DECENTRALIZED ZEROTH-ORDER GRADIENT TRACKING FOR BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In many learning and decision-making problems, the quality of a decision depends on the outcome of an underlying optimization process. For example, selecting a model's regularization strength requires training the model under that choice and assessing its performance on validation data. Model training and hyperparameter selection therefore form two nested optimization problems. When organizations jointly train a shared model while keeping data local and operating without a central coordinator, both problems must be solved through neighbor communication, motivating decentralized bilevel optimization. Such coordination becomes more difficult when participating systems expose only loss values or have insufficient memory for differentiation. These constraints motivate a zeroth-order approach that uses function evaluations at both levels. The challenge extends beyond estimating local gradients: agents must recover how the jointly trained model responds to hyperparameter changes from noisy local evaluations, despite incomplete training and disagreement across the network. To recover this response without derivatives, we propose Decentralized Zeroth-Order Bilevel Optimization (DZBO), which optimizes both levels through stochastic function evaluations and communication between neighbors. For a possibly nonconvex outer objective and uniformly strongly convex local lower-level objectives, DZBO achieves an $\mathcal{O}(K^{-1/2})$ rate for the average expected squared gradient norm at network-averaged iterates after K outer updates. Attaining accuracy ϵ requires $\mathcal{O}(\epsilon^{-3})$ function-value queries per agent and communication rounds, including all inner and auxiliary work, for fixed problem and network parameters. Experiments illustrate the feasibility of our proposed algorithm under heterogeneous data.

1 INTRODUCTION

Bilevel optimization couples an upper-level decision with the solution of a lower-level problem. This structure underlies hyperparameter selection and data reweighting (Franceschi et al., 2018; Ren et al., 2018). When training and validation losses are distributed across agents, *decentralized bilevel optimization* allows them to collaborate through neighbor communication without a central coordinator (Yang et al., 2022; Chen et al., 2023). Consider, for example, a network of hospitals inspired by multicenter histopathology learning (Saldanha et al., 2022). The hospitals could jointly train a classifier on fixed image features while choosing a shared regularization strength to minimize their aggregate validation loss. Training defines the lower level; regularization selection defines the upper level. Changing the regularization strength changes the jointly trained model, so its validation performance depends on training losses throughout the network.

Optimizing this shared objective requires access to more than local loss gradients. The outer gradient, or *hypergradient*, must account for both the direct effect of the upper variable and its effect through the shared lower-level solution. With heterogeneous data, a locally trained model generally has a different response and cannot replace this shared solution. Common hypergradient methods differentiate through training trajectories or use implicit differentiation with Hessian and mixed-derivative products (Franceschi et al., 2018; Chen et al., 2023). The former may require storing or recomputing intermediate training states; the latter still requires local derivative products. Both can be difficult to support when agents have limited memory or use proprietary software and external simulators exposing only loss values. Even for smooth objectives, the derivative oracles assumed by these methods may therefore be unavailable in practice.

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.

Zeroth-order (ZO) optimization provides an alternative interface: it estimates update directions from perturbed function values (Nesterov & Spokoiny, 2017). This interface also offers benefits when derivatives are available but costly to compute. For neural objectives, forward loss evaluations avoid backpropagation and storing activations for the backward pass, potentially reducing local memory requirements. MeZO demonstrates this benefit by fine-tuning language models with memory use comparable to inference (Malladi et al., 2023). Function-value queries also allow agents with different software implementations to collaborate without implementing a common derivative interface. Moreover, evaluations at distinct perturbations can run in parallel when hardware permits, as illustrated by DeepZero (Chen et al., 2024). These properties motivate ZO methods for agents with restricted derivative access or limited memory. They do, however, entail estimation noise and dimension-dependent query costs: lower memory use does not by itself imply lower total computation.

Bringing this interface to decentralized bilevel optimization requires removing derivative access at *both* levels. A ZO outer update alone does not suffice if the lower-level solver uses gradients. Moreover, each agent updates using local losses and messages from neighbors. Centralized ZO bilevel methods assume access to aggregated information, while derivative-based decentralized bilevel methods retain local derivative oracles. Decentralized ZO methods for single-level problems do not capture the response of a shared lower-level solution. These approaches thus leave a need for a method that jointly supports function-value access, a shared lower-level problem, and neighbor-only communication, with guarantees accounting for work at both levels.

The difficulty is recovering *global bilevel sensitivity* from noisy evaluations at local states. First, it depends on the inverse of the aggregate lower-level Hessian. Averaging local inverse-Hessian estimates generally does not yield this inverse, even if each estimate is exact. Neighbor communication must help approximate the global sensitivity. Second, finite differences introduce smoothing bias, while shrinking perturbations can amplify uncancelled evaluation noise. Controlling bias and noise simultaneously requires estimators whose moment bounds remain stable as smoothing decreases. Third, agents evaluate estimators at approximate lower-level solutions and disagreeing local states. Consequently, an accurate estimator at a local state need not provide an accurate direction for the shared outer objective. Estimation error, lower-level optimization error, and network disagreement must be controlled together throughout the decentralized recursions. This coupling determines resource costs: each outer update requires lower-level and auxiliary work, whose queries and communication rounds must be included to turn an outer-iteration rate into a total complexity guarantee.

This leads to the question: *Can agents optimize a shared bilevel objective using only local function values and neighbor communication, with convergence, query, and communication guarantees?*

1.1 Main Contributions

We propose *Decentralized Zeroth-Order Bilevel Optimization* (DZBO), which combines decentralized lower-level optimization with sensitivity tracking using only local stochastic function evaluations and neighbor communication. This design captures the shared lower-level response under heterogeneous data. We consider a uniformly strongly convex local lower-level objectives under the stated smoothness and stochastic oracle assumptions. Our main contributions are as follows.

- **A fully zeroth-order method for decentralized bilevel optimization.** We develop an algorithm that jointly coordinates lower-level optimization, sensitivity estimation, and upper-level updates using only stochastic function values. Compared with centralized zeroth-order bilevel methods and derivative-based decentralized methods, DZBO requires neither centralized aggregation nor gradient or derivative-product oracles. It therefore supports collaborative bilevel learning when agents expose only loss evaluations and communicate locally.
- **Convergence guarantees through joint error control.** We control finite-difference noise amplification through sample reuse and derive second-moment bounds uniform in the smoothing radii. Jointly controlling smoothing bias, lower-level approximation, sensitivity-tracking error, and network disagreement yields an $\mathcal{O}(K^{-1/2})$ rate for the average expected squared gradient norm at network-averaged iterates after K outer updates. Reaching ϵ in this criterion requires $\mathcal{O}(\epsilon^{-3})$ function-value queries per agent and communication rounds.
- **Empirical validation on synthetic and real datasets.** Experiments on synthetic hyperparameter optimization and data hyper-cleaning on MNIST and CIFAR-100 demonstrate the feasibility of

shared bilevel learning using only function values and neighbor communication under heterogeneous data and label corruption. DZBO follows the overall optimization trend of derivative-based MA-DSBO on MNIST across the tested heterogeneity settings, while the synthetic and CIFAR-100 studies characterize the effects of dimension-dependent estimation noise.

2 Related Work

Zeroth-order bilevel optimization. PZOBO (Sow et al., 2022) estimates lower-level responses through perturbed trajectories but retains gradient computations. ZDSBA (Aghasi & Ghadimi, 2025) studies fully ZO stochastic bilevel optimization; later centralized work improves query complexity (Aghasi et al., 2025) or uses single-loop finite differences (Rando & Vaiter, 2026). FedRZO$_{\text{bl}}$ (Qiu et al., 2023) estimates a smoothed implicit objective with server aggregation and an inexact lower-level oracle that may use first-order updates. DZBO addresses the additional challenge of recovering a shared lower-level response using only neighbor communication and function values at both levels.

Decentralized bilevel optimization. Gossip methods (Yang et al., 2022) and DSBO (Chen et al., 2025) estimate global hypergradients from local information; MA-DSBO (Chen et al., 2023) reduces per-iteration costs through vector auxiliaries. DSGDA-GT (Wang et al., 2025) removes second-order oracles, while AdaSDBO (Zhai et al., 2025) adapts stepsizes without problem parameters; both retain derivatives. Resource-allocation methods (Ji & Ying, 2023) allow value-based utility feedback but retain lower-level derivatives. DiZS-GT^{2s} (Ebrahimi et al., 2025) uses local variational-inequality solution maps, rather than a shared global lower-level minimizer. DZBO combines fully ZO access with sensitivity tracking to control estimation and coordination errors under heterogeneity.

Table 1 summarizes representative outer-iteration rates and shows that DZBO achieves an $\mathcal{O}(K^{-1/2})$ rate using only function-value queries at both levels and neighbor-to-neighbor communication. DSBO, ZDSBA, and DZBO use expected squared gradient norms; DSGDA-GT uses the minimum expected unsquared norm; FedRZO$_{\text{bl}}$ uses the squared gradient norm of a smoothed objective; and DiZS-GT^{2s} uses squared distance to an approximate Clarke subdifferential. Thus, DZBO retains the $\mathcal{O}(K^{-1/2})$ outer-iteration order under the combined constraints of function-value-only access and decentralized coordination. Beyond this iteration-level guarantee, our analysis also accounts for the function-query and communication costs incurred by the inner and auxiliary procedures.

Table 1: Outer-iteration rates.

Method	Rate in K
DSBO	$\mathcal{O}(K^{-1/2})$
DSGDA-GT	$\mathcal{O}((nK)^{-1/7})$
ZDSBA	$\mathcal{O}(K^{-1/2})$
FedRZO$_{\text{bl}}$	$\widetilde{\mathcal{O}}((nK)^{-1/2})$
DiZS-GT^{2s}	$\mathcal{O}(K^{-1/2})$
DZBO (ours)	$\boldsymbol{\mathcal{O}}(K^{-1/2})$

3 Problem Formulation and Preliminaries

Bilevel objective. For n agents, let the local stochastic upper- and lower-level objectives be $f_i(x,y) = \mathbb{E}_{\phi_i}[F_i(x,y;\phi_i)]$ and $g_i(x,y) = \mathbb{E}_{\xi_i}[G_i(x,y;\xi_i)]$, where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ denote the upper- and lower-level variables, respectively. The agents collaboratively solve $\min_{x\in\mathbb{R}^p} \Phi(x)$, where $\Phi(x) := n^{-1}\sum_{i=1}^n f_i\big(x, y^*(x)\big)$, $y^*(x) := \arg\min_{y\in\mathbb{R}^q} g(x,y)$, and $g(x,y) := n^{-1}\sum_{i=1}^n g_i(x,y)$. Local objectives may be statistically heterogeneous. Here x specifies a shared learning configuration and $y^*(x)$ is the model trained across agents; the upper objective evaluates that shared model rather than independently trained local models. When $g(x,\cdot)$ is strongly convex, $y^*(x)$ is unique. Under the smoothness conditions stated below, the implicit-function theorem gives $\nabla\Phi(x) = \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(x,y^*(x)) - \nabla^2_{xy} g(x,y^*(x))\big[\nabla^2_{yy} g(x,y^*(x))\big]^{-1}\frac{1}{n}\sum_{i=1}^n \nabla_y f_i(x,y^*(x))$. Here $\nabla^2_{xy} g \in \mathbb{R}^{p\times q}$. The first term captures the direct effect of x; the second accounts for how changing x alters the trained lower-level solution. Recovering this response is the central algorithmic challenge. We seek an ϵ-stationary solution measured by

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big\|\nabla\Phi(\bar{x}_k)\big\|^2 \le \epsilon,\ \bar{x}_k := \frac{1}{n}\sum_{i=1}^n x_{i,k}, \tag{1}$$

using only stochastic function evaluations and communication between neighboring agents. This criterion measures first-order stationarity of the original outer objective. The network average is an analysis quantity; agents need not form it during execution.

Zeroth-order information. At (x, y), agent i observes $F_i(x, y; \phi_i)$ and $G_i(x, y; \xi_i)$ but no derivatives. Each finite-difference stencil reuses one sample at its perturbed points, while samples and Gaussian directions are independent across agents and iterations. For a function $h : \mathbb{R}^d \to \mathbb{R}$ and smoothing radius $\mu > 0$, define its Gaussian smoothing by $h_\mu(z) := \mathbb{E}_u[h(z + \mu u)]$, with $u \sim \mathcal{N}(0, I_d)$. A one-sided zeroth-order gradient estimator is $\widehat{\nabla} h_\mu(z; u, \omega) := \frac{h(z+\mu u;\omega)-h(z;\omega)}{\mu} u$, which is unbiased for the smoothed gradient. The value difference measures change along u, providing a direction estimate from two evaluations; decreasing μ reduces smoothing bias, while sample reuse couples the evaluation noise. SZO-JHIP uses alternating-sign four-point stencils in two y-directions for Hessian terms and joint x–y perturbations for cross-Jacobian terms. Exact estimator formulas, moment bounds, and their proofs are given in Appendix A.2.

Network and standing assumptions. Throughout, $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices, while $\|\cdot\|_F$ denotes the Frobenius norm. The agents communicate over an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Let $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ be its mixing matrix, where $w_{ij} > 0$ only if $i = j$ or $(i, j) \in \mathcal{E}$. We denote $\Pi := n^{-1}\mathbf{1}\mathbf{1}^\top$, $\Pi_\perp := I_n - \Pi$, and $\rho := \|W - \Pi\|_2$.

Assumption 3.1 (Network connectivity)**.** The graph $\mathcal{G}$ is connected. The matrix W is nonnegative and doubly stochastic, i.e., $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top W = \mathbf{1}^\top$. Moreover, $\rho < 1$. The projector Π extracts the network average, while $\Pi_\perp$ isolates disagreement. The factor ρ bounds its contraction under mixing, so the gap $1 - \rho$ quantifies information propagation using only neighbors.

Assumption 3.2 (Lower-level geometry)**.** For every agent i and every x, $g_i(x, \cdot)$ is λ_g-strongly convex, and its lower gradient $\nabla_y g_i$ is L_g-Lipschitz in the joint variables (x, y). Thus $y^*(x)$ is unique and $\|[\nabla^2_{yy} g(x, y)]^{-1}\| \le 1/\lambda_g$; see Boyd & Vandenberghe (2004, Section 9.1.2).

Assumption 3.3 (Deterministic smoothness and derivative bounds)**.** For every agent i, f_i and g_i are twice continuously differentiable. The maps ∇f_i, $\nabla^2_{yy} g_i$, and $\nabla^2_{xy} g_i$ are respectively L_f-, L_H-, and L_J-Lipschitz continuous in the joint variable $z = (x, y)$. Moreover, $\|\nabla f_i(z)\| \le B_f$ and $\|\nabla^2_{xy} g_i(z)\|_F \le B_J$. All five constants are finite, positive, and uniform over i and z.

Assumption 3.4 (Stochastic oracle regularity)**.** For each agent i, the samples have query-independent distributions. The sample functions $F_i(\cdot; \phi_i)$ and $G_i(\cdot; \xi_i)$ are almost surely C^2, with integrable values and respective mean functions f_i and g_i. The second moments of ∇F_i, $\nabla^2_{yy} G_i$, and $\nabla^2_{xy} G_i$ are uniformly bounded by M_F^2, M_H^2, and M_J^2, respectively, over all agents and query points, with finite constants $M_F, M_H, M_J, M_{g,*} > 0$. For every $\mathbf{X} = (x_1, \ldots, x_n)^\top$, let $y^*_{\mathbf{X}} := \arg\min_y n^{-1} \sum_i g_i(x_i, y)$ and require

$$\max_{1\le i\le n} \mathbb{E}_{\xi_i} \|\nabla_y G_i(x_i, y^*_{\mathbf{X}}; \xi_i)\|^2 \le M^2_{g,*}. \tag{2}$$

Samples and Gaussian directions are jointly independent of the preceding history and independent across agents. Within each stencil, its sample and directions are mutually independent, with the sample reused at all query points; directions are also independent across stencils.

Assumption 3.5 (Lower boundedness)**.** The outer objective admits a finite lower bound: there exists a constant $\Phi_{\inf} \in \mathbb{R}$ such that $\Phi(x) \ge \Phi_{\inf}$ for every $x \in \mathbb{R}^p$.

Strong convexity makes the shared response unique and its inverse-curvature term well defined; smoothness and moment bounds control approximation bias and stochastic error. The derivative conditions describe the objectives, not derivative access required by the algorithm. These conditions imply the oracle identities and mean-square smoothness in Lemma A.1, as well as the Lipschitz continuity of y^* and $\nabla\Phi$ in Lemma A.2. Appendix A.1 gives their statements, proofs, and explicit constants.

4 Decentralized Zeroth-Order Bilevel Optimization

Define $\Phi_i(x) := f_i(x, y^*(x))$, so that $\Phi(x) = n^{-1} \sum_{i=1}^n \Phi_i(x)$. At a common point $(x, y^*(x))$, the local contribution to the global hypergradient is

$$\nabla\Phi_i(x) = \nabla_x f_i\big(x, y^*(x)\big) - \nabla^2_{xy} g\big(x, y^*(x)\big) \left[\nabla^2_{yy} g\big(x, y^*(x)\big)\right]^{-1} \nabla_y f_i\big(x, y^*(x)\big). \tag{3}$$

Every local contribution shares the same global inverse product; local counterparts remain biased under heterogeneity because inversion and averaging do not commute. For any common query

point (x, y), define $H(x,y) := n^{-1}\sum_i \nabla^2_{yy} g_i(x,y)$, $J(x,y) := n^{-1}\sum_i \nabla^2_{xy} g_i(x,y)$, $b(x,y) := n^{-1}\sum_i \nabla_y f_i(x,y)$, and $v^*(x,y) := H(x,y)^{-1} b(x,y)$. Equivalently, the global hypergradient averages $\nabla_x f_i - \nabla^2_{xy} g_i\, v^*$. These local vector contributions need not individually equal (3), but their averages agree. The matrix Jacobian–Hessian inverse product (JHIP) is JH^{-1}: H^{-1} accounts for lower-level curvature, and J transfers the response to the upper variables. Solving $Hv = b$ obtains the needed action without explicit inversion. DZBO combines three stages: a decentralized lower-level solve, a zeroth-order sensitivity estimate, and an outer consensus update. We first describe these components; Section 5 gives their convergence and resource guarantees.

At outer iteration k, the fixed local outer variables are $\mathbf{X}_k := (x_{1,k}, \ldots, x_{n,k})^\top$. Define $g_{\mathbf{X}_k}(y) := n^{-1}\sum_{i=1}^n g_i(x_{i,k}, y)$ and $y^*_k := \arg\min_y g_{\mathbf{X}_k}(y)$. Because the local $x_{i,k}$ may disagree, y^*_k need not equal $y^*(\bar{x}_k)$; the analysis must account for this mismatch. Each agent estimates y^*_k using stochastic function-value queries. At each inner step, agents mix their local iterates through W and take a local zeroth-order descent step. This decentralized lower-level solver provides an approximate solution for the shared lower-level problem required by the subsequent sensitivity estimation.

