_k(\eta). \end{aligned} \tag{114}$$

Here $\mathrm{Res}_k(\eta)$ collects the sampling, aggregation, and finite-multiplier terms. The choices in (111) give

$$\begin{aligned} T_1 &= D_I - \frac{2D_I}{t_y \mu_\lambda c_y^\circ} t_y \mu_\lambda c_y^\circ = -D_I, \\ T_2 &= D_J - \frac{2D_J}{t_z \mu_g c_z^\circ} t_z \mu_g c_z^\circ = -D_J. \end{aligned} \tag{115}$$

For the direction term, substituting the same weights and then using the lower bounds on the squared ratios yields

$$\begin{aligned} T_3 &= -d_H + \frac{2D_I Q_y}{t_y \mu_\lambda^2 (c_y^\circ)^2} + \frac{2D_J Q_z}{t_z \mu_g^2 (c_z^\circ)^2} \\ &\le -d_H + \frac{d_H}{4} + \frac{d_H}{4} = -\frac{d_H}{2}. \end{aligned} \tag{116}$$

Thus all three state coefficients are strictly negative. The ratio conditions depend only on the fixed problem parameters, and finite positive ratios satisfying them always exist.

It remains to bound the residual in (114), which is given by

$$\begin{aligned} \mathrm{Res}_k(\eta) := & D_B\,\eta \frac{C_{\mathrm{lag}}^2}{\lambda^2} + \eta\,T_{R_x}(\eta)\,R_{x,k} + \eta\,T_{R_y}(\eta)\,R_{y,k} + \eta\,T_{R_z}(\eta)\,R_{z,k} \\ & + \frac{\eta^2}{n_h}\left[D_2 \Sigma_x + A_1 Q_y (\Sigma_x + (c_y^\circ)^2 \Sigma_y) + A_2 Q_z (\Sigma_x + (c_z^\circ)^2 \Sigma_z)\right], \end{aligned} \tag{117}$$

where the aggregation coefficients are

$$\begin{aligned} T_{R_x}(\eta) &:= D_R + D_2 \eta + \frac{A_1 Q_y}{\mu_\lambda c_y^\circ} + \frac{A_2 Q_z}{\mu_g c_z^\circ}, \\ T_{R_y}(\eta) &:= A_1 Q_y \left(\frac{c_y^\circ}{\mu_\lambda} + (c_y^\circ)^2 \eta\right), \\ T_{R_z}(\eta) &:= A_2 Q_z \left(\frac{c_z^\circ}{\mu_g} + (c_z^\circ)^2 \eta\right). \end{aligned} \tag{118}$$

We first specify a stepsize range on which all preceding recursions hold. Let $a_{\mathrm{t}} > 0$ be a common sufficiently small constant for the two target-drift conditions in Lemmas 7 and 8. With the ratios fixed, take

$$\begin{aligned} \bar{\eta} := \min\Bigg\{ & 1, \eta_{\mathrm{d}}, \eta_{\mathrm{t},y}, \eta_{\mathrm{t},z}, \frac{\eta_{\mathrm{t},y}}{c_y^\circ L_\lambda}, \frac{\eta_{\mathrm{t},z}}{c_z^\circ l_{g,1}}, \\ & \frac{a_{\mathrm{t}} \mu_\lambda c_y^\circ}{\Gamma_{x,\lambda}}, \frac{a_{\mathrm{t}} \mu_g c_z^\circ}{\Gamma_{x,\lambda}} \Bigg\} > 0. \end{aligned} \tag{119}$$

The first six entries enforce the descent and tracking stepsize ceilings. The last two give, for every $0 < \eta \le \bar{\eta}$,

$$\begin{aligned} \Gamma_{x,\lambda} \eta_x^2 = \Gamma_{x,\lambda} \eta^2 \le a_{\mathrm{t}} \mu_\lambda c_y^\circ \eta = a_{\mathrm{t}} \mu_\lambda \eta_y, \\ \Gamma_{x,\lambda} \eta_x^2 \le a_{\mathrm{t}} \mu_g c_z^\circ \eta = a_{\mathrm{t}} \mu_g \eta_z. \end{aligned} \tag{120}$$

Hence the target-drift conditions hold as well. All entries in (119) are fixed before choosing the horizon K.

Each coefficient in (118) is affine and increasing in η. Since $\bar{\eta} \le 1$, define the uniform bounds

$$0 < T_{R_v}(\eta) \le \overline{T}_v := T_{R_v}(1), \qquad \overline{T} := \max_{v\in\{x,y,z\}} \overline{T}_v. \tag{121}$$

Lemma 5 gives $R_{v,k} \le \kappa(\Sigma_v + Z_v)$. To combine the three blocks, use

$$\begin{aligned} \Sigma_x \le 2\Sigma_\lambda, \qquad \Sigma_y = \Sigma_\lambda, \qquad \Sigma_z \le \lambda^{-2}\Sigma_\lambda, \\ Z_x + Z_y + Z_z = 5\zeta_f^2 + (8+\lambda^{-2})\lambda^2\zeta_g^2. \end{aligned} \tag{122}$$

It follows that the aggregation part of the residual satisfies

$$\begin{aligned} \eta \sum_{v\in\{x,y,z\}} T_{R_v}(\eta) R_{v,k} &\le \kappa\eta \sum_{v\in\{x,y,z\}} \overline{T}_v(\Sigma_v + Z_v) \\ &\le \overline{T}\kappa\eta\Big[(3+\lambda^{-2})\Sigma_\lambda + 5\zeta_f^2 + (8+\lambda^{-2})\lambda^2\zeta_g^2\Big] \\ &\le C_{\mathrm{agg}}\kappa\eta(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2), \end{aligned} \tag{123}$$

where $C_{\mathrm{agg}} := \overline{T}(8+\lambda^{-2})$ is independent of k, K, and η. For the honest-sampling part, rearranging its coefficients and applying (122) gives

$$\begin{aligned} &D_2\Sigma_x + A_1Q_y\big(\Sigma_x + (c_y^\circ)^2\Sigma_y\big) + A_2Q_z\big(\Sigma_x + (c_z^\circ)^2\Sigma_z\big) \\ &\quad = (D_2 + A_1Q_y + A_2Q_z)\Sigma_x + A_1Q_y(c_y^\circ)^2\Sigma_y + A_2Q_z(c_z^\circ)^2\Sigma_z \\ &\quad \le \Big[2D_2 + A_1Q_y\big(2 + (c_y^\circ)^2\big) + A_2Q_z\big(2 + (c_z^\circ)^2/\lambda^2\big)\Big]\Sigma_\lambda =: C_{\mathrm{var}}\Sigma_\lambda. \end{aligned} \tag{124}$$

Multiplication by η^2/n_h gives its contribution to (117). Consequently,

$$\mathrm{Res}_k(\eta) \le C_{\mathrm{var}}\eta^2\frac{\Sigma_\lambda}{n_h} + C_{\mathrm{agg}}\kappa\eta(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2) + D_B\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{125}$$

Taking $C \ge \max\{C_{\mathrm{var}}, C_{\mathrm{agg}}, D_B\}$ and substituting (115), (116), and (125) into (114) gives

$$\begin{aligned} V_{k+1} - V_k \le -\eta\left(c_G\,G_k + c_I\,I_k + c_J\,J_k + \frac{d_H}{2}H_k\right) + C\eta^2\frac{\Sigma_\lambda}{n_h} \\ + C\eta\,\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2) + C\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2}, \end{aligned} \tag{126}$$

where $c_G = d_G$, $c_I = D_I$, and $c_J = D_J$ are positive. We retain this stronger inequality for the PL analysis in Appendix D.3. For the present stationarity result, discarding the nonpositive I_k, J_k, and H_k terms gives

$$V_{k+1} - V_k \le -c_G\,\eta\,G_k + C\eta^2\frac{\Sigma_\lambda}{n_h} + C\eta\,\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2) + C\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{127}$$

Because the same stepsize is used throughout the K iterations, the potential increments telescope as

$$\sum_{k=0}^{K-1}(V_{k+1} - V_k) = (V_1 - V_0) + \cdots + (V_K - V_{K-1}) = V_K - V_0. \tag{128}$$

Summing (127) and using (128) therefore yields

$$c_G\,\eta\sum_{k=0}^{K-1}G_k \le V_0 - V_K + CK\eta^2\frac{\Sigma_\lambda}{n_h} + CK\eta\,\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2) + CK\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{129}$$

Since $V_K \ge 0$ and $V_0 = \Delta_0$, division by $c_G\eta K$ gives

$$\frac{1}{K}\sum_{k=0}^{K-1}G_k \le C\left[\frac{\Delta_0}{\eta K} + \frac{\eta\Sigma_\lambda}{n_h} + \kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2}\right]. \tag{130}$$

Set $S := \Sigma_\lambda/n_h > 0$. The only terms on the right-hand side of (130) that depend on η are $\Delta_0/(\eta K) + \eta S$. For every $\eta > 0$, the arithmetic–geometric mean inequality gives

$$\frac{\Delta_0}{\eta K} + \eta S \geq 2\sqrt{\frac{\Delta_0 S}{K}}, \qquad \frac{\Delta_0}{\eta K} = \eta S \iff \eta = \eta_* := \sqrt{\frac{\Delta_0}{KS}}. \tag{131}$$

This unconstrained minimizing stepsize lies in the admissible range precisely when

$$\eta_* \leq \bar{\eta} \iff K \geq \frac{\Delta_0}{\bar{\eta}^2 S} = \frac{n_h \Delta_0}{\bar{\eta}^2 \Sigma_\lambda}. \tag{132}$$

Thus (132) specifies the sufficiently large horizon in Theorem 1. At this stepsize,

$$\frac{\Delta_0}{\eta_* K} + \eta_* S = 2\sqrt{\frac{\Delta_0 S}{K}} = 2\sqrt{\frac{\Delta_0 \Sigma_\lambda}{n_h K}}. \tag{133}$$

Finally, R is uniform on $\{0, \dots, K-1\}$ and independent of the iterates. The law of total expectation gives

$$\begin{aligned} \mathbb{E}\|\nabla\Phi(x^R)\|^2 &= \sum_{k=0}^{K-1} \mathbb{P}(R=k)\mathbb{E}\big[\|\nabla\Phi(x^k)\|^2 \mid R=k\big] \\ &= \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla\Phi(x^k)\|^2 = \frac{1}{K}\sum_{k=0}^{K-1} G_k. \end{aligned} \tag{134}$$

Substituting (133) into (130) therefore proves (15). □

D.3 Proof of Theorem 2: Convergence under the PL Condition

Theorem 2 (Convergence of FOBFBL under the PL condition)**.** *Under the same assumptions on the problem, aggregation rule, and multiplier as in Theorem 1, suppose Assumption 6 also holds. There exists $\gamma_{\mathrm{F}} > 0$ such that, for sufficiently large K, choosing $\eta = \log(n_h K)/(\gamma_{\mathrm{F}} K)$ with the stepsize ratios and zero momentum of Theorem 1 gives*

$$\mathbb{E}[\Phi(x^K) - \Phi_*] = \mathcal{O}\Bigg(\underbrace{\frac{\Delta_0}{n_h K} + \frac{\Sigma_\lambda \log(n_h K)}{n_h K}}_{\text{finite-horizon error}} + \underbrace{\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\mathrm{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}} \Bigg). \tag{16}$$

Proof. Fix the stepsize ratios and Lyapunov weights in (111), and let $0 < \eta \leq \bar{\eta}$, where $\bar{\eta}$ is the admissible ceiling established in (119). Recall the Lyapunov function

$$V_k = \mathbb{E}[\Phi(x^k) - \Phi_*] + A_1 I_k + A_2 J_k, \qquad V_0 = \Delta_0. \tag{135}$$

All three components are nonnegative. We use the PL condition to convert the gradient term in (126) into a bound on the objective gap; the negative tracking terms then give contraction of the remaining components of V_k.

Collect the sampling, aggregation, and finite-multiplier terms per unit stepsize as

$$r_{\mathrm{F}}(\eta) := \frac{\eta\Sigma_\lambda}{n_h} + \kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2\zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{136}$$

Dropping the nonpositive H_k term from (126) yields

$$V_{k+1} - V_k \leq -\eta\big(c_G\, G_k + c_I\, I_k + c_J\, J_k\big) + C\eta\, r_{\mathrm{F}}(\eta). \tag{137}$$

By Assumption 6, the PL inequality holds at each iterate. Taking expectations gives

$$G_k = \mathbb{E}\|\nabla\Phi(x^k)\|^2 \geq 2\mu_{\mathrm{PL}}\, \mathbb{E}[\Phi(x^k) - \Phi_*]. \tag{138}$$

To obtain a common contraction coefficient for all three components, define

$$\gamma_{\mathrm{F}} := \min\left\{ 2\mu_{\mathrm{PL}} c_G, \frac{c_I}{A_1}, \frac{c_J}{A_2} \right\} > 0. \tag{139}$$

This constant is positive and independent of k, K, and η, since the weights and the coefficients c_G, c_I, c_J were fixed in Appendix D.2. Combining (138) with the definition of γ_{F} gives

$$\begin{aligned} c_G G_k &\geq 2\mu_{\mathrm{PL}} c_G \, \mathbb{E}[\Phi(x^k) - \Phi_*] \geq \gamma_{\mathrm{F}} \, \mathbb{E}[\Phi(x^k) - \Phi_*], \\ c_I I_k &= \frac{c_I}{A_1}(A_1 I_k) \geq \gamma_{\mathrm{F}} A_1 I_k, \\ c_J J_k &= \frac{c_J}{A_2}(A_2 J_k) \geq \gamma_{\mathrm{F}} A_2 J_k. \end{aligned} \tag{140}$$

Since all three components are nonnegative, summing the bounds in (140) and using (135) gives

$$c_G \, G_k + c_I \, I_k + c_J \, J_k \geq \gamma_{\mathrm{F}} V_k. \tag{141}$$

Restrict η further to $0 < \eta \leq \min\{\bar{\eta}, 1/\gamma_{\mathrm{F}}\}$. Substituting (141) into (137) gives

$$V_{k+1} \leq (1 - \gamma_{\mathrm{F}}\eta) V_k + C\eta \, r_{\mathrm{F}}(\eta). \tag{142}$$

Let $q_{\mathrm{F}} := 1 - \gamma_{\mathrm{F}}\eta \in [0, 1)$. Because the stepsize is constant over the K iterations, so is $r_{\mathrm{F}}(\eta)$. Iterating (142) gives, for every $1 \leq j \leq K$,

$$V_K \leq q_{\mathrm{F}}^j V_{K-j} + C\eta r_{\mathrm{F}}(\eta) \sum_{t=0}^{j-1} q_{\mathrm{F}}^t. \tag{143}$$

Indeed, $j = 1$ is (142); applying that inequality to V_{K-j} adds the term $C\eta r_{\mathrm{F}}(\eta) q_{\mathrm{F}}^j$ and proves the case $j + 1$. For $j = K$, evaluate the finite geometric series using

$$\sum_{t=0}^{K-1} q_{\mathrm{F}}^t = \frac{1 - q_{\mathrm{F}}^K}{1 - q_{\mathrm{F}}} = \frac{1 - q_{\mathrm{F}}^K}{\gamma_{\mathrm{F}}\eta}. \tag{144}$$

Substitution of (144) into (143), together with $V_0 = \Delta_0$, yields

$$\begin{aligned} V_K &\leq q_{\mathrm{F}}^K V_0 + C\eta \, r_{\mathrm{F}}(\eta) \sum_{t=0}^{K-1} q_{\mathrm{F}}^t \\ &= q_{\mathrm{F}}^K \Delta_0 + \frac{C}{\gamma_{\mathrm{F}}}(1 - q_{\mathrm{F}}^K) r_{\mathrm{F}}(\eta). \end{aligned} \tag{145}$$

Since $\mathbb{E}[\Phi(x^K) - \Phi_*] \leq V_K$ and $1 - q_{\mathrm{F}}^K \leq 1$,

$$\mathbb{E}[\Phi(x^K) - \Phi_*] \leq (1 - \gamma_{\mathrm{F}}\eta)^K \Delta_0 + \frac{C}{\gamma_{\mathrm{F}}} \left[\frac{\eta \Sigma_\lambda}{n_h} + \kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2 \zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2} \right]. \tag{146}$$

Thus the objective gap at the last iterate is bounded by a geometric transient and the accumulated residual. The sampling contribution decreases with η, while the aggregation and finite-multiplier contributions persist at a fixed multiplier.

