FEDBIC: FEDERATED BILEVEL CONSTRAINED OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bilevel optimization has emerged as a fundamental framework for hierarchical machine learning problems, such as hyperparameter optimization, meta-learning, adversarial training, and neural architecture search. Extending this framework to federated settings allows collaborative training between distributed clients without sharing raw data. However, existing federated bilevel methods assume unconstrained lower-level problems, excluding applications with fairness constraints, resource budgets, or safety requirements. We propose FedBiC, the first federated bilevel algorithm that is capable of handling a convex lower-level objective together with general client-specific convex inequality constraints, using a gap function reformulation. FedBiC is single-loop and Hessian-free, supports partial client participation, and requires no bound on client heterogeneity. We establish an $O(1/\sqrt{pK})$ convergence rate to a stationary point of the penalized objective under a fixed penalty and partial participation, where K denotes the number of communication rounds. With an increasing penalty parameter and full participation, FedBiC attains an $O(1/K^{(1-2\rho)/2})$ stationarity rate for the penalized objective, together with an $O(K^{-\rho})$ rate on the clientwise lower-level gap violation. Experiments on federated constrained LASSO, fair few-shot meta-learning, and federated steering under label noise are conducted to validate our approach.

1 INTRODUCTION

Bilevel optimization (BLO) has emerged as a fundamental framework for modeling hierarchical decision-making problems in machine learning, where an upper-level (UL) objective relies on the solution of a lower-level (LL) optimization problem (Dempe, 2002; Colson et al., 2007). This nested structure arises naturally in hyperparameter optimization (Franceschi et al., 2018; Lorraine et al., 2020), meta-learning (Finn et al., 2017; Rajeswaran et al., 2019), neural architecture search (Liu et al., 2019), data reweighting (Ren et al., 2018; Shu et al., 2019), and adversarial training (Madry et al., 2018). BLO algorithms broadly fall into three categories. Iterative differentiation (ITD) methods unroll the LL optimization trajectory to compute hypergradients via backpropagation through the LL iterations (Maclaurin et al., 2015; Franceschi et al., 2018). Approximate implicit differentiation (AID) methods instead differentiate through the LL optimality conditions using the implicit function theorem, avoiding explicit unrolling (Pedregosa, 2016; Lorraine et al., 2020; Rajeswaran et al., 2019). Value function and penalty-based methods reformulate the bilevel problem as a single-level surrogate by characterizing the LL solution through its optimal value function or a penalty term (Liu et al., 2021; Yao et al., 2024). More recently, single-loop and fully first-order methods have eliminated nested optimization and Hessian computations entirely (Hong et al., 2023; Kwon et al., 2023; Chen et al., 2025; Liu et al., 2022), achieving strong computational efficiency in large-scale settings.

Despite these advancements, these BLO methods are designed for centralized settings where all data reside on a single machine. However, in many real-world scenarios, training data is inherently distributed across multiple agents, each holding sensitive private data that cannot be shared due to regulatory, ethical, or communication constraints. Federated learning (FL) addresses this challenge by allowing collaborative model training across distributed clients without sharing raw data (McMahan et al., 2017; Kairouz et al., 2021). The combination of FL and BLO, known as federated bilevel optimization, is therefore a natural and practically important extension. Many federated scenarios inherently involve hierarchical decision structures. Personalized FL, for instance, naturally decomposes into two levels where the UL learns a shared global initialization while the LL adapts to each

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.

client's local distribution (Fallah et al., 2020) (Tarzanagh et al., 2022). Federated hyperparameter tuning, where clients search for optimal model configurations collaboratively without sharing their local data, similarly requires solving nested optimization problems across heterogeneous clients (Li et al., 2023; Huang et al., 2023). This confluence has spurred a growing line of work on federated bilevel algorithms (Tarzanagh et al., 2022; Li et al., 2023; Huang et al., 2023; Huang, 2022; Yang et al., 2023b), inheriting the communication efficiency and privacy guarantees of FL while tackling the algorithmic challenges of BLO.

However, existing federated bilevel approaches assume an unconstrained LL problem, and cannot be applied to a wide range of important applications where the LL optimization is subject to explicit constraints. Prominent examples include fairness-aware learning, where the LL problem seeks model parameters satisfying demographic parity constraints (Hardt et al., 2016; Agarwal et al., 2018); robustness-constrained learning, where the LL model must maintain bounded loss on adversarially perturbed or corrupted inputs (Madry et al., 2018; Cotter et al., 2019); and safe reinforcement learning, where the LL must satisfy safety requirements throughout training (Achiam et al., 2017; Yang et al., 2023a). In all of these settings, existing federated bilevel methods cannot be applied, as they provide no mechanism to enforce LL inequality constraints.

Incorporating LL constraints into federated BLO presents several key challenges. First, the computation of an implicit gradient, which is the crucial part of most bilevel approaches, becomes significantly more complex when the LL problem is constrained, because it requires differentiating through a constrained optimization mapping (Gould et al., 2016). The federated setting makes these difficulties worse. Heterogeneity now acts at four different levels instead of just one. First, clients have different constrained solution maps $\mathcal{S}_i(x)$. Second, clients with similar losses might still have different active constraints and multipliers. Third, an inactive client's state becomes stale as the global variable x changes. This delay silently alters the client's lower-level problem. Finally, the aggregation process must control constraint disagreement and staleness. It cannot just focus on loss dissimilarity alone (Karimireddy et al., 2020; Li et al., 2020). Furthermore, ensuring constraint satisfaction in a distributed manner while maintaining convergence guarantees requires careful design of algorithms.

1.1 Our Approach: FedBiC

We address these challenges with FedBiC, the first federated algorithm for bilevel optimization with inequality-constrained LL problems. FedBiC is single-loop and Hessian-free, with LL primal and dual states retained locally. Building on the gap function and KKT characterization of (Yao et al., 2025), we address two challenges absent from their centralized analysis. First, client-specific constraints and multipliers introduce gradient variation that objective heterogeneity alone cannot capture. A server-side gradient table avoids requiring a bound on this variation by replacing subset sampling error with staleness error, which we control through clientwise smoothness and iterate drift. Second, shared-variable updates change inactive clients' LL problems. We control this drift through a normalized penalty analysis and derive an explicit participation threshold $p_{\min}$. Together with separate inner tracking and outer descent steps, these mechanisms yield an $O(1/\sqrt{pK})$ convergence rate to a stationarity point of the penalized objective with a fixed penalty and under partial participation, where p is the client participation rate. With an increasing penalty and full participation, we obtain a rate of $O(K^{-(1-2\rho)/2})$ to a stationarity point for the penalized objective and an $O(K^{-\rho})$ clientwise gap violation. We additionally show that a limit point of a selected feasible-stationary subsequence is stationary for the original constrained bilevel problem under the additional smoothness, interiority, and regularity conditions, including LICQ, strict complementarity, and the strong second-order sufficient condition. Experiments spanning sparsity, fairness, and label-noise robustness evaluate UL performance and constraint satisfaction against penalty and mixed loss formulations.

1.2 Related Works

There have been multiple recent works related to federated BLO. FedNest (Tarzanagh et al., 2022) proposes a federated alternating gradient method for bilevel, minimax, and compositional optimization but relies on multiple inner-loop iterations and Hessian-vector products. FedBiOAcc (Li et al., 2023) improves communication complexity to $O(\epsilon^{-1})$ through momentum based variance reduction, while FedMBO (Huang et al., 2023) establishes a linear speedup guarantee for non-i.i.d. settings. AdaFBiO (Huang, 2022) incorporates adaptive learning rates, and SimFBO (Yang et al., 2023b) offers a sub-loop-free design with improved communication efficiency. In the decentralized setting, DSBO (Chen et al., 2023) and SPARKLE (Zhu et al., 2024) address BLO over networks with improved per-iteration complexities. However, all of these methods assume an unconstrained LL problem. For variance reduction side, FedVARP (Jhunjhunwala et al., 2022) eliminates partial-participation variance in

Table 1: Comparison of Federated Bilevel Optimization Methods

Method	Setting	LL Constraints	LL Strong Convexity	Hessian-free	Single-loop	Complexity
FedNest Tarzanagh et al. (2022)	Federated	None	Required	✗	✗	$O(\epsilon^{-2})$
AdaFBiO Huang (2022)	Federated	None	Required	✗	✓	$O(\epsilon^{-2})$
FedBiOAcc Li et al. (2023)	Federated	None	Required	✗	✓	$O(\epsilon^{-1})$
SimFBO Yang et al. (2023b)	Federated	None	Required	✗	✓	$O(\epsilon^{-2})$
DSBO Chen et al. (2023)	Decentralized	None	Required	✗	✓	$O(\epsilon^{-2})$
BiC-GAFFA Yao et al. (2025)	Centralized	General	✗	✓	✓	$O(\epsilon^{-2})$
FedBiC (Ours)	**Federated**	**General**	✗	✓	✓	$O(\epsilon^{-2})$

LL = Lower-Level. ✓ = Yes/Not Required. ✗ = No/Required. Complexity is for finding ϵ-stationary point.

single-level FL via server-side gradient storage. We adapt this technique to the federated bilevel constrained setting. On the centralized constrained bilevel side, the paper (Khanduri et al., 2023) tackles linearly constrained LL problems through smoothed implicit gradients, and (Yao et al., 2024) develop a proximal Lagrangian value function approach. (Xiao et al., 2023) address equality-constrained LL problems with federated applications. However, inequality constraints require fundamentally different tools like KKT conditions and gap functions, making the extension non-trivial. Most closely related, BiC-GAFFA (Yao et al., 2025) handles general convex LL inequality constraints via regularized gap functions in a single-loop, Hessian-free manner, but remains limited to the centralized setting. FedBiC is the first to extend this capability to the federated setting. Table 1 presents the extended comparisons between the methods.

2 Problem Formulation

We consider an FL setting with N distributed clients and a central parameter server. Each client holds private local data that is never shared. The goal is to collaboratively solve a BLO problem across all clients, where the UL is coordinated by the server and the LL is solved locally by each client using its own private data. The federated BLO problem is formally stated as follows:

$$\min_{x,\{x_i,y_i\}} \frac{1}{N}\sum_{i=1}^{N} F_i(x_i,y_i) \quad \text{s.t.} \quad x_i = x,\ y_i \in S_i(x_i),\ \forall i \in [N], \tag{1}$$

where for each client i, the LL solution set $S_i(x_i)$ is defined by:

$$S_i(x_i) = \arg\min_{y\in\mathcal{Y}} \{f_i(x_i,y) \mid g_i(x_i,y) \le 0\}. \tag{2}$$

Here, $x \in \mathcal{X} \subseteq \mathbb{R}^d$ is the UL variable shared across all clients, $y_i \in \mathcal{Y} \subseteq \mathbb{R}^m$ is the LL variable local to client i, $F_i : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is the UL objective of client i, $f_i : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is the LL objective of client i, and $g_i : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^{p_i}$ is the LL inequality constraint function of client i. For instance, consider fair personalized FL, where $x \in \mathbb{R}^d$ is the shared model learned by the server and broadcast to all clients. Each client i locally adapts x to its private data by solving a regularized problem. The UL objective F_i evaluates the adapted model on held-out query data, f_i is the local training loss, and $g_i(y_i)$ enforces a demographic parity constraint on the adapted model with tolerance ε, which is jointly optimized by the server.

To enable theoretical analysis and algorithm development, we make the following assumptions, which are standard in the BLO literature and satisfied by many practical machine learning problems.

Assumption 2.1. For each client $i \in [N]$: (i) the sets $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are nonempty, compact, and convex; (ii) for every $x \in \mathcal{X}$, the functions $f_i(x,\cdot)$ and $g_i(x,\cdot)$ are convex on $\mathcal{Y}$; (iii) for every $x \in \mathcal{X}$ with feasible lower-level constraints, the solution set $S_i(x)$ is nonempty; (iv) *(Uniform Slater condition)* There exists $\sigma_i > 0$ such that for every $x \in \mathcal{X}$, there exists a strictly feasible point $\hat{y}_i(x) \in \mathrm{ri}(\mathcal{Y})$ satisfying $g_i(x,\hat{y}_i(x)) \le -\sigma_i \mathbf{1}_{p_i}$ componentwise, where $\mathrm{ri}(\mathcal{Y})$ denotes the relative interior of $\mathcal{Y}$, $\sigma_i > 0$ is the strict feasibility margin, and $\mathbf{1}_{p_i} \in \mathbb{R}^{p_i}$ is the all-ones vector ensuring all p_i constraints are strictly satisfied simultaneously.

Assumption 2.2. For each $i \in [N]$, F_i, f_i, g_i are continuously differentiable on $X \times Y$. Constraint Jacobians are row-indexed by constraint: $\nabla_y g_i(x,y) \in \mathbb{R}^{p_i\times m}$ and $\nabla_x g_i(x,y) \in \mathbb{R}^{p_i\times d}$.