4.1 Zeroth-Order Sensitivity Tracking

There are three compatible JHIP paths. The matrix path estimates $Z^* = H^{-1}J^\top$; the vector path solves $Hv = b$ and applies Jv; the direct vector path in Appendix B replaces sampled matrix actions by direct product estimates. Algorithm 1 implements the matrix and vector paths through a decentralized zeroth-order sensitivity oracle, using residual tracking to recover the aggregate linear-system solution from function-value queries. The vector formulation stores and communicates q auxiliary entries instead of pq, while retaining the outer convergence order. All use the same DZBO recursion. In Algorithm 1, Γ_i is a sampled linear-system residual and D_i tracks its network average. The initialization and doubly stochastic mixing preserve $\sum_i D_i^{(t)} = \sum_i \Gamma_i^{(t)}$, allowing updates to incorporate network-wide residual information. For compactness, denote four-point estimators by

$$\widehat{H}_i^{(t)} := \widehat{\nabla}^2_{yy,\mu_N} G_i(x_i, y_i; \xi_{i,H}^{(t)}), \qquad \widehat{J}_i^{(t)} := \widehat{\nabla}^2_{xy,\eta_N,\mu_N} G_i(x_i, y_i; \xi_{i,J}^{(t)}). \tag{4}$$

Their explicit constructions are given in (12). Alternating signs cancel nuisance terms and yield radius-uniform second moments (Lemma A.3). During decentralized execution, local states may differ. For fixed $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, $\theta_i = (x_i, y_i)$, use the aggregate targets

$$H_{\boldsymbol{\theta}} := \frac{1}{n}\sum_i \nabla^2_{yy} g_i(x_i, y_i), \qquad b_{\boldsymbol{\theta}} := \frac{1}{n}\sum_i \nabla_y f_i(x_i, y_i), \qquad v^*_{\boldsymbol{\theta}} := H_{\boldsymbol{\theta}}^{-1} b_{\boldsymbol{\theta}}. \tag{5}$$

These targets reflect the states actually queried by the agents; they separate auxiliary-solver error from the additional error due to disagreement and incomplete lower-level training. At every auxiliary

Algorithm 1 SZO-JHIP Oracle (Matrix and Vector Modes)

Input: fixed $\theta_i = (x_i, y_i)$, W, γ_N, μ_N, $N \geq 1$, mode $m \in \{\text{full}, \text{vector}\}$; η_N for full mode and ν_N for vector mode.
Use $U_i = Z_i \in \mathbb{R}^{q\times p}$ in full mode and $U_i = v_i \in \mathbb{R}^q$ in vector mode; set $U_i^{(0)} = 0$.
if $m =$ full **then**
  Draw fresh $\widehat{H}_i^{(0)}, \widehat{J}_i^{(0)}$ via (4); set $\Gamma_i^{(0)} = -(\widehat{J}_i^{(0)})^\top$.
else
  Draw $\widehat{b}_i^{(0)}$ via (6); set $\Gamma_i^{(0)} = -\widehat{b}_i^{(0)}$.
end if
Set $D_i^{(0)} = \Gamma_i^{(0)}$.
for $t = 0, \ldots, N-1$ **do**
  $U_i^{(t+1)} \leftarrow \sum_j w_{ij} U_j^{(t)} - \gamma_N D_i^{(t)}$.
  if $t < N-1$ **then**
    With fresh independent samples/Gaussian directions for each stencil, draw $\widehat{H}_i^{(t+1)}$ via (4) and the mode-specific right-hand side: $\widehat{B}_i^{(t+1)} = (\widehat{J}_i^{(t+1)})^\top$ via (4) (full), or $\widehat{B}_i^{(t+1)} = \widehat{b}_i^{(t+1)}$ via (6) (vector).
    $\Gamma_i^{(t+1)} \leftarrow \widehat{H}_i^{(t+1)} U_i^{(t+1)} - \widehat{B}_i^{(t+1)}$; $\quad D_i^{(t+1)} \leftarrow \sum_j w_{ij} D_j^{(t)} + \Gamma_i^{(t+1)} - \Gamma_i^{(t)}$.
  end if
end for
Output: $N^{-1}\sum_{t=1}^N U_i^{(t)}$, denoted $\widetilde{Z}_i^{(N)}$ (full) or $\widetilde{v}_i^{(N)}$ (vector).

residual, refresh the right-hand side using

$$\widehat{b}_i^{(t)} = \frac{F_i(x_i, y_i + \nu_N z_i^{(t)}; \phi_i^{(t)}) - F_i(x_i, y_i; \phi_i^{(t)})}{\nu_N} z_i^{(t)}, \qquad z_i^{(t)} \sim \mathcal{N}(0, I_q). \tag{6}$$

Refreshing $\widehat{b}_i^{(t)}$ prevents a single noisy upper-gradient sample from fixing the auxiliary target; time averaging then supports the true-target error bound in Section 5.1.

4.2 Outer Updates and the Complete Algorithm

With the lower-level solver and sensitivity oracle established, we now combine these components into the complete DZBO procedure. Algorithm 2 alternates decentralized lower-level optimization, zeroth-order sensitivity estimation, and consensus-based outer updates to jointly optimize the bilevel objective. At each outer iteration, agents approximately solve the lower-level problem, use one of the three JHIP formulations to estimate the implicit hypergradient term, and combine it with a fresh zeroth-order x-gradient for a consensus-based outer update. The direction $h_{i,k}$ combines the direct and implicit effects in the hypergradient identity, while mixing keeps agents' configurations aligned. Both effects are estimated through function values.

5 Convergence Analysis

Under the assumptions in Section 3, we bound lower-level and sensitivity errors, then combine them with network disagreement to establish outer convergence. Complete proofs appear in Appendix A.

5.1 Approximation Error Bounds

Theorem 5.1 (Convergence Rate of the Inner Loop)**.** *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. For outer iteration k, let $B_k := n^{-1}\sum_{i=1}^n \mathbb{E}\|y_{i,k}^{(0)} - y_k^*\|^2$. With $\beta_t = \Theta((t + A_2)^{-1})$ and $\mu_2 = \Theta(T^{-1/2})$ satisfying the conditions in Appendix A.4, there is a constant C_{in}, independent of T and k, such that*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|y_{i,k}^{(T)} - y_k^*\|^2 \le \frac{C_{\text{in}}(1 + B_k)}{T}.$$

The factor $1 + B_k$ captures initialization error. Reusing the preceding lower-level iterate avoids restarting training; warm-start stability below keeps this factor bounded under the stated schedules. Diminishing steps and $T^{-1/2}$ smoothing then give uniform $\mathcal{O}(T^{-1})$ accuracy throughout the outer iterations.

Algorithm 2 Decentralized Zeroth-Order Bilevel Optimization (DZBO)

Input: K, W, T, N, steps $\{\alpha_k\}, \{\beta_t\}, \gamma_N$, radii $\nu_x, \mu_2, \mu_N, \eta_N, \nu_N, \delta_N$, and $m \in \{\text{full}, \text{vector}, \text{direct}\}$.
Initialize: $x_{i,0} = x_0$, $y_{i,0}^{(0)}$, and a shared $R \sim \text{Unif}\{0, \dots, K-1\}$ independent of initialization and all oracle randomness.
for $k = 0, \dots, K-1$ **do**
  Store $x_{i,k}$ if $k = R$; set $y_{i,k}^{(0)} = y_{i,k-1}^{(T)}$ for $k > 0$.
  for $t = 0, 1, \dots, T-1$; agents $i = 1, \dots, n$ **in parallel do**
    Draw $v_i^{(t)} \sim \mathcal{N}(0, I_q)$ and $\xi_{i,k}^{(t)}$; form $\widehat{d}_{i,k}^{(t)}$ using the two-point estimator for $y \mapsto G_i(x_{i,k}, y; \xi_{i,k}^{(t)})$ with radius μ_2.
    $y_{i,k}^{(t+1)} = \sum_{j=1}^n w_{ij} y_{j,k}^{(t)} - \beta_t \widehat{d}_{i,k}^{(t)}$.
  end for
  Set $\theta_{i,k} = (x_{i,k}, y_{i,k}^{(T)})$.
  if $m =$ full **then**
    Run Algorithm 1 (full); with a fresh y-gradient stencil of radius ν_N, set $c_{i,k} = [\widetilde{Z}_{i,k}^{(N)}]^\top \widehat{d}_{y,i,k}$.
  else if $m =$ vector **then**
    Run Algorithm 1 (vector); with a fresh cross stencil, set $c_{i,k} = \widehat{J}_{i,k} \widetilde{v}_{i,k}^{(N)}$.
  else
    Run the N-step direct path in Appendix B; with a fresh stencil, set $c_{i,k} = \mathcal{J}_{i,\eta_N,\delta_N}(\widetilde{v}_{i,k}^{(N)})$.
  end if
  At $\theta_{i,k}$, form a fresh x-gradient $\widehat{d}_{x,i,k}$ of F_i using the two-point estimator with radius ν_x; set $h_{i,k} = \widehat{d}_{x,i,k} - c_{i,k}$; $\quad x_{i,k+1} = \sum_j w_{ij} x_{j,k} - \alpha_k h_{i,k}$.
end for
Output: $\{x_{i,R}\}_{i=1}^n$.

Theorem 5.1 controls the error introduced by the inexact lower-level solution. Even with fixed local states, the zeroth-order sensitivity oracle introduces an additional approximation error when estimating the implicit hypergradient term. We next characterize this error for the auxiliary sensitivity tracking procedure. We next characterize the convergence of the auxiliary sensitivity tracking procedure, whose outputs approximate the corresponding aggregate sensitivity targets.

Theorem 5.2 ($\mathcal{O}(1/N)$ Mean-Square Convergence of the Vector Oracle)**.** *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold, and fix admissible local states $\boldsymbol{\theta}$. For prescribed N, use*

$$\gamma_N = c_\gamma N^{-1/2}, \qquad \mu_N = c_\mu N^{-1/2}, \qquad \nu_N = c_\nu N^{-1/2}, \tag{7}$$

with fixed positive constants. There exist finite N_{J} and C_{v}, independent of N and the admissible local states, such that for $N \geq N_{\mathrm{J}}$,

$$\frac{1}{n}\sum_i \mathbb{E}\|\widetilde{v}_i^{(N)} - v_{\boldsymbol{\theta}}^*\|^2 \leq \frac{C_{\mathrm{v}}}{N}. \tag{8}$$

The constants may depend on dimensions, geometry, oracle moments, and the graph. The lower-gradient moment conditions needed only for the inner solver are unnecessary for this auxiliary result.

The finite constants N_{J} and C_{v} are independent of N and uniform over all admissible fixed local states. The same argument also applies to the matrix formulation, which directly tracks the JHIP matrix. We next establish its mean-square convergence. For fixed local states, define $J_{\boldsymbol{\theta}} := n^{-1}\sum_i \nabla^2_{xy} g_i(x_i, y_i)$ and $Z^*_{\boldsymbol{\theta}} := H^{-1}_{\boldsymbol{\theta}} J^\top_{\boldsymbol{\theta}} \in \mathbb{R}^{q\times p}$.

Theorem 5.3 (Matrix JHIP mean-square convergence)**.** *Under Assumptions 3.1–3.4, for $\gamma_N = c_\gamma N^{-1/2}$, $\mu_N = c_\mu N^{-1/2}$, and $\eta_N = c_\eta N^{-1/2}$, the matrix recursion in Algorithm 1 satisfies,*

$$\text{for } N \geq N_{\mathrm{Z}}, \ \ \frac{1}{n}\sum_i \mathbb{E}\|\widetilde{Z}_i^{(N)} - Z^*_{\boldsymbol{\theta}}\|_F^2 \leq \frac{C_Z}{N}. \tag{9}$$

The finite constants N_{Z}, C_Z are independent of N and uniform over all admissible fixed local states.

Both bounds concern time-averaged outputs and unsmoothed aggregate targets. They justify replacing the matrix auxiliary state by a vector without changing the error order; the dimension-dependent constants can still differ. Appendix A.5 proves them through a common biased linear-residual argument. The vector path uses B_f, L_f, M_F because its right-hand side is the upper-level y-gradient.

5.2 Outer Convergence and Resource Complexity

Uniform warm-start stability yields the following stationarity and end-to-end resource guarantees.

Lemma 5.4 (Warm-start stability)**.** *Suppose Algorithm 2 uses $m \in \{\textit{full}, \textit{vector}\}$, the corresponding inner and auxiliary schedules above, $\alpha_k = \alpha \leq 1$, and $N \geq \max\{N_{\mathrm{J}}, N_{\mathrm{Z}}\}$. If $B_0 < \infty$, then for all sufficiently large T,*

$$\sup_{0\leq k\leq K-1} B_k \leq C_{\mathrm{ws}},$$

where C_{ws} is independent of K, T, and N; the explicit threshold is given in the proof.

Theorem 5.5 (Convergence Rate of DZBO)**.** *Suppose Assumptions 3.1, 3.2, 3.3, 3.4, and 3.5 hold. Suppose B_0, as defined in Theorem 5.1, is finite, and use the warm start in Algorithm 2 with $m \in \{\textit{full}, \textit{vector}\}$. With $\alpha = \Theta(K^{-1/2})$, $T, N = \Theta(K^{1/2})$, and $\nu_x = \Theta(K^{-1/4})$, $\eta_N = \Theta(N^{-1/2})$, with the auxiliary schedules above and the detailed conditions in Appendix A.8,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla\Phi(\bar{x}_k)\|^2 \leq \frac{C_{\mathrm{DZBO}}}{\sqrt{K}}.$$

The choices $T, N = \Theta(K^{1/2})$ and $\nu_x = \Theta(K^{-1/4})$ balance lower-level error, auxiliary error, and squared smoothing bias at the outer $K^{-1/2}$ scale. The shared random index R converts the iteration-average bound into a guarantee for a stored output. The alignment bounds in Appendices A.6 and A.7 connect both auxiliary formulations to outer descent. Corollary B.3 gives the direct-mode guarantee.

Corollary 5.6 (End-to-end resource complexity)**.** *Under Theorem 5.5 (Corollary B.3 for direct mode), let R be as in Algorithm 2. For any $0 < \epsilon \leq 1$, suitable choices $K = \Theta(\epsilon^{-2})$ and $T, N = \Theta(\epsilon^{-1})$, with the prescribed parameter schedules, ensure $\mathbb{E}\|\nabla\Phi(\bar{x}_R)\|^2 \leq \epsilon$. All three modes use $\mathcal{O}(\epsilon^{-3})$ function-value queries per agent and $\mathcal{O}(\epsilon^{-3})$ synchronous communication rounds; the total number of queries is $\mathcal{O}(n\epsilon^{-3})$. These orders hold for fixed problem and network parameters.*

The ϵ^{-3} order includes the inner and auxiliary work within each outer iteration, making the cost of value-only coordination explicit. Exact query counts, communication payloads, and storage and arithmetic costs are given in Appendix A for each of the three implementations.

6 EXPERIMENTS

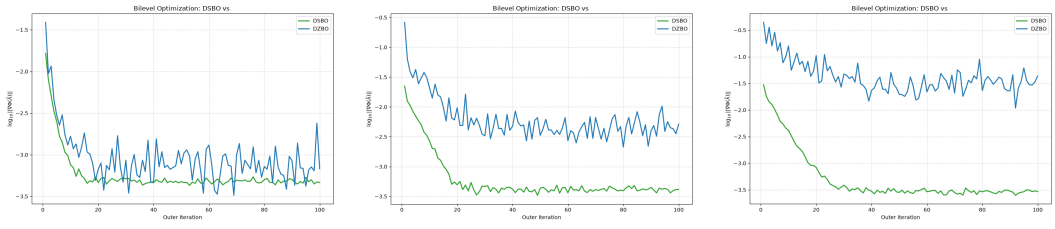
We evaluate DZBO through three complementary studies: synthetic hyperparameter optimization to examine dimension sensitivity, data hyper-cleaning to assess optimization under label corruption and heterogeneous data, and a network-size study to examine behavior as the number of agents increases. The main text presents the synthetic and representative MNIST results; additional corruption levels, CIFAR-100 experiments, and network-size comparisons appear in Appendix C.

Baseline and communication. We compare DZBO with stochastic MA-DSBO (Chen et al., 2023), a decentralized bilevel method that uses derivatives. This comparison tests whether value-only coordination retains useful optimization behavior and exposes the cost of estimating derivative information. All experiments use a cycle with self-weight 0.4 and weight 0.3 on each adjacent edge.

Evaluation metrics. The synthetic study reports the base-10 logarithm of the estimated outer-gradient norm against outer iterations; its formula appears in Appendix C.1. For data hyper-cleaning, we evaluate corrupted-training loss, clean-test loss, and test accuracy using the network-averaged classifier. These metrics are reported against optimization time, with equal time budgets for the two methods and evaluation time excluded. The iteration-based and time-based comparisons respectively characterize optimization progress and practical performance under the chosen implementations.

6.1 SYNTHETIC HYPERPARAMETER OPTIMIZATION

Task and setup. We study heterogeneous logistic regression with coordinate-wise quadratic regularization. The upper problem selects regularization strengths using validation loss, while the lower problem trains the shared classifier (Franceschi et al., 2018). This tests whether function-value queries can recover how regularization affects the trained model. We vary $p = q \in \{20, 50, 100\}$ to examine dimension sensitivity and generate binary labels from a linear score with Gaussian noise of scale 0.1. The objective functions and label-generation rule are specified in Appendix C.1.



(a) $p = q = 20$. (b) $p = q = 50$. (c) $p = q = 100$.

Figure 1: Synthetic problem: stationarity versus dimension.

Results and interpretation. Figure 1 shows that DZBO exhibits greater fluctuations as dimension increases, while MA-DSBO is less affected. This observation is consistent with the dimension dependence of Gaussian zeroth-order estimation (Nesterov & Spokoiny, 2017) and illustrates the practical cost of approximating derivative information from function values. The curves complement the convergence analysis by showing how dimension influences finite-iteration behavior.

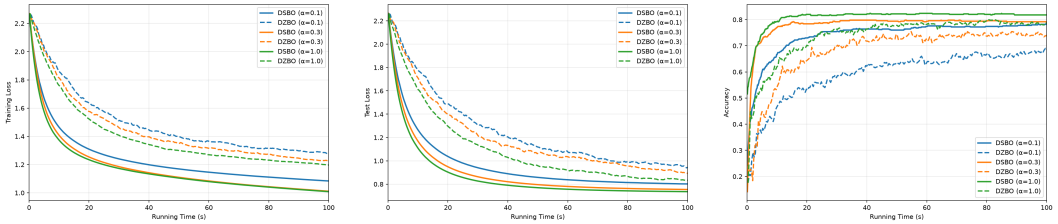
6.2 DATA HYPER-CLEANING

Task and setup. We study data hyper-cleaning on MNIST (LeCun et al., 1998) and CIFAR-100 (Krizhevsky, 2009). The upper problem learns globally aligned example weights from clean validation data; the lower problem trains a regularized softmax classifier on weighted, potentially corrupted training data. Each agent evaluates its own shard and maintains a local classifier. Sigmoid weights let validation feedback adjust each example's influence, while positive quadratic regularization makes the classifier problem strongly convex. This instantiates the shared training response that DZBO estimates without derivative oracles; the exact objectives appear in Appendix C.1. Data are partitioned by clean class before corruption using a Dirichlet distribution with parameter α_{Dir}; larger values yield more homogeneous client datasets (Hsu et al., 2019).

Models and implementation. MNIST uses flattened images as classifier inputs. CIFAR-100 uses fixed features extracted by a pretrained ResNet-18, with only the softmax classifier optimized

at the lower level; details are provided in Appendix C.3. The corresponding matrix and vector implementations are described in Appendix C.1.

MNIST protocol. We consider corruption rates $\chi_{\text{corr}} \in \{0.1, 0.2, 0.4\}$ and report the representative 0.2 setting in the main text. Agents share a zero initialization and, at each outer step, update local classifiers, estimate hypergradients, and mix the globally aligned example weights. One seed controls subsampling, corruption, and partitioning; the curves are single runs with online sampling. These trajectories illustrate individual runs, rather than variability across independently repeated random seeds. Clipping is used only as a logged numerical safeguard. The evaluation uses the time-based metrics and equal budgets described in Section 6.



(a) Train ($\chi_{\text{corr}} = 0.2$) (b) Test ($\chi_{\text{corr}} = 0.2$) (c) Accuracy ($\chi_{\text{corr}} = 0.2$)

Figure 2: MNIST hyper-cleaning ($\chi_{\text{corr}} = 0.2$): MA-DSBO versus DZBO.

Results and interpretation. Figure 2 compares training loss, test loss, and test accuracy across tested heterogeneity settings. Larger α_{Dir} yields better final metrics. MA-DSBO produces smoother trajectories, while DZBO follows the same overall optimization trend. These observations support the feasibility of learning validation-guided weights using only function values and neighbor communication. The additional fluctuations also expose the estimation cost of this interface and the effect of heterogeneity on stability. The comparisons describe behavior under the reported experimental settings; the stationarity and resource guarantees are established separately by theoretical analysis.