Choose the horizon and stepsize so that

$$n_h K > 1, \qquad \eta = \frac{\log(n_h K)}{\gamma_{\mathrm{F}} K} \leq \min\left\{ \bar{\eta}, \frac{1}{\gamma_{\mathrm{F}}} \right\}. \tag{147}$$

Such a horizon exists because $\log(n_h K)/K \to 0$ as $K \to \infty$ and the ceiling is a fixed positive constant. The condition $n_h K > 1$ ensures $\eta > 0$, while the upper bound makes all preceding recursions applicable. Using $1 - u \leq e^{-u}$ for $u \in [0, 1]$, we obtain

$$(1 - \gamma_{\mathrm{F}}\eta)^K \leq \exp(-\gamma_{\mathrm{F}}\eta K) = \exp(-\log(n_h K)) = \frac{1}{n_h K}. \tag{148}$$

The sampling term in (146) becomes

$$\frac{C}{\gamma_{\mathrm{F}}} \frac{\eta \Sigma_\lambda}{n_h} = \frac{C \Sigma_\lambda \log(n_h K)}{\gamma_{\mathrm{F}}^2 n_h K}. \tag{149}$$

Using (148) and (149) in (146) yields

$$\begin{aligned} \mathbb{E}[\Phi(x^K) - \Phi_*] \leq \frac{\Delta_0}{n_h K} &+ \frac{C \Sigma_\lambda \log(n_h K)}{\gamma_{\mathrm{F}}^2 n_h K} \\ &+ \frac{C}{\gamma_{\mathrm{F}}} \left[\kappa(\Sigma_\lambda + \zeta_f^2 + \lambda^2 \zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2} \right]. \end{aligned} \tag{150}$$

Absorbing the fixed factors involving γ_{F} into the implicit constant gives (16). □

D.4 Proofs of the Auxiliary Lemmas

We prove the aggregation-error, descent, and tracking lemmas stated in Appendix D.1, using the history $\mathcal{F}_k$ and the error decomposition introduced there. Recall that every block update has the form

$$v^{k+1} = v^k - \eta_v(\bar{h}_v^k + \epsilon_v^k + r_v^k), \qquad v \in \{x, y, z\}. \tag{151}$$

The blockwise timing condition makes the current honest block noise centered relative to all quantities revealed before that block. No such cancellation is used for r_v^k, which may depend on the full contaminated message set.

D.4.1 Proof of Lemma 5: Aggregation Error

Proof. Fix $k \geq 0$ and $v \in \{x, y, z\}$, and condition on $\mathcal{F}_k$. With zero momentum, the aggregator acts directly on the current reports $h_{v,i}^k$. Definition 2 holds for every realization of those reports, including arbitrary Byzantine values. Hence

$$\|r_v^k\|^2 = \left\|\mathcal{A}(\{h_{v,i}^k\}_{i\in[n]}) - h_v^k\right\|^2 \leq \frac{\kappa}{n_h}\sum_{i\in\mathcal{H}}\|h_{v,i}^k - h_v^k\|^2. \tag{152}$$

The honest empirical mean minimizes the sum of squared distances. More explicitly, $\sum_{i\in\mathcal{H}}(h_{v,i}^k - h_v^k) = 0$ gives the identity

$$\frac{1}{n_h}\sum_{i\in\mathcal{H}}\|h_{v,i}^k - h_v^k\|^2 = \frac{1}{n_h}\sum_{i\in\mathcal{H}}\|h_{v,i}^k - \bar{h}_v^k\|^2 - \|h_v^k - \bar{h}_v^k\|^2. \tag{153}$$

Taking conditional expectations in (152) and discarding the last, nonpositive term in (153) yields

$$\begin{aligned}\mathbb{E}_k\|r_v^k\|^2 &\leq \frac{\kappa}{n_h}\sum_{i\in\mathcal{H}}\mathbb{E}_k\|h_{v,i}^k - h_v^k\|^2 \\ &\leq \frac{\kappa}{n_h}\sum_{i\in\mathcal{H}}\mathbb{E}_k\|h_{v,i}^k - \bar{h}_v^k\|^2.\end{aligned} \tag{154}$$

To bound this dispersion, use the decomposition $h_{v,i}^k - \bar{h}_v^k = \xi_{v,i}^k + d_{v,i}^k$ from (60). The deviation $d_{v,i}^k$ is $\mathcal{F}_k$-measurable and the honest oracle noise satisfies $\mathbb{E}_k\xi_{v,i}^k = 0$, so

$$\begin{aligned}\mathbb{E}_k\|h_{v,i}^k - \bar{h}_v^k\|^2 &= \mathbb{E}_k\|\xi_{v,i}^k\|^2 + 2\langle\mathbb{E}_k\xi_{v,i}^k, d_{v,i}^k\rangle + \|d_{v,i}^k\|^2 \\ &= \mathbb{E}_k\|\xi_{v,i}^k\|^2 + \|d_{v,i}^k\|^2.\end{aligned} \tag{155}$$

Averaging over honest clients and applying the block-moment bounds in Lemma 3 gives

$$\begin{aligned}\frac{1}{n_h}\sum_{i\in\mathcal{H}}\mathbb{E}_k\|h_{v,i}^k - \bar{h}_v^k\|^2 &= \frac{1}{n_h}\sum_{i\in\mathcal{H}}\left(\mathbb{E}_k\|\xi_{v,i}^k\|^2 + \|d_{v,i}^k\|^2\right) \\ &\leq \Sigma_v + Z_v.\end{aligned} \tag{156}$$

Substitution into (154) proves

$$\mathbb{E}_k\|r_v^k\|^2 \leq \kappa(\Sigma_v + Z_v), \qquad R_{v,k} = \mathbb{E}\left[\mathbb{E}_k\|r_v^k\|^2\right] \leq \kappa(\Sigma_v + Z_v). \tag{157}$$

This is (100). Inserting the scales from (55)–(56) yields the three bounds (101)–(103). Only the honest oracle noise was centered in (155); the aggregation error r_v^k may depend on all current reports and need not have zero mean. Finally, the perturbed update (151) gives $v^{k+1} - v_{\mathrm{av}}^{k+1} = -\eta_v r_v^k$, which establishes (104). □

D.4.2 Proof of Lemma 6: Upper-Level Descent

Proof. Condition on $\mathcal{F}_k$ and abbreviate $p = \nabla\Phi(x^k)$, $q = \bar{h}_x^k$, $\epsilon = \epsilon_x^k$, and $r = r_x^k$. The proof separates two effects: the mismatch between the honest direction q and the true hypergradient p, and the perturbation introduced by robust aggregation. Since $\mathbb{E}_k\epsilon = 0$, smoothness gives

$$\mathbb{E}_k\Phi(x^{k+1}) \le \Phi(x^k) - \eta_x\langle p, q\rangle - \eta_x\mathbb{E}_k\langle p, r\rangle + \frac{L_\Phi\eta_x^2}{2}\mathbb{E}_k\|q+\epsilon+r\|^2, \tag{158}$$

where the aggregation error r may depend on the honest messages. We bound the quadratic term without assuming that r is centered or independent of ϵ:

$$\mathbb{E}_k\|q+\epsilon+r\|^2 \le 3\|q\|^2 + 3\frac{\Sigma_x}{n_h} + 3\mathbb{E}_k\|r\|^2. \tag{159}$$

The two linear terms satisfy

$$-\langle p, q\rangle = -\frac{1}{2}\|p\|^2 - \frac{1}{2}\|q\|^2 + \frac{1}{2}\|p-q\|^2, \tag{160}$$

$$-\mathbb{E}_k\langle p, r\rangle \le \frac{1}{8}\|p\|^2 + 2\mathbb{E}_k\|r\|^2. \tag{161}$$

These inequalities retain negative coefficients for both $\|p\|^2$ and $\|q\|^2$. To bound $\|p-q\|^2$, insert the value-gradient direction from (22):

$$\begin{aligned}\|p-q\|^2 &\le 3\|\nabla\Phi(x^k) - \nabla\mathcal{L}_\lambda^\star(x^k)\|^2 \\ &\quad + 3\|\nabla_x f(x^k, y_\lambda^\star) - \nabla_x f(x^k, y^k) + \lambda[\nabla_x g(x^k, y_\lambda^\star) - \nabla_x g(x^k, y^k)]\|^2 \\ &\quad + 3\lambda^2\|\nabla_x g(x^k, z^k) - \nabla_x g(x^k, y^\star)\|^2 \\ &\le 3\frac{C_{\mathrm{lag}}^2}{\lambda^2} + D_y\|y^k - y_\lambda^\star(x^k)\|^2 + D_z\|z^k - y^\star(x^k)\|^2.\end{aligned} \tag{162}$$

Here the optimizer arguments in the intermediate lines are all x^k. The constants are those in (76): gradient Lipschitz continuity gives

$$\begin{aligned}&\|\nabla_x f(x^k, y_\lambda^\star(x^k)) - \nabla_x f(x^k, y^k) + \lambda[\nabla_x g(x^k, y_\lambda^\star(x^k)) - \nabla_x g(x^k, y^k)]\|^2 \\ &\quad \le (l_{f,1} + \lambda l_{g,1})^2\|y^k - y_\lambda^\star(x^k)\|^2 = \frac{D_y}{3}\|y^k - y_\lambda^\star(x^k)\|^2, \\ &\lambda^2\|\nabla_x g(x^k, z^k) - \nabla_x g(x^k, y^\star(x^k))\|^2 \le \frac{D_z}{3}\|z^k - y^\star(x^k)\|^2.\end{aligned} \tag{163}$$

This decomposition is the source of the three deterministic error terms in Lemma 6: $\frac{C_{\mathrm{lag}}^2}{\lambda^2}$ is the finite-multiplier bias, whereas the other two terms measure imperfect tracking of the penalized and original lower-level solutions. In particular, no stochastic or Byzantine quantity is hidden in the direction mismatch. Substitution of equations (159)–(162) into equation (158) yields, before constants are collected,

$$\begin{aligned}\mathbb{E}_k\Phi(x^{k+1}) &\le \Phi(x^k) - \frac{3\eta_x}{8}\|p\|^2 - \left(\frac{\eta_x}{2} - \frac{3L_\Phi\eta_x^2}{2}\right)\|q\|^2 \\ &\quad + \frac{\eta_x}{2}\left(D_y\|y^k - y_\lambda^\star(x^k)\|^2 + D_z\|z^k - y^\star(x^k)\|^2 + 3\frac{C_{\mathrm{lag}}^2}{\lambda^2}\right) \\ &\quad + \left(2\eta_x + \frac{3L_\Phi\eta_x^2}{2}\right)\mathbb{E}_k\|r_x^k\|^2 + \frac{3L_\Phi\eta_x^2\Sigma_x}{2n_h}.\end{aligned} \tag{164}$$

Choose the descent ceiling so that

$$\eta_{\mathrm{d}} \le \frac{1}{6L_\Phi}, \qquad \frac{\eta_x}{2} - \frac{3L_\Phi\eta_x^2}{2} \ge \frac{\eta_x}{4} \quad \text{for } 0 < \eta_x \le \eta_{\mathrm{d}}. \tag{165}$$

Taking total expectations in (164) and using (98)–(99) now yields

$$\begin{aligned}\mathbb{E}\Phi(x^{k+1}) &\le \mathbb{E}\Phi(x^k) - \frac{3\eta_x}{8}G_k - \frac{\eta_x}{4}H_k + \frac{\eta_x}{2}\left(D_y I_k + D_z J_k + 3\frac{C_{\mathrm{lag}}^2}{\lambda^2}\right) \\ &\quad + 2\eta_x R_{x,k} + \frac{3L_\Phi\eta_x^2}{2}\left(\frac{\Sigma_x}{n_h} + R_{x,k}\right).\end{aligned} \tag{166}$$

Thus (105) holds with $c = 1/4$ and any $C \geq \max\{D_y/2, D_z/2, 2, 3L_\Phi/2\}$. In particular, both negative direction terms remain available when the descent and tracking recursions are combined. □

D.4.3 Proof of Lemma 7: Penalized Lower-Level Tracking

Proof. Let $q_y^k = y_\lambda^\star(x^k)$, $d_y^k = y^k - q_y^k$, and $g_y^k = \bar{h}_y^k$. We first establish contraction with x^k fixed, then account for movement of the target.

With $\mu = \mu_\lambda$ and $L = L_\lambda$, the standard interpolation inequality for a μ-strongly convex, L-smooth function is

$$\langle g_y^k, d_y^k \rangle \geq \frac{\mu L}{\mu + L} \|d_y^k\|^2 + \frac{1}{\mu + L} \|g_y^k\|^2. \tag{167}$$

The first term supplies contraction of the distance to q_y^k, and the second supplies a negative direction-norm term that absorbs the quadratic $\eta_y^2 \|g_y^k\|^2$ contribution below. Expanding the y-update around the fixed target and using $\mathbb{E}_k\langle d_y^k, \epsilon_y^k \rangle = 0$ gives

$$\begin{aligned} \mathbb{E}_k\|y^{k+1} - q_y^k\|^2 = \|d_y^k\|^2 - 2\eta_y\langle d_y^k, g_y^k\rangle - 2\eta_y\mathbb{E}_k\langle d_y^k, r_y^k\rangle \\ + \eta_y^2\mathbb{E}_k\|g_y^k + \epsilon_y^k + r_y^k\|^2. \end{aligned} \tag{168}$$

Because the robust error need not be conditionally centered, Young's inequality reserves one quarter of the strong-convexity contraction for its cross term. Together with the three-term squared-norm bound, this gives

$$-2\eta_y\mathbb{E}_k\langle d_y^k, r_y^k\rangle \leq \frac{\mu\eta_y}{4}\|d_y^k\|^2 + \frac{4\eta_y}{\mu}\mathbb{E}_k\|r_y^k\|^2, \tag{169}$$

$$\mathbb{E}_k\|g_y^k + \epsilon_y^k + r_y^k\|^2 \leq 3\|g_y^k\|^2 + 3\frac{\Sigma_y}{n_h} + 3\mathbb{E}_k\|r_y^k\|^2. \tag{170}$$

Substituting equations (167)–(170) into equation (168) gives the pre-absorption inequality

$$\begin{aligned} \mathbb{E}_k\|y^{k+1} - q_y^k\|^2 \leq \left(1 - \frac{2\mu L\eta_y}{\mu + L} + \frac{\mu\eta_y}{4}\right)\|d_y^k\|^2 \\ + \left(-\frac{2\eta_y}{\mu + L} + 3\eta_y^2\right)\|g_y^k\|^2 + 3\eta_y^2\frac{\Sigma_y}{n_h} + \left(\frac{4\eta_y}{\mu} + 3\eta_y^2\right)\mathbb{E}_k\|r_y^k\|^2. \end{aligned} \tag{171}$$

Because $L \geq \mu$, $2\mu L/(\mu + L) \geq \mu$. If $\eta_y \leq [3(\mu + L)]^{-1}$, the coefficient of $\|g_y^k\|^2$ is nonpositive and the distance coefficient is at most $1 - 3\mu\eta_y/4$. Dropping the nonpositive direction term and taking expectations therefore yield

$$\mathbb{E}\|y^{k+1} - q_y^k\|^2 \leq \left(1 - \frac{3\mu_\lambda\eta_y}{4}\right) I_k + 3\eta_y^2\frac{\Sigma_y}{n_h} + \left(\frac{4\eta_y}{\mu_\lambda} + 3\eta_y^2\right) R_{y,k}. \tag{172}$$

Target drift. To obtain a bound on I_{k+1}, expand around the updated target:

$$\begin{aligned} \|y^{k+1} - q_y^{k+1}\|^2 = \|y^{k+1} - q_y^k\|^2 \\ + 2\langle y^{k+1} - q_y^k, q_y^k - q_y^{k+1}\rangle + \|q_y^{k+1} - q_y^k\|^2. \end{aligned} \tag{173}$$

The use of Lemma 4 requires conditioning on the information that determines $u_y^k := y^{k+1} - q_y^k$. Let $\mathcal{K}_{y,k}$ contain $\mathcal{F}_k$, the current y-block reports, and any randomness used to aggregate that block. Then u_y^k is $\mathcal{K}_{y,k}$-measurable. Assumption 3 gives independence of the honest y- and x-block draws; the simultaneous submission protocol in Section 3 prevents the Byzantine y-reports from using the current honest x-reports. As in (81), we therefore have

$$\mathbb{E}[\epsilon_x^k \mid \mathcal{K}_{y,k}] = 0, \qquad \mathbb{E}[\|\epsilon_x^k\|^2 \mid \mathcal{K}_{y,k}] \leq \frac{\Sigma_x}{n_h}. \tag{174}$$

In particular, write $x^{\mathrm{d},k} := x^k - \eta_x\bar{h}_x^k$. For $q = y_\lambda^\star$, the noise term in the expansion around $x^{\mathrm{d},k}$ vanishes by the tower property:

$$\mathbb{E}\langle u_y^k, \nabla q(x^{\mathrm{d},k})\epsilon_x^k\rangle = \mathbb{E}\left\langle u_y^k, \nabla q(x^{\mathrm{d},k})\mathbb{E}[\epsilon_x^k \mid \mathcal{K}_{y,k}]\right\rangle = 0. \tag{175}$$

This cancellation concerns the honest noise only; r_x^k may depend on that noise.