Assumption 2.3. For each client $i \in [N]$, every $x \in X$ and every $y_i \in \mathcal{S}_i(x)$, there exists a Lagrange multiplier $\lambda_i \in \mathbb{R}^{p_i}_+$ satisfying the KKT conditions:

$$0 \in \nabla_y f_i(x,y_i) + \nabla_y g_i(x,y_i)^\top \lambda_i + \mathcal{N}_{\mathcal{Y}}(y_i) \quad \text{and} \quad \lambda_i^\top g_i(x,y_i) = 0, \tag{3}$$

where $\mathcal{N}_{\mathcal{Y}}(y_i)$ is the normal cone to $\mathcal{Y}$ at y_i, $\lambda_i^\top g_i(x,y_i) = 0$ is the complementary slackness condition, and $\mathcal{M}_i(x,y_i)$ denotes the set of all multipliers $\lambda_i \in \mathbb{R}^{p_i}_+$ satisfying equation 3.

The federated bilevel problem equation 1 enforces a shared global model via the consensus constraint $x_i = x$, while client-specific functions (F_i, f_i, g_i) capture data heterogeneity and privacy is preserved by restricting the LL problem to local computation. Compared to the BiC-GAFFA (Yao et al., 2025), which assumes full data access on a single machine, the federated setting introduces three additional challenges. First, partial participation causes client drift, where auxiliary variables become stale while x evolves globally. Second, statistical heterogeneity across clients produces biased gradient estimates due to differing local objectives and constraint functions. Lastly, per-client inequality constraints must be enforced locally without sharing raw data.

3 FEDBIC ALGORITHM

3.1 PRELIMINARIES: THE REGULARIZED GAP FUNCTION

Our algorithm builds on the doubly regularized gap function of (Yao et al., 2025). For each client i,

$$G_{\gamma,i}(x, y_i, z_i) := \max_{\theta_i \in Y,\, \lambda_i \in \mathbb{R}^{p_i}_+} \left\{ \mathcal{L}_i(x, y_i, \lambda_i) - \tfrac{\|\lambda_i - z_i\|^2}{2\gamma_2} - \mathcal{L}_i(x, \theta_i, z_i) - \tfrac{\|\theta_i - y_i\|^2}{2\gamma_1} \right\}, \tag{4}$$

where $\mathcal{L}_i(x, y_i, \lambda_i) := f_i(x, y_i) + \lambda_i^\top g_i(x, y_i)$ is the Lagrangian of client i's LL problem, $z_i \in \mathbb{R}^{p_i}_+$ approximates the optimal KKT multiplier, and $\gamma_1, \gamma_2 > 0$ are proximal parameters ensuring unique maximizers and differentiability. Under Assumption 2.1, $G_{\gamma,i} \geq 0$, with equality if and only if $y_i \in \mathcal{S}_i(x)$ and $z_i \in \mathcal{M}_i(x, y_i)$. The characterization is exact: a zero gap value certifies that y_i solves client i's constrained LL problem and that z_i is a corresponding KKT multiplier. Thus, the gap measures the departure of a client's primal-dual pair from lower-level optimality, accounting for both feasibility and the KKT conditions. Under Assumptions 2.1–2.2, $G_{\gamma,i}$ is continuously differentiable with closed-form gradients requiring only first-order information about f_i and g_i, and no Hessians or implicit differentiation. These gradients are expressed through the unique maximizers of equation 4,

$$\theta_i^* := \operatorname*{argmin}_{\theta \in \mathcal{Y}} \left\{ f_i(x, \theta) + z_i^\top g_i(x, \theta) + \frac{\|\theta - y_i\|^2}{2\gamma_1} \right\}, \quad \lambda_i^* := \mathrm{Proj}_{\mathbb{R}^{p_i}_+} \left(z_i + \gamma_2 g_i(x, y_i) \right). \tag{5}$$

which each client evaluates locally. The implicit constraint $y_i \in \mathcal{S}_i(x)$ can therefore be replaced by the explicit smooth inequality $G_{\gamma,i} \leq 0$, yielding a single-level reformulation equivalent to the original problem equation 1. Under the uniform Slater condition, the lower-level KKT multipliers admit a uniform bound. We denote this bound by

$$r := \max_{i \in [N],\, x \in \mathcal{X}} \|\mu_i^\star(x)\|_\infty < \infty$$

and define $\mathcal{Z}_i := [0, r]^{p_i}$. Since $\mathcal{Z}_i$ is an ℓ_∞ box with radius r, we define

$$p_{\max} := \max_{i \in [N]} p_i, \qquad r_2 := r\sqrt{p_{\max}}.$$

Therefore, $\|z\| \leq r_2$ for every $z \in \mathcal{Z}_i$ and every $i \in [N]$. Penalizing the gap constraint gives

$$\min_{x \in X,\, \{y_i \in Y,\, z_i \in \mathcal{Z}_i\}} \frac{1}{N} \sum_{i=1}^{N} \left[F_i(x, y_i) + c \cdot G_{\gamma,i}(x, y_i, z_i) \right], \tag{6}$$

with penalty parameter $c > 0$. As c grows, global minimizers of equation 6 attain vanishing gap values, and along a schedule $c_k \to \infty$, their accumulation points solve the reformulation exactly. These results concern global minimizers rather than algorithmic iterates. The guarantees for the iterates FedBiC generates are stated in Section 4. Crucially, each $G_{\gamma,i}$ depends only on client i's local data and variables, enabling parallel computation without sharing raw data. The gap construction, its KKT characterization, and its gradient formulas are inherited from the centralized setting. Our contributions are a single-loop and Hessian-free two-timescale algorithm with local primal-dual states, a server-side gradient table that avoids subset-sampling error, and a convergence analysis that controls table staleness and inactive-client drift. This analysis provides an explicit participation threshold and convergence guarantees under both a fixed penalty with partial participation and an increasing penalty with full participation. The supporting proofs and technical details for this subsection are provided in Appendix C.

3.2 THE ALGORITHM

FedBiC solves equation 6 using projected gradient descent with a potentially increasing penalty parameter c_k; it is presented in Algorithm 1. Since computing the exact gradient of $G_{\gamma,i}$ requires solving equation 5 exactly at each iteration, we instead maintain auxiliary tracking variables $\bar{\theta}_i \in \mathcal{Y}$

and $\bar{\lambda}_i \in \mathbb{R}^{p_i}_+$, initialized at $(\theta_i^k, \lambda_i^k)$, that approximate θ_i^* and λ_i^* via E inner gradient steps. The primal and dual iterates $y_i \in \mathcal{Y}$ and $z_i \in [0, r]^{p_i}$ are then updated in a single outer step using the current approximations $(\bar{\theta}_i, \bar{\lambda}_i)$. Consistent with the personalized formulation equation 1, in which the consensus constraint applies to the shared variable alone, each client's LL primal, dual, and auxiliary states remain local at all times. The only quantity communicated to the server is a d-dimensional gradient with respect to x, and the only quantity broadcast is x^{k+1}. Client-specific constraints, multipliers, and active sets are therefore never revealed. This stands in contrast to (Yao et al., 2025), which requires full data access on a single machine.

The inner loop updates $\bar{\theta}_i$ via gradient descent on the θ_i-minimization in equation 5, and $\bar{\lambda}_i$ via gradient ascent on the dual subproblem of the gap function equation 4. A key design choice is that the inner loop uses z_i^k rather than $\bar{\lambda}_i$ in the $\bar{\theta}_i$ update, creating a two-timescale structure where $(\bar{\theta}_i, \bar{\lambda}_i)$ are tracked rapidly in the inner loop while (y_i, z_i) are updated slowly in the outer step. This separation is essential for the contraction analysis in Section 4 and enables provable convergence with a fixed number E of local steps per round, independent of any target accuracy. The outer step substitutes $(\bar{\theta}_i, \bar{\lambda}_i)$ for $(\theta_i^*, \lambda_i^*)$ in the gradient formulas of Lemma C.2 to update (y_i, z_i). Full update directions are provided in Algorithm 1.

Algorithm 1 FedBiC

Require: $x^0 \in X$, $\{y_i^0, z_i^0\}_{i=1}^N$ with $\theta_i^0 = y_i^0$, $\lambda_i^0 = z_i^0$, gradient table $g_{x,i}^0 = 0\ \forall i$, step sizes $\alpha_k, \beta_k, \eta_k, \alpha_x$, penalty c_0, proximal parameters γ_1, γ_2.

for $k = 0, 1, 2, \ldots, K-1$ **do**
  Server selects subset $S_k \subseteq [N]$ with $|S_k| = \lceil pN \rceil$, $p > p_{\min}$; broadcasts x^k to S_k.
  for each $i \in S_k$ **in parallel do**
    $(\bar{\theta}_i, \bar{\lambda}_i) = (\theta_i^k, \lambda_i^k)$
    for $e = 1, \ldots, E$ **do**
      $d_{\bar{\theta}_i} = \nabla_\theta [f_i(x^k, \bar{\theta}_i) + (z_i^k)^\top g_i(x^k, \bar{\theta}_i)] + \frac{1}{\gamma_1}(\bar{\theta}_i - y_i^k)$, $\quad \bar{\theta}_i \leftarrow \mathrm{Proj}_Y(\bar{\theta}_i - \eta_k\, d_{\bar{\theta}_i})$
      $d_{\bar{\lambda}_i} = -(g_i(x^k, y_i^k) - \frac{1}{\gamma_2}(\bar{\lambda}_i - z_i^k))$, $\quad \bar{\lambda}_i \leftarrow \mathrm{Proj}_{\mathbb{R}^{p_i}_+}(\bar{\lambda}_i - \eta_k\, d_{\bar{\lambda}_i})$
    end for
    $d_{y_i} = \nabla_y F_i(x^k, y_i^k) + c_k \big[\nabla_y f_i + \nabla_y g_i^\top \bar{\lambda}_i - \frac{1}{\gamma_1}(y_i^k - \bar{\theta}_i)\big]$, $\quad y_i^{k+1} = \mathrm{Proj}_Y(y_i^k - \alpha_k\, d_{y_i})$
    $d_{z_i} = c_k \big[\frac{1}{\gamma_2}(\bar{\lambda}_i - z_i^k) - g_i(x^k, \bar{\theta}_i)\big]$, $\quad z_i^{k+1} = \mathrm{Proj}_{[0,r]^{p_i}}(z_i^k - \beta_k\, d_{z_i})$
    $(\theta_i^{k+1}, \lambda_i^{k+1}) = (\bar{\theta}_i, \bar{\lambda}_i)$, $\quad$ Compute $d_{x_i}^k$ via equation 7 and send to server.
    $g_{x,i}^{k+1} = d_{x_i}^k / c_k$
  end for
  for each $i \notin S_k$ **do**
    $(y_i^{k+1}, z_i^{k+1}, \theta_i^{k+1}, \lambda_i^{k+1}) = (y_i^k, z_i^k, \theta_i^k, \lambda_i^k)$; $\quad g_{x,i}^{k+1} = g_{x,i}^k$
  end for
  $d_x^k = \frac{c_k}{N} \sum_{i=1}^N g_{x,i}^{k+1}$, $\quad x^{k+1} = \mathrm{Proj}_X(x^k - \alpha_x\, d_x^k)$, $\quad$ Update c_{k+1}
  Server broadcasts x^{k+1} to S_{k+1}.
end for

The server stores each client's last x gradient in a table $\{g_{x,i}\}_{i=1}^N$ (Jhunjhunwala et al., 2022). Participating clients refresh their entry while others remain stale. An inactive client therefore continues to influence the shared update through its stored gradient; the analysis controls the error caused by using this older information. The x update aggregates all N entries, and when $p = 1$ all entries are fresh. The additional server storage is $O(Nd)$, and the per-round communication is one d-dimensional vector in each direction per participating client, no lower-level state is transmitted. Each participating client's contribution to the aggregate is

$$\begin{aligned} d_{x_i}^k = {} & \nabla_x F_i(x^k, y_i^{k+1}) + c_k \big[\nabla_x f_i(x^k, y_i^{k+1}) + \nabla_x g_i(x^k, y_i^{k+1})^\top \lambda_i^{k+1} \\ & - \nabla_x f_i(x^k, \theta_i^{k+1}) - \nabla_x g_i(x^k, \theta_i^{k+1})^\top z_i^{k+1}\big]. \end{aligned} \tag{7}$$

The subset S_k is selected uniformly at random. Non-participating clients retain their local variables unchanged. For our penalty formulation, this staleness yields a non-positive merit change (Lemma D.9), enabling convergence under partial participation when the fraction of selected clients $p = |S_k|/N$ exceeds a threshold $p_{\min}$ that depends on the proximal parameters γ_1, γ_2 and inner step size η (Theorem D.14). This guarantee is absent in existing federated bilevel methods (Tarzanagh et al., 2022; Yang et al., 2023b).