Additional experiments. Appendix C.2 reports MNIST results at $\chi_{\text{corr}} \in \{0.1, 0.4\}$ to complement the corruption setting. Appendix C.3 examines CIFAR-100 with fixed image features, where the gap between MA-DSBO and DZBO becomes more pronounced. Appendix C.4 varies $n \in \{16, 32, 64\}$ while fixing corruption and the Dirichlet parameter to examine sensitivity to network size.

7 Conclusion

We developed DZBO, a fully zeroth-order method for decentralized bilevel optimization using stochastic function evaluations and neighbor communication at both levels. By combining decentralized lower-level updates with auxiliary sensitivity tracking, DZBO approximates the response of the shared lower-level solution without derivative oracles or a central coordinator. Its matrix and vector formulations achieve the same outer convergence order, while vector states reduce auxiliary storage and communication payloads. These features support collaborative model selection when agents expose loss evaluations and communicate locally. Our analysis explains how sample reuse within finite-difference stencils controls estimation noise as the smoothing radii decrease. Under the stated smoothness and oracle assumptions, with uniformly strongly convex local lower-level objectives, we establish $\mathcal{O}(N^{-1})$ mean-square error for time-averaged auxiliary outputs at fixed local query states. Jointly controlling this error, lower-level inexactness, and network disagreement yields an $\mathcal{O}(K^{-1/2})$ rate for the average expected squared gradient norm at network-averaged outer iterates. For fixed problem and network parameters, attaining accuracy ϵ in this criterion requires $\mathcal{O}(\epsilon^{-3})$ function-value queries per agent and communication rounds, including all inner and auxiliary iterations.

Experiments on synthetic hyperparameter tuning and MNIST and CIFAR-100 data hyper-cleaning illustrate the feasibility of function-value-only coordination under heterogeneous data. The observed optimization trends also highlight the practical effects of dimension-dependent estimation noise and data heterogeneity across agents. These findings motivate reducing query variance and improving communication efficiency as complementary directions for further development in decentralized learning settings. Extending the guarantees beyond strongly convex lower-level objectives and to oracles with independently noisy evaluations would broaden the method's scope; Appendix D discusses these limitations and additional network extensions.

AI USE STATEMENT

Generative AI tools assisted with proof checking, manuscript revision, and writing and debugging experimental code. The authors remain responsible for independently verifying AI-assisted revisions and for the final proofs, code, and experimental results.

REPRODUCIBILITY STATEMENT

The problem setting and assumptions are specified in Section 3, and Algorithms 1 and 2 describe the update rules. Parameter schedules and convergence guarantees appear in Section 5, with detailed proofs in Appendix A and direct-product extensions in Appendix B. Section 6 and Appendix C describe the experimental settings and supplementary evaluations.

REFERENCES

Alireza Aghasi and Saeed Ghadimi. Fully zeroth-order bilevel programming via Gaussian smoothing. *Journal of Optimization Theory and Applications*, 205(2), 2025. doi: 10.1007/s10957-025-02647-y. URL https://link.springer.com/article/10.1007/s10957-025-02647-y. Article 31.

Alireza Aghasi, Jeongyeol Kwon, and Saeed Ghadimi. Optimal zeroth-order bilevel optimization. *arXiv preprint arXiv:2510.03646*, 2025. URL https://arxiv.org/abs/2510.03646.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441. URL https://doi.org/10.1017/CBO9780511804441.

Aochuan Chen, Yimeng Zhang, Jinghan Jia, James Diffenderfer, Konstantinos Parasyris, Jiancheng Liu, Yihua Zhang, Zheng Zhang, Bhavya Kailkhura, and Sijia Liu. DeepZero: Scaling up zeroth-order optimization for deep model training. In *International Conference on Learning Representations*, pp. 50185–50206, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d9dd691fa4eeb4741f16ac42ed5d197e-Abstract-Conference.html.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishna Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4641–4671. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23n.html.

Xuxing Chen, Minhui Huang, and Shiqian Ma. Decentralized bilevel optimization. *Optimization Letters*, 19(7):1249–1313, 2025. doi: 10.1007/s11590-024-02101-4. URL https://doi.org/10.1007/s11590-024-02101-4.

Mohammadjavad Ebrahimi, Uday V. Shanbhag, and Farzad Yousefian. On the resolution of stochastic MPECs over networks: Distributed implicit zeroth-order gradient tracking methods. *arXiv preprint arXiv:2505.22916*, 2025. URL https://arxiv.org/abs/2505.22916v2. Version 2, revised May 2026.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018. URL https://arxiv.org/abs/1802.02246.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. URL https://arxiv.org/abs/1909.06335.

Kaiyi Ji and Lei Ying. Network utility maximization with unknown utility functions: A distributed, data-driven bilevel optimization approach. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 131–140, 2023. doi: 10.1145/3565287.3610260. URL https://doi.org/10.1145/3565287.3610260.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4882–4892. PMLR, 2021. URL https://proceedings.mlr.press/v139/ji21c.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53038–53075, 2023. doi: 10.52202/075280-2308. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a627810151be4d13f907ac898ff7e948-Abstract-Conference.html.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2. URL https://doi.org/10.1007/s10208-015-9296-2.

Yuyang Qiu, Uday V. Shanbhag, and Farzad Yousefian. Zeroth-order methods for nondifferentiable, nonconvex, and hierarchical federated optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3425–3438. Curran Associates, Inc., 2023. doi: 10.52202/075280-0151. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0a70c9cd8179fe6f8f6135fafa2a8798-Paper-Conference.pdf.

Marco Rando and Samuel Vaiter. ZOBA: An efficient single-loop zeroth-order bilevel optimization algorithm. *arXiv preprint arXiv:2601.21836*, 2026. URL https://arxiv.org/abs/2601.21836.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Oliver Lester Saldanha, Philip Quirke, Nicholas P. West, Jacqueline A. James, Maurice B. Loughrey, Heike I. Grabsch, Manuel Salto-Tellez, Elizabeth Alwers, Didem Cifci, Narmin Ghaffari Laleh, Tobias Seibel, Richard Gray, Gordon G. A. Hutchins, Hermann Brenner, Marko van Treeck, Tanwei Yuan, Titus J. Brinker, Jenny Chang-Claude, Firas Khader, Andreas Schuppert, Tom Luedde, Christian Trautwein, Hannah Sophie Muti, Sebastian Foersch, Michael Hoffmeister, Daniel Truhn, and Jakob Nikolas Kather. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nature Medicine*, 28(6):1232–1239, 2022. doi: 10.1038/s41591-022-01768-5. URL https://www.nature.com/articles/s41591-022-01768-5.

Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for Hessian-free bilevel algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4136–4149. Curran Associates, Inc., 2022. doi: 10.52202/068431-0299. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/1a82986c9f321217f2ed407a14dcfa0b-Paper-Conference.pdf.

Xiaoyu Wang, Xuxing Chen, Shiqian Ma, and Tong Zhang. Fully first-order methods for decentralized bilevel optimization. *IEEE Transactions on Signal Processing*, 73:4734–4747, 2025. doi: 10.1109/TSP.2025.3624427. URL https://doi.org/10.1109/TSP.2025.3624427.

Shuoguang Yang, Xuezhou Zhang, and Mengdi Wang. Decentralized gossip-based stochastic bilevel optimization over communication networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 238–252. Curran Associates, Inc., 2022. doi: 10.52202/068431-0018. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/01db36a646c07c64dd39a92b4eceb417-Paper-Conference.pdf.

Zhiwei Zhai, Wenjing Yan, and Ying-Jun Zhang. Problem-parameter-free decentralized bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 144986–145053. Curran Associates, Inc., 2025. doi: 10.52202/085713-4853. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d5a8e37f38a08c68162452dcba89ae9c-Paper-Conference.pdf.

A Technical Lemmas and Convergence Proofs

This appendix separates the proof into four layers. First, we derive the oracle and outer-objective regularity consequences and establish the second-moment and bias properties of the four-point estimators. Second, we collect deterministic recursion and target-perturbation lemmas. Third, we prove the inner-loop and SZO-JHIP rates by decomposing each error into a network average and a disagreement component. Finally, the warm-start and alignment lemmas connect the two subroutines to the outer descent argument, which proves Theorem 5.5. Each main theorem proof begins with a short description of its error decomposition and the role of the preceding lemmas.

Assumptions used by each result. The vector auxiliary theorem uses network contraction, the lower Hessian bounds, and the upper-gradient bounds B_f, L_f, M_F already present in the full assumptions. The direct extension uses the same deterministic second-order Lipschitz bounds and stochastic second moments. Uniform iterate moments are proved, not imposed as assumptions.

A.1 Regularity and Smoothness Lemmas

Lemma A.1 (Oracle consequences)**.** *Under Assumptions 3.2 and 3.4, differentiation and expectation commute for ∇F_i, $\nabla_y G_i$, $\nabla^2_{yy} G_i$, and $\nabla^2_{xy} G_i$. The sample lower gradients $\nabla_y G_i$ are L_G-Lipschitz continuous in mean square in the joint variable $z = (x, y)$, uniformly in i, where $L_G := \sqrt{M_H^2 + M_J^2}$; the input and output norms are Euclidean.*

Lemma A.2 (Outer smoothness)**.** *Under Assumptions 3.2 and 3.3, y^* and $\nabla\Phi$ are Lipschitz continuous on $\mathbb{R}^p$ with constants L_g/λ_g and L_Φ, respectively, in the Euclidean norm, for a finite $L_\Phi > 0$.*

The outer smoothness result follows the standard analysis of Ghadimi & Wang (2018); Ji et al. (2021).

Proof of Lemma A.1. The derivative identities claimed in the lemma are

$$\begin{aligned} \mathbb{E}_{\phi_i}[\nabla F_i(z;\phi_i)] &= \nabla f_i(z), & \mathbb{E}_{\xi_i}[\nabla_y G_i(z;\xi_i)] &= \nabla_y g_i(z), \\ \mathbb{E}_{\xi_i}[\nabla^2_{yy} G_i(z;\xi_i)] &= \nabla^2_{yy} g_i(z), & \mathbb{E}_{\xi_i}[\nabla^2_{xy} G_i(z;\xi_i)] &= \nabla^2_{xy} g_i(z). \end{aligned} \tag{10}$$

Fix an agent i. Write $\Delta x = x - x'$, $\Delta y = y - y'$, and $z_t = z' + t(z - z')$. Almost-sure sample smoothness gives

$$\begin{aligned} &\nabla_y G_i(z;\xi_i) - \nabla_y G_i(z';\xi_i) \\ &\quad = \int_0^1 \left[\nabla^2_{xy} G_i(z_t;\xi_i)^\top \Delta x + \nabla^2_{yy} G_i(z_t;\xi_i)\Delta y\right] dt. \end{aligned}$$

By Jensen's inequality, Cauchy–Schwarz, and the second-order moment bounds in Assumption 3.4,

$$\begin{aligned} &\mathbb{E}_{\xi_i}\|\nabla_y G_i(z;\xi_i) - \nabla_y G_i(z';\xi_i)\|^2 \\ &\quad \le \int_0^1 \mathbb{E}_{\xi_i}\left[\|\nabla^2_{xy} G_i(z_t;\xi_i)\|^2 + \|\nabla^2_{yy} G_i(z_t;\xi_i)\|^2\right] dt\, \|z - z'\|^2 \\ &\quad \le (M_J^2 + M_H^2)\|z - z'\|^2. \end{aligned}$$

This proves the mean-square Lipschitz claim without using derivative unbiasedness.

To justify differentiation under the expectation, choose the deterministic instance $\mathbf{X}^0 = (0, \dots, 0)^\top$ and let $z_0 = (0, y^*_{\mathbf{X}^0})$. Assumption 3.2 ensures that this solution exists and is unique. The solution moment bound (2) and the inequality just proved imply

$$\mathbb{E}_{\xi_i}\|\nabla_y G_i(z;\xi_i)\|^2 \le 2M^2_{g,*} + 2L_G^2\|z - z_0\|^2.$$

Hence the lower gradients have uniformly bounded second moments on every bounded neighborhood of a fixed query point.

For a joint coordinate vector e_j and $h \neq 0$, define

$$Q_{F,h} := \frac{F_i(z + he_j;\phi_i) - F_i(z;\phi_i)}{h} = \int_0^1 \partial_j F_i(z + the_j;\phi_i)\, dt.$$

The upper-gradient moment bound in Assumption 3.4 gives $\sup_{h\neq 0}\mathbb{E}|Q_{F,h}|^2 \le M_F^2$. These difference quotients are therefore uniformly integrable and converge almost surely to $\partial_j F_i(z;\phi_i)$ as $h \to 0$. Their expectations converge to the expectation of this derivative. Function-value unbiasedness identifies those expectations with the difference quotients of f_i, giving $\nabla f_i = \mathbb{E}_{\phi_i}\nabla F_i$. The same argument for G_i in each y-coordinate uses the local second-moment bound above and gives $\nabla_y g_i = \mathbb{E}_{\xi_i}\nabla_y G_i$.

Finally, for every joint coordinate vector e_j,

$$\mathbb{E}_{\xi_i}\left\|\frac{\nabla_y G_i(z+he_j;\xi_i)-\nabla_y G_i(z;\xi_i)}{h}\right\|^2 \le L_G^2.$$

These vector difference quotients are uniformly integrable and converge almost surely to the corresponding derivative of $\nabla_y G_i$. Passing to expectations and differentiating the first-order identity yields the yy- and xy-block identities in (10). □

Proof of Lemma A.2. Following the standard argument of Ghadimi & Wang (2018, Lemma 2.2(c)) and Ji et al. (2021, Supplementary Material, Lemma 2), we show that a valid constant in our notation is

$$L_\Phi := \sqrt{1+\left(\frac{L_g}{\lambda_g}\right)^2}\left[L_f\left(1+\frac{B_J}{\lambda_g}\right)+\frac{B_f L_J}{\lambda_g}+\frac{B_J B_f L_H}{\lambda_g^2}\right]. \tag{11}$$

Let $f(z) := n^{-1}\sum_i f_i(z)$. Averaging preserves the derivative bounds and Lipschitz constants in Assumptions 3.2 and 3.3; Frobenius-norm bounds also bound spectral norms. Strong monotonicity and the optimality conditions for $y^*(x)$ and $y^*(x')$ give

$$\begin{aligned}\lambda_g\|y^*(x)-y^*(x')\| &\le \|\nabla_y g(x,y^*(x'))-\nabla_y g(x',y^*(x'))\|\\ &\le L_g\|x-x'\|.\end{aligned}$$

Consequently, for $z=(x,y^*(x))$ and $z'=(x',y^*(x'))$, $\|z-z'\| \le \sqrt{1+(L_g/\lambda_g)^2}\,\|x-x'\|$.

Write $H(z) := \nabla^2_{yy} g(z)$ and $J(z) := \nabla^2_{xy} g(z)$. Since $H(z) \succeq \lambda_g I_q$, the inverse identity gives

$$\begin{aligned}\|H(z)^{-1}-H(z')^{-1}\| &= \|H(z)^{-1}[H(z')-H(z)]H(z')^{-1}\|\\ &\le \frac{L_H}{\lambda_g^2}\|z-z'\|.\end{aligned}$$

For $v(z) := H(z)^{-1}\nabla_y f(z)$, it follows that

$$\begin{aligned}\|v(z)\| &\le \frac{B_f}{\lambda_g},\\ \|v(z)-v(z')\| &\le \left(\frac{L_f}{\lambda_g}+\frac{B_f L_H}{\lambda_g^2}\right)\|z-z'\|.\end{aligned}$$

Using the hypergradient identity in Section 3,

$$\begin{aligned}\|\nabla\Phi(x)-\nabla\Phi(x')\| &\le \|\nabla_x f(z)-\nabla_x f(z')\| + \|J(z)-J(z')\|\,\|v(z)\|\\ &\quad + \|J(z')\|\,\|v(z)-v(z')\|\\ &\le \left[L_f+\frac{B_f L_J + B_J L_f}{\lambda_g}+\frac{B_J B_f L_H}{\lambda_g^2}\right]\|z-z'\|.\end{aligned}$$

Combining the last bound with the bound for $\|z-z'\|$ proves Lemma A.2 with (11). □

A.2 Zeroth-Order Second-Order Estimators

For completeness, we give the explicit four-query constructions underlying the compact notation in (4). Let $\mathcal{S}=\{-1,1\}$, and let $u \sim \mathcal{N}(0,I_p)$ and $v_1,v_2,v_3 \sim \mathcal{N}(0,I_q)$ be independent. For a fixed

state (x_i, y_i) and stochastic samples $\xi_{i,H}$ and $\xi_{i,J}$, define

$$\widehat{\nabla}^2_{yy,\mu} G_i(x_i, y_i; \xi_{i,H}) := \frac{1}{8\mu^2} \left[\sum_{a,b\in\mathcal{S}} ab\, G_i\big(x_i, y_i + \mu(av_1 + bv_2); \xi_{i,H}\big) \right] (v_1 v_2^\top + v_2 v_1^\top),$$

$$\widehat{\nabla}^2_{xy,\eta,\mu} G_i(x_i, y_i; \xi_{i,J}) := \frac{1}{4\eta\mu} \left[\sum_{a,b\in\mathcal{S}} ab\, G_i\big(x_i + a\eta u, y_i + b\mu v_3; \xi_{i,J}\big) \right] u v_3^\top. \tag{12}$$

The same sample $\xi_{i,H}$ or $\xi_{i,J}$ is reused at all four query points of the corresponding estimator.

Lemma A.3 (Second moments of the four-point estimators)**.** *Under Assumption 3.4, the four-point estimators satisfy, uniformly in $\mu, \eta > 0$,*

$$\mathbb{E}\|\widehat{H}_i\|_F^2 \le M_H^2 \big[q(q+2)\big]^2 =: V_H(q), \quad \mathbb{E}\|\widehat{J}_i\|_F^2 \le M_J^2 p(p+2)q(q+2) =: V_J(p,q). \tag{13}$$

Proof. The integral forms of the estimators in (12) are

$$\widehat{H}_i = \frac{1}{8\mu^2} \int_{-\mu}^{\mu}\int_{-\mu}^{\mu} v_1^\top \nabla^2_{yy} G_i(x_i, y_i + sv_1 + rv_2; \xi) v_2 \, dr\, ds\, (v_1 v_2^\top + v_2 v_1^\top),$$

$$\widehat{J}_i = \frac{1}{4\eta\mu} \int_{-\eta}^{\eta}\int_{-\mu}^{\mu} u^\top \nabla^2_{xy} G_i(x_i + su, y_i + rv; \xi) v \, dr\, ds\, uv^\top.$$

Since $\|v_1 v_2^\top + v_2 v_1^\top\|_F \le 2\|v_1\|\|v_2\|$, conditional Jensen's inequality and the sample second-order moment bounds imply

$$\mathbb{E}_\xi[\|\widehat{H}_i\|_F^2 \mid v_1, v_2] \le M_H^2 \|v_1\|^4 \|v_2\|^4, \qquad \mathbb{E}_\xi[\|\widehat{J}_i\|_F^2 \mid u, v] \le M_J^2 \|u\|^4 \|v\|^4.$$

Finally, for $a \sim \mathcal{N}(0, I_d)$, $\mathbb{E}\|a\|^4 = d(d+2)$. Independence of the directions proves (13). □

Lemma A.4 (Expectation and smoothing bias of the four-point estimators)**.** *Under Assumptions 3.2, 3.3, and 3.4, fix (x_i, y_i) and define*

$$H^{\mathrm{sm}}_{i,\mu} := \mathbb{E}_{v_1,v_2} \nabla^2_{yy} g_i(x_i, y_i + \mu v_1 + \mu v_2),$$
$$J^{\mathrm{sm}}_{i,\eta,\mu} := \mathbb{E}_{u,v} \nabla^2_{xy} g_i(x_i + \eta u, y_i + \mu v),$$

where $u \sim \mathcal{N}(0, I_p)$ and $v, v_1, v_2 \sim \mathcal{N}(0, I_q)$ are independent. For the four-point estimators in the vector branch of Algorithm 1,

$$\mathbb{E}[\widehat{H}_i] = H^{\mathrm{sm}}_{i,\mu}, \tag{14}$$

$$\mathbb{E}[(\widehat{J}_i)^\top] = (J^{\mathrm{sm}}_{i,\eta,\mu})^\top. \tag{15}$$

Moreover, with $H_i = \nabla^2_{yy} g_i(x_i, y_i)$ and $J_i = \nabla^2_{xy} g_i(x_i, y_i)$,

$$\|H^{\mathrm{sm}}_{i,\mu} - H_i\|_F^2 \le 2L_H^2 q\mu^2, \qquad \|J^{\mathrm{sm}}_{i,\eta,\mu} - J_i\|_F^2 \le L_J^2(p\eta^2 + q\mu^2). \tag{16}$$

With the Gaussian smoothing defined in Section 3, $H^{\mathrm{sm}}_{i,\mu}$ is the Hessian of the y-smoothed objective with radius $\sqrt{2}\mu$.