Write $a_y := \mu_\lambda \eta_y$. Choose the stepsizes to satisfy

$$0 < a_y \le 1, \qquad \chi_y := \frac{3a_y}{8} \in (0,1], \qquad \Gamma_{x,\lambda}\eta_x^2 \le \frac{3c_{q_y}}{8} a_y = c_{q_y}\chi_y, \tag{176}$$

where c_{q_y} is the constant in Lemma 4 for $q = y_\lambda^\star$. The x-update has the decomposition (77) with $x = x^k$ and $x^+ = x^{k+1}$. Applying the lemma with $u = u_y^k$ and $\chi = \chi_y$, and taking total expectations, gives

$$\begin{aligned} &2\mathbb{E}\langle y^{k+1} - q_y^k, q_y^k - q_y^{k+1}\rangle + \mathbb{E}\|q_y^{k+1} - q_y^k\|^2 \\ &\quad \le \frac{3\mu_\lambda\eta_y}{8}\mathbb{E}\|y^{k+1} - q_y^k\|^2 + \frac{8C_{q_y}\eta_x^2}{3a_y}(H_k + R_{x,k}) + C_{q_y}\eta_x^2\left(\frac{\Sigma_x}{n_h} + R_{x,k}\right). \end{aligned} \tag{177}$$

Here C_{q_y} denotes the corresponding drift constant. Taking expectations in (173) and substituting the fixed-target and drift bounds yields

$$\begin{aligned} I_{k+1} \le \left(1 + \frac{3a_y}{8}\right)\left[\left(1 - \frac{3a_y}{4}\right) I_k + 3\eta_y^2\frac{\Sigma_y}{n_h} + \left(\frac{4\eta_y}{\mu_\lambda} + 3\eta_y^2\right) R_{y,k}\right] \\ + \frac{8C_{q_y}\eta_x^2}{3a_y}(H_k + R_{x,k}) + C_{q_y}\eta_x^2\left(\frac{\Sigma_x}{n_h} + R_{x,k}\right). \end{aligned} \tag{178}$$

The distance coefficient retains a strict contraction:

$$\left(1 + \frac{3\mu_\lambda\eta_y}{8}\right)\left(1 - \frac{3\mu_\lambda\eta_y}{4}\right) = 1 - \frac{3a_y}{8} - \frac{9a_y^2}{32} \le 1 - \frac{3\mu_\lambda\eta_y}{8}. \tag{179}$$

Since $a_y \le 1$, the remaining coefficients satisfy

$$\begin{aligned} 1 + \frac{3a_y}{8} &\le 2, \\ \frac{8C_{q_y}\eta_x^2}{3a_y}(H_k + R_{x,k}) + C_{q_y}\eta_x^2 R_{x,k} &\le \frac{11C_{q_y}\eta_x^2}{3a_y}(H_k + R_{x,k}). \end{aligned} \tag{180}$$

Substitution into (178) therefore gives the explicit recursion

$$\begin{aligned} I_{k+1} \le \left(1 - \frac{3\mu_\lambda\eta_y}{8}\right) I_k + \frac{11C_{q_y}\eta_x^2}{3\mu_\lambda\eta_y}(H_k + R_{x,k}) \\ + C_{q_y}\eta_x^2\frac{\Sigma_x}{n_h} + 6\eta_y^2\frac{\Sigma_y}{n_h} + \left(\frac{8\eta_y}{\mu_\lambda} + 6\eta_y^2\right) R_{y,k}. \end{aligned} \tag{181}$$

This is (106) with $c_y = 3/8$ and $C_y = \max\{8, 11C_{q_y}/3\}$. To match the stepsize conditions in the lemma statement, one may take

$$\eta_{\mathrm{t},y} := \min\left\{1, \frac{L_\lambda}{3(\mu_\lambda + L_\lambda)}\right\}, \qquad c \le \frac{3c_{q_y}}{8}. \tag{182}$$

Indeed, $\eta_y L_\lambda \le \eta_{\mathrm{t},y}$ implies both $\eta_y \le [3(\mu_\lambda + L_\lambda)]^{-1}$ and $a_y \le 1$, while $\Gamma_{x,\lambda}\eta_x^2 \le c\mu_\lambda\eta_y$ gives (176). □

D.4.4 Proof of Lemma 8: Original Lower-Level Tracking

Proof. Let $q_z^k := y^\star(x^k)$, $d_z^k := z^k - q_z^k$, and $g_z^k := \bar{h}_z^k = \nabla_y g(x^k, z^k)$. As in the y-analysis, the proof has a fixed-target contraction step and a moving-target correction. The constants now involve μ_g and $l_{g,1}$ because z^k tracks the minimizer of the original lower-level objective rather than that of the penalized objective.

The fixed-target expansion is

$$\begin{aligned} \mathbb{E}_k\|z^{k+1} - q_z^k\|^2 = \|d_z^k\|^2 - 2\eta_z\langle d_z^k, g_z^k\rangle - 2\eta_z\mathbb{E}_k\langle d_z^k, r_z^k\rangle \\ + \eta_z^2\mathbb{E}_k\|g_z^k + \epsilon_z^k + r_z^k\|^2. \end{aligned} \tag{183}$$

Strong convexity and smoothness of $g(x^k, \cdot)$ give

$$\langle g_z^k, d_z^k \rangle \geq \frac{\mu_g l_{g,1}}{\mu_g + l_{g,1}} \|d_z^k\|^2 + \frac{1}{\mu_g + l_{g,1}} \|g_z^k\|^2. \tag{184}$$

As before, interpolation simultaneously provides distance contraction and a negative direction-norm term. The latter is what allows the squared update to be absorbed by a small stepsize. Using Young's inequality on the aggregation cross term and bounding the squared update by three squared norms yields

$$-2\eta_z \mathbb{E}_k \langle d_z^k, r_z^k \rangle \leq \frac{\mu_g \eta_z}{4} \|d_z^k\|^2 + \frac{4\eta_z}{\mu_g} \mathbb{E}_k \|r_z^k\|^2, \tag{185}$$

$$\mathbb{E}_k \|g_z^k + \epsilon_z^k + r_z^k\|^2 \leq 3\|g_z^k\|^2 + 3\frac{\Sigma_z}{n_h} + 3\mathbb{E}_k \|r_z^k\|^2. \tag{186}$$

Substituting equations (184)–(186) into equation (183) gives

$$\begin{aligned} \mathbb{E}_k \|z^{k+1} - q_z^k\|^2 \leq & \left(1 - \frac{2\mu_g l_{g,1} \eta_z}{\mu_g + l_{g,1}} + \frac{\mu_g \eta_z}{4}\right) \|d_z^k\|^2 \\ & + \left(-\frac{2\eta_z}{\mu_g + l_{g,1}} + 3\eta_z^2\right) \|g_z^k\|^2 + 3\eta_z^2 \frac{\Sigma_z}{n_h} + \left(\frac{4\eta_z}{\mu_g} + 3\eta_z^2\right) \mathbb{E}_k \|r_z^k\|^2. \end{aligned} \tag{187}$$

Since $l_{g,1} \geq \mu_g$, the first coefficient is at most $1 - 3\mu_g \eta_z / 4$. If $\eta_z \leq [3(\mu_g + l_{g,1})]^{-1}$, the coefficient of $\|g_z^k\|^2$ is nonpositive. Dropping that term and taking total expectations give

$$\mathbb{E}\|z^{k+1} - q_z^k\|^2 \leq \left(1 - \frac{3\mu_g \eta_z}{4}\right) J_k + 3\eta_z^2 \frac{\Sigma_z}{n_h} + \left(\frac{4\eta_z}{\mu_g} + 3\eta_z^2\right) R_{z,k}. \tag{188}$$

This controls tracking relative to q_z^k. To obtain J_{k+1}, we must additionally account for the displacement of $y^\star(x)$ caused by the current x-update.

Target drift. Set $u_z^k := z^{k+1} - q_z^k$. The updated tracking error satisfies the exact identity

$$J_{k+1} = \mathbb{E}\|u_z^k\|^2 + 2\mathbb{E}\langle u_z^k, q_z^k - q_z^{k+1} \rangle + \mathbb{E}\|q_z^{k+1} - q_z^k\|^2. \tag{189}$$

Let $\mathcal{K}_{z,k}$ contain $\mathcal{F}_k$, the current z-block reports, and any randomness used in their aggregation. Then u_z^k is $\mathcal{K}_{z,k}$-measurable. The same block independence and submission protocol used in (174) give

$$\mathbb{E}[\epsilon_x^k \mid \mathcal{K}_{z,k}] = 0, \qquad \mathbb{E}[\|\epsilon_x^k\|^2 \mid \mathcal{K}_{z,k}] \leq \frac{\Sigma_x}{n_h}. \tag{190}$$

Thus the tower-property argument in (175) also applies with u_z^k and $q = y^\star$. Write $a_z := \mu_g \eta_z$ and impose

$$0 < a_z \leq 1, \qquad \chi_z := \frac{3a_z}{8} \in (0, 1], \qquad \Gamma_{x,\lambda} \eta_x^2 \leq \frac{3c_{q_z}}{8} a_z = c_{q_z} \chi_z, \tag{191}$$

where c_{q_z} is the constant in Lemma 4 for $q = y^\star$. Applying that lemma with $u = u_z^k$, $\chi = \chi_z$, and $x^+ = x^{k+1}$ gives

$$\begin{aligned} & 2\mathbb{E}\langle z^{k+1} - q_z^k, q_z^k - q_z^{k+1} \rangle + \mathbb{E}\|q_z^{k+1} - q_z^k\|^2 \\ & \quad \leq \frac{3\mu_g \eta_z}{8} \mathbb{E}\|z^{k+1} - q_z^k\|^2 + \frac{8C_{q_z} \eta_x^2}{3a_z} (H_k + R_{x,k}) + C_{q_z} \eta_x^2 \left(\frac{\Sigma_x}{n_h} + R_{x,k}\right). \end{aligned} \tag{192}$$

Here C_{q_z} is the drift constant for the original optimizer map. Substituting (188) and (192) into (189) yields

$$\begin{aligned} J_{k+1} \leq & \left(1 + \frac{3a_z}{8}\right) \left[\left(1 - \frac{3a_z}{4}\right) J_k + 3\eta_z^2 \frac{\Sigma_z}{n_h} + \left(\frac{4\eta_z}{\mu_g} + 3\eta_z^2\right) R_{z,k}\right] \\ & + \frac{8C_{q_z} \eta_x^2}{3a_z} (H_k + R_{x,k}) + C_{q_z} \eta_x^2 \left(\frac{\Sigma_x}{n_h} + R_{x,k}\right). \end{aligned} \tag{193}$$

The contraction coefficient satisfies

$$\left(1 + \frac{3\mu_g \eta_z}{8}\right) \left(1 - \frac{3\mu_g \eta_z}{4}\right) = 1 - \frac{3a_z}{8} - \frac{9a_z^2}{32} \leq 1 - \frac{3\mu_g \eta_z}{8}. \tag{194}$$

For the residual terms, $a_z \le 1$ gives

$$1 + \frac{3a_z}{8} \le 2,$$
$$\frac{8C_{q_z}\eta_x^2}{3a_z}(H_k + R_{x,k}) + C_{q_z}\eta_x^2 R_{x,k} \le \frac{11C_{q_z}\eta_x^2}{3a_z}(H_k + R_{x,k}). \tag{195}$$

Consequently,

$$\begin{aligned} J_{k+1} \le \left(1 - \frac{3\mu_g\eta_z}{8}\right) J_k + \frac{11C_{q_z}\eta_x^2}{3\mu_g\eta_z}(H_k + R_{x,k}) \\ + C_{q_z}\eta_x^2\frac{\Sigma_x}{n_h} + 6\eta_z^2\frac{\Sigma_z}{n_h} + \left(\frac{8\eta_z}{\mu_g} + 6\eta_z^2\right) R_{z,k}. \end{aligned} \tag{196}$$

This proves (107) with $c_z = 3/8$ and $C_z = \max\{8, 11C_{q_z}/3\}$. An admissible choice of the lemma's stepsize constants is

$$\eta_{\mathrm{t},z} := \min\left\{1, \frac{l_{g,1}}{3(\mu_g + l_{g,1})}\right\}, \qquad c \le \frac{3c_{q_z}}{8}. \tag{197}$$

Indeed, $\eta_z l_{g,1} \le \eta_{\mathrm{t},z}$ ensures the fixed-target stepsize bound and $a_z \le 1$, and the target-drift condition implies (191). The H_k term in (196) is retained for absorption by upper-level descent in Appendix D.2; the $R_{x,k}$ term is controlled there by Lemma 5. □

E CONVERGENCE ANALYSIS OF FOBFBL−M

We prove Theorems 3 and 4 using aggregation, momentum-deviation, descent, and tracking bounds. Throughout this section, the assumptions of Theorem 3 hold, with fixed $\lambda > 0$, $\lambda \geq 2l_{f,1}/\mu_g$, and zero initial momentum buffers. The Lyapunov function includes the three honest momentum deviations in addition to the objective gap and two tracking errors used in Appendix D. The PL condition is imposed in Appendix E.3.

E.1 AUXILIARY LEMMAS

We state the aggregation, momentum-deviation, descent, and tracking estimates used in both convergence proofs. Their proofs are given in Appendix E.4.