Compared to the centralized BiC-GAFFA (Yao et al., 2025), FedBiC introduces three algorithmic modifications. First, each client maintains its own lower-level primal, dual, and auxiliary states, so that client-specific constraints, multipliers, and active sets remain local; BiC-GAFFA tracks a single lower-level primal-dual pair from current centralized information. Second, the server maintains a gradient table that reuses each inactive client's most recent contribution, so that the x-update aggregates all N clients even when only a fraction participates. Third, outer step sizes scale as $\alpha_k = \alpha/c_k$ while the inner step size η remains constant, explicitly separating the two timescales and preventing divergence under a growing penalty.

4 Convergence Analysis

In this section, we establish non-asymptotic convergence guarantees for FedBiC. Table 5 in Appendix A collects the notation. Beyond Assumptions 2.1–2.3, we require three additional assumptions: the standard smoothness and lower-boundedness conditions of nonconvex first-order analysis, which supply the descent lemma, the Lipschitz auxiliary mappings, and the inner-loop contraction. We assume no bound on client heterogeneity. Definition 4.4 introduces the usual heterogeneity constants only so that the effect of the server-side gradient table can be quantified, and neither rate below depends on them. Proofs for this section are shown in Appendix D.

Assumption 4.1 (UL Smoothness)**.** For each client $i \in [N]$: F_i is L_F smooth on $X \times Y$ and F_i is bounded below, i.e., $\underline{F}_i := \inf_{(x,y)} F_i(x,y) > -\infty$.

Assumption 4.2 (Lower-Level Smoothness)**.** For each client $i \in [N]$: (1) for every $x \in \mathcal{X}$, $f_i(x,\cdot)$ is convex on $\mathcal{Y}$; (2) f_i is L_f-smooth on $\mathcal{X} \times \mathcal{Y}$.

Assumption 4.3 (Lower-Level Constraints)**.** For each client $i \in [N]$: (1) for every $x \in \mathcal{X}$, $g_i(x,\cdot)$ is convex on $\mathcal{Y}$; (2) g_i is L_g-Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$; (3) $\nabla_x g_i$ and $\nabla_y g_i$ are L_{gx}- and L_{gy}-Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$, respectively.

Definition 4.4 (Data Heterogeneity)**.** Let F, f, g, and G_γ denote the corresponding client averages. We define

$$\zeta_F^2 := \sup_w \frac{1}{N}\sum_{i=1}^N \|\nabla F_i(w) - \nabla F(w)\|^2, \qquad \zeta_f^2 := \sup_{x,\theta} \frac{1}{N}\sum_{i=1}^N \|\nabla f_i(x,\theta) - \nabla f(x,\theta)\|^2,$$

$$\zeta_g^2 := \sup_{x,\theta} \frac{1}{N}\sum_{i=1}^N \|g_i(x,\theta) - g(x,\theta)\|^2, \qquad \zeta_G^2 := \sup_w \frac{1}{N}\sum_{i=1}^N \|\nabla G_{\gamma,i}(w) - \nabla G_\gamma(w)\|^2.$$

Each supremum is taken over the relevant compact domain. Under Assumptions 2.1 and 2.2 and Lemma C.2, the functions inside these suprema are continuous, so all four constants are finite. These constants are used only in the sampling-error comparison of Lemma D.11. The server-side gradient table avoids this sampling term.

We measure our algorithmic progress using a per-client Lyapunov-style merit function that combines the penalized objective with auxiliary tracking errors. Unlike the centralized merit function of (Yao et al., 2025), our formulation decomposes across clients, so that each client's contribution is evaluated locally and no averaging across clients is required. We define a penalized objective $\phi_{c,i}$ for each client:

$$\phi_{c,i}(x, y_i, z_i) := \tfrac{1}{c}(F_i(x, y_i) - \underline{F}_i) + G_{\gamma,i}(x, y_i, z_i), \quad \text{where } \underline{F}_i := \inf_{x,y} F_i(x,y). \tag{8}$$

Based on this, we define the local merit function $V_{k,i}$ and the global merit function V_k^G:

$$V_{k,i} := \phi_{c_k,i}(w_i^k) + C_\theta\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + C_\lambda\|\lambda_i^k - \lambda_i^*(w_i^k)\|^2, \tag{9}$$

$$V_k^G := \tfrac{1}{N}\textstyle\sum_{i=1}^N V_{k,i}, \tag{10}$$

where $w_i^k := (x^k, y_i^k, z_i^k)$ is the local iterate and the constants are $C_\theta = L_f + r_2 L_{g_y} + \frac{1}{\gamma_1}$, $C_\lambda = L_g + \frac{1}{\gamma_2}$. Since Algorithm 1 broadcasts a single shared x^k at each round, $x_i^k \equiv x^k$ for all clients. Because the local variables (y_i, z_i) are never averaged across clients, the same merit function serves both the participation analyses. The only additional quantity required under partial participation is the gradient-table staleness introduced in Appendix D.1. For the convergence analysis we write

$$w^k := (x^k, u_1^k, \ldots, u_N^k) \in \mathcal{W} := \mathcal{X} \times \textstyle\prod_{i=1}^N (\mathcal{Y} \times \mathcal{Z}_i), \qquad u_i^k := (y_i^k, z_i^k),$$

for the concatenated iterate, and

$$\Theta^k := \frac{1}{N}\sum_{i=1}^N \|\theta_i^k - \theta_i^*(w_i^k)\|^2, \qquad \Lambda^k := \frac{1}{N}\sum_{i=1}^N \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2$$

for the averaged auxiliary tracking errors. The merit function V_k^G serves as a Lyapunov function, where a successful iteration reduces V_k^G in expectation, combining progress in the objective value, constraint satisfaction, and auxiliary variable accuracy. Since Algorithm 1 takes projected steps on $\mathcal{W}$, we measure stationarity by the residual,

$$R_k := \operatorname{dist}\big(0, \nabla\psi_{c_k}(w^{k+1}) + \mathcal{N}_{\mathcal{W}}(w^{k+1})\big),\ \psi_c(w) := \tfrac{1}{N}\textstyle\sum_{i=1}^{N}[F_i(x,y_i) + cG_{\gamma,i}(w_i)], \quad (11)$$

which is zero exactly at a stationary point of ψ_{c_k} over $\mathcal{W}$. Under partial participation an inactive client's blocks are not produced by a projected step, so equation 11 is unavailable and Theorem 4.5 reports a gradient mapping instead. We write $\phi_c(w) := \frac{1}{N}\sum_{i=1}^{N}\phi_{c,i}(x,u_i)$ for the aggregate of the normalized objectives equation 8 and let $D := \operatorname{diag}(I_d, NI, \ldots, NI)$ be the block scaling that is the identity on the shared block and N on each local block, so that $D\nabla\phi_c(w)$ carries exactly the directions along which Algorithm 1 moves the corresponding blocks. The block-scaled gradient mapping is

$$\mathcal{G}_\alpha(w^k) \ := \ \tfrac{1}{\alpha}\big(w^k - \operatorname{Proj}_{\mathcal{W}}(w^k - \alpha D\nabla\phi_{c_k}(w^k))\big). \quad (12)$$

The factor N compensates for the averaging in the global objective, so each local block uses its own client's gradient. Since D is positive definite, $\mathcal{G}_\alpha(w^k) = 0$ if and only if $-\nabla\phi_{c_k}(w^k) \in \mathcal{N}_{\mathcal{W}}(w^k)$, and since $\phi_c = \frac{1}{c}\psi_c - \frac{1}{cN}\sum_i \underline{F}_i$ gives $\nabla\phi_c = \frac{1}{c}\nabla\psi_c$, the two measures vanish at the same points; $\|\mathcal{G}_\alpha\|$ moreover controls equation 11 wherever the latter applies.

Theorem 4.5 (Fixed penalty). *Under Assumptions 4.1–4.3, with fixed penalty $c_k = c > 0$, participation rate $p \geq p_0$ for some fixed $p_0 > p_{\min}$, step sizes $\alpha_k = \beta_k = \alpha_x = \alpha/c$ and $\eta_k = \eta \in (0, \bar{\eta}]$, and $E \geq E_0$ inner iterations, there is $\bar{\alpha}_{\mathrm{pp}} > 0$ depending on the problem constants and on p_0 such that for every $\alpha \in (0, \bar{\alpha}_{\mathrm{pp}}]$ FedBiC achieves:*

$$\min_{0\leq k\leq K-1} \mathbb{E}\big[\|\mathcal{G}_\alpha(w^k)\|\big] \ = \ O\left(\frac{1}{\sqrt{pK}}\right), \quad (13)$$

where $\mathcal{G}_\alpha$ is the gradient mapping equation 12 at the concatenated iterates on $\mathcal{W}$, and $p_{\min} \in (0,1)$ is the explicit participation threshold of Lemma D.9, determined by γ_1, γ_2 and η. A small $\|\mathcal{G}_\alpha(w^k)\|$ means the projected step barely moves the iterate, so w^k is close to a fixed point of the projected-gradient update on $\mathcal{W}$.

Remark 4.6. The rate improves with the participation rate, reflecting a linear speedup in the number of participating clients, and reduces to $O(1/\sqrt{K})$ under full participation. Despite allowing general convex lower-level inequality constraints, it matches the rates of existing unconstrained federated bilevel methods (Tarzanagh et al., 2022; Li et al., 2023; Yang et al., 2023b) and the centralized BiC-GAFFA rate (Yao et al., 2025).

Theorem 4.7 (Increasing penalty). *Under Assumptions 4.1–4.3, consider full participation and the increasing penalty $c_k = c_0(k+1)^\rho$, where $\rho \in (0, 1/2)$. With suitable step sizes and sufficiently many inner iterations, FedBiC satisfies*

$$\min_{0\leq k\leq K-1} \mathbb{E}[R_k] = O\Big(K^{-(1-2\rho)/2}\Big). \quad (14)$$

Furthermore, if $\psi_{c_k}(w^k) \leq M$ for every k and some $M > 0$, then

$$G_{\gamma,i}(w_i^K) = O(K^{-\rho}), \qquad i \in [N]. \quad (15)$$

The precise parameter conditions are provided in Appendix D.

Remark 4.8. The parameter ρ controls the tradeoff between penalized stationarity and lower-level gap reduction: a larger ρ improves the gap bound but weakens the stationarity bound. Choosing $\rho = 1/4$ balances their exponents, giving $O(K^{-1/4})$ for both quantities. The stationarity guarantee holds for the best iterate among the first K rounds, while the gap guarantee holds at the final iterate under the stated boundedness condition. Stationarity for the original bilevel problem additionally requires the regularity conditions in Appendix E.

Remark 4.9 (Relation to the original problem). By Lemma C.1, $G_{\gamma,i} = 0$ holds if and only if y_i solves client i's constrained lower-level problem and z_i is a corresponding multiplier, so the feasibility rate above is a rate on exact lower-level optimality rather than on a surrogate. Stationarity in x does not follow immediately, since the bilevel structure is carried by the product $c_k\nabla_x G_{\gamma,i}$ in which $c_k \to \infty$ while $\nabla_x G_{\gamma,i} \to 0$. Appendix E identifies that limit, which the centralized analysis does not. Rescaling by c_k turns the stationarity conditions of the penalized problem into the adjoint system of the original problem, and If the lower-level problems are nondegenerate at a limit point and satisfy

LICQ, strict complementarity, and the strong second-order sufficient condition, then that limit point is stationary for the original constrained bilevel problem equation 1. Thus, the growth of the penalty parameter is the mechanism through which the hypergradient's sensitivity correction is recovered, rather than an obstruction to be tolerated.

5 Numerical Experiments

We conduct extensive experiments to validate our FedBiC algorithm. We compare FedBiC with various related algorithms such as FedNest Tarzanagh et al. (2022), SimFBO Yang et al. (2023b), FedBiOAcc (Li et al., 2023) , BiC-GAFFA (Yao et al., 2025), FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020) and SCAFFOLD (Karimireddy et al., 2020) on several tasks. The comparisons examine both predictive performance and the effect of explicitly modeling lower-level constraints. Experiments details can be found in appendix B.

5.1 Federated Constrained LASSO

We formulate federated sparse regression as bilevel hyperparameter optimization. The shared upper-level variable is a vector of per-feature budgets $x \in \mathbb{R}^d_+$. Each client's lower level fits its own coefficients y_i to local data under the box constraint $|y_{i,j}| \le x_j$, which instantiates the client-specific convex constraint $g_i(x, y_i) \le 0$ of equation 2. The upper level minimizes $\Phi(x)$, the average validation MSE plus the budget penalty $\nu\|x\|_1$ at $\nu = 0.05$, so lower is better and a method is credited only for predicting well while spending little budget. The constraint is what makes the budget do the selecting: $x_j = 0$ forces $y_{i,j} = 0$ exactly, for every client. We use synthetic data with $d = 20$ features (8 relevant, 12 irrelevant) and $n = 2000$ samples across $N = 10$ clients.