Proof. Lemma A.3 gives finite second moments for both complete stencils. We may therefore condition on the Gaussian directions and first average each complete stencil over its stochastic sample, replacing G_i by g_i at the query points. All remaining expectations in this proof are over Gaussian directions.

At fixed x_i, lower-gradient Lipschitz continuity gives at most quadratic growth in y for $g_i(x_i, y)$, so the following Gaussian moments are integrable. Sign symmetry gives

$$\mathbb{E}[\widehat{H}_i] = \frac{1}{2\mu^2} \mathbb{E}\left[g_i(x_i, y_i + \mu v_1 + \mu v_2)(v_1 v_2^\top + v_2 v_1^\top) \right].$$

Applying Stein's identity successively to v_1 and v_2 gives

$$\mathbb{E}\left[g_i(x_i, y_i + \mu v_1 + \mu v_2) v_1 v_2^\top \right] = \mu^2 \mathbb{E} \nabla^2_{yy} g_i(x_i, y_i + \mu v_1 + \mu v_2).$$

The term with $v_2 v_1^\top$ gives the transpose of the same Hessian, which is equal to it. This proves (14).

For the cross estimator, set $r_i(u,v) := g_i(x_i+\eta u, y_i+\mu v) - g_i(x_i+\eta u, y_i)$. The subtracted term cancels in the signed stencil. Moreover, the fundamental theorem of calculus and lower-gradient Lipschitz continuity yield

$$|r_i(u,v)| \le \mu\|v\| \left(\|\nabla_y g_i(x_i,y_i)\| + L_g\eta\|u\| + L_g\mu\|v\|\right).$$

Thus $r_i(u,v)vu^\top$ is integrable, and sign symmetry can be applied to this centered difference to obtain

$$\mathbb{E}[(\widehat{J}_i)^\top] = \frac{1}{\eta\mu}\mathbb{E}\left[r_i(u,v)vu^\top\right].$$

Stein's identity first in v and then in u gives

$$\begin{aligned}\mathbb{E}\left[r_i(u,v)vu^\top\right] &= \mu\,\mathbb{E}\left[\nabla_y g_i(x_i+\eta u, y_i+\mu v)u^\top\right] \\ &= \eta\mu\,\mathbb{E}\nabla^2_{yx} g_i(x_i+\eta u, y_i+\mu v),\end{aligned}$$

where lower-gradient Lipschitz continuity and the cross-derivative bound justify the integrability required by these identities. This proves (15). As a normalization check, if $g(x,y) = x^\top A y$, then the signed four-point numerator equals $4\eta\mu\, u^\top A v$, so $\mathbb{E}[\widehat{J}_i] = A$ and $\mathbb{E}[(\widehat{J}_i)^\top] = A^\top$.

Finally, Lipschitz continuity of the two second-order derivative blocks and Jensen's inequality give

$$\begin{aligned}\|H^{\mathrm{sm}}_{i,\mu} - H_i\|_F^2 &\le L_H^2\mu^2\mathbb{E}\|v_1+v_2\|^2 = 2L_H^2 q\mu^2, \\ \|J^{\mathrm{sm}}_{i,\eta,\mu} - J_i\|_F^2 &\le L_J^2\left(\eta^2\mathbb{E}\|u\|^2 + \mu^2\mathbb{E}\|v\|^2\right) = L_J^2(p\eta^2 + q\mu^2),\end{aligned}$$

which proves (16). No classical third derivative is needed; Lipschitz continuity of the second-order blocks is sufficient. □

Exact factorization and arithmetic cost. Suppress the agent and iteration indices, and let a_H and a_J denote the scalar stencil prefactors in (12). Then $\widehat{H} = a_H(v_1 v_2^\top + v_2 v_1^\top)$ and $\widehat{J} = a_J u v_3^\top$. For every $Z \in \mathbb{R}^{q\times p}$,

$$\widehat{H}Z - \widehat{J}^\top = a_H\left[v_1(Z^\top v_2)^\top + v_2(Z^\top v_1)^\top\right] - a_J v_3 u^\top. \tag{17}$$

The identity follows from associativity of matrix multiplication. It uses two matrix–vector products and three rank-one updates, costing $\mathcal{O}(pq)$ arithmetic operations and requiring no $q\times q$ Hessian storage. With the same samples and directions, it produces exactly the same $\Gamma_i^{(t)}$, and hence the same $Z_i^{(t)}$, $D_i^{(t)}$, and averaged output, as the matrix recursion in Algorithm 1. In comparison, direct dense evaluation of the sampled Hessian–matrix product in the DSBO JHIP recursion of Chen et al. (2025) costs $\mathcal{O}(pq^2)$ operations. The improvement comes from the rank-two structure of the zeroth-order Hessian estimate; merely splitting an arbitrary dense product into columns does not reduce its arithmetic order. This comparison excludes oracle-evaluation and neighbor-aggregation costs and does not assert a speedup over every structured or Hessian-vector-product implementation of DSBO. The matrix recursion maintains full Z_i, D_i states and requires $\mathcal{O}(pq)$ auxiliary storage and communication.

Lemma A.5 (Vector actions and the refreshed right-hand side)**.** *Under Assumptions 3.2, 3.3, and 3.4, conditionally on a supplied history-measurable v independent of the current stencil,*

$$\begin{aligned}\mathbb{E}\|\widehat{H}_i v\|^2 &\le V_{Hv}(q)\|v\|^2, & V_{Hv}(q) &:= M_H^2 q(q+2)^2, \\ \mathbb{E}\|\widehat{J}_i v\|^2 &\le V_{Jv}(p,q)\|v\|^2, & V_{Jv}(p,q) &:= M_J^2 p(p+2)(q+2), \\ \|\mathbb{E}\widehat{H}_i v - H_i v\|^2 &\le 2L_H^2 q\mu_N^2\|v\|^2, & & \\ \|\mathbb{E}\widehat{J}_i v - J_i v\|^2 &\le L_J^2(p\eta_N^2 + q\mu_N^2)\|v\|^2. & & \end{aligned} \tag{18}$$

Also, (6) satisfies

$$\mathbb{E}\|\widehat{b}_i\|^2 \le V_b(q) := M_F^2 q(q+2), \qquad \|\mathbb{E}\widehat{b}_i - b_i\|^2 \le L_f^2 q\nu_N^2. \tag{19}$$

These are second-moment bounds, not sharp variance identities.

Proof. Let a_H, a_J denote the scalar coefficients of the four-point matrices. Their integral representations, conditional on the directions, give

$$\mathbb{E}_\xi a_H^2 \le \tfrac{1}{4} M_H^2 \|v_1\|^2 \|v_2\|^2, \qquad \mathbb{E}_\xi a_J^2 \le M_J^2 \|u\|^2 \|v_3\|^2.$$

Use the Gaussian identities $\mathbb{E}\|r\|^4 = d(d+2)$ and $\mathbb{E}[\|r\|^2 (r^\top v)^2] = (d+2)\|v\|^2$, with $r \sim \mathcal{N}(0, I_d)$. Independence of the directions yields

$$\mathbb{E}\|\widehat{H}_i v\|^2 \le \frac{M_H^2}{2} \mathbb{E}\big[\|v_1\|^4 \|v_2\|^2 (v_2^\top v)^2 + \|v_2\|^4 \|v_1\|^2 (v_1^\top v)^2\big] = V_{Hv}(q)\|v\|^2,$$
$$\mathbb{E}\|\widehat{J}_i v\|^2 \le M_J^2 \mathbb{E}\|u\|^4 \mathbb{E}[\|v_3\|^2 (v_3^\top v)^2] = V_{Jv}(p,q)\|v\|^2.$$

The action biases follow by multiplying (16) by v.

For either block of F_i, the integral form of the two-point estimator in Section 3 and the uniform sample-gradient moment bound give, in block dimension d,

$$\mathbb{E}\|\widehat{\nabla} F_{i,\tau}\|^2 \le M_F^2 d(d+2), \qquad \|\mathbb{E}\widehat{\nabla} F_{i,\tau} - \nabla f_i\| \le L_f \tau \sqrt{d}. \tag{20}$$

For the first bound, condition on the direction, apply Jensen to the integral, and use its fourth moment. For the second, Gaussian integration by parts gives the smoothed gradient, whose bias is bounded by the L_f-Lipschitz property. Taking $d = q$ and $\tau = \nu_N$ proves (19). Freshness makes all these bounds valid conditionally on the complete state history. □

A.3 Auxiliary Bounds and Target Perturbations

The following recursion and lower-solution bounds support the inner and auxiliary analyses. The vector target perturbation is used for the vector branch of Algorithm 1; matrix alignment is established in Appendix A.7.

Lemma A.6 (Polynomial and geometric decay recursions)**.** *Let $\{a_t\}$ and $\{c_t\}$ be nonnegative sequences, with $b > 0$ and $B \ge 0$.*

1. *If $a_{t+1} \le (1 - a/(t+b))a_t + B/(t+b)^2$, $a > 1$, and $b \ge a$, then $a_t \le C_a/(t+b)$ for some $C_a > 0$.*
2. *If $c_{t+1} \le r c_t + B/(t+b)^2$, $0 < r < 1$, and $r < (b/(b+1))^2$, then $c_t \le C_c/(t+b)^2$ for some $C_c > 0$.*

Proof. For the first claim, induction with $C_a \ge \max\{b a_0, B/(a-1)\}$ gives $a_t \le C_a/(t+b)$. For the second, the assumed strict inequality makes the denominator below positive. Choose $C_c \ge \max\{b^2 c_0, B/((b/(b+1))^2 - r)\}$; the same induction gives $c_t \le C_c/(t+b)^2$. □

Lemma A.7 (Coupled polynomial–geometric recursion)**.** *Let $\{e_t\}$ and $\{c_t\}$ be nonnegative sequences satisfying*

$$e_{t+1} \le \left(1 - \frac{a}{t+b}\right) e_t + \frac{A}{t+b} c_t + \frac{B}{(t+b)^2},$$
$$c_{t+1} \le \left(r + \frac{D}{(t+b)^2}\right) c_t + \frac{E}{(t+b)^2} e_t + \frac{F}{(t+b)^2},$$

where $a > 1$, $0 < r < 1$, and $A, B, D, E, F \ge 0$. Let $U \ge 1$ and assume $e_0 + c_0 \le C_0 U$. For $b \ge a$, define $\delta_b := (b/(b+1))^2 - r - D/b^2$, $m_b := A/((a-1)b)$, and $n_b := E/(\delta_b b)$. If $\delta_b > 0$ and $m_b n_b < 1$, then there are constants C_e, C_c, independent of U and t, such that

$$e_t \le \frac{C_e U}{t+b}, \qquad c_t \le \frac{C_c U}{(t+b)^2}.$$

Proof. Define $\overline{C}_e := \max\{b C_0, B/(a-1)\}$ and $\overline{C}_c := \max\{b^2 C_0, F/\delta_b\}$, and choose $C_e := (\overline{C}_e + m_b \overline{C}_c)/(1 - m_b n_b)$ and $C_c := \overline{C}_c + n_b C_e$. These constants satisfy

$$(a-1) C_e \ge B + \frac{A}{b} C_c, \qquad \delta_b C_c \ge F + \frac{E}{b} C_e, \tag{21}$$

and the desired bounds hold at $t = 0$.

Suppose they hold at iteration t and write $s = t + b$. The first recursion and (21) give

$$\begin{aligned} e_{t+1} &\le \frac{C_e U}{s} - \frac{aC_e U}{s^2} + \frac{AC_c U}{s^3} + \frac{BU}{s^2} \\ &\le \frac{C_e U}{s} - \frac{C_e U}{s^2} \le \frac{C_e U}{s+1}. \end{aligned}$$

For the second recursion, since $s^2/(s+1)^2 \ge b^2/(b+1)^2$, we obtain

$$\begin{aligned} c_{t+1} &\le \frac{U}{s^2}\left(rC_c + \frac{DC_c}{s^2} + \frac{EC_e}{s} + F\right) \\ &\le \frac{C_c U}{(s+1)^2}, \end{aligned}$$

where the last step again uses (21). Induction proves the result. □

Lemma A.8 (Lower-solution perturbation)**.** *Under Assumption 3.2, for fixed* $\mathbf{X} = (x_1, \dots, x_n)^\top$*, define* $g_{\mathbf{X}}(y) := n^{-1}\sum_{i=1}^n g_i(x_i, y)$*,* $y^*_{\mathbf{X}} := \arg\min_y g_{\mathbf{X}}(y)$*, and* $\bar{x} := n^{-1}\sum_{i=1}^n x_i$*. The solution perturbation satisfies*

$$\|y^*_{\mathbf{X}} - y^*(\bar{x})\|^2 \le \frac{L_g^2}{\lambda_g^2}\frac{1}{n}\sum_{i=1}^n \|x_i - \bar{x}\|^2. \tag{22}$$

Moreover, for any $\mathbf{X}' = (x'_1, \dots, x'_n)^\top$*,*

$$\|y^*_{\mathbf{X}'} - y^*_{\mathbf{X}}\|^2 \le \frac{L_g^2}{\lambda_g^2}\frac{1}{n}\sum_{i=1}^n \|x'_i - x_i\|^2. \tag{23}$$

Proof. Strong monotonicity of $\nabla_y g_{\mathbf{X}}$ and the two optimality conditions give

$$\begin{aligned} \lambda_g \|y^*_{\mathbf{X}} - y^*(\bar{x})\| &\le \left\| \frac{1}{n}\sum_{i=1}^n [\nabla_y g_i(x_i, y^*(\bar{x})) - \nabla_y g_i(\bar{x}, y^*(\bar{x}))] \right\| \\ &\le L_g \frac{1}{n}\sum_{i=1}^n \|x_i - \bar{x}\|. \end{aligned}$$

Squaring and applying Jensen's inequality proves the claim.

For the second inequality, strong monotonicity of $\nabla_y g_{\mathbf{X}'}$, together with $\nabla_y g_{\mathbf{X}'}(y^*_{\mathbf{X}'}) = 0$ and $\nabla_y g_{\mathbf{X}}(y^*_{\mathbf{X}}) = 0$, gives

$$\begin{aligned} \lambda_g \|y^*_{\mathbf{X}'} - y^*_{\mathbf{X}}\| &\le \|\nabla_y g_{\mathbf{X}'}(y^*_{\mathbf{X}})\| \\ &= \left\| \frac{1}{n}\sum_{i=1}^n [\nabla_y g_i(x'_i, y^*_{\mathbf{X}}) - \nabla_y g_i(x_i, y^*_{\mathbf{X}})] \right\| \\ &\le \frac{L_g}{n}\sum_{i=1}^n \|x'_i - x_i\|. \end{aligned}$$

Squaring and applying Jensen's inequality proves (23). □

Lemma A.9 (Inverse-vector target perturbation)**.** *Under Assumptions 3.2 and 3.3, let* $\theta_i = (x_i, y_i)$*,* $y^\circ = y^*(\bar{x})$*, and* $v^\circ = v^*(\bar{x}, y^\circ)$*. Then*

$$\|v^*_{\boldsymbol{\theta}} - v^\circ\|^2 \le C_{\mathrm{pert}}\left(\frac{1}{n}\sum_i \|x_i - \bar{x}\|^2 + \frac{1}{n}\sum_i \|y_i - y^\circ\|^2\right), \tag{24}$$

where $C_{\mathrm{pert}} = (L_f + L_H R_v)^2/\lambda_g^2$*.*

Proof. Set $H^\circ = H(\bar{x}, y^\circ)$ and $b^\circ = b(\bar{x}, y^\circ)$. The exact identity

$$v^*_{\boldsymbol{\theta}} - v^\circ = H_{\boldsymbol{\theta}}^{-1}[(b_{\boldsymbol{\theta}} - b^\circ) - (H_{\boldsymbol{\theta}} - H^\circ)v^\circ]$$

uses two global averaged systems. Bound the inverse norm by $1/\lambda_g$, $\|v^\circ\|$ by R_v, and the two block differences by their L_f, L_H Lipschitz constants and Jensen's inequality over agents. This proves the stated bound without averaging local inverses. □

A.4 Proof of Theorem 5.1

Parameter conditions. Fix $A_1 > 2/\lambda_g$ and $\mu_0 > 0$, and set

$$\beta_t = \frac{A_1}{t + A_2}, \qquad \mu_2 = \frac{\mu_0}{\sqrt{T}}. \tag{25}$$

The offset must satisfy $A_2 \geq A_{2,\min}$, where the finite threshold $A_{2,\min}$ is specified in (31) below.

Proof of Theorem 5.1. Proof strategy. We compare the decentralized iterates with the minimizer of the Gaussian-smoothed aggregate lower objective. The total error is decomposed into the network-average optimization error, the disagreement among agents, and the smoothing bias between the smoothed and original minimizers. Conditional second-moment bounds yield a coupled polynomial–geometric recursion for the first two components; Lemma A.7 then gives their respective $\mathcal{O}(t^{-1})$ and $\mathcal{O}(t^{-2})$ rates.

Fix k, write $\mathbf{X} = \mathbf{X}_k$, and define $g_{i,\mu_2}(x, y) := \mathbb{E}_v[g_i(x, y+\mu_2 v)]$ and $g_{\mathbf{X},\mu_2}(y) := n^{-1}\sum_{i=1}^n \mathbb{E}_v[g_i(x_{i,k}, y+\mu_2 v)]$. Let $\widetilde{y}^*_{\mathbf{X}}$ be its unique minimizer. Gaussian smoothing preserves λ_g-strong convexity and L_g-smoothness. The smoothed gradient satisfies

$$\|\nabla_y g_{\mathbf{X},\mu_2}(y) - \nabla_y g_{\mathbf{X}}(y)\| \leq L_g \mu_2 \mathbb{E}\|v\| \leq L_g \mu_2 \sqrt{q}.$$

Strong monotonicity therefore gives

$$\|\widetilde{y}^*_{\mathbf{X}} - y^*_{\mathbf{X}}\| \leq \frac{L_g\sqrt{q}}{\lambda_g}\mu_2. \tag{26}$$

Let $\mathbf{Y}_t, \widehat{\mathbf{d}}_t \in \mathbb{R}^{n\times q}$ have rows $(y_{i,k}^{(t)})^\top$ and $(\widehat{d}_{i,k}^{(t)})^\top$, respectively. Define $\bar{y}_t := n^{-1}\sum_{i=1}^n y_{i,k}^{(t)}$, $e_{y,t} := \mathbb{E}\|\bar{y}_t - \widetilde{y}^*_{\mathbf{X}}\|^2$, and $c_{y,t} := n^{-1}\mathbb{E}\|\Pi_\perp \mathbf{Y}_t\|_F^2 = n^{-1}\sum_{i=1}^n \mathbb{E}\|y_{i,k}^{(t)} - \bar{y}_t\|^2$. Let $\mathcal{F}_t$ contain $\mathbf{X}_k$, the initialization, and all history before the samples at inner iteration t are drawn. Thus the target $\widetilde{y}^*_{\mathbf{X}}$ is $\mathcal{F}_t$-measurable; unconditioned expectations also average over the outer history. The common-random-number construction gives $\mathbb{E}[\widehat{d}_{i,k}^{(t)} \mid \mathcal{F}_t] = \nabla_y g_{i,\mu_2}(x_{i,k}, y_{i,k}^{(t)})$. The disagreement term

$$\delta_t := \frac{1}{n}\sum_{i=1}^n \Big[\nabla_y g_{i,\mu_2}(x_{i,k}, y_{i,k}^{(t)}) - \nabla_y g_{i,\mu_2}(x_{i,k}, \bar{y}_t)\Big]$$

satisfies $\mathbb{E}\|\delta_t\|^2 \leq L_g^2 c_{y,t}$.