E.1.1 MOMENTUM ERROR DECOMPOSITION

We retain the notation and filtration of Appendix D.1, with I_k, J_k, G_k, and H_k evaluated along the iterates of FOBFBL−M. For $v \in \{x, y, z\}$, define the honest mean momentum, its aggregation error, and its deviation from the current honest direction by

$$m_v^k := \frac{1}{n_h} \sum_{i \in \mathcal{H}} m_{v,i}^k, \qquad s_v^k := \widehat{m}_v^k - m_v^k, \qquad e_v^k := m_v^k - \bar{h}_v^k, \tag{198}$$

where $\bar{h}_v^k$ is the conditional-mean honest direction in (54). Their error second moments are

$$M_{v,k} := \mathbb{E}\|e_v^k\|^2, \qquad R_{v,k}^{\mathrm{m}} := \mathbb{E}\|s_v^k\|^2. \tag{199}$$

For a run with K updates, take virtual fresh honest directions $h_{v,i}^K$ at the terminal iterates and set $m_{v,i}^K := \beta_v m_{v,i}^{K-1} + (1-\beta_v) h_{v,i}^K$. Their honest mean defines m_v^K and e_v^K as above, solely for the terminal Lyapunov function; Algorithm 1 need not compute them or perform another update. For $k = 0, \ldots, K-1$, the server update is

$$\widehat{m}_v^k = \bar{h}_v^k + e_v^k + s_v^k, \qquad v^{k+1} = v^k - \eta_v(\bar{h}_v^k + e_v^k + s_v^k). \tag{200}$$

The deviation e_v^k is controlled through the momentum recursion, while s_v^k is bounded through the robustness of $\mathcal{A}$. We use $\eta_x = \eta > 0$, $\eta_y = c_y^\circ \eta$, $\eta_z = c_z^\circ \eta$ with fixed positive ratios, and $\beta_v = \sqrt{1 - \rho_v \eta}$ with $\rho_v > 0$ to control both contributions. All sums indexed by v below range over $\{x, y, z\}$. Constants in the deviation bounds depend on the problem parameters and fixed stepsize ratios, but not on k, K, η, or the momentum constants ρ_v.

E.1.2 AGGREGATION ERROR

The following lemma is the momentum counterpart of Lemma 5. It separates stochastic variance from client heterogeneity.

Lemma 9 (Aggregation error with momentum)**.** *Suppose Assumptions 3 and 4 hold and $\mathcal{A}$ satisfies Definition 2 almost surely. With $m_{v,i}^{-1} = 0$ and fixed $\beta_v \in [0, 1)$, every $v \in \{x, y, z\}$ and $k \geq 0$ satisfy*

$$R_{v,k}^{\mathrm{m}} \leq 2\kappa \left(\frac{1-\beta_v}{1+\beta_v} \Sigma_v + Z_v \right), \tag{201}$$

where the block scales are defined in Lemma 3. Equivalently, the three block bounds are

$$R_{x,k}^{\mathrm{m}} \leq 2\kappa \left[\frac{1-\beta_x}{1+\beta_x} (\sigma_f^2 + 2\lambda^2 \sigma_g^2) + 3\zeta_f^2 + 6\lambda^2 \zeta_g^2 \right], \tag{202}$$

$$R_{y,k}^{\mathrm{m}} \leq 2\kappa \left[\frac{1-\beta_y}{1+\beta_y} (\sigma_f^2 + \lambda^2 \sigma_g^2) + 2\zeta_f^2 + 2\lambda^2 \zeta_g^2 \right], \tag{203}$$

$$R_{z,k}^{\mathrm{m}} \leq 2\kappa \left[\frac{1-\beta_z}{1+\beta_z} \sigma_g^2 + \zeta_g^2 \right]. \tag{204}$$

Proof. See Appendix E.4.1.

If $v_{\mathrm{av}}^{k+1} := v^k - \eta_v m_v^k$ denotes the update using the honest mean momentum, the corresponding error in the iterate is

$$\mathbb{E}\|v^{k+1} - v_{\mathrm{av}}^{k+1}\|^2 = \eta_v^2 R_{v,k}^{\mathrm{m}} \le 2\kappa\eta_v^2 \left(\frac{1-\beta_v}{1+\beta_v}\Sigma_v + Z_v\right). \tag{205}$$

Under $\beta_v^2 = 1 - \rho_v\eta$, the ratio $(1-\beta_v)/(1+\beta_v)$ is $\mathcal{O}(\eta)$. Thus sampling noise contributes $\mathcal{O}(\eta^2)$ to $\eta R_{v,k}^{\mathrm{m}}$, whereas client heterogeneity contributes $\mathcal{O}(\eta)$. Appendix E.2 tracks both terms in the final residual.

E.1.3 Momentum Deviation

We first bound the momentum deviations in the two lower-level blocks.

Lemma 10 (Momentum deviation: y block)**.** *Suppose $\beta_y = \sqrt{1-\rho_y\eta}$, $0 < \rho_y\eta \le 1/2$, and the block stepsizes have fixed ratios to η. Then*

$$M_{y,k+1} \le \left(1 - \frac{\rho_y\eta}{2}\right) M_{y,k} + \frac{A_y\eta}{\rho_y}\left(H_k + I_k + M_{x,k} + M_{y,k} + R_{x,k}^{\mathrm{m}} + R_{y,k}^{\mathrm{m}}\right) + \rho_y^2\eta^2\frac{\Sigma_y}{n_h}, \tag{206}$$

for a fixed $A_y > 0$.

Proof. See Appendix E.4.2.

Lemma 11 (Momentum deviation: z block)**.** *Suppose $\beta_z = \sqrt{1-\rho_z\eta}$, $0 < \rho_z\eta \le 1/2$, and the block stepsizes have fixed ratios to η. Then*

$$M_{z,k+1} \le \left(1 - \frac{\rho_z\eta}{2}\right) M_{z,k} + \frac{A_z\eta}{\rho_z}\left(H_k + J_k + M_{x,k} + M_{z,k} + R_{x,k}^{\mathrm{m}} + R_{z,k}^{\mathrm{m}}\right) + \rho_z^2\eta^2\frac{\Sigma_z}{n_h}, \tag{207}$$

for a fixed $A_z > 0$.

Proof. See Appendix E.4.3.

For the x-block, the deterministic direction depends on all three iterates. Its deviation bound therefore involves both tracking errors and all three momentum errors.

Lemma 12 (Momentum deviation: x block)**.** *Suppose $\beta_x = \sqrt{1-\rho_x\eta}$, $0 < \rho_x\eta \le 1/2$, and the block stepsizes have fixed ratios to η. Then*

$$M_{x,k+1} \le \left(1 - \frac{\rho_x\eta}{2}\right) M_{x,k} + \frac{A_x\eta}{\rho_x}\left(H_k + I_k + J_k + \sum_v M_{v,k} + \sum_v R_{v,k}^{\mathrm{m}}\right) + \rho_x^2\eta^2\frac{\Sigma_x}{n_h}, \tag{208}$$

for a fixed $A_x > 0$.

Proof. See Appendix E.4.4.

Each deviation bound consists of contraction, direction drift, and fresh sampling noise.

E.1.4 Descent and Tracking Bounds

The descent and tracking bounds use the smoothness of Φ and the Lipschitz optimizer maps from Lemma 2. They do not use its Jacobian-Lipschitz bound for $y_\lambda^\star$, so no Lipschitz Hessian assumption on f is needed here.

Lemma 13 (Upper-level descent with momentum)**.** *If $0 < \eta_x \le (4L_\Phi)^{-1}$,*

$$\mathbb{E}\Phi(x^{k+1}) \le \mathbb{E}\Phi(x^k) - c_{\mathrm{m}}\eta_x G_k - c_{\mathrm{m}}\eta_x H_k + C_{\mathrm{m}}\eta_x(I_k + J_k + M_{x,k} + R_{x,k}^{\mathrm{m}} + \frac{C_{\mathrm{lag}}^2}{\lambda^2}). \tag{209}$$

Proof. See Appendix E.4.5.

The negative H_k term absorbs target drift in the following tracking estimates.

Lemma 14 (Penalized tracking with momentum). *If* $0 < \eta_y \le [3(\mu_\lambda + L_\lambda)]^{-1}$,

$$I_{k+1} \le (1 - c_{\mathrm{m},y}\mu_\lambda\eta_y)I_k + \frac{C_{\mathrm{m},y}\eta_y}{\mu_\lambda}(M_{y,k} + R^{\mathrm{m}}_{y,k}) + \frac{C_{\mathrm{m},y}\eta_x^2}{\mu_\lambda\eta_y}(H_k + M_{x,k} + R^{\mathrm{m}}_{x,k}). \tag{210}$$

Proof. See Appendix E.4.6.

For the original lower-level solution, the corresponding bound is as follows.

Lemma 15 (Original tracking with momentum). *If* $0 < \eta_z \le [3(\mu_g + l_{g,1})]^{-1}$,

$$J_{k+1} \le (1 - c_{\mathrm{m},z}\mu_g\eta_z)J_k + \frac{C_{\mathrm{m},z}\eta_z}{\mu_g}(M_{z,k} + R^{\mathrm{m}}_{z,k}) + \frac{C_{\mathrm{m},z}\eta_x^2}{\mu_g\eta_z}(H_k + M_{x,k} + R^{\mathrm{m}}_{x,k}). \tag{211}$$

Proof. See Appendix E.4.7.

Combining these estimates with the momentum-deviation recursions yields the Lyapunov inequality used in the proofs of Theorems 3 and 4.

E.2 Proof of Theorem 3: Nonconvex Convergence

Theorem 3 (Convergence of FOBFBL-M for nonconvex upper-level objectives). *Suppose Assumptions 1–4 hold, g has an $l_{g,2}$-Lipschitz Hessian, $\Phi_* > -\infty$, and $\mathcal{A}$ satisfies Definition 2 almost surely. Fix $\lambda \ge 2l_{f,1}/\mu_g$ and assume $\Sigma_\lambda > 0$. Define the effective aggregation factor $\chi_\kappa := 1 + n_h\kappa$ and let $\overline{\Delta}_0^{\mathrm{m}} > 0$ denote the uniform initial-gap bound in Appendix E. For sufficiently large K, choose $\eta = \sqrt{n_h\overline{\Delta}_0^{\mathrm{m}}/(K\chi_\kappa\Sigma_\lambda)}$, $\eta_v = c_v^\circ\eta$, and $\beta_v = \sqrt{1-\rho_v\eta}$ for $v \in \{x,y,z\}$, where $c_x^\circ = 1$ and $c_v^\circ, \rho_v > 0$ are problem-dependent constants. With deterministic initialization, x^R satisfies*

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 = \mathcal{O}\Bigg(\underbrace{\sqrt{\frac{\overline{\Delta}_0^{\mathrm{m}}\chi_\kappa\Sigma_\lambda}{n_hK}}}_{\text{finite-horizon error}} + \underbrace{\kappa(\zeta_f^2 + \lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\mathrm{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}}\Bigg). \tag{17}$$

Proof. Set $\eta_x = \eta$, $\eta_y = c_y^\circ\eta$, and $\eta_z = c_z^\circ\eta$. Write $\Delta\Phi_k := \mathbb{E}[\Phi(x^{k+1}) - \Phi(x^k)]$, $\Delta I_k := I_{k+1} - I_k$, and $\Delta J_k := J_{k+1} - J_k$. For now, take η sufficiently small for the descent and tracking bounds to hold. After substituting these ratios, Lemmas 13–15 give

$$\Delta\Phi_k \le -d_G^{\mathrm{m}}\eta G_k - d_H^{\mathrm{m}}\eta H_k + \eta(D_I^{\mathrm{m}}I_k + D_J^{\mathrm{m}}J_k + D_B^{\mathrm{m}}\frac{C_{\mathrm{lag}}^2}{\lambda^2}) + \eta P_x^{(0)}(M_{x,k} + R^{\mathrm{m}}_{x,k}), \tag{212}$$

$$\Delta I_k \le -t_y^{\mathrm{m}}\mu_\lambda c_y^\circ\eta I_k + Q_y^{\mathrm{m}}\frac{c_y^\circ\eta}{\mu_\lambda}(M_{y,k} + R^{\mathrm{m}}_{y,k}) + \frac{Q_y^{\mathrm{m}}\eta}{\mu_\lambda c_y^\circ}(H_k + M_{x,k} + R^{\mathrm{m}}_{x,k}), \tag{213}$$

$$\Delta J_k \le -t_z^{\mathrm{m}}\mu_g c_z^\circ\eta J_k + Q_z^{\mathrm{m}}\frac{c_z^\circ\eta}{\mu_g}(M_{z,k} + R^{\mathrm{m}}_{z,k}) + \frac{Q_z^{\mathrm{m}}\eta}{\mu_g c_z^\circ}(H_k + M_{x,k} + R^{\mathrm{m}}_{x,k}). \tag{214}$$

The positive coefficients above depend only on fixed problem parameters, including λ; they are independent of $c_y^\circ, c_z^\circ, \rho_v, \eta, K, n_h$, and κ.

To combine the objective and tracking estimates, choose the stepsize ratios and Lyapunov weights as

$$(c_y^\circ)^2 \geq \frac{8D_I^{\mathrm{m}}Q_y^{\mathrm{m}}}{t_y^{\mathrm{m}}\mu_\lambda^2 d_H^{\mathrm{m}}}, \qquad (c_z^\circ)^2 \geq \frac{8D_J^{\mathrm{m}}Q_z^{\mathrm{m}}}{t_z^{\mathrm{m}}\mu_g^2 d_H^{\mathrm{m}}}, \tag{215}$$

$$a := \frac{2D_I^{\mathrm{m}}}{t_y^{\mathrm{m}}\mu_\lambda c_y^\circ}, \qquad b := \frac{2D_J^{\mathrm{m}}}{t_z^{\mathrm{m}}\mu_g c_z^\circ}. \tag{216}$$

The tracking coefficients after weighting satisfy

$$\begin{aligned} D_I^{\mathrm{m}} - at_y^{\mathrm{m}}\mu_\lambda c_y^\circ &= -D_I^{\mathrm{m}}, \\ D_J^{\mathrm{m}} - bt_z^{\mathrm{m}}\mu_g c_z^\circ &= -D_J^{\mathrm{m}}, \\ -d_H^{\mathrm{m}} + \frac{aQ_y^{\mathrm{m}}}{\mu_\lambda c_y^\circ} + \frac{bQ_z^{\mathrm{m}}}{\mu_g c_z^\circ} &= -d_H^{\mathrm{m}} + \frac{2D_I^{\mathrm{m}}Q_y^{\mathrm{m}}}{t_y^{\mathrm{m}}\mu_\lambda^2 (c_y^\circ)^2} + \frac{2D_J^{\mathrm{m}}Q_z^{\mathrm{m}}}{t_z^{\mathrm{m}}\mu_g^2 (c_z^\circ)^2} \\ &\leq -d_H^{\mathrm{m}} + \frac{d_H^{\mathrm{m}}}{4} + \frac{d_H^{\mathrm{m}}}{4} = -\frac{d_H^{\mathrm{m}}}{2}. \end{aligned} \tag{217}$$

Thus half of the upper-level direction decrease remains after absorbing target drift. Define the base Lyapunov function

$$U_k := \mathbb{E}[\Phi(x^k) - \Phi_*] + aI_k + bJ_k. \tag{218}$$

Adding (212), a times (213), and b times (214) gives

$$\begin{aligned} U_{k+1} - U_k \leq{}& -d_G^{\mathrm{m}}\eta G_k - \frac{d_H^{\mathrm{m}}\eta}{2}H_k - D_I^{\mathrm{m}}\eta I_k - D_J^{\mathrm{m}}\eta J_k \\ &+ \eta\left(P_x M_{x,k} + P_y M_{y,k} + P_z M_{z,k}\right) \\ &+ \eta\left(P_x R_{x,k}^{\mathrm{m}} + P_y R_{y,k}^{\mathrm{m}} + P_z R_{z,k}^{\mathrm{m}}\right) + D_B^{\mathrm{m}}\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2}, \end{aligned} \tag{219}$$

where the positive coefficients are explicitly

$$\begin{aligned} P_x &:= P_x^{(0)} + \frac{aQ_y^{\mathrm{m}}}{\mu_\lambda c_y^\circ} + \frac{bQ_z^{\mathrm{m}}}{\mu_g c_z^\circ}, \\ P_y &:= \frac{aQ_y^{\mathrm{m}}c_y^\circ}{\mu_\lambda}, \qquad P_z := \frac{bQ_z^{\mathrm{m}}c_z^\circ}{\mu_g}. \end{aligned} \tag{220}$$

The same coefficients multiply $M_{v,k}$ and $R_{v,k}^{\mathrm{m}}$ because both enter through the perturbed momentum message. We retain $M_{v,k}$ as states and bound $R_{v,k}^{\mathrm{m}}$ as residuals.