None of the three federated bilevel baselines supports lower-level inequality constraints. We therefore drop the box constraint for these methods and solve an unconstrained, strongly convex LL problem. BiC-GAFFA enforces the constraint but is centralized, so we run it on pooled data only as a reference. It learns a single LL model shared by all clients, so its gap to FedBiC reflects per-client personalization rather than optimizer quality. All methods minimize the same Φ at the same ν and receive the same tuning budget. As shown in Table 2, FedBiC achieves the lowest Φ. Although FedNest obtains a lower raw MSE, it uses several times the budget and still produces false positives. Without the box constraint, x_j no longer bounds $|y_{i,j}|$, so irrelevant coordinates are not driven to zero. Thus, FedNest improves accuracy by increasing $\|x\|_1$ rather than by allocating the budget selectively.

Table 2: Federated constrained LASSO

Method	$\Phi \downarrow$	Val MSE	$\|x\|_1$	True	False	LL Constr.
FedBiC (Ours)	$\mathbf{0.3170 \pm 0.0085}$	0.2118 ± 0.0078	2.104	7.8/8	0.0/12	✓
BiC-GAFFA (pooled)	0.4009 ± 0.0120	0.2644 ± 0.0141	2.730	7.8/8	0.0/12	✓
FedBiOAcc	0.4645 ± 0.0104	0.3340 ± 0.0121	2.610	7.8/8	0.0/12	✗
SimFBO	0.4788 ± 0.0094	0.3570 ± 0.0104	2.436	7.6/8	0.0/12	✗
FedNest	0.5615 ± 0.0092	$\mathbf{0.1669 \pm 0.0119}$	7.891	8.0/8	0.6/12	✗

5.2 Fair Few-Shot Meta-Learning

We evaluate FedBiC in the personalized FL paradigm (Fallah et al., 2020) using EMNIST digits (Cohen et al., 2017). Each client adapts a shared initialization $x \in \mathbb{R}^d$ on 20 support examples, subject to a demographic parity constraint (Hardt et al., 2016; Zafar et al., 2017), imposed as the affine surrogate $|c_i^\top y_i| \le \varepsilon$, a widely adopted group-fairness criterion that requires model predictions to be statistically independent of a sensitive attribute. Similar constraint-based formulations have been studied in both centralized and federated cases (Agarwal et al., 2018; Cotter et al., 2019) (Papadaki et al., 2022). For this experiment, the upper level jointly learns the shared initialization x and fairness tolerance ε.

Table 3: Fair few-shot meta-learning

Method	Test Acc $\uparrow$	Test DP Gap $\downarrow$	LL Constr.
FedBiC (Ours)	0.7112 ± 0.0143	$\mathbf{0.0734 \pm 0.0223}$	✓
SCAFFOLD	0.7191 ± 0.0069	0.2459 ± 0.0717	✗
FedAvg	0.7259 ± 0.0167	0.2605 ± 0.0554	✗
FedProx	0.7243 ± 0.0060	0.2626 ± 0.0729	✗
FedNest	$\mathbf{0.7305 \pm 0.0086}$	0.2830 ± 0.0227	✗

Table 3 shows a clear accuracy-fairness trade-off across methods. None of the baselines has a fairness mechanism, which is exactly the capability the lower-level constraint supplies, so their parity gaps reflect the natural behaviour of the converged models. All methods are adapted and scored identically on a held-out test split the upper level never sees, and each is tuned on its own grid, extended until the selected configuration is interior. FedBiC attains the lowest parity gap, a reduction of 70–74% over every baseline. Against SCAFFOLD, FedAvg and FedProx the accuracy differences fall within one standard deviation; FedNest is more accurate, by 1.9 points, but carries almost four times the parity gap, since without a constraint its adapted models are free to exploit the stroke-density direction.

5.3 Federated Steering: Robust Learning from Noisy Labels

We formulate noise robustness as a client-specific lower-level inequality constraint and validate FedBiC on a federated hyperparameter optimization task using CIFAR-10. Each client holds a large noisy training set and a small clean steering set. The shared upper-level variable is a single noise tolerance $c \geq 0$, chosen to minimize the average clean validation loss plus a penalty νc on loose tolerance; each client's lower level fits its model w_i to its noisy data subject to $\mathcal{L}(w_i; \mathcal{D}_{i,\text{steer}}) \leq c$, a bound on the loss over that client's own clean examples, which instantiates the constraint $g_i(x, y_i) \leq 0$ of equation 2. Because the steering sets differ across clients, so does each constraint. Penalizing c at the upper level makes the tolerance a quantity the algorithm discovers rather than one the user tunes by hand (Cotter et al., 2019).

Method	**IID**	**non-IID**	**LL Constr.**
FedBiC (Ours)	$\mathbf{80.8 \pm 0.7}$	$\mathbf{80.2 \pm 0.6}$	✓
FedProx	78.6 ± 0.3	74.6 ± 1.3	✗
SCAFFOLD	78.1 ± 0.4	72.9 ± 1.9	✗
FedNest	77.0 ± 0.2	71.9 ± 2.2	✗

Table 4: Federated steering performance under 30% label noise.

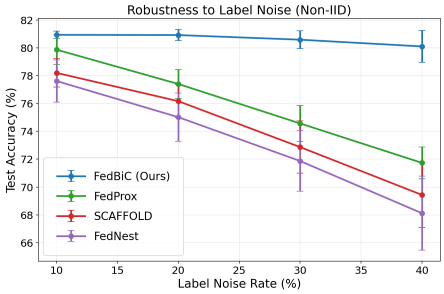

Figure 1: Robustness to label noise.

All methods receive the same clean steering examples. FedBiC imposes them through the lower-level constraint, whereas the baselines mix them into their local training loss, since they cannot handle constraints. The clean validation split plays different roles in the two cases. FedBiC's upper level trains on it, while the baselines use it only for model selection. Table 4 shows that, at a 30% label noise rate, FedBiC achieves the highest test accuracy in both the IID and non-IID settings. This advantage reflects FedBiC's use of the clean validation data through its upper level, together with the steering constraint. Figure 1 presents a noise sensitivity analysis across corruption rates of 10–40%. FedBiC's accuracy remains nearly flat, while the baselines degrade as the noise rate increases.

6 Conclusion

We present FedBiC as the first federated BLO algorithm capable of handling general LL inequality constraints. FedBiC transforms the bilevel problem into a penalty based single level objective that can be optimized in a federated setting by reformulating the constrained LL problem using a gap function. FedBiC achieves several desirable properties simultaneously, such as a single loop structure for communication efficiency, Hessian-free gradient computation for scalability, partial client participation, and robustness to data heterogeneity. All of them are achieved without requiring strong convexity of the LL objective. Our theoretical analysis establishes that FedBiC achieves $O(1/\sqrt{pK})$ convergence to a stationary point of the penalized objective under fixed penalty with partial participation, exhibiting linear speedup in the participation rate p. With an increasing penalty schedule and full participation, FedBiC attains an $O(1/K^{(1-2\rho)/2})$ stationarity rate for the iteration-dependent penalized objective, together with an $O(K^{-\rho})$ rate on the clientwise lower-level gap violation; Appendix E identifies conditions under which a limit point of these iterates is stationary for the original constrained bilevel problem. These rates match the best known results for the constrained BLO in centralized settings.

AI USE STATEMENT

In this work, we used a large language model for writing assistance, proofreading, rephrasing, and LaTeX formatting, for editing and debugging experiment codes. We did not use it to generate the research idea, the problem formulation, or the algorithm, which originate with the authors. All AI-assisted material was verified by the authors: proof steps were checked against the surrounding argument, code was run and its outputs confirmed against independently computed quantities, and all text was read and edited by the authors. We take full responsibility for all content in the manuscript.

ETHICS STATEMENT

This work develops an optimization algorithm and its convergence analysis. It involves no human subjects, and all experiments use synthetic data or publicly available datasets (EMNIST and CIFAR-10). The fair meta-learning experiment constrains an affine surrogate of demographic parity with respect to a sensitive attribute derived from image statistics; this surrogate bounds a difference in mean scores between groups and does not certify demographic parity or other fairness criteria, which should be assessed separately in any deployment. FedBiC keeps raw data and lower-level states on the clients and transmits only upper-level gradient vectors, but we do not provide a formal privacy guarantee: transmitted gradients may still leak information, and combining FedBiC with secure aggregation or differential privacy is left to future work.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pp. 22–31. PMLR, 2017.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. SIAM, 2017.

Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109):1–56, 2025.

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

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019. URL `http://jmlr.org/papers/v20/18-616.html`.

Stephan Dempe. *Foundations of Bilevel Programming*. Springer Science & Business Media, 2002.

Stephan Dempe and Joydeep Dutta. Is bilevel programming a special case of a mathematical program with complementarity constraints? *Mathematical Programming*, 131(1):37–48, 2012.

Asen L Dontchev and R Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer Monographs in Mathematics. Springer, 2009.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3557–3568, 2020.

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

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

Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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

Feihu Huang. Fast adaptive federated bilevel optimization. *arXiv preprint arXiv:2211.01122*, 2022.

Minhui Huang, Dewei Zhang, and Kaiyi Ji. Achieving linear speedup in non-iid federated bilevel learning. In *International Conference on Machine Learning*, pp. 13955–13983. PMLR, 2023.

Divyansh Jhunjhunwala, PRANAY SHARMA, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL https://openreview.net/forum?id=HlWLLdUocx5.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, 2020.

Prashant Khanduri, Ioannis Tsaknakis, Yihua Zhang, Jia Liu, Sijia Liu, Jiawei Zhang, and Mingyi Hong. Linearly constrained bilevel optimization: A smoothed implicit gradient approach. In *International Conference on Machine Learning*, pp. 16291–16325. PMLR, 2023.

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

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

Junyi Li, Feihu Huang, and Heng Huang. Communication-efficient federated bilevel optimization with global and local lower level problems. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/04bd683d5428d91c5fbb5a7d2c27064d-Abstract-Conference.html.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 2, pp. 429–450, 2020.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. BOME! bilevel optimization made easy: A simple first-order approach. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17248–17262, 2022.

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

Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1540–1552. PMLR, 2020.

Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pp. 2113–2122. PMLR, 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 142–159, 2022.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, pp. 737–746. PMLR, 2016.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

Stephen M Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. In *ICML*, 2022.

Quan Xiao, Han Shen, Wotao Yin, and Tianyi Chen. Alternating projected sgd for equality-constrained bilevel optimization. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/xiao23a.html.

Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112(3):859–887, 2023a.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 33027–33040. Curran Associates, Inc., 2023b. doi: 10.52202/075280-1433. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/686a3f32067838c8dbb68da6e9e3cf69-Paper-Conference.pdf.

Wei Yao, Chengming Yu, Shangzhi Zeng, and Jin Zhang. Constrained bi-level optimization: Proximal Lagrangian value function approach and Hessian-free algorithm. In *International Conference on Learning Representations*, 2024.

Wei Yao, Haian Yin, Shangzhi Zeng, and Jin Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cyPMEXdqQ2.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 962–970, 2017.

Shuchen Zhu, Boao Kong, Songtao Lu, Xinmeng Huang, and Kun Yuan. SPARKLE: A unified single-loop primal-dual framework for decentralized bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Appendix

A Notation

Table 5: Notation used throughout the paper.

x	shared upper-level variable	$\mathcal{A}_k$	clients participating in round k
y_i	client i's lower-level variable	p	participation rate, $\lvert\mathcal{A}_k\rvert/N$
z_i	client i's dual variable	$g_{x,i}$	client i's entry in the server gradient table
$\bar{\theta}_i, \bar{\lambda}_i$	inner-loop auxiliaries tracking θ_i^*, λ_i^*	S^k	staleness of the gradient table
$G_{\gamma,i}$	regularized gap function of client i	Θ^k, Λ^k	averaged auxiliary tracking errors
ψ_c	penalized objective, $\frac{1}{N}\sum_i[F_i + cG_{\gamma,i}]$	$V_{k,i}, V_k^G$	local and global merit functions
$\phi_{c,i}$	normalized objective, $\frac{1}{c}(F_i - \underline{F}_i) + G_{\gamma,i}$	R_k	stationarity residual
c_k	penalty parameter	$\mathcal{G}_\alpha$	gradient mapping
γ_1, γ_2	proximal parameters of the gap function	E	inner-loop steps per round

B Experiments Details

This appendix provides complete details for the experiments in Section 5, including dataset construction, the mathematical formulation for each baseline, hyperparameter settings, and per-feature analysis.

All numerical results in Tables are reported as mean $\pm$ standard deviation across five independent runs with random seeds (three for the steering experiment). The reported variability captures randomness from data partitioning, support-query splits, label-noise corruption (where applicable), and stochastic client sampling, with all other settings held fixed across runs.

Compute resources. All experiments run on a single workstation (AMD Ryzen 7 7700, 32 GB RAM, NVIDIA RTX 5070 12 GB) within a few CPU-hours per experiment. Total compute across all experiments, including hyperparameter sweeps and five-seed replication, is approximately 80 CPU-hours.