The fundamental theorem of calculus gives, conditionally on $\mathcal{F}_t$,

$$\widehat{d}_{i,k}^{(t)} = \int_0^1 \Big\langle \nabla_y G_i(x_{i,k}, y_{i,k}^{(t)} + s\mu_2 v_i^{(t)}; \xi_{i,k}^{(t)}), v_i^{(t)} \Big\rangle v_i^{(t)}\, ds.$$

Using the mean-square Lipschitz bound in Lemma A.1, the solution moment bound (2), (26), and $\mathbb{E}\|v\|^4 = q(q+2)$, $\mathbb{E}\|v\|^6 = q(q+2)(q+4)$, we obtain constants $V_{y,0}(q), V_{y,1}(q) < \infty$ such that

$$\mathbb{E}[\|\widehat{d}_{i,k}^{(t)}\|^2 \mid \mathcal{F}_t] \leq V_{y,0}(q) + V_{y,1}(q)\|y_{i,k}^{(t)} - \widetilde{y}^*_{\mathbf{X}}\|^2. \tag{27}$$

Here $V_{y,1}(q)$ is proportional to $L_G^2 q(q+2)$, while $V_{y,0}(q)$ is bounded by a constant multiple of

$$M_{g,*}^2 q(q+2) + L_G^2\mu_2^2 q(q+2)(q+4) + \frac{L_G^2 L_g^2}{\lambda_g^2}\mu_2^2 q^2(q+2).$$

Thus these constants are uniform in k, t and in T whenever $\mu_2 \leq \mu_0$.

Averaging the update and using double stochasticity yields $\bar{y}_{t+1} = \bar{y}_t - \beta_t n^{-1}\sum_{i=1}^n \widehat{d}_{i,k}^{(t)}$. Let $a_t := \bar{y}_t - \widetilde{y}^*_{\mathbf{X}}$ and $\zeta_t := n^{-1}\sum_{i=1}^n(\widehat{d}_{i,k}^{(t)} - \mathbb{E}[\widehat{d}_{i,k}^{(t)} \mid \mathcal{F}_t])$. Conditionally on $\mathcal{F}_t$, the average direction is $\nabla g_{\mathbf{X},\mu_2}(\bar{y}_t) + \delta_t + \zeta_t$ and $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$. Hence

$$\mathbb{E}[\|a_{t+1}\|^2 \mid \mathcal{F}_t] = \|a_t - \beta_t(\nabla g_{\mathbf{X},\mu_2}(\bar{y}_t) + \delta_t)\|^2 + \beta_t^2\mathbb{E}[\|\zeta_t\|^2 \mid \mathcal{F}_t].$$

Use $\langle a_t, \nabla g_{\mathbf{X},\mu_2}(\bar{y}_t)\rangle \geq \lambda_g \|a_t\|^2$, $\|\nabla g_{\mathbf{X},\mu_2}(\bar{y}_t)\| \leq L_g \|a_t\|$, and

$$2|\langle a_t, \delta_t\rangle| \leq \frac{\lambda_g}{2}\|a_t\|^2 + \frac{2}{\lambda_g}\|\delta_t\|^2.$$

Conditional independence and (27) give, after taking total expectation,

$$\mathbb{E}\|\zeta_t\|^2 \leq \frac{1}{n}\left[V_{y,0}(q) + V_{y,1}(q)(e_{y,t} + c_{y,t})\right],$$

where we used $n^{-1}\sum_i \mathbb{E}\|y_{i,k}^{(t)} - \widetilde{y}_{\mathbf{X}}^*\|^2 = e_{y,t} + c_{y,t}$. Let $\bar{\beta}_{\text{in}} > 0$ be a constant, depending only on λ_g, L_g and $V_{y,1}(q)$, such that the quadratic term is at most $(\lambda_g/2)\beta_t e_{y,t}$ whenever $0 < \beta_t \leq \bar{\beta}_{\text{in}}$. Under this condition, the resulting recursion is

$$e_{y,t+1} \leq \left(1 - \frac{\lambda_g}{2}\beta_t\right) e_{y,t} + C_1\beta_t c_{y,t} + C_2 V_{y,0}(q)\beta_t^2. \tag{28}$$

For the disagreement,

$$\Pi_\perp \mathbf{Y}_{t+1} = (W - \Pi)\Pi_\perp \mathbf{Y}_t - \beta_t \Pi_\perp \widehat{\mathbf{d}}_t.$$

Fix any $r_\rho \in (\rho^2, 1)$. Choosing the Young parameter so that the network term contracts by r_ρ, and using (27), gives

$$c_{y,t+1} \leq \left(r_\rho + C_3\beta_t^2\right) c_{y,t} + C_3\beta_t^2 e_{y,t} + C_4\beta_t^2. \tag{29}$$

Set $\beta_t = A_1/(t + A_2)$. The variance decomposition and (26) give

$$\begin{aligned} e_{y,0} + c_{y,0} &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\|y_{i,k}^{(0)} - \widetilde{y}_{\mathbf{X}}^*\|^2 \\ &\leq 2B_k + 2\mathbb{E}\|\widetilde{y}_{\mathbf{X}}^* - y_{\mathbf{X}}^*\|^2 \leq C_0(1 + B_k), \end{aligned} \tag{30}$$

where C_0 is independent of k and T.

We next retain this initialization dependence in the coupled recursion. Let $\overline{V}_{y,0}(q)$ be the uniform upper bound on $V_{y,0}(q)$ established above. Apply Lemma A.7 to (28) and (29) with $U = 1 + B_k$, $b = A_2$, $a = \lambda_g A_1/2 > 1$, $r = r_\rho$, $A = C_1 A_1$, $B = C_2\overline{V}_{y,0}(q)A_1^2$, $D = E = C_3 A_1^2$, and $F = C_4 A_1^2$. For $b \geq 1$, define $\delta_b := (b/(b+1))^2 - r_\rho - D/b^2$, $m_b := A/((a-1)b)$, and $n_b := E/(\delta_b b)$. Choose a finite $A_{2,\min}$ satisfying

$$\begin{aligned} A_{2,\min} &\geq \max\left\{a, \frac{A_1}{\bar{\beta}_{\text{in}}}\right\}, \\ r_\rho + \frac{C_3 A_1^2}{A_{2,\min}^2} < 1, &\qquad \delta_{A_{2,\min}} > 0, \qquad m_{A_{2,\min}} n_{A_{2,\min}} < 1. \end{aligned} \tag{31}$$

Such a threshold exists because $r_\rho < 1$ and all inverse powers of $A_{2,\min}$ vanish as $A_{2,\min} \to \infty$. For every $A_2 \geq A_{2,\min}$, Lemma A.7 applies. Consequently, there are constants $\widehat{C}_e, \widehat{C}_c$, depending only on the problem and schedule constants, such that

$$e_{y,t} \leq \frac{\widehat{C}_e(1 + B_k)}{t + A_2}, \qquad c_{y,t} \leq \frac{\widehat{C}_c(1 + B_k)}{(t + A_2)^2}.$$

None of these constants depends on k, T, or B_k.

Combining the average and disagreement bounds with (26) gives

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|y_{i,k}^{(T)} - y_k^*\|^2 \leq 2e_{y,T} + 2c_{y,T} + \frac{2L_g^2 q}{\lambda_g^2}\mu_2^2.$$

Since $\mu_2 = \mu_0/\sqrt{T}$ and $1 + B_k \geq 1$, the right-hand side is at most $C_{\text{in}}(1 + B_k)/T$. □

A.5 Proofs of the Auxiliary Convergence Theorems

We prove a finite-dimensional Hilbert-space result so that vector and matrix states are covered without identifying their different targets. For matrix states, the inner product is the Frobenius inner product. The result also allows a nonlinear, state-dependent oracle bias.

Theorem A.10 (Biased stochastic linear residuals with gradient tracking)**.** *Let W satisfy Assumption 3.1. Let $\mathcal{F}_t$ contain the states and history immediately before residual t is sampled. Let A_i be deterministic self-adjoint operators on a finite-dimensional real Hilbert space $\mathcal{H}$, with $\|A_i\| \le L$, and suppose $A := n^{-1}\sum_i A_i \succeq \lambda_g I$ for $\lambda_g > 0$. Let $\|B_i\| \le B$, $B_{\mathrm{av}} = n^{-1}\sum_i B_i$, and $U^* = A^{-1}B_{\mathrm{av}}$. For a prescribed horizon N, a fresh oracle at $U_i^{(t)}$ returns*

$$\Gamma_i^{(t)} = A_i U_i^{(t)} - B_i + \mathfrak{b}_{i,N}(U_i^{(t)}) + \zeta_i^{(t)}. \tag{32}$$

Conditionally on $\mathcal{F}_t$, the noises have zero mean and are independent across agents. Suppose, uniformly in the local states, iterations, and horizons,

$$\begin{aligned} \|\mathfrak{b}_{i,N}(U)\|^2 &\le s_N^2(K_0 + K_1\|U\|^2), \\ \mathbb{E}[\|\zeta_i^{(t)}\|^2 \mid \mathcal{F}_t] &\le K_2 + K_3\|U_i^{(t)}\|^2, \qquad s_N^2 \le S/N. \end{aligned} \tag{33}$$

The constants K_j, S, L, B, λ_g are finite and independent of N. Starting with $U_i^{(0)} = 0$ and $D_i^{(0)} = \Gamma_i^{(0)}$, use

$$U_i^{(t+1)} = \sum_j w_{ij} U_j^{(t)} - \gamma_N D_i^{(t)}, \qquad D_i^{(t+1)} = \sum_j w_{ij} D_j^{(t)} + \Gamma_i^{(t+1)} - \Gamma_i^{(t)}. \tag{34}$$

For $\gamma_N = c_\gamma/\sqrt{N}$, there exist finite N_0, C such that, for $N \ge N_0$,

$$\sup_{t\ge 0} \frac{1}{n}\sum_i \mathbb{E}\|U_i^{(t)}\|^2 \le C, \qquad \frac{1}{n}\sum_i \mathbb{E}\left\| \frac{1}{N}\sum_{t=1}^{N} U_i^{(t)} - U^* \right\|^2 \le C/N. \tag{35}$$

The virtual extension to $t \ge N$ is used only to prove stability; the algorithm requires just the first N updates and N residuals.

Proof. We give the stability argument before averaging, so that bounded iterates are established rather than assumed. All constants below depend only on the constants in the theorem and the fixed graph. The proof is uniform in the choice of operators and right-hand sides satisfying these bounds.

Average and disagreement recursions. Write $\bar{U}_t = n^{-1}\sum_i U_i^{(t)}$ and similarly for other variables. Define $e_t = \mathbb{E}\|\bar{U}_t - U^*\|^2$, $c_t = n^{-1}\sum_i \mathbb{E}\|U_i^{(t)} - \bar{U}_t\|^2$, and $d_t = n^{-1}\sum_i \mathbb{E}\|D_i^{(t)} - \bar{D}_t\|^2$. The initialization and double stochasticity imply exactly $\bar{D}_t = \bar{\Gamma}_t$. Since $\|U^*\| \le R := B/\lambda_g$,

$$\frac{1}{n}\sum_i \mathbb{E}\|U_i^{(t)}\|^2 \le 2e_t + c_t + 2R^2. \tag{36}$$

Let $R_t = n^{-1}\sum_i A_i(U_i^{(t)} - \bar{U}_t)$, $\bar{\mathfrak{b}}_t = n^{-1}\sum_i \mathfrak{b}_{i,N}(U_i^{(t)})$, and $\bar{\zeta}_t = n^{-1}\sum_i \zeta_i^{(t)}$. Then

$$\bar{\Gamma}_t = A(\bar{U}_t - U^*) + R_t + \bar{\mathfrak{b}}_t + \bar{\zeta}_t. \tag{37}$$

Jensen's inequality and conditional independence give

$$\begin{aligned} \mathbb{E}\|R_t\|^2 &\le L^2 c_t, \\ \mathbb{E}\|\bar{\mathfrak{b}}_t\|^2 &\le C s_N^2(1 + e_t + c_t), \\ \mathbb{E}\|\bar{\zeta}_t\|^2 &\le C(1 + e_t + c_t)/n. \end{aligned} \tag{38}$$

The same assumptions bound the uncentered local residuals:

$$\frac{1}{n}\sum_i \mathbb{E}\|\Gamma_i^{(t)}\|^2 \le C(1 + e_t + c_t), \quad \text{provided } s_N^2 \le 1. \tag{39}$$

For $u_t = \bar{U}_t - U^*$, the conditional expansion before applying any inequality is

$$\mathbb{E}[\|u_{t+1}\|^2 \mid \mathcal{F}_t] = \|u_t - \gamma_N(Au_t + R_t + \bar{\mathfrak{b}}_t)\|^2 + \gamma_N^2 \mathbb{E}[\|\bar{\zeta}_t\|^2 \mid \mathcal{F}_t].$$

Use $\langle u, Au\rangle \geq \lambda_g \|u\|^2$, $\|A\| \leq L$, and, for $\gamma_N \leq 1$,

$$2\gamma_N |\langle u - \gamma_N Au, R_t + \bar{\mathfrak{b}}_t\rangle| \leq \frac{\lambda_g \gamma_N}{2}\|u\|^2 + \frac{2(1+L)^2\gamma_N}{\lambda_g}\|R_t + \bar{\mathfrak{b}}_t\|^2.$$

Together with (38), this first yields an e_t coefficient of the form $1 - 3\lambda_g\gamma_N/2 + C\gamma_N^2 + C\gamma_N s_N^2$. For sufficiently small γ_N, s_N^2, fix $a = \lambda_g/2$ and constants L_1, L_2, L_3 so that

$$e_{t+1} \leq (1 - a\gamma_N)e_t + L_1\gamma_N c_t + L_2\gamma_N s_N^2 + L_3\gamma_N^2. \tag{40}$$

Choose the threshold also so that $0 < a\gamma_N \leq 1$.

Let $r_\rho = (1+\rho^2)/2 < 1$ and $\chi_\rho = (1+\rho^2)/(1-\rho^2)$. Projecting the state update onto the disagreement space and using weighted Young's inequality gives

$$c_{t+1} \leq r_\rho c_t + \chi_\rho \gamma_N^2 d_t. \tag{41}$$

For the tracker, the analogous inequality gives

$$d_{t+1} \leq r_\rho d_t + \frac{\chi_\rho}{n}\mathbb{E}\|\Pi_\perp(\mathbf{\Gamma}_{t+1} - \mathbf{\Gamma}_t)\|^2.$$

Here the projection acts only on the agent index. It is not legitimate to discard the noise difference between two freshly sampled residuals. By (39), its second moment is bounded by $C(1 + e_t + c_t + e_{t+1} + c_{t+1})$. Directly from the state update, $\bar{D}_t = \bar{\Gamma}_t$, and (39),

$$e_{t+1} + c_{t+1} \leq C(1 + e_t + c_t + \gamma_N^2 d_t).$$

Substitution eliminates future terms and yields fixed L_4, L_5 such that

$$d_{t+1} \leq (r_\rho + L_4\gamma_N^2)d_t + L_5(1 + e_t + c_t). \tag{42}$$

In particular, the additive noise term in this tracker bound is not multiplied by γ_N.

Closing the stability bounds without circularity. For a finite m, let E_m, C_m, D_m be the maxima of e_t, c_t, d_t over $0 \leq t \leq m$. Since $c_0 = 0$, (41) gives

$$C_m \leq K_c\gamma_N^2 D_m, \qquad K_c = \chi_\rho/(1 - r_\rho).$$

The initial tracker has a uniformly bounded second moment by (39). If $L_4\gamma_N^2 \leq (1 - r_\rho)/2$, summing the geometric recursion (42) gives constants D_*, K_d with

$$D_m \leq D_* + K_d(E_m + C_m).$$

Summing (40) gives

$$E_m \leq e_0 + K_e C_m + K_s s_N^2 + K_\gamma \gamma_N, \quad K_e = L_1/a,\ K_s = L_2/a,\ K_\gamma = L_3/a.$$

Impose the additional explicit smallness conditions

$$K_d K_c \gamma_N^2 \leq \tfrac{1}{2}, \qquad 2K_e K_c K_d \gamma_N^2 \leq \tfrac{1}{2}. \tag{43}$$

The first yields $D_m \leq 2D_* + 2K_d E_m$. Using this in the bound on E_m and applying the second gives

$$E_m \leq 2\left[e_0 + K_s s_N^2 + K_\gamma \gamma_N + 2K_e K_c D_* \gamma_N^2\right].$$

This is bounded independently of m, N when $\gamma_N, s_N^2 \leq 1$. It then bounds D_m uniformly and C_m by $C\gamma_N^2$. Letting m increase proves

$$\sup_t e_t \leq C, \qquad \sup_t d_t \leq C, \qquad \sup_t c_t \leq C\gamma_N^2. \tag{44}$$

All imposed restrictions are satisfied for sufficiently large N because $\gamma_N = c_\gamma/\sqrt{N}$ and $s_N^2 \leq S/N$. This constructs a finite threshold rather than assuming a stable trajectory.

Averaging and the $1/N$ bound. Let $\widehat{\bar{U}}_N = N^{-1}\sum_{t=0}^{N-1}\bar{U}_t$. Summing the average state update using (37) gives the exact identity

$$A(\widehat{\bar{U}}_N - U^*) = \frac{\bar{U}_0 - \bar{U}_N}{N\gamma_N} - \frac{1}{N}\sum_{t=0}^{N-1}(R_t + \bar{\mathfrak{b}}_t + \bar{\zeta}_t). \tag{45}$$

Uniform stability bounds the squared boundary term by $C/(N^2\gamma_N^2)$. Jensen's inequality and (44) bound the squared averaged disagreement term by $C\gamma_N^2$, and the squared averaged bias by Cs_N^2. For $s < t$, $\bar{\zeta}_s$ is $\mathcal{F}_t$-measurable and $\mathbb{E}[\bar{\zeta}_t \mid \mathcal{F}_t] = 0$. Thus

$$\mathbb{E}\left\|\frac{1}{N}\sum_{t=0}^{N-1}\bar{\zeta}_t\right\|^2 = \frac{1}{N^2}\sum_{t=0}^{N-1}\mathbb{E}\|\bar{\zeta}_t\|^2 \le C/N. \tag{46}$$

Using $\|A^{-1}\| \le 1/\lambda_g$ in (45) yields the explicit form

$$\mathbb{E}\|\widehat{\bar{U}}_N - U^*\|^2 \le C\left[\frac{1}{N^2\gamma_N^2} + \gamma_N^2 + s_N^2 + \frac{1}{N}\right] \le C/N. \tag{47}$$

Replacing the average over $t = 0, \dots, N-1$ by the reported average over $t = 1, \dots, N$ adds $(\bar{U}_N - \bar{U}_0)/N$, whose squared moment is $O(N^{-2})$. Finally,

$$\frac{1}{n}\sum_i \mathbb{E}\left\|\frac{1}{N}\sum_{t=1}^{N}(U_i^{(t)} - \bar{U}_t)\right\|^2 \le \frac{1}{N}\sum_{t=1}^{N} c_t \le C\gamma_N^2 = C/N.$$

The orthogonal decomposition into network average and disagreement proves (35). For any fixed positive schedule constants, the finitely many horizons $N < N_0$ also have finite second moments by (33) and the finite recursion. Enlarging C can therefore extend the output bound to those horizons, without asserting infinite-time stability at their larger step sizes. □

Horizon condition. Let $\bar{\gamma}, \bar{s}^2 > 0$ satisfy all smallness conditions established in the preceding proof. For the vector schedules, it suffices to choose

$$N \ge N_{\mathrm{J}} := \left\lceil \max\left\{1, (c_\gamma/\bar{\gamma})^2, \frac{q(c_\mu^2 + c_\nu^2)}{\bar{s}^2}\right\}\right\rceil. \tag{48}$$

The thresholds are uniform over admissible fixed local states.