To control the temporal deviations, we augment U_k with their second moments. Fix positive weights $\vartheta_x, \vartheta_y, \vartheta_z$ independently of η, K, n_h, and κ (for example, all equal to one), and define

$$V_k^{\mathrm{m}} := U_k + \sum_{v\in\{x,y,z\}} \vartheta_v M_{v,k}. \tag{221}$$

Adding the weighted recursions of Lemmas 10–12 to (219) gives the following coefficients of the state quantities:

$$\begin{aligned} T_G &= -d_G^{\mathrm{m}}, \\ T_H &= -\frac{d_H^{\mathrm{m}}}{2} + \frac{\vartheta_y A_y}{\rho_y} + \frac{\vartheta_z A_z}{\rho_z} + \frac{\vartheta_x A_x}{\rho_x}, \\ T_I &= -D_I^{\mathrm{m}} + \frac{\vartheta_y A_y}{\rho_y} + \frac{\vartheta_x A_x}{\rho_x}, \\ T_J &= -D_J^{\mathrm{m}} + \frac{\vartheta_z A_z}{\rho_z} + \frac{\vartheta_x A_x}{\rho_x}, \end{aligned} \tag{222}$$

$$\begin{aligned} T_{M_x} &= P_x - \frac{\vartheta_x\rho_x}{2} + \frac{\vartheta_y A_y}{\rho_y} + \frac{\vartheta_z A_z}{\rho_z} + \frac{\vartheta_x A_x}{\rho_x}, \\ T_{M_y} &= P_y - \frac{\vartheta_y\rho_y}{2} + \frac{\vartheta_y A_y}{\rho_y} + \frac{\vartheta_x A_x}{\rho_x}, \\ T_{M_z} &= P_z - \frac{\vartheta_z\rho_z}{2} + \frac{\vartheta_z A_z}{\rho_z} + \frac{\vartheta_x A_x}{\rho_x}. \end{aligned} \tag{223}$$

The aggregation coefficients are the corresponding positive coefficients without the momentum-contraction terms:

$$\begin{aligned}\widetilde{P}_x &:= P_x + \sum_v \frac{\vartheta_v A_v}{\rho_v},\\ \widetilde{P}_y &:= P_y + \frac{\vartheta_y A_y}{\rho_y} + \frac{\vartheta_x A_x}{\rho_x},\\ \widetilde{P}_z &:= P_z + \frac{\vartheta_z A_z}{\rho_z} + \frac{\vartheta_x A_x}{\rho_x}.\end{aligned} \tag{224}$$

In particular, the complete weighted recursion before eliminating state terms is

$$\begin{aligned}V_{k+1}^{\mathrm{m}} - V_k^{\mathrm{m}} \le \eta \left(T_G G_k + T_H H_k + T_I I_k + T_J J_k + \sum_v T_{M_v} M_{v,k} \right)\\ + \eta \sum_v \widetilde{P}_v R_{v,k}^{\mathrm{m}} + \frac{\eta^2}{n_h} \sum_v \vartheta_v \rho_v^2 \Sigma_v + D_B^{\mathrm{m}} \eta \frac{C_{\mathrm{lag}}^2}{\lambda^2}.\end{aligned} \tag{225}$$

The first coefficient group contains the coefficients of H_k, I_k, and J_k; the second contains those of the momentum deviations. After fixing the weights, choose ρ_x, ρ_y, ρ_z large enough that

$$\sum_v \frac{\vartheta_v A_v}{\rho_v} \le \frac{1}{4} \min\{d_H^{\mathrm{m}}, D_I^{\mathrm{m}}, D_J^{\mathrm{m}}\}, \tag{226}$$

$$\frac{\vartheta_v \rho_v}{4} \ge P_v + \sum_u \frac{\vartheta_u A_u}{\rho_u}, \qquad v \in \{x, y, z\}. \tag{227}$$

To exhibit a finite choice, set

$$\begin{aligned}m_0 := \min\{d_H^{\mathrm{m}}, D_I^{\mathrm{m}}, D_J^{\mathrm{m}}\} > 0, \qquad B_0 := \sum_v \vartheta_v A_v,\\ \rho_x = \rho_y = \rho_z = \bar{\rho}, \qquad \bar{\rho} \ge \max\left\{ \frac{4B_0}{m_0}, \max_v \frac{4(P_v + m_0/4)}{\vartheta_v} \right\}.\end{aligned} \tag{228}$$

Then $\sum_v \vartheta_v A_v / \rho_v = B_0/\bar{\rho} \le m_0/4$ and $\vartheta_v \rho_v / 4 \ge P_v + m_0/4$, which verify both requirements. All quantities on the right were fixed before choosing the momentum constants, so this construction is independent of η, K, n_h, and κ. Substituting (226) and (227) into the displayed coefficient groups gives

$$T_G = -d_G^{\mathrm{m}}, \qquad T_H \le -\frac{d_H^{\mathrm{m}}}{4},$$

$$T_I \le -\frac{3D_I^{\mathrm{m}}}{4}, \qquad T_J \le -\frac{3D_J^{\mathrm{m}}}{4}, \tag{229}$$

$$T_{M_v} \le -\frac{\vartheta_v \rho_v}{4}, \qquad v \in \{x, y, z\}. \tag{230}$$

All state coefficients are now negative. The H_k term has absorbed target motion and may be dropped, since it is not a component of the potential.

The remaining terms arise from sampling noise, robust aggregation, and the finite-multiplier bias. The fresh honest innovations in the three deviation recursions contribute

$$\eta^2 \sum_v \vartheta_v \rho_v^2 \frac{\Sigma_v}{n_h} \le \left(2\vartheta_x \rho_x^2 + \vartheta_y \rho_y^2 + \lambda^{-2} \vartheta_z \rho_z^2\right) \eta^2 \frac{\Sigma_\lambda}{n_h} =: C_{\mathrm{noise}} \eta^2 \frac{\Sigma_\lambda}{n_h}. \tag{231}$$

For the aggregation terms, $\beta_v^2 = 1 - \rho_v \eta$ implies

$$\frac{1 - \beta_v}{1 + \beta_v} = \frac{1 - \beta_v^2}{(1 + \beta_v)^2} = \frac{\rho_v \eta}{(1 + \beta_v)^2} \le \rho_v \eta. \tag{232}$$

Let $\overline{P} := \max_v \widetilde{P}_v$. Lemma 9 and (232) give the following bound on the aggregation part of (225):

$$\begin{aligned}\eta \sum_v \widetilde{P}_v R_{v,k}^{\mathrm{m}} &\le 2\overline{P} \kappa \eta \sum_v (\rho_v \eta \Sigma_v + Z_v)\\ &\le 2\overline{P} \kappa \eta \left[(2\rho_x + \rho_y + \lambda^{-2} \rho_z) \eta \Sigma_\lambda + 5\zeta_f^2 + (8 + \lambda^{-2}) \lambda^2 \zeta_g^2 \right]\\ &\le C_{\mathrm{agg}} \kappa \eta^2 \Sigma_\lambda + C_{\mathrm{agg}} \kappa \eta (\zeta_f^2 + \lambda^2 \zeta_g^2).\end{aligned} \tag{233}$$

Here the second inequality uses (122), and one can take

$$C_{\mathrm{agg}} := 2\overline{P}\max\{2\rho_x + \rho_y + \lambda^{-2}\rho_z, 8 + \lambda^{-2}\}. \tag{234}$$

The variance contribution in (233) is of order η^2, while the heterogeneity contribution is of order η and yields the Byzantine-induced floor after telescoping. All recursions hold on the explicit common stepsize range

$$\begin{aligned}\bar{\eta}_{\mathrm{m}} := \min\Big\{1, (4L_\Phi)^{-1}, [3c_y^\circ(\mu_\lambda + L_\lambda)]^{-1},\\ [3c_z^\circ(\mu_g + l_{g,1})]^{-1}, \min_v(2\rho_v)^{-1}\Big\} > 0.\end{aligned} \tag{235}$$

Indeed, $\eta \le \bar{\eta}_{\mathrm{m}}$ implies the three descent/tracking ceilings in Lemmas 13–15 and $\rho_v\eta \le 1/2$, so $\beta_v = \sqrt{1-\rho_v\eta} \in [1/\sqrt{2}, 1)$. Choosing $C \ge \max\{C_{\mathrm{noise}}, C_{\mathrm{agg}}, D_B^{\mathrm{m}}\}$ and substituting the negative coefficients and residual bounds into (225) gives

$$\begin{aligned}V_{k+1}^{\mathrm{m}} - V_k^{\mathrm{m}} \le -\eta\left(c_G^{\mathrm{m}}G_k + c_I^{\mathrm{m}}I_k + c_J^{\mathrm{m}}J_k + \sum_v c_{M,v}^{\mathrm{m}}M_{v,k}\right)\\ + C\eta^2\left(\frac{1}{n_h} + \kappa\right)\Sigma_\lambda + C\eta\kappa(\zeta_f^2 + \lambda^2\zeta_g^2) + C\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2},\end{aligned} \tag{236}$$

where the fixed positive coefficients may be chosen as

$$c_G^{\mathrm{m}} := d_G^{\mathrm{m}}, \quad c_I^{\mathrm{m}} := \frac{3D_I^{\mathrm{m}}}{4}, \quad c_J^{\mathrm{m}} := \frac{3D_J^{\mathrm{m}}}{4}, \quad c_{M,v}^{\mathrm{m}} := \frac{\vartheta_v\rho_v}{4}. \tag{237}$$

Dropping every contraction except G_k gives

$$V_{k+1}^{\mathrm{m}} - V_k^{\mathrm{m}} \le -c_G^{\mathrm{m}}\eta G_k + C\eta^2\left(\frac{1}{n_h} + \kappa\right)\Sigma_\lambda + C\eta\kappa(\zeta_f^2 + \lambda^2\zeta_g^2) + C\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{238}$$

We retain (236) for the PL proof, where G_k controls the objective gap. Equation (238) suffices for averaged stationarity.

Before summing (238), we bound the initial potential uniformly over the admissible stepsizes, since β_v depends on η. Using $m_v^{-1} = 0$ gives $m_v^0 = (1-\beta_v)(\bar{h}_v^0 + \epsilon_v^0)$. Since the initial iterates are deterministic and the honest noise is centered, $\mathbb{E}\langle\bar{h}_v^0, \epsilon_v^0\rangle = 0$. Consequently,

$$\begin{aligned}e_v^0 &= -\beta_v\bar{h}_v^0 + (1-\beta_v)\epsilon_v^0,\\ M_{v,0} &= \beta_v^2\mathbb{E}\|\bar{h}_v^0\|^2 + (1-\beta_v)^2\mathbb{E}\|\epsilon_v^0\|^2 \le \mathbb{E}\|\bar{h}_v^0\|^2 + \frac{\Sigma_v}{n_h}.\end{aligned} \tag{239}$$

Accordingly, define the initial energy and its horizon-independent upper bound by

$$\Delta_0^{\mathrm{m}} := V_0^{\mathrm{m}} = \mathbb{E}[\Phi(x^0) - \Phi_*] + aI_0 + bJ_0 + \sum_v \vartheta_v M_{v,0}, \tag{240}$$

$$\Delta_0^{\mathrm{m}} \le \overline{\Delta}_0^{\mathrm{m}} := \mathbb{E}[\Phi(x^0) - \Phi_*] + aI_0 + bJ_0 + \sum_v \vartheta_v\left(\mathbb{E}\|\bar{h}_v^0\|^2 + \frac{\Sigma_v}{n_h}\right). \tag{241}$$

The bound $\overline{\Delta}_0^{\mathrm{m}}$ depends on the initial state and fixed problem parameters and is uniform over the admissible stepsizes.

The potential increments telescope as

$$\sum_{k=0}^{K-1}(V_{k+1}^{\mathrm{m}} - V_k^{\mathrm{m}}) = V_K^{\mathrm{m}} - V_0^{\mathrm{m}}. \tag{242}$$

Summing (238) over $k = 0, \dots, K-1$ therefore yields

$$c_G^{\mathrm{m}}\eta\sum_{k=0}^{K-1}G_k \le \Delta_0^{\mathrm{m}} - V_K^{\mathrm{m}} + CK\eta^2\left(\frac{1}{n_h} + \kappa\right)\Sigma_\lambda + CK\eta\kappa(\zeta_f^2 + \lambda^2\zeta_g^2) + CK\eta\frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{243}$$

Using $V_K^{\mathrm{m}} \geq 0$ and dividing by $c_G^{\mathrm{m}} \eta K$ yields

$$\frac{1}{K}\sum_{k=0}^{K-1} G_k \leq C\left[\frac{\Delta_0^{\mathrm{m}}}{\eta K} + \eta\left(\frac{1}{n_h}+\kappa\right)\Sigma_\lambda + \kappa(\zeta_f^2+\lambda^2\zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2}\right]. \tag{244}$$

We now choose the stepsize to balance the first two terms in (244), using the uniform initial-gap bound. Define

$$S_{\mathrm{m}} := \left(\frac{1}{n_h}+\kappa\right)\Sigma_\lambda = \frac{\chi_\kappa \Sigma_\lambda}{n_h}. \tag{245}$$

Using $\Delta_0^{\mathrm{m}} \leq \overline{\Delta}_0^{\mathrm{m}}$, apply the arithmetic–geometric mean inequality to the two transient terms:

$$\begin{aligned} &\frac{\overline{\Delta}_0^{\mathrm{m}}}{\eta K} + \eta S_{\mathrm{m}} \geq 2\sqrt{\frac{\overline{\Delta}_0^{\mathrm{m}} S_{\mathrm{m}}}{K}}, \\ &\frac{\overline{\Delta}_0^{\mathrm{m}}}{\eta K} = \eta S_{\mathrm{m}} \quad \Longleftrightarrow \quad \eta = \eta_*^{\mathrm{m}} := \sqrt{\frac{\overline{\Delta}_0^{\mathrm{m}}}{K S_{\mathrm{m}}}}. \end{aligned} \tag{246}$$

This positive minimizing stepsize is admissible precisely when

$$\eta_*^{\mathrm{m}} \leq \bar{\eta}_{\mathrm{m}} \quad \Longleftrightarrow \quad K \geq \frac{\overline{\Delta}_0^{\mathrm{m}}}{\bar{\eta}_{\mathrm{m}}^2 S_{\mathrm{m}}} = \frac{n_h \overline{\Delta}_0^{\mathrm{m}}}{\bar{\eta}_{\mathrm{m}}^2 \chi_\kappa \Sigma_\lambda}. \tag{247}$$

This specifies the sufficiently large horizon in Theorem 3. Substitution gives

$$\frac{\overline{\Delta}_0^{\mathrm{m}}}{\eta_*^{\mathrm{m}} K} + \eta_*^{\mathrm{m}} S_{\mathrm{m}} = 2\sqrt{\frac{\overline{\Delta}_0^{\mathrm{m}} S_{\mathrm{m}}}{K}} = 2\sqrt{\frac{\overline{\Delta}_0^{\mathrm{m}} \chi_\kappa \Sigma_\lambda}{n_h K}}. \tag{248}$$