Evaluation protocol. The reported metrics are computed on per-client validation or query sets held out from lower-level training. These sets provide the upper-level objective signal during bilevel optimization; hence, the results primarily measure how effectively FedBiC optimizes held-out lower-level performance, rather than generalization to a fully disjoint test set. The steering experiment additionally includes a separate clean test set, providing evidence of generalization beyond the upper-level data. This protocol follows standard meta-learning evaluation practice (Fallah et al., 2020) and isolates the optimization quality of the bilevel mechanism from unrelated data-distribution effects.

B.1 Federated Constrained LASSO: Additional Details

Dataset Construction

Data generation. We generate a synthetic sparse regression dataset with known ground truth:

- $n = 2000$ samples, $d = 20$ features.
- Covariates $A \in \mathbb{R}^{n\times d}$ drawn from $\mathcal{N}(0, \Sigma)$ with $\Sigma_{jj} = 1$ and $\Sigma_{j,j+1} = \Sigma_{j+1,j} = 0.3$ (tridiagonal correlation).
- True coefficients: $\beta_j^* \sim \mathrm{Uniform}(0.5, 2.0) \times \mathrm{Rademacher}$ for $j = 1, \ldots, 8$ (relevant); $\beta_j^* = 0$ for $j = 9, \ldots, 20$ (irrelevant).
- Targets: $b = A\beta^* + \xi$, $\xi \sim \mathcal{N}(0, 0.3^2 I)$.
- Both A and b are standardized (zero mean, unit variance) after generation.

- Seeds: hyperparameters are selected on $\{42, 123\}$ and all reported results are averaged over the disjoint seeds $\{7, 2024, 11, 3, 99\}$. Each seed draws its own β^*, covariates and client partition.

The realization on evaluation seed 7, the seed reported in Table 7, is

$$\beta^* = [\underbrace{0.601,\ 1.384,\ 0.960,\ -1.872,\ 1.711,\ 0.713,\ 0.932,\ -1.222}_{\text{8 relevant}},\ \underbrace{0,\ \ldots,\ 0}_{\text{12 irrelevant}}]^\top. \tag{16}$$

To simulate statistical heterogeneity, we sort all samples by target value b before partitioning, so that each client receives data from a contiguous range of the response distribution. The sorted data is split into $N = 10$ equal shards of 200 samples each. Within each shard, samples are randomly shuffled (using the shared seed) and split into 140 training samples (used by the lower-level objective f_i) and 60 validation samples (used by the upper-level objective F_i); the per-client validation MSE is then averaged across the 10 clients to produce the "Val MSE" values. This creates substantial heterogeneity: clients at opposite ends of the target distribution observe very different input–output relationships.

Problem Formulation Details

All bilevel methods optimize the same two-level objective. The upper level seeks a sparsity-inducing bound vector $x \in \mathbb{R}^d_+$:

$$\Phi(x) = \frac{1}{N}\sum_{i=1}^{N} F_i\big(x, y_i^*(x)\big) + \nu \sum_{j=1}^{d} x_j\,, \quad F_i(x,y) = \frac{1}{2n_v}\|A_i^{\text{val}} y - b_i^{\text{val}}\|^2, \tag{17}$$

with sparsity regularizer $\nu = 0.05$ throughout the main text; the sensitivity of every result to this choice is reported in Table 8. The LL, per client i, is:

$$y_i^*(x) = \arg\min_{y_i} \underbrace{\frac{1}{2n_t}\|A_i y_i - b_i\|^2 + \frac{r}{2}\sum_{j=1}^{d}\frac{y_{i,j}^2}{x_j}}_{f_i(x,y_i)} \quad \text{s.t.} \quad |y_{i,j}| \le x_j,\ \forall\, j, \tag{18}$$

with $r = 0.5$. For FedBiC, the inequality constraints $g(x, y) = [y - x;\ -y - x] \le 0$ are handled via the gap function. For the baselines, these constraints are dropped, and the LL is solved as an unconstrained strongly convex problem.

Hyperparameters

All bilevel methods share the same problem parameters ($\nu = 0.05$, $r = 0.5$, $x_{\min} = 0.001$, $x_{\max} = 5.0$) and run for $K = 2000$ communication rounds. Table 6 lists the centre of each method's search grid. Every method is searched over the same number of configurations: the two parameters identified in Appendix B.1 are each scaled by $\{0.5, 1, 2\}$, giving nine configurations per method, selected on the tuning seeds by Φ and reported on the disjoint evaluation seeds.

All bilevel methods initialize $x = 0.5 \cdot \mathbf{1}_d$ and warm-start y using a ridge regression solution: $y_{\text{init}} = (A^\top A/n + (r/\bar{x})I)^{-1} A^\top b/n$, where $\bar{x} = 0.5$ is the initial mean of x. FedBiC additionally initializes the dual variables $\tilde{z} = 0.1 \cdot \mathbf{1}_{2d}$ and $\bar{\lambda} = 0.01 \cdot \mathbf{1}_{2d}$.

The cross-derivative $\nabla^2_{xy} f_i = -r \cdot y/x^2$ becomes ill-conditioned as $x_j \to x_{\min}$. FedBiC uses the safe-division threshold $\max(x, 0.001)$, matching the lower bound $x_{\min} = 0.001$, with x-gradients clipped to $[-10d, 10d]$. The baselines diverge at this threshold because their Neumann-series and STORM-style hypergradient estimators amplify near-zero denominators; we therefore use the looser threshold $\max(x, 0.01)$ for all baselines, together with cross-derivative clipping to $[-100, 100]$, Neumann vectors to $[-100, 100]$, and STORM momenta to $[-100, 100]$. The reported numbers use, for each method, the configuration its own nine-point grid selected on the tuning seeds.

Table 6: Hyperparameter settings for all methods.

Method	Parameter	Value
FedBiC (Ours)	Penalty parameters c_0, ρ	$1.0,\ 0.25$
	Step size α (y-update, scaled by $1/c_k$)	0.003
	Step size α_x (x-update)	0.005
	Step size β (z-update)	$\beta = \alpha$
	Inner tracking step size η	0.02
	Gap parameters γ_1, γ_2	$1.0,\ 1.0$
	Inner tracking steps E	10
FedNest	Upper-level step size α_x	0.003
	Lower-level step size β_y	0.005
	Lower-level local steps K_{inner}	20
	Neumann iterations Q / learning rate η_N	5 / 0.05
BiC-GAFFA	Same as FedBiC (centralized, pooled data)	
	Penalty parameters c_0, ρ	$1.0,\ 0.25$
	Step size α (y-update, scaled by $1/c_k$)	0.003
	Step size α_x (x-update)	0.005
	Step size β (z-update)	$\beta = \alpha$
	Inner tracking step size η	0.02
	Inner tracking steps E	10
FedBiOAcc	Step sizes γ, η, τ	$0.5,\ 0.2,\ 0.3$
	Schedule parameters δ, u_0	$0.5,\ 10.0$
	STORM constant c	0.1
	Communication interval I_{comm}	5
	u projection radius	10.0
SimFBO	Upper-level step size α_x	0.003
	Lower-level step size β_y	0.01
	Auxiliary step size τ_u	0.01
	Server momentum	0.9
	Local steps E / Comm. interval I_{comm}	5 / 5

PER-FEATURE ANALYSIS

Table 7 reports the final learned sparsity bounds x_j for each feature across all bilevel methods, alongside the true coefficient magnitude $|\beta_j^*|$.

FedBiC retains all eight relevant features above the 0.1 threshold and drives eleven of the twelve irrelevant ones to exactly $x_{\min} = 10^{-3}$, the twelfth to 0.015 — an order of magnitude below the smallest retained relevant bound. The support is therefore readable without a tuned threshold.

FedNest also retains all eight, but at bounds two to four times larger (0.615–1.338 against FedBiC's 0.119–0.398), and it does not prune cleanly: $x_9 = 0.122$ remains above the threshold, with x_{10}, x_{14} and x_{20} an order of magnitude above $x_{\min}$. Without the box constraint the penalty $\nu \sum_j x_j$ shrinks every coordinate at the same rate, so nothing ties x_j to the magnitude of $y_{i,j}$; the budget is spent, not allocated. This is the mechanism behind the $\|x\|_1$ gap in Table 2: 7.937 against 2.075 on this seed.

Pooled BiC-GAFFA recovers the same support as FedBiC and prunes just as cleanly, but assigns uniformly larger bounds (0.175–0.509), giving $\|x\|_1 = 2.744$ against FedBiC's 2.075 for the same eight features. Both enforce the same constraints with the same algorithm; the difference is that per-client y_i adapt to local distributions while a single shared y over pooled data must compromise across them.

SimFBO prunes cleanly but assigns the smallest bound to a relevant feature ($x_7 = 0.097$, just below the threshold), losing one true positive. FedBiOAcc prunes cleanly as well and recovers the same support as FedBiC, at $\|x\|_1 = 2.620$ with uniformly larger relevant bounds and the second-highest validation MSE in Table 2.

Table 7: Per-feature sparsity bounds x_j at K=2000 and the common $\nu = 0.05$, on evaluation seed 7. Features marked $*$ are relevant. Sparsity threshold $x_j < 0.1$; bold marks a retained irrelevant feature.

Feature	FedBiC	BiC-GAFFA	FedNest	SimFBO	FedBiOAcc	$\lvert\beta_j^*\rvert$
x_1^*	0.177	0.223	0.724	0.174	0.188	0.601
x_2^*	0.398	0.500	1.338	0.519	0.543	1.384
x_3^*	0.151	0.224	0.729	0.157	0.177	0.960
x_4^*	0.386	0.509	1.306	0.475	0.504	1.872
x_5^*	0.343	0.461	1.201	0.445	0.471	1.711
x_6^*	0.230	0.306	0.902	0.288	0.304	0.713
x_7^*	0.119	0.175	0.615	0.097	0.117	0.932
x_8^*	0.246	0.334	0.961	0.283	0.305	1.222
x_9	0.015	0.001	**0.122**	0.001	0.001	0.000
x_{10}	0.001	0.001	0.011	0.001	0.001	0.000
x_{11}	0.001	0.001	0.001	0.001	0.001	0.000
x_{12}	0.001	0.001	0.001	0.001	0.001	0.000
x_{13}	0.001	0.001	0.001	0.001	0.001	0.000
x_{14}	0.001	0.001	0.010	0.001	0.001	0.000
x_{15}	0.001	0.001	0.001	0.001	0.001	0.000
x_{16}	0.001	0.001	0.001	0.001	0.001	0.000
x_{17}	0.001	0.001	0.001	0.001	0.001	0.000
x_{18}	0.001	0.001	0.001	0.001	0.001	0.000
x_{19}	0.001	0.001	0.001	0.001	0.001	0.000
x_{20}	0.001	0.001	0.010	0.001	0.001	0.000

For each method the grid spans the two parameters that govern the upper-level trajectory: (α, α_x) for FedBiC and BiC-GAFFA, (α_x, β_y) for FedNest and SimFBO, and (η, δ) for FedBiOAcc, whose step schedule $\delta/(u_0 + t)^{1/3}$ bounds the total distance x can travel and is therefore binding where γ is not.

Sensitivity to the sparsity coefficient

Table 8 repeats the study at $\nu \in \{0.05, 0.1, 0.2\}$, with the equal tuning budget re-run independently at each ν.

Table 8: Φ (top) and retained relevant features out of eight (bottom) as a function of ν. Equal budget at every ν.

Method	$\nu = 0.05$	$\nu = 0.1$	$\nu = 0.2$
FedBiC (Ours)	**0.3170**	**0.4223**	**0.5640**
BiC-GAFFA (pooled)	0.4009	0.5392	0.7063
FedBiOAcc	0.4645	0.6314	0.7763
SimFBO	0.4788	0.6314	0.7841
FedNest	0.5114	0.7536	1.0865
FedBiC (Ours)	7.8	6.8	3.2
BiC-GAFFA (pooled)	7.8	7.8	4.2
FedBiOAcc	7.8	7.8	2.6
SimFBO	7.6	5.8	3.4
FedNest	8.0	8.0	8.0

FedBiC attains the lowest Φ at every ν, and its margin over pooled BiC-GAFFA is 20.9%, 21.7% and 20.1% — essentially constant, so the choice of operating point does not produce the ranking. (SimFBO and FedBiOAcc happen to coincide to four decimals at $\nu = 0.1$.) The operating point does determine whether the instance is informative: at $\nu \geq 0.1$ the penalty begins removing true signal for every method that prunes at all, and by $\nu = 0.2$ FedBiC, SimFBO and FedBiOAcc retain fewer than half the relevant features. We report $\nu = 0.05$ in the main text for that reason. FedNest retains all eight throughout only because it does not prune: its budget is 7.891 at $\nu = 0.05$, falling to 4.344 at $\nu = 0.2$, against 2.104 and 0.680 for FedBiC.

PARTICIPATION RATE

Table 9 sweeps the participation rate, retuning FedBiC independently at each p under the same nine-configuration budget.