Application to the vector inverse theorem.

Proof of Theorem 5.2. Apply Theorem A.10 with $\mathcal{H} = \mathbb{R}^q$, $A_i = H_i$, $B_i = b_i$, and $U_i = v_i$. The exact average operator is $H_{\boldsymbol{\theta}}$ and the target is precisely (5). The norm bounds $\|H_i\| \le L_g$ and $\|b_i\| \le B_f$ are uniform in the local states.

At a history-measurable v, the residual bias is $\mathfrak{b}_{i,N}(v) = (H^{\mathrm{sm}}_{i,\mu_N} - H_i)v - (\mathbb{E}\widehat{b}_i - b_i)$. Lemma A.5, including (19), implies

$$\begin{aligned}\|\mathfrak{b}_{i,N}(v)\|^2 &\le 4L_H^2 q\mu_N^2\|v\|^2 + 2L_f^2 q\nu_N^2,\\ \mathbb{E}\|\widehat{H}_i v - \widehat{b}_i\|^2 &\le 2V_{Hv}(q)\|v\|^2 + 2V_b(q).\end{aligned} \tag{49}$$

Centering a random vector cannot increase its second moment. Thus the second line is also a bound for the centered residual noise. With $s_N^2 = q(\mu_N^2 + \nu_N^2)$, all hypotheses of Theorem A.10 hold. The zero initial vector requires no Hessian query and gives $\Gamma_i^{(0)} = -\widehat{b}_i^{(0)}$ exactly as in the vector branch of Algorithm 1. The theorem proves (8). □

Application to the matrix JHIP theorem.

Proof of Theorem 5.3. Apply Theorem A.10 on $\mathcal{H} = \mathbb{R}^{q\times p}$ with the Frobenius inner product, $A_i(Z) = H_i Z$, and $B_i = J_i^\top$. Symmetry of H_i makes A_i self-adjoint, $\|A_i\| \le L_g$, and

$$\left\langle Z, \frac{1}{n}\sum_i A_i(Z)\right\rangle_F = \mathrm{tr}(Z^\top H_{\boldsymbol{\theta}} Z) \ge \lambda_g\|Z\|_F^2.$$

Moreover, $\|B_i\|_F \le B_J$, and the exact target is $Z^*_{\boldsymbol{\theta}} = H^{-1}_{\boldsymbol{\theta}} J^\top_{\boldsymbol{\theta}}$, with $\|Z^*_{\boldsymbol{\theta}}\|_F \le R_Z := B_J/\lambda_g$. Its transpose is the global JHIP at the fixed local states.

For a history-measurable matrix Z, the residual bias equals $\mathfrak{b}^Z_{i,N}(Z) = (H^{\mathrm{sm}}_{i,\mu_N} - H_i)Z - (J^{\mathrm{sm}}_{i,\eta_N,\mu_N} - J_i)^\top$. Lemmas A.3 and A.4 give

$$\begin{aligned} \|\mathfrak{b}^Z_{i,N}(Z)\|^2_F &\le 4L^2_H q\mu^2_N \|Z\|^2_F + 2L^2_J(p\eta^2_N + q\mu^2_N), \\ \mathbb{E}\|\widehat{H}_i Z - \widehat{J}_i^\top\|^2_F &\le 2V_H(q)\|Z\|^2_F + 2V_J(p,q). \end{aligned} \tag{50}$$

The second line follows from $\|\widehat{H}_i Z\|_F \le \|\widehat{H}_i\|_F \|Z\|_F$. Centering the residual only decreases its conditional second moment. Thus the generic bias and noise conditions hold with $s^2_N = q\mu^2_N + p\eta^2_N$.

Let $\bar\gamma_Z, \bar s^2_Z > 0$ satisfy the generic smallness conditions for these matrix-oracle constants. A sufficient horizon is

$$N_{\mathrm{Z}} := \left\lceil \max\left\{1, (c_\gamma/\bar\gamma_Z)^2, \frac{qc^2_\mu + pc^2_\eta}{\bar s^2_Z}\right\}\right\rceil. \tag{51}$$

The initialization in Algorithm 1 is exactly $Z^{(0)}_i = 0$, $D^{(0)}_i = \Gamma^{(0)}_i = -(\widehat{J}^{(0)}_i)^\top$. Theorem A.10 therefore proves both uniform iterate second moments and (9). All constants are independent of the fixed local query states. □

Assumptions for the vector formulation. The vector residual uses the bounds B_f, L_f, M_F for its upper-gradient right-hand side; the matrix residual uses the corresponding cross-block bounds. All are included in the stated assumptions. The vector guarantee requires a fresh right-hand side at each residual and the prescribed smoothing schedule. The two formulations have distinct targets and constants.

Proof of Lemma 5.4. We first bound the output of the chosen auxiliary formulation and its terminal hypergradient, then control outer consensus and the drift of the lower solution before applying the inner-loop bound. For vector mode, Theorem 5.2, conditional on the local states, gives

$$\frac{1}{n}\sum_i \mathbb{E}\|\widetilde{v}^{(N)}_{i,k}\|^2 \le \frac{2C_{\mathrm{v}}}{N} + \frac{2B^2_f}{\lambda^2_g} \le M^2_{v,0}. \tag{52}$$

For full mode, Theorem 5.3 similarly gives

$$\frac{1}{n}\sum_i \mathbb{E}\|\widetilde{Z}^{(N)}_{i,k}\|^2_F \le \frac{2C_Z}{N} + \frac{2B^2_J}{\lambda^2_g} \le M^2_{Z,0}. \tag{53}$$

These constants are uniform for $N \ge \max\{N_{\mathrm{J}}, N_{\mathrm{Z}}\}$ and all fixed local states. The full-mode terminal y-gradient is formed as in (6), with radius ν_N and fresh samples independent of the matrix output. Lemma A.5 and $\|Z^\top d\| \le \|Z\|_F\|d\|$ imply, for either mode,

$$\frac{1}{n}\sum_i \mathbb{E}\|h_{i,k}\|^2 \le 2M^2_F p(p+2) + 2\max\{V_{Jv}(p,q)M^2_{v,0}, V_b(q)M^2_{Z,0}\} =: H^2_0. \tag{54}$$

The terminal x-gradient is also sampled independently after the auxiliary solve. This bound uses neither inner-loop accuracy nor warm-start stability, so the argument below is not circular.

Let $\mathcal{E}_{x,k} := n^{-1}\mathbb{E}\|\Pi_\perp \mathbf{X}_k\|^2_F$. The outer recursion gives

$$\Pi_\perp \mathbf{X}_{k+1} = (W - \Pi)\Pi_\perp \mathbf{X}_k - \alpha \Pi_\perp \mathbf{h}_k, \qquad \mathbf{h}_k := \mathrm{col}(h^\top_{1,k}, \dots, h^\top_{n,k}) \in \mathbb{R}^{n\times p}.$$

Weighted Young's inequality and $\|W - \Pi\|_2 = \rho < 1$ give constants $a_\rho = (1+\rho^2)/2 < 1$ and $C_\rho < \infty$ such that

$$\mathcal{E}_{x,k+1} \le a_\rho \mathcal{E}_{x,k} + C_\rho \alpha^2 H^2_0.$$

Because $\mathcal{E}_{x,0} = 0$, iteration yields

$$\sup_{0\le k\le K} \mathcal{E}_{x,k} \le \frac{C_\rho H^2_0}{1 - a_\rho}\alpha^2 =: C_x \alpha^2. \tag{55}$$

Moreover, $\mathbf{X}_{k+1} - \mathbf{X}_k = (W - I)\Pi_\perp \mathbf{X}_k - \alpha \mathbf{h}_k$. Since $\|W - I\|_2 \le 2$, (54) and (55) imply

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|x_{i,k+1} - x_{i,k}\|^2 \le 8\mathcal{E}_{x,k} + 2\alpha^2 H_0^2 \le C_\Delta \alpha^2. \tag{56}$$

Applying (23) with $\mathbf{X} = \mathbf{X}_k$ and $\mathbf{X}' = \mathbf{X}_{k+1}$ gives

$$\mathbb{E}\|y^*_{\mathbf{X}_{k+1}} - y^*_{\mathbf{X}_k}\|^2 \le \frac{L_g^2 C_\Delta}{\lambda_g^2}\alpha^2. \tag{57}$$

For $0 \le k \le K-2$, the warm start $y^{(0)}_{i,k+1} = y^{(T)}_{i,k}$, Theorem 5.1, and (57) yield

$$\begin{aligned} B_{k+1} &\le \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\|y^{(T)}_{i,k} - y^*_{\mathbf{X}_k}\|^2 + 2\mathbb{E}\|y^*_{\mathbf{X}_{k+1}} - y^*_{\mathbf{X}_k}\|^2 \\ &\le \frac{2C_{\mathrm{in}}}{T} B_k + C_B\left(\frac{1}{T} + \alpha^2\right). \end{aligned} \tag{58}$$

Choose $T_{\mathrm{ws}} \ge \max\{1, 4C_{\mathrm{in}}\}$. For $T \ge T_{\mathrm{ws}}$, the coefficient of B_k in (58) is at most $1/2$, and therefore

$$B_k \le 2^{-k} B_0 + 2C_B\left(\frac{1}{T} + \alpha^2\right) \le B_0 + 2C_B\left(T_{\mathrm{ws}}^{-1} + 1\right) =: C_{\mathrm{ws}},$$

where we used $\alpha \le 1$. This proves the claim. □

A.6 Local-State Alignment for the Outer Loop

Throughout this subsection, use vector mode and assume Assumptions 3.1, 3.2, 3.3, and 3.4. Use the schedules, initialization, and warm start of Lemma 5.4, with $B_0 < \infty$ and $T \ge T_{\mathrm{ws}}$. The constants below are independent of k, K, T, and N.

Define $\mathcal{E}_{x,k} := n^{-1}\sum_{i=1}^{n} \mathbb{E}\|x_{i,k} - \bar{x}_k\|^2 = n^{-1}\mathbb{E}\|\Pi_\perp \mathbf{X}_k\|_F^2$.

Lemma A.11 (Outer-state alignment)**.** *At outer iteration k,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|y^{(T)}_{i,k} - y^*(\bar{x}_k)\|^2 \le C_y\left(\frac{1}{T} + \mathcal{E}_{x,k}\right), \tag{59}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\widetilde{v}^{(N)}_{i,k} - v^*(\bar{x}_k, y^*(\bar{x}_k))\right\|^2 \le C_{\mathrm{a},v}\left(\frac{1}{N} + \frac{1}{T} + \mathcal{E}_{x,k}\right). \tag{60}$$

Proof. Theorem 5.1 and Lemma 5.4 control the error relative to $y^*_{\mathbf{X}_k}$ by $C(1+B_k)/T \le C/T$, while Lemma A.8 controls the squared distance from $y^*_{\mathbf{X}_k}$ to $y^*(\bar{x}_k)$ by $C\mathcal{E}_{x,k}$. This proves (59).

Conditionally on $\boldsymbol{\theta}_k = (\theta_{1,k}, \dots, \theta_{n,k})$ with $\theta_{i,k} = (x_{i,k}, y^{(T)}_{i,k})$, Theorem 5.2 gives

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\widetilde{v}^{(N)}_{i,k} - v^*_{\boldsymbol{\theta}_k}\|^2 \mid \boldsymbol{\theta}_k\right] \le \frac{C}{N}.$$

Lemma A.9 and the first part of the lemma then prove (60). □

Define $\mathcal{T}_i(x, y, v) := \nabla_x f_i(x, y) - \nabla^2_{xy} g_i(x, y) v$. At a common point,

$$\nabla\Phi(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{T}_i\left(x, y^*(x), v^*(x, y^*(x))\right). \tag{61}$$

Lemma A.12 (Outer zeroth-order bias and variance)**.** *Let $0 < \nu_x, \eta_N, \mu_N \le 1$. Let $\mathcal{F}_k$ contain the outer iterates and all inner and vector-oracle randomness before the fresh terminal stencils at*

iteration k, and let $\bar{h}_k = n^{-1}\sum_i h_{i,k}$. Then

$$\mathbb{E}\left\|\nabla\Phi(\bar{x}_k) - \mathbb{E}[\bar{h}_k \mid \mathcal{F}_k]\right\|^2 \le C_h\left(\mathcal{E}_{x,k} + \frac{1}{T} + \frac{1}{N} + p\nu_x^2 + p\eta_N^2 + q\mu_N^2\right), \tag{62}$$

$$\mathbb{E}\|\bar{h}_k - \mathbb{E}[\bar{h}_k \mid \mathcal{F}_k]\|^2 \le \frac{C_{\mathrm{h},2}}{n}\left(1 + \mathcal{E}_{x,k} + \frac{1}{T} + \frac{1}{N}\right), \tag{63}$$

The local directions also satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|h_{i,k}\|^2 \le C_{\mathrm{h},2}\left(1 + \mathcal{E}_{x,k} + \frac{1}{T} + \frac{1}{N}\right). \tag{64}$$

Proof. Conditionally on $\mathcal{F}_k$, all local states and vector outputs are fixed. The fresh terminal stencils satisfy

$$\mathbb{E}[h_{i,k} \mid \mathcal{F}_k] = \nabla_x f^x_{i,\nu_x}(x_{i,k}, y^{(T)}_{i,k}) - J^{\mathrm{sm}}_{i,\eta_N,\mu_N}\widetilde{v}^{(N)}_{i,k},$$

where $f^x_{i,\nu_x}(x,y) := \mathbb{E}_u f_i(x + \nu_x u, y)$. Using (20) and (16), removal of this terminal smoothing costs at most

$$C\left[p\nu_x^2 + (p\eta_N^2 + q\mu_N^2)\frac{1}{n}\sum_i \mathbb{E}\|\widetilde{v}^{(N)}_{i,k}\|^2\right] \tag{65}$$

in squared error after network averaging and expectation. The vector output moment is uniformly bounded by (52).

Let $x^\circ = \bar{x}_k$, $y^\circ = y^*(\bar{x}_k)$, and $v^\circ = v^*(\bar{x}_k, y^\circ)$. The exact local mapping is $\mathcal{T}_i(x,y,v) = \nabla_x f_i(x,y) - \nabla^2_{xy} g_i(x,y)v$. The decomposition

$$\begin{aligned}\mathcal{T}_i(x_i,y_i,v_i) - \mathcal{T}_i(x^\circ,y^\circ,v^\circ) = {}& \nabla_x f_i(x_i,y_i) - \nabla_x f_i(x^\circ,y^\circ)\\ & - \nabla^2_{xy} g_i(x_i,y_i)(v_i - v^\circ)\\ & - [\nabla^2_{xy} g_i(x_i,y_i) - \nabla^2_{xy} g_i(x^\circ,y^\circ)]v^\circ\end{aligned}$$

and $\|\nabla^2_{xy} g_i\| \le B_J$, $\|v^\circ\| \le R_v$ imply

$$\|\mathcal{T}_i(x_i,y_i,v_i) - \mathcal{T}_i(x^\circ,y^\circ,v^\circ)\|^2 \le C(\|x_i - x^\circ\|^2 + \|y_i - y^\circ\|^2 + \|v_i - v^\circ\|^2).$$

Combining Lemma A.11, (61), and (65) proves (62). No independence between the communicating agents' vector states is used.

The fresh terminal product and Lemma A.5 give

$$\mathbb{E}\|h_{i,k}\|^2 \le 2M_F^2 p(p+2) + 2V_{J_v}(p,q)\mathbb{E}\|\widetilde{v}^{(N)}_{i,k}\|^2. \tag{66}$$

The uniform moment (52) proves (64), with a constant independent of the horizons. Finally, conditional independence of the terminal samples across agents yields

$$\begin{aligned}\mathbb{E}\|\bar{h}_k - \mathbb{E}[\bar{h}_k \mid \mathcal{F}_k]\|^2 &= \frac{1}{n^2}\sum_i \mathbb{E}\|h_{i,k} - \mathbb{E}[h_{i,k} \mid \mathcal{F}_k]\|^2\\ &\le \frac{1}{n^2}\sum_i \mathbb{E}\|h_{i,k}\|^2 \le \frac{C_{\mathrm{h},2}}{n}\left(1 + \mathcal{E}_{x,k} + T^{-1} + N^{-1}\right).\end{aligned}$$

This proves (63). The variance of a single terminal hypergradient sample need not vanish as N grows; the proof bounds its conditional bias separately and retains its variance in the outer descent inequality. □

A.7 Matrix-State Alignment for the Outer Loop

We establish the corresponding outer bounds for full mode without identifying its local hypergradient contributions with the vector ones. Write $\theta^\circ_k = (\bar{x}_k, y^*(\bar{x}_k))$ and $Z^\circ_k = H(\theta^\circ_k)^{-1}J(\theta^\circ_k)^\top$.

Lemma A.13 (Matrix alignment and outer oracle bounds)**.** *Under the assumptions, initialization, and schedules of Lemma 5.4, use full mode and $T \geq T_{\mathrm{ws}}$. Let the terminal y-gradient radius be $\nu_N = c_\nu N^{-1/2}$, and let $0 < \nu_x, \nu_N, \eta_N, \mu_N \leq 1$. The matrix error satisfies*

$$\frac{1}{n}\sum_i \mathbb{E}\|\widetilde{Z}_{i,k}^{(N)} - Z_k^\circ\|_F^2 \leq C_{\mathrm{a},Z}\left(N^{-1} + T^{-1} + \mathcal{E}_{x,k}\right). \tag{67}$$

Let $\mathcal{F}_k$ contain the outer iterates and all inner and matrix-oracle randomness before the fresh terminal stencils. The bias, variance, and second-moment bounds (62)–(64) also hold in full mode, with constants independent of k, K, T, N.

Proof. Set $H^\circ = H(\theta_k^\circ)$ and $J^\circ = J(\theta_k^\circ)$. The target perturbation identity is

$$Z_{\boldsymbol{\theta}_k}^* - Z_k^\circ = H_{\boldsymbol{\theta}_k}^{-1}\left[(J_{\boldsymbol{\theta}_k} - J^\circ)^\top - (H_{\boldsymbol{\theta}_k} - H^\circ)Z_k^\circ\right]. \tag{68}$$

Since $\|Z_k^\circ\|_F \leq R_Z = B_J/\lambda_g$, the deterministic block Lipschitz bounds and Jensen's inequality give, pathwise,

$$\|Z_{\boldsymbol{\theta}_k}^* - Z_k^\circ\|_F^2 \leq \frac{(L_J + L_H R_Z)^2}{\lambda_g^2}\frac{1}{n}\sum_i \|\theta_{i,k} - \theta_k^\circ\|^2. \tag{69}$$

Theorem 5.1, Lemma 5.4, and (22) imply, in full mode as well,

$$\frac{1}{n}\sum_i \mathbb{E}\|y_{i,k}^{(T)} - y^*(\bar{x}_k)\|^2 \leq C(T^{-1} + \mathcal{E}_{x,k}).$$

Consequently, the expectation of the right-hand side of (69) is $O(T^{-1} + \mathcal{E}_{x,k})$. Conditioning Theorem 5.3 on $\boldsymbol{\theta}_k$ supplies C_Z/N for the error around $Z_{\boldsymbol{\theta}_k}^*$. Combining the two bounds proves (67).