Finally, R is uniform on $\{0,\dots,K-1\}$ and independent of the algorithmic randomness. The law of total expectation gives

$$\begin{aligned} \mathbb{E}\|\nabla\Phi(x^R)\|^2 &= \sum_{k=0}^{K-1} \mathbb{P}(R=k)\,\mathbb{E}[\|\nabla\Phi(x^k)\|^2 \mid R=k] \\ &= \frac{1}{K}\sum_{k=0}^{K-1} G_k. \end{aligned} \tag{249}$$

Combining (248) and (249) with (244) proves (17). □

E.3 Proof of Theorem 4: Convergence under the PL Condition

Theorem 4 (Convergence of FOBFBL-M under the PL condition)**.** *Under the same assumptions on the problem, aggregation rule, and multiplier as in Theorem 3, suppose Assumption 6 also holds. There exists $\gamma_{\mathrm{M}} > 0$ such that, for sufficiently large K, choosing $\eta = \log(n_h K)/(\gamma_{\mathrm{M}} K)$ with the stepsize ratios and momentum parameters of Theorem 3 gives*

$$\mathbb{E}[\Phi(x^K) - \Phi_*] = \mathcal{O}\Bigg(\underbrace{\frac{\overline{\Delta}_0^{\mathrm{m}}}{n_h K} + \frac{\chi_\kappa \Sigma_\lambda \log(n_h K)}{n_h K}}_{\text{finite-horizon error}} + \underbrace{\kappa(\zeta_f^2+\lambda^2\zeta_g^2)}_{\text{Byzantine-induced floor}} + \underbrace{\frac{C_{\mathrm{lag}}^2}{\lambda^2}}_{\text{Lagrangian bias}}\Bigg). \tag{18}$$

Proof. Fix the stepsize ratios, Lyapunov weights, and momentum constants constructed in Appendix E.2. Let $\bar{\eta}_{\mathrm{m}}$ be the ceiling in (235). Recall the nonnegative Lyapunov function

$$V_k^{\mathrm{m}} = \mathbb{E}[\Phi(x^k) - \Phi_*] + aI_k + bJ_k + \sum_v \vartheta_v M_{v,k}, \qquad V_0^{\mathrm{m}} \leq \overline{\Delta}_0^{\mathrm{m}}. \tag{250}$$

Collect the sampling, aggregation, and finite-multiplier terms per unit stepsize as

$$r_{\mathrm{M}}(\eta) := \frac{\eta \chi_\kappa \Sigma_\lambda}{n_h} + \kappa(\zeta_f^2+\lambda^2\zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2}. \tag{251}$$

The strict recursion (236) becomes

$$V_{k+1}^{\mathrm{m}} - V_k^{\mathrm{m}} \le -\eta \left(c_G^{\mathrm{m}} G_k + c_I^{\mathrm{m}} I_k + c_J^{\mathrm{m}} J_k + \sum_v c_{M,v}^{\mathrm{m}} M_{v,k} \right) + C\eta r_{\mathrm{M}}(\eta). \tag{252}$$

By Assumption 6, the PL inequality holds at each iterate. Taking expectations gives

$$G_k = \mathbb{E}\|\nabla\Phi(x^k)\|^2 \ge 2\mu_{\mathrm{PL}} \mathbb{E}[\Phi(x^k) - \Phi_*]. \tag{253}$$

Define the common contraction coefficient by

$$\gamma_{\mathrm{M}} := \min\left\{ 2\mu_{\mathrm{PL}} c_G^{\mathrm{m}}, \frac{c_I^{\mathrm{m}}}{a}, \frac{c_J^{\mathrm{m}}}{b}, \min_v \frac{c_{M,v}^{\mathrm{m}}}{\vartheta_v} \right\} > 0. \tag{254}$$

All weights and coefficients were fixed independently of η and K, so the same is true of γ_{M}. Its definition and (253) imply

$$\begin{aligned} c_G^{\mathrm{m}} G_k &\ge 2\mu_{\mathrm{PL}} c_G^{\mathrm{m}} \mathbb{E}[\Phi(x^k) - \Phi_*] \ge \gamma_{\mathrm{M}} \mathbb{E}[\Phi(x^k) - \Phi_*], \\ c_I^{\mathrm{m}} I_k &= \frac{c_I^{\mathrm{m}}}{a}(aI_k) \ge \gamma_{\mathrm{M}} a I_k, \\ c_J^{\mathrm{m}} J_k &= \frac{c_J^{\mathrm{m}}}{b}(bJ_k) \ge \gamma_{\mathrm{M}} b J_k, \\ c_{M,v}^{\mathrm{m}} M_{v,k} &= \frac{c_{M,v}^{\mathrm{m}}}{\vartheta_v}(\vartheta_v M_{v,k}) \ge \gamma_{\mathrm{M}} \vartheta_v M_{v,k}, \qquad v \in \{x, y, z\}. \end{aligned} \tag{255}$$

Because all six components of V_k^{m} are nonnegative, summing these bounds gives

$$c_G^{\mathrm{m}} G_k + c_I^{\mathrm{m}} I_k + c_J^{\mathrm{m}} J_k + \sum_v c_{M,v}^{\mathrm{m}} M_{v,k} \ge \gamma_{\mathrm{M}} V_k^{\mathrm{m}}. \tag{256}$$

For $0 < \eta \le \min\{\bar{\eta}_{\mathrm{m}}, 1/\gamma_{\mathrm{M}}\}$, substitution into (252) yields

$$V_{k+1}^{\mathrm{m}} \le (1 - \gamma_{\mathrm{M}}\eta) V_k^{\mathrm{m}} + C\eta r_{\mathrm{M}}(\eta). \tag{257}$$

Let $q_{\mathrm{M}} := 1 - \gamma_{\mathrm{M}}\eta \in [0, 1)$. The parameters are constant over the K iterations, so $r_{\mathrm{M}}(\eta)$ is constant as well. For every $1 \le j \le K$,

$$V_K^{\mathrm{m}} \le q_{\mathrm{M}}^j V_{K-j}^{\mathrm{m}} + C\eta r_{\mathrm{M}}(\eta) \sum_{t=0}^{j-1} q_{\mathrm{M}}^t. \tag{258}$$

Indeed, $j = 1$ is (257); applying the same inequality to V_{K-j}^{m} proves the case $j + 1$. At $j = K$, the geometric sum is

$$\sum_{t=0}^{K-1} q_{\mathrm{M}}^t = \frac{1 - q_{\mathrm{M}}^K}{1 - q_{\mathrm{M}}} = \frac{1 - q_{\mathrm{M}}^K}{\gamma_{\mathrm{M}}\eta}. \tag{259}$$

Thus

$$\begin{aligned} V_K^{\mathrm{m}} &\le q_{\mathrm{M}}^K V_0^{\mathrm{m}} + C\eta r_{\mathrm{M}}(\eta) \sum_{t=0}^{K-1} q_{\mathrm{M}}^t \\ &= q_{\mathrm{M}}^K V_0^{\mathrm{m}} + \frac{C}{\gamma_{\mathrm{M}}}(1 - q_{\mathrm{M}}^K) r_{\mathrm{M}}(\eta). \end{aligned} \tag{260}$$

Using $\mathbb{E}[\Phi(x^K) - \Phi_*] \le V_K^{\mathrm{m}}$, $V_0^{\mathrm{m}} \le \overline{\Delta}_0^{\mathrm{m}}$, and $1 - q_{\mathrm{M}}^K \le 1$, we obtain

$$\begin{aligned} \mathbb{E}[\Phi(x^K) - \Phi_*] \le{}& (1 - \gamma_{\mathrm{M}}\eta)^K \overline{\Delta}_0^{\mathrm{m}} \\ &+ \frac{C}{\gamma_{\mathrm{M}}} \left[\frac{\eta \chi_\kappa \Sigma_\lambda}{n_h} + \kappa(\zeta_f^2 + \lambda^2 \zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2} \right]. \end{aligned} \tag{261}$$

The first term is the geometric transient. The sampling term decreases with η; the heterogeneity and finite-multiplier terms remain in the residual.

Choose the horizon and stepsize so that

$$n_h K > 1, \qquad \eta = \frac{\log(n_h K)}{\gamma_{\mathrm{M}} K} \le \min\{\bar{\eta}_{\mathrm{m}}, 1/\gamma_{\mathrm{M}}\}. \tag{262}$$

Such a horizon exists because $\log(n_h K)/K \to 0$ as $K \to \infty$ and the ceiling is fixed and positive. The first condition ensures $\eta > 0$. The inequality $1 - u \le e^{-u}$ for $u \in [0, 1]$ gives

$$(1 - \gamma_{\mathrm{M}}\eta)^K \le \exp(-\gamma_{\mathrm{M}}\eta K) = \exp(-\log(n_h K)) = \frac{1}{n_h K}. \tag{263}$$

The sampling term in (261) becomes

$$\frac{C}{\gamma_{\mathrm{M}}} \frac{\eta \chi_\kappa \Sigma_\lambda}{n_h} = \frac{C \chi_\kappa \Sigma_\lambda \log(n_h K)}{\gamma_{\mathrm{M}}^2 n_h K}. \tag{264}$$

Substituting both estimates into (261) yields

$$\begin{aligned} \mathbb{E}[\Phi(x^K) - \Phi_*] \le \frac{\overline{\Delta}_0^{\mathrm{m}}}{n_h K} + \frac{C \chi_\kappa \Sigma_\lambda \log(n_h K)}{\gamma_{\mathrm{M}}^2 n_h K} \\ + \frac{C}{\gamma_{\mathrm{M}}} \left[\kappa(\zeta_f^2 + \lambda^2 \zeta_g^2) + \frac{C_{\mathrm{lag}}^2}{\lambda^2} \right]. \end{aligned} \tag{265}$$

Absorbing the fixed factors involving γ_{M} into the implicit constant proves (18). □

E.4 Proofs of the Auxiliary Lemmas

We prove the lemmas stated in Appendix E.1, beginning with aggregation error and momentum deviation, followed by upper-level descent and lower-level tracking. We use the error decomposition introduced there throughout.

E.4.1 Proof of Lemma 9: Aggregation Error with Momentum

Proof. Fix $v \in \{x, y, z\}$ and write $\beta = \beta_v$. All client sums in this proof are over $i \in \mathcal{H}$. Define the current oracle noise and predictable client deviation by

$$\xi_{v,i}^t := h_{v,i}^t - \bar{h}_{v,i}^t, \qquad d_{v,i}^t := \bar{h}_{v,i}^t - \bar{h}_v^t, \qquad \epsilon_v^t := \frac{1}{n_h} \sum_i \xi_{v,i}^t. \tag{266}$$

The zero initialization of the momentum buffers gives

$$m_{v,i}^k = (1-\beta) \sum_{t=0}^k \beta^{k-t} h_{v,i}^t, \qquad m_v^k = (1-\beta) \sum_{t=0}^k \beta^{k-t} h_v^t. \tag{267}$$

Subtracting these identities yields

$$m_{v,i}^k - m_v^k = (1-\beta) \sum_{t=0}^k \beta^{k-t} \big[(\xi_{v,i}^t - \epsilon_v^t) + d_{v,i}^t \big] =: N_{v,i}^k + D_{v,i}^k, \tag{268}$$

where $N_{v,i}^k$ and $D_{v,i}^k$ denote the sums of the noise and client-deviation terms, respectively. We control their second moments separately because future client deviations may depend on earlier oracle noise.

For $t < s$, $\xi_{v,i}^t - \epsilon_v^t$ is $\mathcal{F}_s$-measurable, whereas the s-th oracle noise has zero conditional mean. The tower property therefore gives

$$\begin{aligned} &\mathbb{E}\langle \xi_{v,i}^t - \epsilon_v^t, \xi_{v,i}^s - \epsilon_v^s \rangle \\ &\quad = \mathbb{E}\left\langle \xi_{v,i}^t - \epsilon_v^t, \mathbb{E}[\xi_{v,i}^s - \epsilon_v^s \mid \mathcal{F}_s] \right\rangle = 0. \end{aligned} \tag{269}$$

At a fixed round, the empirical-mean identity and Lemma 3 imply

$$\frac{1}{n_h} \sum_i \mathbb{E}\|\xi_{v,i}^t - \epsilon_v^t\|^2 = \frac{1}{n_h} \sum_i \mathbb{E}\|\xi_{v,i}^t\|^2 - \mathbb{E}\|\epsilon_v^t\|^2 \le \Sigma_v. \tag{270}$$

Expanding the squared temporal sum in $N_{v,i}^k$ and using these two identities yields

$$\begin{aligned} \frac{1}{n_h} \sum_i \mathbb{E}\|N_{v,i}^k\|^2 &= (1-\beta)^2 \sum_{t=0}^k \beta^{2(k-t)} \frac{1}{n_h} \sum_i \mathbb{E}\|\xi_{v,i}^t - \epsilon_v^t\|^2 \\ &\le (1-\beta)^2 \sum_{j=0}^k \beta^{2j} \Sigma_v. \end{aligned} \tag{271}$$

The squared geometric weights have the exact sum

$$(1-\beta)^2 \sum_{j=0}^k \beta^{2j} = \frac{1-\beta}{1+\beta} \big(1 - \beta^{2(k+1)}\big). \tag{272}$$

Consequently,

$$\mathcal{D}_{v,k}^{\text{stoch}} := \frac{1}{n_h} \sum_i \mathbb{E}\|N_{v,i}^k\|^2 \le \frac{1-\beta}{1+\beta} \big(1 - \beta^{2(k+1)}\big) \Sigma_v. \tag{273}$$

For the client-deviation part, let $w_t = (1-\beta)\beta^{k-t}$ and $a_{v,k} := \sum_{t=0}^k w_t = 1 - \beta^{k+1} \le 1$. Weighted Cauchy–Schwarz, followed by the heterogeneity bound in (58), gives

$$\begin{aligned} \frac{1}{n_h} \sum_i \left\| \sum_{t=0}^k w_t d_{v,i}^t \right\|^2 &\le a_{v,k} \sum_{t=0}^k w_t \frac{1}{n_h} \sum_i \|d_{v,i}^t\|^2 \\ &\le a_{v,k}^2 Z_v. \end{aligned} \tag{274}$$

This bound holds for every realization of the iterates. Taking expectations gives

$$\mathcal{D}_{v,k}^{\mathrm{het}} := \frac{1}{n_h}\sum_i \mathbb{E}\|D_{v,i}^k\|^2 \le (1-\beta^{k+1})^2 Z_v. \tag{275}$$

Apply Definition 2 to the momentum messages and use $\|N+D\|^2 \le 2\|N\|^2 + 2\|D\|^2$:

$$\begin{aligned} R_{v,k}^{\mathrm{m}} &\le \frac{\kappa}{n_h}\sum_i \mathbb{E}\|m_{v,i}^k - m_v^k\|^2 \\ &\le 2\kappa\big(\mathcal{D}_{v,k}^{\mathrm{stoch}} + \mathcal{D}_{v,k}^{\mathrm{het}}\big) \\ &\le 2\kappa\left[\frac{1-\beta}{1+\beta}(1-\beta^{2(k+1)})\Sigma_v + (1-\beta^{k+1})^2 Z_v\right] \\ &\le 2\kappa\left[\frac{1-\beta}{1+\beta}\Sigma_v + Z_v\right]. \end{aligned} \tag{276}$$

This proves (201). Inserting the three block scales gives (202)–(204). The stochastic factor comes from squared temporal weights, while heterogeneity is controlled by their total mass. The proof uses no centering of the aggregation error s_v^k and no cancellation between $N_{v,i}^k$ and $D_{v,i}^k$. Finally, $v^{k+1} - v_{\mathrm{av}}^{k+1} = -\eta_v s_v^k$ proves (205). □