Table 9: FedBiC under partial participation, $\nu = 0.2$, five seeds.

p	0.2	0.4	0.6	0.8	1.0
Φ	0.5650 $\pm$0.0156	0.5634 $\pm$0.0168	0.5632 $\pm$0.0157	0.5632 $\pm$0.0168	0.5640 $\pm$0.0173
Violation	0.069	0.076	0.073	0.076	0.077

The objective is flat to within one standard deviation across the whole range. At $\gamma_1 = \gamma_2 = 1$, $\eta = 0.02$ and $\tau = 1$ the threshold of Lemma D.9 evaluates to $p_{\min}(1) = 0.995$, so every setting except $p = 1$ lies outside the regime that instantiation covers. The threshold is sufficient rather than necessary, and Remark D.15 shows that any fixed $p > 0$ becomes admissible for τ small enough, at the price of a smaller admissible step size; we report this sweep as an empirical probe of rates below the explicitly instantiated $\tau = 1$ bound rather than as a claim covered by it.

B.2 FAIR FEW-SHOT META-LEARNING: ADDITIONAL DETAILS

We use the digits split of EMNIST (Cohen et al., 2017) (280,000 grayscale 28×28 images of handwritten digits 0–9). Raw images are flattened to 784 dimensions, standardized, and reduced to $d = 40$ via PCA, followed by a second standardization. A bias term is appended, giving $d = 41$.

The binary classification task is digits 0–4 (positive, $+1$) versus digits 5–9 (negative, -1). The protected attribute is stroke density: for each image, we compute the fraction of pixels exceeding intensity 0.1; images above the dataset median are labeled $a = 1$ (thick stroke) and those below are labeled $a = 0$ (thin stroke). Thick-stroke writers tend to produce more filled digit shapes, which can introduce spurious correlations between writing style and digit category.

We construct $N = 5$ clients with heterogeneous demographic compositions: thick-stroke fractions of 15%, 30%, 50%, 70%, and 85%. Each client receives $k = 20$ support examples for adaptation and 500 query examples for evaluation.

This experiment instantiates the personalized federated learning paradigm (Fallah et al., 2020), where the upper-level variable x is a shared model initialization of the same dimensionality as the per-client model y_i. The upper level learns x and a fairness tolerance ε:

$$\min_{x,\, \varepsilon \geq \varepsilon_{\min}} \ \frac{1}{N} \sum_{i=1}^{N} L_i^{\text{query}}(y_i^*(x, \varepsilon)) + \nu\, \varepsilon, \tag{19}$$

where L_i^{query} is the logistic loss on client i's clean query set, and $\nu\, \varepsilon$ penalizes loose fairness tolerance. The lower level adapts from x on the support set subject to a demographic parity constraint:

$$y_i^*(x, \varepsilon) = \arg\min_{y_i} \ L_i^{\text{support}}(y_i) + \frac{r}{2}\|y_i - x\|^2 \quad \text{s.t.} \quad |c_i^\top y_i| \leq \varepsilon, \tag{20}$$

where the proximal term $\frac{r}{2}\|y_i - x\|^2$ anchors adaptation near the shared initialization (the meta-learning mechanism), and c_i is the normalized mean feature difference between demographic groups at client i:

$$c_i = \frac{\bar{x}_i^{(a=1)} - \bar{x}_i^{(a=0)}}{\|\bar{x}_i^{(a=1)} - \bar{x}_i^{(a=0)}\|}, \tag{21}$$

with the bias dimension zeroed out. Each client has a different c_i due to heterogeneous demographic compositions, exercising FedBiC's ability to handle client-specific constraint functions.

Started at $\varepsilon = 0.15$, strictly interior to $[\varepsilon_{\min}, \varepsilon_{\max}] = [0.001, 0.5]$, the bilevel optimizer settles at $\varepsilon^\star \approx 0.001$ in every seed, driven to the lower bound $\varepsilon_{\min}$ by the upper-level penalty $\nu\varepsilon$. Saturation at the lower bound is an informative outcome rather than a degeneracy. The upper-level objective trades the penalty $\nu\varepsilon$ against the query loss at the adapted models, whose dependence on ε runs through the lower-level solutions $y_i^\star(\varepsilon)$; that $\varepsilon^\star$ reaches $\varepsilon_{\min}$ in every seed at $\nu = 2.0$ says that on this task, at this penalty weight, the trajectory is driven to the imposed lower bound $\varepsilon_{\min}$ rather than to an interior

equilibrium. What the experiment automates is the search itself: ε is set by the optimization rather than guessed and tuned by hand. Without the bilevel formulation, ε would have to be guessed and tuned manually. Here, the algorithm discovers it from data. Simultaneously, the lower-level solver enforces $|c_i^\top y_i| \approx \varepsilon^*$ at every participating client, confirming that the constraint is active and binding at the solver's equilibrium. The residual demographic parity gap of 0.073 reflects the approximation gap between the linear constraint direction c_i and the non-linear demographic parity metric rather than constraint violation: perfect fairness ($\mathrm{DP} = 0$) would require the decision boundary to be entirely orthogonal to the demographic direction in the original classifier space, whereas the linear constraint only projects out the mean-feature difference.

Table 10: FedBiC hyperparameters for the fair meta-learning experiment; c_0 and α_x are the values selected by the grid search.

Parameter	**Symbol**	**Value**
Communication rounds	K	800
Inner tracking steps	E	10
Fairness penalty	ν	2.0
Penalty parameter (fixed; $\rho = 0$)	c_0	16.0
Gap parameters	γ_1, γ_2	1.0, 5.0
Proximal strength	r	0.05
Step size (y-update, scaled by $1/c_k$)	α_y	0.1
Step size (x-update)	α_x	0.08
Step size (ε-update)	α_ε	0.01
Step size (z-update)	β	0.05
Inner tracking step size	η	0.05
ε bounds	$[\varepsilon_{\min}, \varepsilon_{\max}]$	[0.001, 0.5]
Initial ε	$\varepsilon_{\text{init}}$	0.15
Participation rate	p	1.0

All methods are warm-started with a global logistic regression model (300 gradient steps, $\mathrm{lr} = 0.1$, $\ell_2 = 0.01$). FedBiC initializes $\varepsilon = 0.15$ and dual variables $\tilde{z}_i = 0.1 \cdot \mathbf{1}_2$, $\bar{\lambda}_i = 0.01 \cdot \mathbf{1}_2$.

Table 11: Baseline hyperparameters for the fair meta-learning experiment, as selected by each method's own search.

Method	**Parameter**	**Value**
FedAvg	Learning rate	1.2
	ℓ_2 regularization	0.01
	Local SGD steps E	1
FedProx	Learning rate	0.01
	ℓ_2 regularization	0.01
	Proximal coefficient μ	3.0
	Local steps E	10
SCAFFOLD	Learning rate	0.1
	ℓ_2 regularization	0.01
	Local steps E	5
FedNest	Inner learning rate $\mathrm{lr}_{\text{inner}}$	0.6
	Outer learning rate $\mathrm{lr}_{\text{outer}}$	0.1
	ℓ_2 regularization	0.01
	Inner steps K_{inner}	10
	Neumann iterations Q / FD radius ϵ_{HVP}	5 / 0.01

No baseline includes a fairness mechanism, which is precisely the capability the lower-level constraint provides; their demographic parity gaps reflect the natural behaviour of the converged models. All methods run for $K = 800$ rounds at full participation and are scored identically: each method's output is adapted with ten support-set gradient steps per client, and accuracy and the demographic parity gap are measured on a held-out test split that the upper level never sees. Hyperparameters are selected on seeds $\{2026, 7\}$ and results reported on the disjoint seeds $\{42, 123, 456, 789, 1024\}$, so

no method is evaluated on a seed it was tuned on. Each method is searched over the two parameters that govern its trajectory — (c_0, α_x) for FedBiC, $(\mathrm{lr}_{\mathrm{inner}}, \mathrm{lr}_{\mathrm{outer}})$ for FedNest, (lr, E) for FedAvg and SCAFFOLD, (lr, μ) for FedProx — and each grid is extended until the selected configuration is interior to it, so no method is reported from a truncated search. This terminates after 18 configurations for FedBiC, SCAFFOLD and FedProx, 55 for FedNest, and 133 for FedAvg, whose selection reaches $E = 1$, the smallest admissible number of local steps.

B.3 Federated Steering: Additional Details

Federated clients often generate massive but noisy training data. In many practical settings, however, a small clean steering subset is also available and can be used to guide learning. The challenge is to leverage the large noisy dataset for capacity while being anchored by the small clean dataset. All methods in this experiment receive the same clean steering subset. FedBiC uses it through a lower-level inequality constraint, whereas the baselines use it through direct loss mixing in their local training objective. The clean validation split plays different roles. FedBiC's upper level trains on it, while the baselines use it only for model selection.

Each client i holds noisy training data $\mathcal{D}_{i,\text{noisy}}$ and a small clean steering set $\mathcal{D}_{i,\text{steer}}$ (5% of local data, with a minimum of 20 examples). A clean validation split is also distributed across clients and used by the upper-level objective. FedBiC optimizes a shared noise tolerance $c \geq 0$:

$$\Phi(c) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{val}}^{\mathrm{clean}}(w_i^*(c)) + \nu \cdot c, \tag{22}$$

where $\mathcal{L}$ is the logistic loss

$$\mathcal{L}(w; X, y) = \frac{1}{n} \sum_{j} \log(1 + \exp(-y_j w^\top x_j)),$$

and ν penalizes overly loose tolerance. The lower level for client i is

$$w_i^*(c) = \arg\min_{w_i} \ \mathcal{L}_{\mathrm{train}}^{\mathrm{noisy}}(w_i) + \frac{r}{2}\|w_i\|^2 \quad \text{s.t.} \quad \mathcal{L}(w_i; \mathcal{D}_{i,\mathrm{steer}}) - c \leq 0, \tag{23}$$

with $r = 0.01$. The constraint

$$g_i(c, w_i) = \mathcal{L}(w_i; \mathcal{D}_{i,\mathrm{steer}}) - c \leq 0$$

enforces that the learned client model maintains at most c loss on its clean steering examples.

This is an instance of the main problem with shared upper-level variable $x = c$ (noise tolerance), local lower-level variables $y_i = w_i$ (model weights), and client-specific lower-level inequality constraints $g_i(c, w_i) \leq 0$. The term νc drives the solution toward tighter tolerance, while the lower level resists when the steering constraint becomes too restrictive. The bilevel optimizer therefore automatically discovers the equilibrium tolerance c^*.

For the baselines, we use the same clean steering subset through a mixed local objective of the form

$$\mathcal{L}_i^{\mathrm{mix}}(w) = \mathcal{L}_{\mathrm{train}}^{\mathrm{noisy}}(w) + \lambda\, \mathcal{L}(w; \mathcal{D}_{i,\mathrm{steer}}), \tag{24}$$

where λ is tuned per method. Thus, all methods use the same clean steering subset, and FedBiC additionally trains on the clean validation split through its upper level.

Dataset and preprocessing. We use CIFAR-10 Krizhevsky (2009), restricted to a binary classification task (airplane vs. automobile). The raw $32 \times 32 \times 3$ images are flattened to 3072 dimensions, standardized, reduced to $d = 200$ via PCA (retaining approximately 95% of the variance), and standardized again. The resulting binary dataset contains 10 000 training examples and 2 000 test examples.

From the 10 000 binary training examples, we use an 80/20 train/validation split (8 000 training, 2 000 validation). In the IID setting, the 8 000 training examples are divided equally across $N = 5$ clients (1 600 samples each). Each client then reserves 5% of its local data as a clean steering subset (approximately 80 examples in the IID case), and uses the remaining 95% as noisy training data.

We corrupt the training labels by independently flipping each label with probability $q \in \{0.1, 0.2, 0.3, 0.4\}$, while keeping the steering sets, validation set, and test set clean. Table 4 uses $q = 0.3$, under which a fraction 0.3 of the noisy training labels is corrupted in expectation.

For the non-IID setting, we partition the training data across clients using a Dirichlet distribution with concentration parameter $\alpha = 0.1$. For each class, client proportions are sampled independently from a Dirichlet distribution, which induces heterogeneous class compositions across clients. To avoid pathological tiny-client regimes, the partition is repeated until every client receives at least a minimum number of training examples.

All methods use $E = 5$ local steps and $K = 400$ communication rounds.[1] FedBiC is tuned over 7 configurations, FedProx over 6, SCAFFOLD over 5, and FedNest over 5, drawn from the value sets in Table 12. For each seed and noise rate, the configuration with the lowest final validation loss is selected.

Table 12: Hyperparameter grids for the federated steering experiment. Best configuration selected by validation loss.