Define the full-mode local mapping $\mathcal{T}_i^{\mathrm{mat}}(x, y, Z) := \nabla_x f_i(x, y) - Z^\top \nabla_y f_i(x, y)$. Symmetry of H° and the global hypergradient identity give

$$\frac{1}{n}\sum_i \mathcal{T}_i^{\mathrm{mat}}(\theta_k^\circ, Z_k^\circ) = \nabla\Phi(\bar{x}_k).$$

For any local state $\theta_i = (x_i, y_i)$ and matrix Z_i,

$$\begin{aligned}\mathcal{T}_i^{\mathrm{mat}}(\theta_i, Z_i) - \mathcal{T}_i^{\mathrm{mat}}(\theta_k^\circ, Z_k^\circ) = {}& \nabla_x f_i(\theta_i) - \nabla_x f_i(\theta_k^\circ) \\ & - (Z_i - Z_k^\circ)^\top \nabla_y f_i(\theta_i) \\ & - (Z_k^\circ)^\top[\nabla_y f_i(\theta_i) - \nabla_y f_i(\theta_k^\circ)].\end{aligned}$$

Thus, using the bounded true gradient and true target,

$$\begin{aligned}&\|\mathcal{T}_i^{\mathrm{mat}}(\theta_i, Z_i) - \mathcal{T}_i^{\mathrm{mat}}(\theta_k^\circ, Z_k^\circ)\|^2 \\ &\quad \leq 2L_f^2(1 + R_Z)^2\|\theta_i - \theta_k^\circ\|^2 + 2B_f^2\|Z_i - Z_k^\circ\|_F^2.\end{aligned}$$

This decomposition requires no fourth moment of the estimated matrix.

Conditionally on $\mathcal{F}_k$, the terminal stencils have expectation

$$\mathbb{E}[h_{i,k} \mid \mathcal{F}_k] = \nabla_x f_{i,\nu_x}^x(\theta_{i,k}) - [\widetilde{Z}_{i,k}^{(N)}]^\top \nabla_y f_{i,\nu_N}^y(\theta_{i,k}),$$

where $f_{i,\nu_N}^y(x, y) = \mathbb{E}_z f_i(x, y + \nu_N z)$ and f_{i,ν_x}^x is defined as above. By (20), removing this smoothing costs at most

$$C\left[p\nu_x^2 + q\nu_N^2\frac{1}{n}\sum_i \mathbb{E}\|\widetilde{Z}_{i,k}^{(N)}\|_F^2\right] \leq C(p\nu_x^2 + q\nu_N^2)$$

in squared error after averaging, where (53) gives the last inequality. The mapping and alignment bounds therefore yield

$$\mathbb{E}\|\nabla\Phi(\bar{x}_k) - \mathbb{E}[\bar{h}_k \mid \mathcal{F}_k]\|^2 \leq C\left[\mathcal{E}_{x,k} + T^{-1} + N^{-1} + p\nu_x^2 + q\nu_N^2\right]. \tag{70}$$

Since $q\nu_N^2 = qc_\nu^2/N$, this implies (62), with the nonnegative $p\eta_N^2 + q\mu_N^2$ terms included and a possibly larger constant. Matrix-oracle smoothing is already controlled by its $1/N$ error bound.

Finally, independence of the terminal samples from the matrix output gives

$$\mathbb{E}\|h_{i,k}\|^2 \le 2M_F^2 p(p+2) + 2V_b(q)\mathbb{E}\|\widetilde{Z}_{i,k}^{(N)}\|_F^2. \tag{71}$$

Equation (53) proves the claimed averaged second-moment bound. Given $\mathcal{F}_k$, the terminal noises are independent across agents, even though their matrix states need not be. Their average therefore satisfies

$$\mathbb{E}\|\bar{h}_k - \mathbb{E}[\bar{h}_k \mid \mathcal{F}_k]\|^2 = \frac{1}{n^2}\sum_i \mathbb{E}\|h_{i,k} - \mathbb{E}[h_{i,k} \mid \mathcal{F}_k]\|^2 \le \frac{H_0^2}{n},$$

which implies (63). Taking the larger constants from the two mode-specific analyses gives common $C_h, C_{\mathrm{h},2}$ for both formulations. □

A.8 Proof of Theorem 5.5

Parameter conditions. For a common constant $C_{\mathrm{h},2}$ satisfying Lemmas A.12 and A.13, define $\bar{\alpha}_{\mathrm{out}} := \min\{1, 1/L_\Phi, (1-\rho^2)/\sqrt{8C_{\mathrm{h},2}}\}$. Choose $c_\alpha \in (0, \bar{\alpha}_{\mathrm{out}}]$ and positive constants c_T, c_N, c_x, c_η, and set

$$\alpha = c_\alpha K^{-1/2}, \quad T = \lceil c_T K^{1/2} \rceil, \quad N = \lceil c_N K^{1/2} \rceil, \quad \nu_x = c_x K^{-1/4}, \quad \eta_N = c_\eta N^{-1/2}. \tag{72}$$

Here $T \ge T_{\mathrm{ws}}$, $N \ge \max\{N_{\mathrm{J}}, N_{\mathrm{Z}}\}$, and $\nu_x, \eta_N, \mu_N, \nu_N \le 1$. The inner parameters satisfy (25); γ_N, μ_N, ν_N follow (7). The auxiliary horizon satisfies (48) and (51). In full mode, ν_N is the radius of the fresh terminal y-gradient. All thresholds are finite for fixed problem and schedule parameters.

Proof of Theorem 5.5. Proof strategy. We first average the decentralized outer update and apply smoothness of Φ. Lemmas A.12 and A.13 give common conditional bias and variance bounds for vector and full modes, while a second recursion controls the outer consensus error. Summing the descent inequality and substituting the inner, JHIP, smoothing, and consensus accuracies yields the master bound (76); the stated parameter schedules balance all its terms at order $K^{-1/2}$.

Averaging the outer update and using double stochasticity gives $\bar{x}_{k+1} = \bar{x}_k - \alpha\bar{h}_k$. Let $m_k := \mathbb{E}[\bar{h}_k \mid \mathcal{F}_k]$. The smoothness derived in Lemma A.2 and the identity $-\langle a, b\rangle = -\|a\|^2/2 - \|b\|^2/2 + \|a-b\|^2/2$ give, after conditioning on $\mathcal{F}_k$ and then taking total expectation,

$$\begin{aligned}\mathbb{E}\Phi(\bar{x}_{k+1}) \le \mathbb{E}\Phi(\bar{x}_k) &- \alpha\mathbb{E}\langle\nabla\Phi(\bar{x}_k), m_k\rangle \\ &+ \frac{L_\Phi\alpha^2}{2}\mathbb{E}\|m_k\|^2 + \frac{L_\Phi\alpha^2}{2}\mathbb{E}\|\bar{h}_k - m_k\|^2.\end{aligned}$$

When $\alpha \le 1/L_\Phi$, the coefficient of $\|m_k\|^2$ is nonpositive after applying the identity. Using the corresponding bounds from Lemmas A.12 and A.13 therefore yields

$$\begin{aligned}\frac{\alpha}{2}\mathbb{E}\|\nabla\Phi(\bar{x}_k)\|^2 \le \mathbb{E}\left[\Phi(\bar{x}_k) - \Phi(\bar{x}_{k+1})\right] &\\ + \frac{\alpha}{2}\mathbb{E}\left\|\nabla\Phi(\bar{x}_k) - \mathbb{E}[\bar{h}_k \mid \mathcal{F}_k]\right\|^2 &\\ + \frac{L_\Phi\alpha^2 C_{\mathrm{h},2}}{2n}\left(1 + \mathcal{E}_{x,k} + \frac{1}{T} + \frac{1}{N}\right).&\end{aligned} \tag{73}$$

For the consensus error,

$$\Pi_\perp \mathbf{X}_{k+1} = (W - \Pi)\Pi_\perp \mathbf{X}_k - \alpha\Pi_\perp \mathbf{h}_k.$$

Weighted Young's inequality and (64) give

$$\mathcal{E}_{x,k+1} \le \frac{1+\rho^2}{2}\mathcal{E}_{x,k} + \frac{2C_{\mathrm{h},2}\alpha^2}{1-\rho^2}\left(1 + \mathcal{E}_{x,k} + \frac{1}{T} + \frac{1}{N}\right). \tag{74}$$

Choose the constant in α so that $2C_{\mathrm{h},2}\alpha^2/(1-\rho^2) \le (1-\rho^2)/4$. The coefficient of $\mathcal{E}_{x,k}$ in (74) is then at most $(3+\rho^2)/4 < 1$. Since $\mathcal{E}_{x,0} = 0$,

$$\sup_{0\le k\le K}\mathcal{E}_{x,k} \le \frac{8C_{\mathrm{h},2}\alpha^2}{(1-\rho^2)^2}\left(1 + \frac{1}{T} + \frac{1}{N}\right). \tag{75}$$

Thus (64) also yields a second-moment bound for the outer estimates that is uniform in k and K; no separate bounded-oracle assumption is needed.

Summing (73), using $\Phi \geq \Phi_{\inf}$, and substituting (62) and (75) yield

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(\bar{x}_k)\|^2 \leq \frac{2(\Phi(\bar{x}_0)-\Phi_{\inf})}{K\alpha} + C\left(\alpha + \frac{\alpha^2}{(1-\rho^2)^2} + \frac{1}{T} + \frac{1}{N} + p\nu_x^2 + p\eta_N^2 + q\mu_N^2\right). \tag{76}$$

With the schedules in (72), every term in (76) is $\mathcal{O}(K^{-1/2})$. Therefore,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(\bar{x}_k)\|^2 \leq \frac{C_{\mathrm{DZBO}}}{\sqrt{K}}.$$

This proves the theorem for either mode. The shared index R affects only which iterate is stored; it does not enter any update and is independent of the entire iterate trajectory. Thus the reported outputs satisfy

$$\mathbb{E}\|\nabla\Phi(\bar{x}_R)\|^2 = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(\bar{x}_k)\|^2 \leq C_{\mathrm{DZBO}}/\sqrt{K}.$$

Their local stationarity also obeys

$$\frac{1}{n}\sum_i \mathbb{E}\|\nabla\Phi(x_{i,R})\|^2 \leq 2\mathbb{E}\|\nabla\Phi(\bar{x}_R)\|^2 + \frac{2L_\Phi^2}{K}\sum_{k=0}^{K-1}\mathcal{E}_{x,k} = \mathcal{O}(K^{-1/2}),$$

by (75). The network average $\bar{x}_R$ is an analysis quantity; forming it is not required by the local output rule. □

Proof of Corollary 5.6. Theorem 5.5 and Corollary B.3 require $K = \mathcal{O}(\epsilon^{-2})$ and $T, N = \mathcal{O}(\epsilon^{-1})$ for the stated squared-gradient criterion. Each inner step uses two lower-level values. In vector and direct modes, the auxiliary solve uses two upper values initially and six values for each of the remaining $N-1$ residuals. The terminal x-gradient and cross product use two upper and four lower values. Full mode, as written in Algorithm 1, forms both four-point matrices at each of its N residuals, and its two terminal gradient blocks use four upper values. Consequently,

$$Q_{\mathrm{node}} = \begin{cases} K(2T+6N+2), & m \in \{\text{vector}, \text{direct}\}, \\ K(2T+8N+4), & m = \text{full}, \end{cases} \qquad Q_{\mathrm{net}} = nQ_{\mathrm{node}}.$$

These counts imply the mode-uniform upper bound in the corollary. If f_i is independent of x, the terminal x-gradient vanishes and its two queries can be omitted.

An auxiliary round exchanges the current state ($v_i^{(t)}$ or $Z_i^{(t)}$) and its tracker $D_i^{(t)}$ jointly. After receiving them, each node updates its state, samples its new residual, and updates its tracker using the received old trackers and its stored old residual. One such exchange suffices per auxiliary step, giving $R_{\mathrm{comm}} = K(T+N+1)$. With r_m as defined in the corollary, the payloads are $q, 2r_m, p$ for inner, auxiliary, and outer rounds. Hence $S_i = d_iK(qT+2r_mN+p)$ and $S_{\mathrm{net}} = 2|\mathcal{E}|K(qT+2r_mN+p)$. The counts exchange both states at every auxiliary step; omitting the unused final tracker payload can only reduce them.

The factored matrix actions cost $\mathcal{O}(q)$ for a vector residual, and (17) costs $\mathcal{O}(pq)$ for a matrix residual. The terminal work is $\mathcal{O}(p+q)$ in vector/direct modes and $\mathcal{O}(pq+p+q)$ in full mode. Including streamed neighbor aggregation gives

$$\mathcal{O}(K[T(d_i+1)q + N(d_i+1)r_m + p + q + r_m + d_ip])$$

arithmetic operations. Perturbations, accumulators, the stored output, and sparse mixing weights use $\mathcal{O}(p+q+r_m+d_i)$ memory. Thus full mode requires $\mathcal{O}(pq+p+q+d_i)$ storage, while vector/direct modes require $\mathcal{O}(p+q+d_i)$. The shared output index is generated from a preshared independent seed. Storing one outer iterate adds no function queries or communication rounds. These counts exclude dataset storage and the internal cost of evaluating a sample function. The displayed ϵ-orders hold for fixed problem and schedule parameters; convergence constants retain their dimension and graph dependence. □

B DZBO-DIRECT: DIRECT PRODUCTS AND THEIR GUARANTEES

This extension estimates the required products directly, motivating smaller direction-noise bounds without constructing a random matrix action. The supplied direction depends on the current vector, so its bias is analyzed explicitly before applying the common convergence theorem. The product bounds below quantify this advantage without asserting a proportional improvement in complete bilevel complexity.

B.1 DIRECT FOUR-POINT ESTIMATORS AND CONVERGENCE

For a supplied, history-measurable vector $v \neq 0$, write $e_v = v/\|v\|$. Draw fresh $r \sim \mathcal{N}(0, I_q)$ and $u \sim \mathcal{N}(0, I_p)$, independently for the two stencils, and set

$$\mathcal{H}_{i,\mu,\delta}(v) = \frac{\|v\|}{4\mu\delta} \sum_{a,b\in\mathcal{S}} ab\, G_i(x_i, y_i + a\mu r + b\delta e_v; \xi_H)\, r,$$

$$\mathcal{J}_{i,\eta,\delta}(v) = \frac{\|v\|}{4\eta\delta} \sum_{a,b\in\mathcal{S}} ab\, G_i(x_i + a\eta u, y_i + b\delta e_v; \xi_J)\, u. \tag{77}$$

Both outputs are defined as zero at $v = 0$. Each stencil uses the same sample at its four points. Only one direction is random within each stencil; the other is the supplied normalized vector. The estimator $\mathcal{H}(v)$ is not assumed to equal a fixed random matrix times v at positive δ.

Lemma B.1 (Direct-product moments and bias)**.** *Under Assumptions 3.2, 3.3, and 3.4, conditionally on a supplied v,*

$$\begin{aligned}
\mathbb{E}\|\mathcal{H}_{i,\mu,\delta}(v)\|^2 &\le M_H^2 q(q+2)\|v\|^2, \\
\mathbb{E}\|\mathcal{J}_{i,\eta,\delta}(v)\|^2 &\le M_J^2 p(p+2)\|v\|^2, \\
\|\mathbb{E}\mathcal{H}_{i,\mu,\delta}(v) - H_i v\| &\le L_H\|v\|(\mu\sqrt{q} + \delta/2), \\
\|\mathbb{E}\mathcal{J}_{i,\eta,\delta}(v) - J_i v\| &\le L_J\|v\|(\eta\sqrt{p} + \delta/2).
\end{aligned} \tag{78}$$

No third derivative, fourth derivative, or sample-Hessian Lipschitz assumption is required for these bounds.

Proof. The claims at $v = 0$ are immediate. For $v \neq 0$, condition on its value, so e_v is a fixed unit vector independent of the current randomness. The integral form of the Hessian estimator is

$$\mathcal{H}(v) = \|v\| \operatorname*{avg}_{s\in[-\mu,\mu],t\in[-\delta,\delta]} [r^\top \nabla^2_{yy} G_i(x_i, y_i + sr + te_v; \xi_H) e_v] r.$$

Conditioning on r and applying Jensen's inequality bounds its second moment by $M_H^2\|v\|^2\|r\|^4$. Averaging over r proves the first line of (78). Replacing the bilinear form by $u^\top \nabla^2_{xy} G_i e_v$ proves the second line.

For the bias, average over the sample before integrating by parts in the *random* direction. Gaussian symmetry gives

$$\begin{aligned}
\mathbb{E}\mathcal{H}(v) &= \frac{\|v\|}{2\delta}\mathbb{E}_r[\nabla_y g_i(x_i, y_i + \mu r + \delta e_v) - \nabla_y g_i(x_i, y_i + \mu r - \delta e_v)] \\
&= \frac{1}{2\delta}\int_{-\delta}^{\delta} \mathbb{E}_r[\nabla^2_{yy} g_i(x_i, y_i + \mu r + te_v)] v\, dt.
\end{aligned}$$

The Hessian Lipschitz bound now yields

$$\|\mathbb{E}\mathcal{H}(v) - H_i v\| \le \frac{L_H\|v\|}{2\delta}\int_{-\delta}^{\delta} (\mu\mathbb{E}\|r\| + |t|)\, dt \le L_H\|v\|(\mu\sqrt{q} + \delta/2).$$

For the cross product, integration by parts in u instead yields

$$\mathbb{E}\mathcal{J}(v) = \frac{1}{2\delta}\int_{-\delta}^{\delta} \mathbb{E}_u[\nabla^2_{xy} g_i(x_i + \eta u, y_i + te_v)] v\, dt.$$

The cross-block Lipschitz bound proves the last line. These calculations use deterministic Hessian and cross-block Lipschitz continuity after sample averaging. They do not differentiate the sample Hessian or differentiate e_v as a function of v. A state-dependent bias is retained explicitly. □

Theorem B.2 (Direct inverse-vector convergence)**.** *Under the assumptions of Theorem 5.2, choose $\gamma_N = c_\gamma N^{-1/2}$, $\mu_N = c_\mu N^{-1/2}$, $\nu_N = c_\nu N^{-1/2}$, and $\delta_N = c_\delta N^{-1/2}$. For all sufficiently large N, the direct vector path satisfies*

$$\frac{1}{n}\sum_i \mathbb{E}\|\widetilde{v}_i^{(N)} - v_{\boldsymbol{\theta}}^*\|^2 \le C_{\mathrm{dir}}/N, \tag{79}$$

uniformly over admissible fixed local states. No additional function regularity assumption is required relative to Theorem 5.2.