Shared temporal-deviation recursion. The three blockwise deviation proofs use the same temporal-deviation identity. For $v \in \{x, y, z\}$, express the honest stochastic mean direction as $\bar{h}_v^{k+1} + \epsilon_v^{k+1}$. Conditional unbiasedness and the blockwise variance bound give

$$\mathbb{E}[\epsilon_v^{k+1} \mid \mathcal{F}_{k+1}] = 0, \qquad \mathbb{E}[\|\epsilon_v^{k+1}\|^2 \mid \mathcal{F}_{k+1}] \le \frac{\Sigma_v}{n_h}. \tag{277}$$

Subtracting $\bar{h}_v^{k+1}$ from the honest mean-momentum update yields the exact identity

$$e_v^{k+1} = \beta_v e_v^k + \beta_v(\bar{h}_v^k - \bar{h}_v^{k+1}) + (1-\beta_v)\epsilon_v^{k+1}. \tag{278}$$

Both e_v^k and $\bar{h}_v^{k+1} - \bar{h}_v^k$ are $\mathcal{F}_{k+1}$-measurable. Thus (277) removes the fresh-noise cross term:

$$\mathbb{E}\left\langle e_v^k + \bar{h}_v^k - \bar{h}_v^{k+1}, \epsilon_v^{k+1}\right\rangle = \mathbb{E}\left\langle e_v^k + \bar{h}_v^k - \bar{h}_v^{k+1}, \mathbb{E}[\epsilon_v^{k+1} \mid \mathcal{F}_{k+1}]\right\rangle = 0. \tag{279}$$

Squaring (278) therefore gives

$$M_{v,k+1} = \beta_v^2 \mathbb{E}\|e_v^k + \bar{h}_v^k - \bar{h}_v^{k+1}\|^2 + (1-\beta_v)^2\mathbb{E}\|\epsilon_v^{k+1}\|^2. \tag{280}$$

Young's inequality bounds the cross term between the preceding deviation and the direction drift. For every $\alpha_v > 0$,

$$\begin{aligned} M_{v,k+1} \le{}& \beta_v^2(1+\alpha_v)M_{v,k} \\ &+ \beta_v^2(1+\alpha_v^{-1})\mathbb{E}\|\bar{h}_v^{k+1} - \bar{h}_v^k\|^2 + \frac{(1-\beta_v)^2\Sigma_v}{n_h}. \end{aligned} \tag{281}$$

Under $0 < \rho_v\eta \le 1/2$ and $\beta_v = \sqrt{1-\rho_v\eta}$, choose $\alpha_v = (1-\beta_v^2)/(2\beta_v^2) > 0$ and use $1-\beta_v^2 = \rho_v\eta$. Then

$$\beta_v^2(1+\alpha_v) = 1 - \frac{\rho_v\eta}{2}, \tag{282}$$

$$\beta_v^2(1+\alpha_v^{-1}) = \beta_v^2 + \frac{2\beta_v^4}{\rho_v\eta} = \frac{(1-\rho_v\eta)(2-\rho_v\eta)}{\rho_v\eta} \le \frac{2}{\rho_v\eta},$$

$$(1-\beta_v)^2 = \frac{\rho_v^2\eta^2}{(1+\beta_v)^2} \le \rho_v^2\eta^2. \tag{283}$$

This choice retains a contraction of $\rho_v\eta/2$ and introduces a factor $1/(\rho_v\eta)$ in the direction-drift term. Substitution in (281) gives

$$M_{v,k+1} \le \left(1 - \frac{\rho_v\eta}{2}\right)M_{v,k} + \frac{2}{\rho_v\eta}\mathbb{E}\|\bar{h}_v^{k+1} - \bar{h}_v^k\|^2 + \frac{\rho_v^2\eta^2\Sigma_v}{n_h}. \tag{284}$$

It remains to bound the change in each deterministic block direction under the joint server update.

Second-moment bounds. For the lower-level directions, optimality of the two targets and gradient Lipschitz continuity give

$$\|\bar{h}_y^k\|^2 \le L_\lambda^2 \|y^k - y_\lambda^\star(x^k)\|^2, \qquad \|\bar{h}_z^k\|^2 \le l_{g,1}^2 \|z^k - y^\star(x^k)\|^2. \tag{285}$$

We now record the second moments used in the subsequent proofs. The decomposition $\widehat{m}_v^k = \bar{h}_v^k + e_v^k + s_v^k$ gives

$$\mathbb{E}\|\widehat{m}_x^k\|^2 \le 3H_k + 3M_{x,k} + 3R_{x,k}^{\mathrm{m}}, \tag{286}$$

$$\mathbb{E}\|\widehat{m}_y^k\|^2 \le 3L_\lambda^2 I_k + 3M_{y,k} + 3R_{y,k}^{\mathrm{m}}, \tag{287}$$

$$\mathbb{E}\|\widehat{m}_z^k\|^2 \le 3l_{g,1}^2 J_k + 3M_{z,k} + 3R_{z,k}^{\mathrm{m}}. \tag{288}$$

For the last two bounds, the deterministic directions vanish at $y_\lambda^\star(x^k)$ and $y^\star(x^k)$, respectively. Lipschitz continuity then turns their squared norms into I_k and J_k. Thus no bounded-iterate assumption is used.

E.4.2 Proof of Lemma 10: Momentum Deviation in the y Block

Proof. By (284), it suffices to control motion of the deterministic y-direction. This direction changes through both coordinates updated in the round. Insert the intermediate direction $\bar{h}_y(x^{k+1}, y^k)$ to separate the y-update from the x-update, and use joint L_λ-Lipschitz continuity:

$$\begin{aligned}\|\bar{h}_y^{k+1} - \bar{h}_y^k\|^2 &\le 2\|\bar{h}_y(x^{k+1}, y^{k+1}) - \bar{h}_y(x^{k+1}, y^k)\|^2 \\ &\quad + 2\|\bar{h}_y(x^{k+1}, y^k) - \bar{h}_y(x^k, y^k)\|^2 \\ &\le 2L_\lambda^2 \left(\eta_y^2 \|\widehat{m}_y^k\|^2 + \eta_x^2 \|\widehat{m}_x^k\|^2\right).\end{aligned} \tag{289}$$

Taking expectations and substituting (286) and (287) gives

$$\begin{aligned}\mathbb{E}\|\bar{h}_y^{k+1} - \bar{h}_y^k\|^2 &\le 6L_\lambda^2 \eta^2 \Big[H_k + M_{x,k} + R_{x,k}^{\mathrm{m}} + (c_y^\circ)^2 (L_\lambda^2 I_k + M_{y,k} + R_{y,k}^{\mathrm{m}})\Big] \\ &\le B_y \eta^2 \big(H_k + I_k + M_{x,k} + M_{y,k} + R_{x,k}^{\mathrm{m}} + R_{y,k}^{\mathrm{m}}\big),\end{aligned} \tag{290}$$

where a valid coefficient, independent of the momentum constants, is

$$B_y := 6L_\lambda^2 \max\{1, (c_y^\circ)^2 L_\lambda^2, (c_y^\circ)^2\}. \tag{291}$$

Here H_k is the second moment of the conditional mean x-direction, while I_k is the squared tracking error controlling the y-direction norm. The other four terms are temporal-deviation and aggregation perturbations. Putting (290) into (284) yields

$$\begin{aligned}M_{y,k+1} &\le \left(1 - \frac{\rho_y \eta}{2}\right) M_{y,k} \\ &\quad + \frac{A_y \eta}{\rho_y} \big(H_k + I_k + M_{x,k} + M_{y,k} + R_{x,k}^{\mathrm{m}} + R_{y,k}^{\mathrm{m}}\big) + \rho_y^2 \eta^2 \frac{\Sigma_y}{n_h},\end{aligned} \tag{292}$$

with $A_y := 2B_y$. The factor η/ρ_y comes from multiplying the $\mathcal{O}(\eta^2)$ direction drift by the $1/(\rho_y \eta)$ factor in the common recursion, which proves (206). □

E.4.3 Proof of Lemma 11: Momentum Deviation in the z Block

Proof. The argument is parallel to the y-block, but the original lower direction depends only on (x, z). Insert $\bar{h}_z(x^{k+1}, z^k)$ to separate those two coordinate changes and apply $l_{g,1}$-Lipschitz continuity:

$$\begin{aligned}\|\bar{h}_z^{k+1} - \bar{h}_z^k\|^2 &\le 2\|\bar{h}_z(x^{k+1}, z^{k+1}) - \bar{h}_z(x^{k+1}, z^k)\|^2 \\ &\quad + 2\|\bar{h}_z(x^{k+1}, z^k) - \bar{h}_z(x^k, z^k)\|^2 \\ &\le 2l_{g,1}^2 \left(\eta_z^2 \|\widehat{m}_z^k\|^2 + \eta_x^2 \|\widehat{m}_x^k\|^2\right).\end{aligned} \tag{293}$$

Using (286) and (288), followed by the fixed stepsize ratios, gives

$$\begin{aligned}\mathbb{E}\|\bar{h}_z^{k+1} - \bar{h}_z^k\|^2 &\le 6l_{g,1}^2 \eta^2 \Big[H_k + M_{x,k} + R_{x,k}^{\mathrm{m}} + (c_z^\circ)^2 (l_{g,1}^2 J_k + M_{z,k} + R_{z,k}^{\mathrm{m}})\Big] \\ &\le B_z \eta^2 \big(H_k + J_k + M_{x,k} + M_{z,k} + R_{x,k}^{\mathrm{m}} + R_{z,k}^{\mathrm{m}}\big).\end{aligned} \tag{294}$$

Here we may choose

$$B_z := 6l_{g,1}^2 \max\{1, (c_z^\circ)^2 l_{g,1}^2, (c_z^\circ)^2\}. \tag{295}$$

Substitution into (284) gives

$$\begin{aligned} M_{z,k+1} \le &\left(1 - \frac{\rho_z \eta}{2}\right) M_{z,k} \\ &+ \frac{A_z \eta}{\rho_z}\left(H_k + J_k + M_{x,k} + M_{z,k} + R_{x,k}^{\mathrm{m}} + R_{z,k}^{\mathrm{m}}\right) + \rho_z^2 \eta^2 \frac{\Sigma_z}{n_h}, \end{aligned} \tag{296}$$

with $A_z := 2B_z$. Here J_k replaces I_k because the deterministic z-direction vanishes at $y^\star(x^k)$. This proves (207). □

E.4.4 Proof of Lemma 12: Momentum Deviation in the x Block

Proof. Unlike the two lower directions, $\bar{h}_x$ depends on all three iterates. Its drift must therefore be split along all three coordinate updates. Insert the intermediate points

$$(x^{k+1}, y^{k+1}, z^{k+1}), \quad (x^{k+1}, y^k, z^{k+1}), \quad (x^{k+1}, y^k, z^k), \quad (x^k, y^k, z^k). \tag{297}$$

For the changes in x, y, and z, respectively, the direction in (54) has Lipschitz constants

$$\ell_{x,x} := l_{f,1} + 2\lambda l_{g,1}, \qquad \ell_{x,y} := L_\lambda, \qquad \ell_{x,z} := \lambda l_{g,1}. \tag{298}$$

Telescoping along the intermediate points and applying the three-term squared-norm bound gives

$$\|\bar{h}_x^{k+1} - \bar{h}_x^k\|^2 \le 3\left(\ell_{x,x}^2 \eta_x^2 \|\widehat{m}_x^k\|^2 + \ell_{x,y}^2 \eta_y^2 \|\widehat{m}_y^k\|^2 + \ell_{x,z}^2 \eta_z^2 \|\widehat{m}_z^k\|^2\right). \tag{299}$$

Taking expectations, using all three inequalities in equations (286)–(288), and collecting fixed ratios yield

$$\begin{aligned} \mathbb{E}\|\bar{h}_x^{k+1} - \bar{h}_x^k\|^2 &\le 9\eta^2 \Big[\ell_{x,x}^2 (H_k + M_{x,k} + R_{x,k}^{\mathrm{m}}) \\ &\qquad + (c_y^\circ)^2 \ell_{x,y}^2 (L_\lambda^2 I_k + M_{y,k} + R_{y,k}^{\mathrm{m}}) + (c_z^\circ)^2 \ell_{x,z}^2 (l_{g,1}^2 J_k + M_{z,k} + R_{z,k}^{\mathrm{m}})\Big] \\ &\le B_x \eta^2 \left(H_k + I_k + J_k + \sum_{v \in \{x,y,z\}} M_{v,k} + \sum_{v \in \{x,y,z\}} R_{v,k}^{\mathrm{m}}\right), \end{aligned} \tag{300}$$

where one valid choice is

$$\begin{aligned} B_x := 9 \max\big\{&\ell_{x,x}^2, (c_y^\circ)^2 \ell_{x,y}^2 \max\{1, L_\lambda^2\}, \\ &(c_z^\circ)^2 \ell_{x,z}^2 \max\{1, l_{g,1}^2\}\big\}. \end{aligned} \tag{301}$$

Putting this estimate into (284) gives

$$\begin{aligned} M_{x,k+1} \le &\left(1 - \frac{\rho_x \eta}{2}\right) M_{x,k} \\ &+ \frac{A_x \eta}{\rho_x}\left(H_k + I_k + J_k + \sum_v M_{v,k} + \sum_v R_{v,k}^{\mathrm{m}}\right) + \rho_x^2 \eta^2 \frac{\Sigma_x}{n_h}, \end{aligned} \tag{302}$$

with $A_x := 2B_x$. Both tracking errors and all three momentum errors appear because the hypergradient direction depends on every coordinate. This proves (208). □

E.4.5 Proof of Lemma 13: Upper-Level Descent

Proof. Set $p_k = \nabla\Phi(x^k)$, $q_k = \bar{h}_x^k$, and $r_k = e_x^k + s_x^k$. The proof separates the deterministic mismatch $p_k - q_k$ from the two message perturbations collected in r_k. Since $x^{k+1} = x^k - \eta_x(q_k + r_k)$, L_Φ-smoothness gives

$$\mathbb{E}\Phi(x^{k+1}) \le \mathbb{E}\Phi(x^k) - \eta_x \mathbb{E}\langle p_k, q_k\rangle - \eta_x \mathbb{E}\langle p_k, r_k\rangle + \frac{L_\Phi \eta_x^2}{2}\mathbb{E}\|q_k + r_k\|^2. \tag{303}$$

The polarization identity and Young's inequality give, respectively,

$$-\langle p_k, q_k\rangle = -\frac{1}{2}\|p_k\|^2 - \frac{1}{2}\|q_k\|^2 + \frac{1}{2}\|p_k - q_k\|^2, \tag{304}$$

$$-\mathbb{E}\langle p_k, r_k\rangle \le \frac{1}{8}G_k + 2\mathbb{E}\|r_k\|^2 \le \frac{1}{8}G_k + 4(M_{x,k} + R^{\mathrm{m}}_{x,k}), \tag{305}$$

$$\mathbb{E}\|q_k + r_k\|^2 \le 2H_k + 4M_{x,k} + 4R^{\mathrm{m}}_{x,k}. \tag{306}$$

These bounds retain negative coefficients for both G_k and H_k. The latter absorbs the target-drift terms in the tracking recursions. To bound the remaining direction bias, insert the penalized exact direction:

$$p_k - q_k = \big[p_k - \nabla\mathcal{L}^\star_\lambda(x^k)\big] + \big[\nabla\mathcal{L}^\star_\lambda(x^k) - q_k\big]. \tag{307}$$

Lemma 1 controls the first bracket. For the second bracket, the two tracking components in (73)–(74) give

$$\mathbb{E}\|\nabla\mathcal{L}^\star_\lambda(x^k) - q_k\|^2 \le 2L_\lambda^2 I_k + 2\lambda^2 l_{g,1}^2 J_k. \tag{308}$$