Method	**Grid**
FedBiC	$c_{\text{init}} \in \{0.3, 0.5, 0.8, 1.0\}$, $\nu \in \{0.005, 0.01, 0.02, 0.05\}$, $\alpha_w \in \{0.005, 0.01\}$, $\alpha_c \in \{0.01, 0.02\}$, $\gamma_1, \gamma_2 \in \{0.5, 1.0\}$, $r = 0.01$
FedProx	$\text{lr} \in \{0.01, 0.02, 0.05\}$, $\mu \in \{0.01, 0.05, 0.10\}$, $r \in \{0.001, 0.01\}$, $\lambda \in \{0.1, 0.2, 0.5, 1.0\}$
SCAFFOLD	$\text{lr} \in \{0.0005, 0.001, 0.002\}$, $r \in \{0.001, 0.01\}$, $\lambda \in \{0.05, 0.1, 0.2, 0.5\}$, control clipping $\in \{5, 10\}$
FedNest	$\text{lr}_{\text{inner}} \in \{0.01, 0.02\}$, $\text{lr}_{\text{outer}} \in \{0.005, 0.01\}$, $r \in \{0.001, 0.01\}$, $\lambda \in \{0.1, 0.2, 0.5\}$

For FedBiC, the learned tolerance converges to a non-trivial value: $c^* \approx 0.46$ in the IID setting and $c^* \approx 0.43$ in the non-IID setting. The corresponding mean residual violation on the local steering sets remains small (approximately 0.05 in both settings): the steering constraint shapes the solution without being overly restrictive, as expected of a penalty method, which drives violation down rather than enforcing feasibility exactly. This is the intended behavior: the clean steering set acts as a guardrail rather than an objective to be perfectly fit.

Figure 1 presents a noise sensitivity analysis under the non-IID setting for corruption rates ranging from 10% to 40%. FedBiC's accuracy remains nearly flat, while the baselines degrade as the noise rate increases. Every method is re-tuned at each noise rate with the procedure above.

C The Single-Level Reformulation

This appendix states and proves the properties of the doubly regularized gap function $G_{\gamma,i}$ introduced in Section 3, together with the equivalence between the original federated bilevel problem equation 1 and its single-level reformulation. Section C.1 establishes that $G_{\gamma,i} \leq 0$ characterizes the KKT points of the lower-level problem; Section C.2 derives closed-form expressions for $\nabla G_{\gamma,i}$ and verifies its continuity; Section C.3 proves that the reformulation preserves both feasibility and optimality; Section C.4 establishes the uniform multiplier bound that makes $\mathcal{Z}_i$ a valid dual domain; and Sections C.5 and C.6 establish the penalty relaxation and asymptotic exactness guarantees that justify the penalty-based formulation. The latter two concern global minimizers of the penalized problems; guarantees for the iterates generated by Algorithm 1 are established separately in Appendix D.

C.1 Characterization of the Gap Function

Lemma C.1. *Under Assumption 2.1 and $\gamma_1, \gamma_2 > 0$, for each $i \in [N]$:*

1. $G_{\gamma,i}(x, y_i, z_i) \geq 0$ *for all* $(x, y_i, z_i) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^{p_i}_+$.
2. $G_{\gamma,i}(x, y_i, z_i) = 0$ *if and only if* $y_i \in \mathcal{S}_i(x)$ *and* $z_i \in \mathcal{M}_i(x, y_i)$.

Consequently, $G_{\gamma,i}(x, y_i, z_i) \leq 0$ is equivalent to $y_i \in \mathcal{S}_i(x)$ and $z_i \in \mathcal{M}_i(x, y_i)$.

[1] We use $K = 400$ rounds to ensure convergence of all baselines. $K = 200$ yields comparable trends but with slightly higher variance for some methods.

Proof. Fix a client $i \in [N]$ and suppress the subscript i for readability.

For any $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^p_+$, the choice $(\theta, \lambda) = (y, z)$ is feasible in the definition of G_γ. Hence,

$$G_\gamma(x, y, z) \geq \mathcal{L}(x, y, z) - \mathcal{L}(x, y, z) = 0.$$

It remains to characterize when equality holds. Define

$$G_1 := \max_{\lambda \geq 0} \left\{ \mathcal{L}(x, y, \lambda) - \frac{\|\lambda - z\|^2}{2\gamma_2} \right\}, \qquad G_2 := \min_{\theta \in \mathcal{Y}} \left\{ \mathcal{L}(x, \theta, z) + \frac{\|\theta - y\|^2}{2\gamma_1} \right\}.$$

Then $G_\gamma = G_1 - G_2$.

Suppose first that $y \in \mathcal{S}(x)$ and $z \in \mathcal{M}(x, y)$. Then

$$0 \in \nabla_y f(x, y) + \nabla_y g(x, y)^\top z + N_{\mathcal{Y}}(y), \qquad g(x, y) \leq 0, \quad z \geq 0, \quad z^\top g(x, y) = 0.$$

The maximizer of G_1 is

$$\lambda^* = \mathrm{Proj}_{\mathbb{R}^p_+}\big(z + \gamma_2 g(x, y)\big).$$

Primal feasibility and complementary slackness imply $\lambda^* = z$. Thus,

$$G_1 = \mathcal{L}(x, y, z) = f(x, y).$$

The normal-cone stationarity condition is also the first-order optimality condition for $\theta = y$ in the strongly convex problem defining G_2. Hence,

$$G_2 = \mathcal{L}(x, y, z) = f(x, y).$$

Therefore, $G_\gamma(x, y, z) = 0$.

Conversely, suppose that $G_\gamma(x, y, z) = 0$. Since the feasible choice $(\theta, \lambda) = (y, z)$ gives value zero, we have

$$G_1 \geq \mathcal{L}(x, y, z), \qquad G_2 \leq \mathcal{L}(x, y, z).$$

Together with $G_1 - G_2 = 0$, this implies

$$G_1 = \mathcal{L}(x, y, z) = G_2.$$

Thus, $\lambda = z$ maximizes the problem defining G_1, and $\theta = y$ minimizes the problem defining G_2. The optimality of $\theta = y$ gives

$$0 \in \nabla_y f(x, y) + \nabla_y g(x, y)^\top z + N_{\mathcal{Y}}(y).$$

The optimality of $\lambda = z$ gives

$$z = \mathrm{Proj}_{\mathbb{R}^p_+}\big(z + \gamma_2 g(x, y)\big).$$

Equivalently, $z_j = \max\{0, z_j + \gamma_2 g_j(x, y)\}$ for every j, which implies

$$g(x, y) \leq 0, \qquad z \geq 0, \qquad z^\top g(x, y) = 0.$$

Therefore, the KKT conditions hold. Since the lower-level problem is convex in y, these conditions are sufficient, and hence $y \in \mathcal{S}(x)$ and $z \in \mathcal{M}(x, y)$.

Finally, because $G_\gamma \geq 0$, the condition $G_\gamma \leq 0$ is equivalent to $G_\gamma = 0$. This proves the claim. □

C.2 Gradients of the Gap Function

Lemma C.2. *Under Assumptions 2.1–2.2, $G_{\gamma,i}$ is continuously differentiable with*

$$\nabla_x G_{\gamma,i} = \nabla_x f_i(x, y_i) + \nabla_x g_i(x, y_i)^\top \lambda_i^* - \nabla_x f_i(x, \theta_i^*) - \nabla_x g_i(x, \theta_i^*)^\top z_i, \tag{25}$$

$$\nabla_{y_i} G_{\gamma,i} = \nabla_{y_i} f_i(x, y_i) + \nabla_{y_i} g_i(x, y_i)^\top \lambda_i^* - \tfrac{1}{\gamma_1}(y_i - \theta_i^*), \tag{26}$$

$$\nabla_{z_i} G_{\gamma,i} = \tfrac{1}{\gamma_2}(\lambda_i^* - z_i) - g_i(x, \theta_i^*), \tag{27}$$

where θ_i^* and λ_i^* are the unique maximizers of equation 4, given in closed form by equation 5.

Proof. Decompose $G_{\gamma,i} = \Phi_1 - \Phi_2$, where

$$\Phi_1(x, y, z) := \max_{\lambda \geq 0} \left\{ f_i(x, y) + \lambda^\top g_i(x, y) - \frac{\|\lambda - z\|^2}{2\gamma_2} \right\},$$

$$\Phi_2(x, y, z) := \min_{\theta \in \mathcal{Y}} \left\{ f_i(x, \theta) + z^\top g_i(x, \theta) + \frac{\|\theta - y\|^2}{2\gamma_1} \right\}.$$

We compute the gradients of Φ_1 and Φ_2 separately, then combine them.

Gradients of Φ_1. The inner maximization is strongly concave in λ with unique maximizer $\lambda_i^* = \mathrm{Proj}_{\mathbb{R}_+^{p_i}}(z + \gamma_2 g_i(x,y))$. Applying Danskin's theorem yields

$$\begin{aligned}\nabla_x \Phi_1(x,y,z) &= \nabla_x f_i(x,y) + \nabla_x g_i(x,y)^\top \lambda_i^*,\\ \nabla_y \Phi_1(x,y,z) &= \nabla_y f_i(x,y) + \nabla_y g_i(x,y)^\top \lambda_i^*,\\ \nabla_z \Phi_1(x,y,z) &= \frac{1}{\gamma_2}(\lambda_i^* - z).\end{aligned}$$

Gradients of Φ_2. The inner minimization is strongly convex in θ with unique minimizer θ_i^*. Applying Danskin's theorem yields

$$\begin{aligned}\nabla_x \Phi_2(x,y,z) &= \nabla_x f_i(x,\theta_i^*) + \nabla_x g_i(x,\theta_i^*)^\top z,\\ \nabla_y \Phi_2(x,y,z) &= -\frac{1}{\gamma_1}(\theta_i^* - y),\\ \nabla_z \Phi_2(x,y,z) &= g_i(x,\theta_i^*).\end{aligned}$$

Combining the gradients. Since $G_{\gamma,i} = \Phi_1 - \Phi_2$, subtracting the corresponding components yields equation 25–equation 27. Continuity of $\nabla G_{\gamma,i}$ follows from the continuity of the mappings $(x,y,z) \mapsto \theta_i^*(x,y,z)$ and $(x,y,z) \mapsto \lambda_i^*(x,y,z)$, a standard consequence of parametric strongly convex optimization, combined with Assumption 2.2. □

C.3 Equivalence of the Single-Level Reformulation

Replacing the implicit constraint $y_i \in \mathcal{S}_i(x)$ in equation 1 with the explicit inequality $G_{\gamma,i} \le 0$ gives the single-level reformulation

$$\min_{x,\{y_i,z_i\}} \tfrac{1}{N}\textstyle\sum_{i=1}^N F_i(x,y_i) \quad \text{s.t.} \quad G_{\gamma,i}(x,y_i,z_i) \le 0,\ z_i \in \mathcal{Z}_i,\ \forall i, \tag{28}$$

whose penalized counterpart equation 6 is the problem solved by Algorithm 1.

Theorem C.3. *Under Assumptions 2.1–2.3 and $\gamma > 0$, the problem equation 28 is equivalent to equation 1: their feasible sets project identically onto $(x, \{y_i\})$, and their optimal values coincide.*

Proof. We establish the equivalence by verifying that feasibility projects identically in both directions.

($\Rightarrow$) From original to reformulation. Let $(x^*, \{y_i^*\})$ be feasible for the original problem equation 1, so that $y_i^* \in \mathcal{S}_i(x^*)$ for all i. By Assumption 2.3 and strong duality, there exist multipliers $\mu_i^* \in \mathcal{M}_i(x^*, y_i^*)$. Set $z_i^* = \mu_i^* \in [0,r]^{p_i}$, which is feasible by Lemma C.4. Lemma C.1(ii) then gives $G_{\gamma,i}(x^*, y_i^*, z_i^*) = 0 \le 0$, so $(x^*, \{y_i^*\}, \{z_i^*\})$ is feasible for the reformulation equation 28.

($\Leftarrow$) From reformulation to original. Let $(x^*, \{y_i\}, \{z_i\})$ be feasible for the reformulation, so that $G_{\gamma,i}(x^*, y_i, z_i) \le 0$ for all i. Combining this inequality with Lemma C.1(i) forces $G_{\gamma,i}(x^*, y_i, z_i) = 0$, which by Lemma C.1(ii) implies $y_i \in \mathcal{S}_i(x^*)$.

Since both formulations share the same objective over the common feasible projection on $(x, \{y_i\})$, their optimal values coincide. □

C.4 Uniform Bound on the Lower-Level Multipliers

The dual domain $\mathcal{Z}_i = [0,r]^{p_i}$ used in equation 28 and in Algorithm 1 is a valid restriction only if every lower-level multiplier lies in it. The following lemma establishes this and identifies r explicitly.