Proof. Use the same $A_i = H_i$, $B_i = b_i$, and target as in Theorem 5.2. Define the residual bias by $\mathfrak{b}_{i,N}(v) = \mathbb{E}\mathcal{H}_{i,\mu_N,\delta_N}(v) - H_i v - (\mathbb{E}\widehat{b}_i - b_i)$. Lemma B.1 and (19) give

$$\begin{aligned}\|\mathfrak{b}_{i,N}(v)\|^2 &\le 4L_H^2(q\mu_N^2 + \delta_N^2/4)\|v\|^2 + 2L_f^2 q\nu_N^2,\\ \mathbb{E}\|\mathcal{H}_{i,\mu_N,\delta_N}(v) - \widehat{b}_i\|^2 &\le 2M_H^2 q(q+2)\|v\|^2 + 2V_b(q).\end{aligned} \tag{80}$$

Set $s_N^2 = q\mu_N^2 + \delta_N^2 + q\nu_N^2 = O(N^{-1})$. These are exactly the bias and noise bounds required by Theorem A.10. That theorem does not assume that $\mathfrak{b}_{i,N}(v)$ is linear, or that the oracle is the gradient of a single smoothed quadratic. Its stability argument and averaging identity use the true deterministic H_i plus the displayed bias. They therefore apply to the normalized-direction estimator and prove (79). □

Corollary B.3 (Outer guarantee for DZBO-direct)**.** *Under the assumptions and schedules of Theorem 5.5, with $\delta_N = \Theta(N^{-1/2})$ and the stability thresholds adjusted to the direct oracle's constants, Algorithm 2 with mode direct has the same $\mathcal{O}(K^{-1/2})$ average squared-stationarity order and the same query and communication orders as DZBO.*

Proof. The direct inverse theorem supplies the uniform output moment and $1/N$ accuracy used in Appendix A.6. The fresh terminal cross oracle has conditional bias at most $L_J(\eta_N\sqrt{p} + \delta_N/2)\|\widetilde{v}_i\|$ and second moment at most $M_J^2 p(p+2)\|\widetilde{v}_i\|^2$. Consequently, the proof of (62) applies with the cross smoothing term $p\eta_N^2 + q\mu_N^2$ replaced by $p\eta_N^2 + \delta_N^2$. This is $O(1/N)$ under the stated schedule. The same warm-start and descent arguments apply, including the guarantee for the stored random output. Each direct product uses four function evaluations, so the total query count is $K(2T + 6N + 2)$ per node. The inner, auxiliary, and outer stages exchange $q, 2q, p$ scalars per directed edge and use $K(T + N + 1)$ rounds in total. □

B.2 Dimension Dependence and Quadratic Comparisons

The conditional second-moment bounds proved above are

	randomized matrix action	direct product
Hv	$M_H^2 q(q+2)^2\|v\|^2$	$M_H^2 q(q+2)\|v\|^2$
Jv	$M_J^2 p(p+2)(q+2)\|v\|^2$	$M_J^2 p(p+2)\|v\|^2$.

(81)

The comparison uses action bounds on both sides, rather than comparing a full-matrix Frobenius bound with a vector bound. The constants controlling the true blocks are the same. These bounds are not asserted to be sharp.

Proposition B.4 (Fixed quadratic comparison)**.** *Consider a deterministic quadratic sample function, so its blocks H, J are constant, and condition on a fixed vector v. The direct estimators are unbiased, with exact direction variances*

$$\mathbb{E}\|\mathcal{H}(v) - Hv\|^2 = (q+1)\|Hv\|^2, \qquad \mathbb{E}\|\mathcal{J}(v) - Jv\|^2 = (p+1)\|Jv\|^2. \tag{82}$$

For the randomized matrix-action Hessian estimator, if $H = \kappa I_q$,

$$\mathbb{E}\|\widehat{H}v - Hv\|^2 = \tfrac{1}{2}(q^2 + 7q + 8)\|Hv\|^2. \tag{83}$$

For an arbitrary fixed cross block J,

$$\mathbb{E}\|\widehat{J}v - Jv\|^2 = (p+2)\|J\|_F^2\|v\|^2 + (2p+3)\|Jv\|^2. \tag{84}$$

These statements concern Gaussian direction noise only, not data noise, finite-precision error, or the inverse solve's output error.

Proof. The four-point remainder vanishes for a quadratic, giving $\mathcal{H}(v) = rr^\top Hv$ and $\mathcal{J}(v) = uu^\top Jv$. For a standard Gaussian $r \in \mathbb{R}^d$ and fixed a, $\mathbb{E}(rr^\top a) = a$ and $\mathbb{E}\|rr^\top a\|^2 = (d+2)\|a\|^2$. Subtracting the squared mean proves (82).

For $H = \kappa I_q$, the matrix-action estimator equals $\kappa(r^\top s)[r(s^\top v) + s(r^\top v)]/2$. Rotational invariance and Gaussian fourth and sixth moments give

$$\mathbb{E}[(r^\top s)^2\|r\|^2(s^\top v)^2] = (q+2)^2\|v\|^2, \quad \mathbb{E}[(r^\top s)^3(r^\top v)(s^\top v)] = 3(q+2)\|v\|^2.$$

Its second moment is therefore $\kappa^2(q+2)(q+5)\|v\|^2/2$. Subtracting $\kappa^2\|v\|^2$ proves (83). Finally, $\widehat{J}v = u(u^\top Jw)(w^\top v)$. First averaging over u, then over w, gives the second moment $(p+2)[\|J\|_F^2\|v\|^2 + 2\|Jv\|^2]$. Its mean is Jv, which proves (84). □

C Supplementary Experiments

This section first specifies the experimental objectives and evaluation formula, then extends the main studies along three axes: label corruption, the image-classification task and representation, and the number of agents. The studies retain the baseline and cycle topology described in Section 6; task-specific settings are given below.

C.1 Experimental Formulations and Evaluation

This subsection gives the objective functions, synthetic label-generation rule, and evaluation formula used in Section 6.

Synthetic hyperparameter optimization. We study heterogeneous logistic regression with element-wise ℓ^2 regularization, where the upper variable selects regularization strengths and the lower variable is the classifier:

$$f_i(\lambda, \vartheta) = \frac{1}{|\mathcal{D}_i'|} \sum_{(a_e, b_e) \in \mathcal{D}_i'} \psi(b_e a_e^\top \vartheta),$$

$$g_i(\lambda, \vartheta) = \frac{1}{|\mathcal{D}_i|} \sum_{(a_e, b_e) \in \mathcal{D}_i} \psi(b_e a_e^\top \vartheta) + \frac{1}{2}\vartheta^\top \operatorname{diag}(\exp(\lambda))\vartheta,$$

where $\psi(s) = \log(1 + \exp(-s))$, $\mathcal{D}_i'$ and $\mathcal{D}_i$ are the local validation and training sets, and $\vartheta^*(\lambda)$ minimizes the aggregate lower objective (Franceschi et al., 2018). Each positive coefficient $\exp(\lambda_j)$ controls shrinkage of one classifier coordinate; validation performance guides these choices through their effect on the trained classifier. Labels are generated as $b_e = \operatorname{sign}(a_e^\top \vartheta_{\text{true}} + \delta_{\text{noise}}\zeta_e)$, where $\zeta_e \sim \mathcal{N}(0,1)$ and $\delta_{\text{noise}} = 0.1$. We vary $p = q \in \{20, 50, 100\}$ to examine how problem dimension affects the optimization trajectories.

Data hyper-cleaning. We next study data hyper-cleaning on MNIST (LeCun et al., 1998) and CIFAR-100 (Krizhevsky, 2009). The upper problem learns example weights from clean validation data, while the lower problem fits a classifier to weighted, potentially corrupted training data. Each training example has a globally aligned weight λ_j, mapped to $(0,1)$ by a sigmoid. Agent i evaluates only its shard $\mathcal{I}_i$ and maintains a local softmax classifier W_i:

$$g_i(\lambda_i, W_i) = \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} \sigma(\lambda_{i,j}) \ell_{\text{ce}}(W_i; a_j, \widetilde{b}_j) + \frac{c_r}{2}\|W_i\|_F^2,$$

$$f_i(W_i) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(a,b) \in \mathcal{D}_{\text{val}}} \ell_{\text{ce}}(W_i; a, b),$$

where $\widetilde{b}_j$ may be corrupted, while validation and test labels remain clean. Data are partitioned by clean class before corruption using a Dirichlet distribution with parameter α_{Dir}; larger values produce more homogeneous client datasets (Hsu et al., 2019). The sigmoid weights let validation feedback adjust each training example's influence, while positive quadratic regularization makes the lower-level classifier problem strongly convex. This instantiates the shared training response that DZBO estimates without derivative oracles.

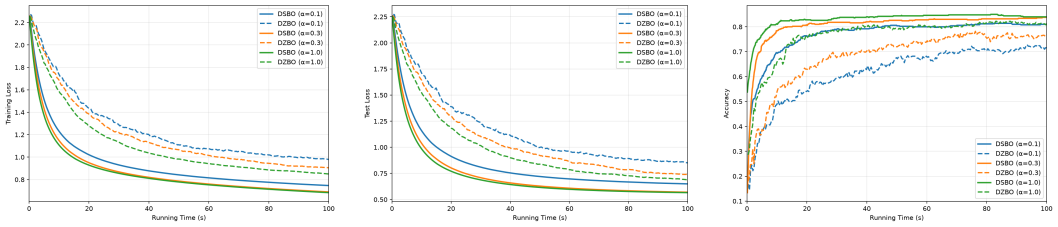
For $y = \mathrm{vec}(W)$, the matrix JHIP path solves $HZ = J^\top$, while the vector path solves $H_{\boldsymbol{\theta}} v = b_{\boldsymbol{\theta}}$ and uses $-\widehat{J}_{i,k}\widetilde{v}_{i,k}^{(N)}$. The corresponding global products agree, although their stochastic recursions differ.

Stationarity metric. The synthetic study reports the stationarity proxy $\log_{10} \|\widetilde{\nabla}\Phi(\bar{\lambda})\|$ against outer iterations. Here $\bar{\lambda}$ is the network-averaged upper variable and $\widetilde{\nabla}\Phi(\bar{\lambda})$ denotes the estimated outer gradient. The remaining evaluation metrics and timing protocol are described in Section 6.

C.2 Additional MNIST Corruption Levels

Purpose and setup. To complement the representative $\chi_{\mathrm{corr}} = 0.2$ setting in Figure 2, we evaluate MNIST hyper-cleaning at $\chi_{\mathrm{corr}} \in \{0.1, 0.4\}$. The formulation, classifier, initialization, and evaluation protocol follow Section 6.2; the corruption rate is the factor varied in this additional comparison.

Reported metrics. Figure 3 presents the training loss, clean-test loss, and test accuracy for MA-DSBO and DZBO. The two rows correspond to corruption rates 0.1 and 0.4, respectively, allowing the optimization trajectories to be compared with the main-text setting using the same metrics.



(a) Train ($\chi_{\mathrm{corr}} = 0.1$) (b) Test ($\chi_{\mathrm{corr}} = 0.1$) (c) Accuracy ($\chi_{\mathrm{corr}} = 0.1$)

(d) Train ($\chi_{\mathrm{corr}} = 0.4$) (e) Test ($\chi_{\mathrm{corr}} = 0.4$) (f) Accuracy ($\chi_{\mathrm{corr}} = 0.4$)

Figure 3: Additional MNIST hyper-cleaning results at $\chi_{\mathrm{corr}} \in \{0.1, 0.4\}$. Each row reports training loss, test loss, and test accuracy for MA-DSBO and DZBO.

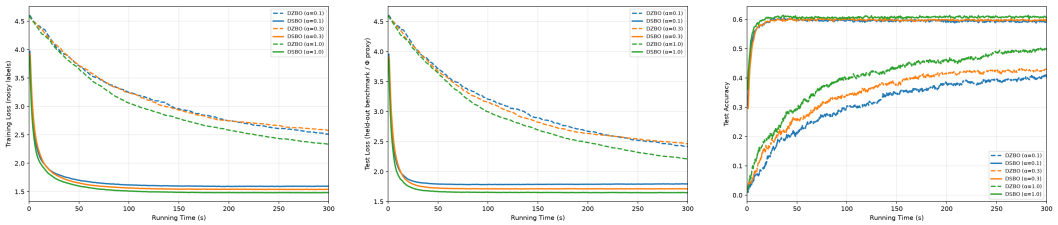
C.3 CIFAR-100 with Fixed Image Features

Purpose and representation. We extend the hyper-cleaning evaluation to CIFAR-100. Each image is mapped to a fixed feature representation using the penultimate layer of a pretrained ResNet-18 (He et al., 2016), with the normalization expected by the pretrained network. Features are extracted before bilevel optimization and remain frozen. Thus, the lower-level variable is the linear softmax classifier on the stored features, and neither gradients nor zeroth-order perturbations are propagated through ResNet-18.

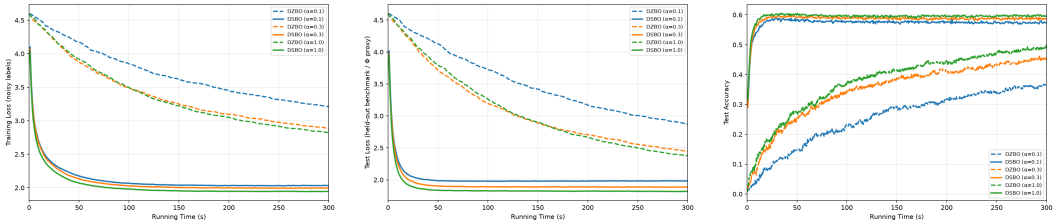
Formulation and implementation. The experiment uses the hyper-cleaning formulation in Section 6.2. Per-example sigmoid weights form the upper variable, corrupted labels enter only the weighted lower-level cross-entropy loss, and clean validation loss defines the upper objective. Agents receive Dirichlet shards formed from clean class labels and maintain globally aligned copies of the example-weight vector. Classifier, tracker, and upper-variable states are exchanged only with neighbors. MA-DSBO uses analytic derivative products, while DZBO uses zeroth-order inner estimates and the matrix JHIP path in Algorithm 1, implemented through (17).

Settings and evaluation. We consider $\chi_{\mathrm{corr}} \in \{0.1, 0.2\}$ and $\alpha_{\mathrm{Dir}} \in \{0.1, 0.3, 1.0\}$. For each corruption rate, Figure 4 reports training loss, held-out test loss, and test accuracy against optimization

time, with all three heterogeneity levels shown in each panel. Evaluation uses the network-averaged classifier, and evaluation time is excluded from optimization time.



(a) Training loss, $\chi_{\text{corr}} = 0.1$. (b) Test loss, $\chi_{\text{corr}} = 0.1$. (c) Test accuracy, $\chi_{\text{corr}} = 0.1$.



(d) Training loss, $\chi_{\text{corr}} = 0.2$. (e) Test loss, $\chi_{\text{corr}} = 0.2$. (f) Test accuracy, $\chi_{\text{corr}} = 0.2$.

Figure 4: CIFAR-100 hyper-cleaning experiments using fixed ResNet-18 features. The first row corresponds to corruption rate $\chi_{\text{corr}} = 0.1$, and the second row corresponds to $\chi_{\text{corr}} = 0.2$. From left to right, the columns show training loss, held-out test loss, and test accuracy. Each panel compares Dirichlet heterogeneity parameters $\alpha_{\text{Dir}} \in \{0.1, 0.3, 1.0\}$.

Results and interpretation. Compared with MNIST, the performance gap between MA-DSBO and DZBO is more pronounced on CIFAR-100. This observation is consistent with the sensitivity of Gaussian zeroth-order estimators to problem dimension (Nesterov & Spokoiny, 2017): a fixed number of random function evaluations can provide a noisier approximation in a larger parameter space. The controlled dimension comparison is given by the synthetic study; the CIFAR-100 study also changes the dataset and feature representation and therefore complements that comparison at the task level.

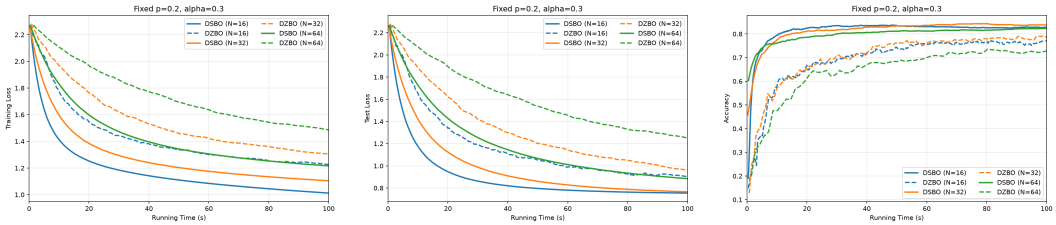
C.4 Effect of the Number of Agents

Purpose and setup. We examine sensitivity to network size by varying $n \in \{16, 32, 64\}$ while fixing the corruption rate at $\chi_{\text{corr}} = 0.2$ and the Dirichlet parameter at $\alpha_{\text{Dir}} = 0.3$. All other experimental settings are kept unchanged. At each network size, MA-DSBO and DZBO receive the same optimization-time budget.

Metrics and interpretation. Figures 5a–5c report training loss, test loss, and test accuracy, respectively. Increasing the number of agents changes both the amount of local data per node and the consensus process on the cycle. The comparison therefore examines their combined effect on optimization under a fixed time budget.

D Limitations and Future Work

Assumptions and model scope. Our analysis allows the upper-level objective to be nonconvex, but it assumes that every local lower-level objective $g_i(x, \cdot)$ is uniformly strongly convex; consequently, the aggregate lower-level problem is also strongly convex. The stated smoothness and stochastic-oracle regularity conditions are also imposed. Assumption 3.2 requires $\nabla_y g_i$ to be L_g-Lipschitz continuous in the joint variable (x, y); a globally Lipschitz full joint gradient ∇g_i is not required. The L_H, L_J bounds separately control the variation of the Hessian and cross blocks. The solution-gradient moment condition (2) is an additional uniform assumption over every local-state instance $\mathbf{X}$. It bounds local gradient heterogeneity and oracle noise at $y^*_{\mathbf{X}}$ and does not follow from strong convexity alone. Together with Lemma A.1, it gives the inner solver's state-uniform moment bounds and



(a) Training loss. (b) Test loss. (c) Test accuracy.

Figure 5: Performance of MA-DSBO and DZBO with different numbers of agents $n \in \{16, 32, 64\}$, while fixing $\chi_{\text{corr}} = 0.2$ and $\alpha_{\text{Dir}} = 0.3$.

supports warm-start stability; the auxiliary-oracle theorems do not require the solution-gradient moment bound. Lower-level strong convexity ensures that $y^*(x)$ is uniquely defined and makes the auxiliary JHIP problem strongly convex, while the Lipschitz and moment conditions permit explicit control of the zeroth-order bias and variance. These assumptions provide sufficient conditions for the nonasymptotic guarantees, but they do not cover end-to-end training of overparameterized neural networks or lower-level problems with multiple solutions. Extending the analysis to nonconvex or locally regular lower-level objectives, possibly in terms of stationarity or solution-set stability, is an important direction for future work.

Vector states and dimension dependence. As in MA-DSBO (Chen et al., 2023), vector states reduce auxiliary storage and payload from order pq to order q, while the matrix formulation is useful when a reusable inverse map is required. Gaussian direction noise still depends on dimension. Appendix B proves smaller product second-moment bounds for direct products under the same regularity, but this does not by itself establish a proportionate improvement in the complete bilevel complexity. Structured perturbations, randomized sketching, and adaptive query allocation remain useful directions. The true-target vector bound also requires the specified right-hand-side refreshing and smoothing schedules; it does not transfer to a single frozen stochastic right-hand side.

Common randomness and oracle noise. The smoothing-radius-uniform moment bounds use common random numbers within each finite-difference stencil: the same stochastic realization is evaluated at its paired perturbation points. This coupling is natural when a simulator can be replayed with a fixed seed or when the stochastic sample can be held fixed across queries, but it may be unavailable in black-box services that return independently noisy evaluations. With independent query noise, the noise terms can be amplified as the smoothing radii decrease, and the present moment bounds no longer apply without modification. Future work may combine batching, control variates, multilevel sampling, or noise-aware smoothing schedules to obtain guarantees under weaker oracle couplings.

Conditional moments, momentum, and network extensions. The proof assumes that oracle samples and perturbation directions are fresh conditionally on the algorithmic history and uses conditional first- and second-moment bounds. The analyzed recursions do not include momentum; therefore, the current theory does not establish conditional-on-history convergence for momentum-accelerated variants. Such a guarantee would require a Lyapunov analysis that couples the momentum state with lower-level error, JHIP tracking error, and network disagreement. It is also of interest to study directed, time-varying, or asynchronous networks, as well as communication compression, privacy mechanisms, and robustness to unreliable or adversarial function-value queries. These extensions would broaden the applicability of decentralized zeroth-order bilevel optimization beyond the fixed, connected, undirected network and oracle model considered here.