Keeping all three components separate, the full direction-bias bound (59) yields

$$\mathbb{E}\|p_k - q_k\|^2 \le 3\frac{C_{\mathrm{lag}}^2}{\lambda^2} + D_y I_k + D_z J_k. \tag{309}$$

Hence $\frac{C_{\mathrm{lag}}^2}{\lambda^2}$ is the finite-multiplier approximation, while I_k and J_k measure imperfect tracking of the two lower solutions. Neither momentum noise nor aggregation error is hidden in this deterministic mismatch. Substitute equations (304)–(309) into (303). Before the stepsize restriction is imposed, this gives

$$\begin{aligned}\mathbb{E}\Phi(x^{k+1}) \le \mathbb{E}\Phi(x^k) &- \frac{3\eta_x}{8}G_k - \left(\frac{\eta_x}{2} - L_\Phi\eta_x^2\right)H_k \\ &+ \frac{\eta_x}{2}\left(D_y I_k + D_z J_k + 3\frac{C_{\mathrm{lag}}^2}{\lambda^2}\right) + (4\eta_x + 2L_\Phi\eta_x^2)(M_{x,k} + R^{\mathrm{m}}_{x,k}).\end{aligned} \tag{310}$$

For $0 < \eta_x \le 1/(4L_\Phi)$, the two stepsize-dependent coefficients satisfy

$$\frac{\eta_x}{2} - L_\Phi\eta_x^2 \ge \frac{\eta_x}{4}, \qquad 4\eta_x + 2L_\Phi\eta_x^2 \le \frac{9\eta_x}{2}. \tag{311}$$

Hence

$$\begin{aligned}\mathbb{E}\Phi(x^{k+1}) \le \mathbb{E}\Phi(x^k) &- \frac{3\eta_x}{8}G_k - \frac{\eta_x}{4}H_k \\ &+ \frac{\eta_x}{2}\left(D_y I_k + D_z J_k + 3\frac{C_{\mathrm{lag}}^2}{\lambda^2}\right) + \frac{9\eta_x}{2}(M_{x,k} + R^{\mathrm{m}}_{x,k}).\end{aligned} \tag{312}$$

This proves (209) with $c_{\mathrm{m}} = 1/4$ and $C_{\mathrm{m}} = \max\{D_y/2, D_z/2, 9/2\}$. Both components of r_k are controlled through second moments throughout the proof. □

E.4.6 Proof of Lemma 14: Penalized Lower-Level Tracking

Proof. Let $q_y^k := y^\star_\lambda(x^k)$, $d_y^k := y^k - q_y^k$, and $g_y^k := \bar{h}_y^k$. The direction g_y^k is the gradient of a μ_λ-strongly convex, L_λ-smooth lower-level objective and vanishes at q_y^k. We first expand the update around this fixed target:

$$\begin{aligned}\mathbb{E}\|y^{k+1} - q_y^k\|^2 = I_k &- 2\eta_y\mathbb{E}\langle d_y^k, g_y^k\rangle - 2\eta_y\mathbb{E}\langle d_y^k, e_y^k + s_y^k\rangle \\ &+ \eta_y^2\mathbb{E}\|g_y^k + e_y^k + s_y^k\|^2.\end{aligned} \tag{313}$$

Interpolation for a strongly convex, smooth function gives

$$\langle d_y^k, g_y^k\rangle \ge \frac{\mu_\lambda L_\lambda}{\mu_\lambda + L_\lambda}\|d_y^k\|^2 + \frac{1}{\mu_\lambda + L_\lambda}\|g_y^k\|^2. \tag{314}$$

Neither the momentum deviation nor the aggregation error is assumed to be centered. Young's inequality and $\|e+s\|^2 \le 2\|e\|^2 + 2\|s\|^2$ give

$$-2\eta_y \mathbb{E}\langle d_y^k, e_y^k + s_y^k\rangle \le \frac{\mu_\lambda \eta_y}{4} I_k + \frac{8\eta_y}{\mu_\lambda}(M_{y,k} + R_{y,k}^{\mathrm{m}}), \tag{315}$$

$$\mathbb{E}\|g_y^k + e_y^k + s_y^k\|^2 \le 3\mathbb{E}\|g_y^k\|^2 + 3M_{y,k} + 3R_{y,k}^{\mathrm{m}}. \tag{316}$$

Substituting these bounds into (313) yields

$$\begin{aligned}\mathbb{E}\|y^{k+1} - q_y^k\|^2 \le{}& \left(1 - \frac{2\mu_\lambda L_\lambda \eta_y}{\mu_\lambda + L_\lambda} + \frac{\mu_\lambda \eta_y}{4}\right) I_k \\ &+ \left(-\frac{2\eta_y}{\mu_\lambda + L_\lambda} + 3\eta_y^2\right)\mathbb{E}\|g_y^k\|^2 + \left(\frac{8\eta_y}{\mu_\lambda} + 3\eta_y^2\right)(M_{y,k} + R_{y,k}^{\mathrm{m}}).\end{aligned} \tag{317}$$

Assume the stepsize bound in the lemma and write $a_y := \mu_\lambda \eta_y$. Since $L_\lambda \ge \mu_\lambda$,

$$\begin{aligned}&\frac{2\mu_\lambda L_\lambda}{\mu_\lambda + L_\lambda} \ge \mu_\lambda, \qquad -\frac{2\eta_y}{\mu_\lambda + L_\lambda} + 3\eta_y^2 \le -\frac{\eta_y}{\mu_\lambda + L_\lambda} \le 0, \\ &0 < a_y \le 1, \qquad \frac{8\eta_y}{\mu_\lambda} + 3\eta_y^2 \le \frac{11\eta_y}{\mu_\lambda}.\end{aligned} \tag{318}$$

The distance coefficient is at most $1 - 3a_y/4$. Retaining the weaker coefficient $1 - a_y/4$ and discarding the nonpositive direction term gives

$$\mathbb{E}\|y^{k+1} - q_y^k\|^2 \le \left(1 - \frac{a_y}{4}\right) I_k + \frac{11\eta_y}{\mu_\lambda}(M_{y,k} + R_{y,k}^{\mathrm{m}}). \tag{319}$$

Target drift. To replace q_y^k by q_y^{k+1}, choose $\chi_y := a_y/8$. The optimizer-map bound in Lemma 2 gives

$$\mathbb{E}\|q_y^{k+1} - q_y^k\|^2 \le L_{\lambda,*}^2 \eta_x^2 \mathbb{E}\|\widehat{m}_x^k\|^2. \tag{320}$$

Expanding $y^{k+1} - q_y^{k+1} = (y^{k+1} - q_y^k) + (q_y^k - q_y^{k+1})$ and applying Young's inequality yields

$$\begin{aligned}I_{k+1} &\le (1+\chi_y)\mathbb{E}\|y^{k+1} - q_y^k\|^2 + (1 + \chi_y^{-1})\mathbb{E}\|q_y^{k+1} - q_y^k\|^2 \\ &\le \left(1 + \frac{a_y}{8}\right)\mathbb{E}\|y^{k+1} - q_y^k\|^2 + \left(1 + \frac{8}{a_y}\right) L_{\lambda,*}^2 \eta_x^2 \mathbb{E}\|\widehat{m}_x^k\|^2.\end{aligned} \tag{321}$$

Thus this drift estimate uses only Lipschitz continuity of the optimizer map and second moments of the server direction. Substituting (319) and (286) gives

$$\begin{aligned}I_{k+1} \le{}& \left(1 + \frac{a_y}{8}\right)\left(1 - \frac{a_y}{4}\right) I_k \\ &+ \left(1 + \frac{a_y}{8}\right)\frac{11\eta_y}{\mu_\lambda}(M_{y,k} + R_{y,k}^{\mathrm{m}}) \\ &+ 3\left(1 + \frac{8}{a_y}\right) L_{\lambda,*}^2 \eta_x^2 (H_k + M_{x,k} + R_{x,k}^{\mathrm{m}}).\end{aligned} \tag{322}$$

For $0 < a_y \le 1$, the three coefficients satisfy

$$\begin{aligned}&\left(1 + \frac{a_y}{8}\right)\left(1 - \frac{a_y}{4}\right) = 1 - \frac{a_y}{8} - \frac{a_y^2}{32} \le 1 - \frac{a_y}{8}, \\ &1 + \frac{a_y}{8} \le 2, \qquad 3\left(1 + \frac{8}{a_y}\right) \le \frac{27}{a_y}.\end{aligned} \tag{323}$$

Consequently,

$$\begin{aligned}I_{k+1} \le{}& \left(1 - \frac{\mu_\lambda \eta_y}{8}\right) I_k + \frac{22\eta_y}{\mu_\lambda}(M_{y,k} + R_{y,k}^{\mathrm{m}}) \\ &+ \frac{27 L_{\lambda,*}^2 \eta_x^2}{\mu_\lambda \eta_y}(H_k + M_{x,k} + R_{x,k}^{\mathrm{m}}).\end{aligned} \tag{324}$$

This proves (210) with $c_{\mathrm{m},y} = 1/8$ and $C_{\mathrm{m},y} = \max\{22, 27L_{\lambda,*}^2\}$. The first perturbation is due to the y-block update, and the second is due to movement of its target under the x-update. □

E.4.7 Proof of Lemma 15: Original Lower-Level Tracking

Proof. Let $q_z^k := y^\star(x^k)$, $d_z^k := z^k - q_z^k$, and $g_z^k := \bar{h}_z^k$. The direction g_z^k is the gradient of a μ_g-strongly convex, $l_{g,1}$-smooth lower-level objective and vanishes at q_z^k. We first expand the update around this fixed target:

$$\begin{aligned}\mathbb{E}\|z^{k+1} - q_z^k\|^2 = J_k - 2\eta_z\mathbb{E}\langle d_z^k, g_z^k\rangle - 2\eta_z\mathbb{E}\langle d_z^k, e_z^k + s_z^k\rangle \\ + \eta_z^2\mathbb{E}\|g_z^k + e_z^k + s_z^k\|^2.\end{aligned} \tag{325}$$

Interpolation for a strongly convex, smooth function gives

$$\langle d_z^k, g_z^k\rangle \geq \frac{\mu_g l_{g,1}}{\mu_g + l_{g,1}}\|d_z^k\|^2 + \frac{1}{\mu_g + l_{g,1}}\|g_z^k\|^2. \tag{326}$$

Neither the momentum deviation nor the aggregation error is assumed to be centered. Young's inequality and $\|e+s\|^2 \leq 2\|e\|^2 + 2\|s\|^2$ give

$$-2\eta_z\mathbb{E}\langle d_z^k, e_z^k + s_z^k\rangle \leq \frac{\mu_g\eta_z}{4}J_k + \frac{8\eta_z}{\mu_g}(M_{z,k} + R_{z,k}^{\mathrm{m}}), \tag{327}$$

$$\mathbb{E}\|g_z^k + e_z^k + s_z^k\|^2 \leq 3\mathbb{E}\|g_z^k\|^2 + 3M_{z,k} + 3R_{z,k}^{\mathrm{m}}. \tag{328}$$

Substituting these bounds into (325) yields

$$\begin{aligned}\mathbb{E}\|z^{k+1} - q_z^k\|^2 \leq \left(1 - \frac{2\mu_g l_{g,1}\eta_z}{\mu_g + l_{g,1}} + \frac{\mu_g\eta_z}{4}\right) J_k \\ + \left(-\frac{2\eta_z}{\mu_g + l_{g,1}} + 3\eta_z^2\right)\mathbb{E}\|g_z^k\|^2 + \left(\frac{8\eta_z}{\mu_g} + 3\eta_z^2\right)(M_{z,k} + R_{z,k}^{\mathrm{m}}).\end{aligned} \tag{329}$$

Assume the stepsize bound in the lemma and write $a_z := \mu_g\eta_z$. Since $l_{g,1} \geq \mu_g$,

$$\begin{aligned}\frac{2\mu_g l_{g,1}}{\mu_g + l_{g,1}} \geq \mu_g, \qquad -\frac{2\eta_z}{\mu_g + l_{g,1}} + 3\eta_z^2 \leq -\frac{\eta_z}{\mu_g + l_{g,1}} \leq 0, \\ 0 < a_z \leq 1, \qquad \frac{8\eta_z}{\mu_g} + 3\eta_z^2 \leq \frac{11\eta_z}{\mu_g}.\end{aligned} \tag{330}$$

The distance coefficient is at most $1 - 3a_z/4$. Retaining the weaker coefficient $1 - a_z/4$ and discarding the nonpositive direction term gives

$$\mathbb{E}\|z^{k+1} - q_z^k\|^2 \leq \left(1 - \frac{a_z}{4}\right) J_k + \frac{11\eta_z}{\mu_g}(M_{z,k} + R_{z,k}^{\mathrm{m}}). \tag{331}$$

Target drift. To replace q_z^k by q_z^{k+1}, choose $\chi_z := a_z/8$. The optimizer-map bound in Lemma 2 gives

$$\mathbb{E}\|q_z^{k+1} - q_z^k\|^2 \leq L_*^2\eta_x^2\mathbb{E}\|\widehat{m}_x^k\|^2. \tag{332}$$

Expanding $z^{k+1} - q_z^{k+1} = (z^{k+1} - q_z^k) + (q_z^k - q_z^{k+1})$ and applying Young's inequality yields

$$\begin{aligned}J_{k+1} &\leq (1 + \chi_z)\mathbb{E}\|z^{k+1} - q_z^k\|^2 + (1 + \chi_z^{-1})\mathbb{E}\|q_z^{k+1} - q_z^k\|^2 \\ &\leq \left(1 + \frac{a_z}{8}\right)\mathbb{E}\|z^{k+1} - q_z^k\|^2 + \left(1 + \frac{8}{a_z}\right)L_*^2\eta_x^2\mathbb{E}\|\widehat{m}_x^k\|^2.\end{aligned} \tag{333}$$

Thus this drift estimate uses only Lipschitz continuity of the optimizer map and second moments of the server direction. Substituting (331) and (286) gives

$$\begin{aligned}J_{k+1} \leq \left(1 + \frac{a_z}{8}\right)\left(1 - \frac{a_z}{4}\right) J_k \\ + \left(1 + \frac{a_z}{8}\right)\frac{11\eta_z}{\mu_g}(M_{z,k} + R_{z,k}^{\mathrm{m}}) \\ + 3\left(1 + \frac{8}{a_z}\right)L_*^2\eta_x^2(H_k + M_{x,k} + R_{x,k}^{\mathrm{m}}).\end{aligned} \tag{334}$$

For $0 < a_z \le 1$, the three coefficients satisfy

$$\begin{aligned} \left(1+\frac{a_z}{8}\right)\left(1-\frac{a_z}{4}\right) = 1-\frac{a_z}{8}-\frac{a_z^2}{32} \le 1-\frac{a_z}{8}, \\ 1+\frac{a_z}{8} \le 2, \qquad 3\left(1+\frac{8}{a_z}\right) \le \frac{27}{a_z}. \end{aligned} \tag{335}$$

Consequently,

$$\begin{aligned} J_{k+1} \le \left(1-\frac{\mu_g \eta_z}{8}\right) J_k + \frac{22\eta_z}{\mu_g}(M_{z,k} + R^{\mathrm{m}}_{z,k}) \\ + \frac{27 L_*^2 \eta_x^2}{\mu_g \eta_z}(H_k + M_{x,k} + R^{\mathrm{m}}_{x,k}). \end{aligned} \tag{336}$$

This proves (211) with $c_{\mathrm{m},z} = 1/8$ and $C_{\mathrm{m},z} = \max\{22, 27L_*^2\}$. The first perturbation is due to the z-block update, and the second is due to movement of its target under the x-update. □