Lemma C.4 (Uniform multiplier bound)**.** *Under Assumptions 2.1–2.3, for every $i \in [N]$, every $x \in \mathcal{X}$, every $y \in \mathcal{S}_i(x)$ and every $\mu \in \mathcal{M}_i(x,y)$,*

$$\|\mu\|_\infty \le \|\mu\|_1 \le \frac{M_{f,i}}{\sigma_i}, \qquad M_{f,i} := \sup_{\mathcal{X}\times\mathcal{Y}} f_i - \inf_{\mathcal{X}\times\mathcal{Y}} f_i, \tag{29}$$

where $\sigma_i > 0$ is the uniform Slater margin of Assumption 2.1. Since $\mathcal{X}$ and $\mathcal{Y}$ are compact and f_i is continuous (Assumption 2.2), $M_{f,i} < \infty$. Consequently $r := \max_{i\in[N]} M_{f,i}/\sigma_i$ is a finite constant bounding every lower-level multiplier in the ℓ_∞ norm, and $\mathcal{Z}_i := [0,r]^{p_i}$ contains $\mathcal{M}_i(x,y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{S}_i(x)$.

Proof. Fix i, x, y, μ as stated. Since $f_i(x,\cdot)$ and $g_i(x,\cdot)$ are convex (Assumption 2.1) and $\mu \geq 0$, the map $\mathcal{L}_i(x,\cdot,\mu)$ is convex on $\mathcal{Y}$; the KKT condition $0 \in \nabla_y \mathcal{L}_i(x,y,\mu) + N_{\mathcal{Y}}(y)$ therefore says that y minimizes $\mathcal{L}_i(x,\cdot,\mu)$ over $\mathcal{Y}$. Let $\hat{y}_i(x) \in \mathrm{ri}(\mathcal{Y})$ be the Slater point, so that $g_i(x,\hat{y}_i(x)) \leq -\sigma_i \mathbf{1}$. Then

$$f_i(x,y) = \mathcal{L}_i(x,y,\mu) \;\leq\; \mathcal{L}_i(x,\hat{y}_i(x),\mu) = f_i(x,\hat{y}_i(x)) + \mu^\top g_i(x,\hat{y}_i(x)) \;\leq\; f_i(x,\hat{y}_i(x)) - \sigma_i \|\mu\|_1,$$

where the first equality uses complementary slackness $\mu^\top g_i(x,y) = 0$ and the last step uses $\mu \geq 0$. Rearranging gives $\|\mu\|_1 \leq [f_i(x,\hat{y}_i(x)) - f_i(x,y)]/\sigma_i \leq M_{f,i}/\sigma_i$, and $\|\mu\|_\infty \leq \|\mu\|_1$ always. □

C.5 Penalty Relaxation

Proposition C.5 (Penalty relaxation)**.** *Assume that each $F_i(x,y)$ is bounded below by $\underline{F}_i \in \mathbb{R}$ on $\mathcal{X} \times \mathcal{Y}$. For any $\varepsilon > 0$, there exists $\bar{c} > 0$ such that every global solution $(x_c, \{y_{i,c}\}, \{z_{i,c}\})$ of the penalty formulation equation 6 with $c \geq \bar{c}$ satisfies*

$$G_{\gamma,i}(x_c, y_{i,c}, z_{i,c}) \leq \varepsilon, \qquad \forall i \in [N].$$

Proof. Let $(\bar{x}, \{\bar{y}_i\}, \{\bar{z}_i\})$ be any feasible point of the reformulation equation 28. Then $G_{\gamma,i}(\bar{x}, \bar{y}_i, \bar{z}_i) = 0$ for all i. By global optimality of $(x_c, \{y_{i,c}\}, \{z_{i,c}\})$ for the penalty formulation,

$$\frac{1}{N}\sum_{i=1}^{N} [F_i(x_c, y_{i,c}) + c G_{\gamma,i}(x_c, y_{i,c}, z_{i,c})] \leq \frac{1}{N}\sum_{i=1}^{N} F_i(\bar{x}, \bar{y}_i).$$

Using $F_i \geq \underline{F}_i$ and $G_{\gamma,i} \geq 0$, we obtain

$$\frac{c}{N}\sum_{i=1}^{N} G_{\gamma,i}(x_c, y_{i,c}, z_{i,c}) \leq \frac{1}{N}\sum_{i=1}^{N} F_i(\bar{x}, \bar{y}_i) - \frac{1}{N}\sum_{i=1}^{N} \underline{F}_i =: M.$$

Hence,

$$\sum_{i=1}^{N} G_{\gamma,i}(x_c, y_{i,c}, z_{i,c}) \leq \frac{NM}{c}.$$

Since each gap value is nonnegative, it follows that

$$\max_{i \in [N]} G_{\gamma,i}(x_c, y_{i,c}, z_{i,c}) \leq \frac{NM}{c}.$$

Therefore, choosing $\bar{c} = NM/\varepsilon$ ensures that $G_{\gamma,i}(x_c, y_{i,c}, z_{i,c}) \leq \varepsilon$ for every $i \in [N]$ whenever $c \geq \bar{c}$. □

C.6 Asymptotic Exactness

Theorem C.6 (Asymptotic exactness)**.** *Suppose Assumption 2.1 holds, $\mathcal{Z}_i$ is as in Lemma C.4, and F_i, f_i, g_i are continuous on $\mathcal{X} \times \mathcal{Y}$. Assume $c_k \to \infty$ and let $(x^k, \{y_i^k\}, \{z_i^k\})$ be a global solution of the penalty formulation equation 6 with penalty c_k. Then every accumulation point $(\bar{x}, \{\bar{y}_i\}, \{\bar{z}_i\})$ is a solution of equation 28.*

Proof. The product domain $\mathcal{W} := \mathcal{X} \times \mathcal{Y}^N \times \prod_i \mathcal{Z}_i$ is compact by Assumption 2.1 together with $\mathcal{Z}_i = [0,r]^{p_i}$. Consequently, the sequence $\{(x^k, \{y_i^k\}, \{z_i^k\})\} \subseteq \mathcal{W}$ admits a convergent subsequence with limit $(\bar{x}, \{\bar{y}_i\}, \{\bar{z}_i\})$.

Let $(\hat{x}, \{\hat{y}_i\}, \{\hat{z}_i\})$ be any feasible point of equation 28, so that $G_{\gamma,i}(\hat{x}, \hat{y}_i, \hat{z}_i) = 0$. Global optimality of $(x^k, \{y_i^k\}, \{z_i^k\})$ gives

$$\frac{c_k}{N}\sum_{i=1}^{N} G_{\gamma,i}(x^k, y_i^k, z_i^k) \;\leq\; \frac{1}{N}\sum_{i=1}^{N} F_i(\hat{x}, \hat{y}_i) - \frac{1}{N}\sum_{i=1}^{N} F_i(x^k, y_i^k).$$

The right-hand side is bounded by continuity on a compact domain, while $c_k \to \infty$ and $G_{\gamma,i} \geq 0$. Dividing both sides by c_k and passing to the limit therefore gives $G_{\gamma,i}(x^k, y_i^k, z_i^k) \to 0$ for all i.

By lower semicontinuity of $G_{\gamma,i}$ (Yao et al., 2025, Lemma 2),

$$0 \;\leq\; G_{\gamma,i}(\bar{x}, \bar{y}_i, \bar{z}_i) \;\leq\; \liminf_{k \to \infty} G_{\gamma,i}(x^k, y_i^k, z_i^k) \;=\; 0,$$

which gives $G_{\gamma,i}(\bar{x}, \bar{y}_i, \bar{z}_i) = 0$ and, by Lemma C.1(ii), $\bar{y}_i \in \mathcal{S}_i(\bar{x})$.

For optimality, observe that any feasible $(x, \{y_i\}, \{z_i\})$ of equation 28 satisfies $G_{\gamma,i} = 0$, so global optimality of $(x^k, \{y_i^k\}, \{z_i^k\})$ gives

$$\frac{1}{N}\sum_i F_i(x^k, y_i^k) \;\le\; \frac{1}{N}\sum_i \left[F_i(x^k, y_i^k) + c_k G_{\gamma,i}(x^k, y_i^k, z_i^k)\right] \;\le\; \frac{1}{N}\sum_i F_i(x, y_i).$$

Passing to the limit by continuity yields

$$\frac{1}{N}\sum_i F_i(\bar{x}, \bar{y}_i) \;\le\; \frac{1}{N}\sum_i F_i(x, y_i).$$

Since this inequality holds for every feasible point, $(\bar{x}, \{\bar{y}_i\}, \{\bar{z}_i\})$ is a solution of equation 28. □

D Convergence Proofs

The non-asymptotic convergence analysis of FedBiC rests on establishing a sufficient descent property for a global merit function that accounts for partial participation and stale client states. Writing out the merit function V_k^G of Section 4,

$$\begin{aligned} V_k^G &= \frac{1}{N}\sum_{i=1}^N V_{k,i} \\ &= \frac{1}{N}\sum_{i=1}^N \phi_{c_k,i}(w_i^k) + C_\theta \frac{1}{N}\sum_{i=1}^N \|\theta_i^k - \theta_i^*(w_i^k)\|^2 \\ &\quad + C_\lambda \frac{1}{N}\sum_{i=1}^N \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2, \end{aligned} \tag{30}$$

where the normalized per-client objective is defined by

$$\phi_{c_k,i}(x, y_i, z_i) \;:=\; \frac{1}{c_k}\big(F_i(x, y_i) - \underline{F}_i\big) + G_{\gamma,i}(x, y_i, z_i), \tag{31}$$

and the constants are

$$C_\theta \;:=\; L_f + r_2 L_{g_y} + \frac{1}{\gamma_1}, \qquad C_\lambda \;:=\; L_g + \frac{1}{\gamma_2}. \tag{32}$$

Because the local variables (y_i, z_i) are never averaged across clients, the same merit function serves both the full- and partial-participation analyses; the only additional quantity required under partial participation is the gradient-table staleness S^k introduced in Appendix D.1.

The round-local variables $\bar{\theta}_i$ and $\bar{\lambda}_i$, together with the associated directions $d_{\bar{\theta}_i}, d_{\bar{\lambda}_i}, d_{y_i}, d_{z_i}$, are reinitialized at the start of each round and therefore do not carry a k superscript. In the convergence proofs throughout Appendix D, we write $\tilde{y}_i$ and $\tilde{z}_i$ for the values (y_i^k, z_i^k) held fixed by client i throughout the inner loop of round k. To establish the sufficient descent property, we first derive several auxiliary lemmas that extend the centralized results to the federated setting.

Lemma D.1 (Lipschitz continuity of $\theta^*(w)$ and $\lambda^*(w)$)**.** *Under Assumptions 4.2 and 4.3, let $\gamma_1, \gamma_2 > 0$. Then, for any $w = (x, y, z)$ and $w' = (x', y', z')$ in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, the auxiliary mappings satisfy*

$$\|\theta^*(w) - \theta^*(w')\| \le L_\theta \|w - w'\|, \qquad \|\lambda^*(w) - \lambda^*(w')\| \le L_\lambda \|w - w'\|,$$

where

$$L_\theta := \sqrt{3}\max\left\{\gamma_1(L_f + r_2 L_{gy}),\, 1,\, \gamma_1 L_g\right\}, \qquad L_\lambda := \sqrt{2}\max\left\{1,\, \gamma_2 L_g\right\}.$$

Proof. We establish the bound for $\theta^*(w)$ first and then for $\lambda^*(w)$.

Bound for $\theta^*(w)$. Recall that

$$\theta^*(w) \;=\; \arg\min_{\theta\in\mathcal{Y}} \left\{ f(x,\theta) + z^\top g(x,\theta) + \frac{1}{2\gamma_1}\|\theta - y\|^2 \right\} \;=:\; \arg\min_{\theta\in\mathcal{Y}} \Phi(w,\theta). \tag{33}$$

For fixed $w = (x, y, z)$, the mapping $\theta \mapsto \Phi(w,\theta)$ is $\frac{1}{\gamma_1}$-strongly convex on $\mathcal{Y}$: $f(x,\cdot)$ is convex by Assumption 4.2, $g(x,\cdot)$ is convex by Assumption 4.3 and $z \in \mathcal{Z} \subseteq \mathbb{R}^p_+$ implies that $z^\top g(x,\cdot)$ is convex, and the proximal term $\frac{1}{2\gamma_1}\|\theta - y\|^2$ is $\frac{1}{\gamma_1}$-strongly convex.

Since $\theta^*(w)$ minimizes $\Phi(w,\cdot)$ over the convex set $\mathcal{Y}$, there exists $\nu \in \mathcal{N}_{\mathcal{Y}}(\theta^*(w))$ such that

$$\nabla_\theta f(x,\theta^*(w)) + z^\top \nabla_\theta g(x,\theta^*(w)) + \frac{1}{\gamma_1}\big(\theta^*(w) - y\big) + \nu \;=\; 0. \tag{34}$$

Similarly, for $w' = (x', y', z')$, there exists $\nu' \in \mathcal{N}_{\mathcal{Y}}(\theta^*(w'))$ such that

$$\nabla_\theta f(x',\theta^*(w')) + (z')^\top \nabla_\theta g(x',\theta^*(w')) + \frac{1}{\gamma_1}\big(\theta^*(w') - y'\big) + \nu' \;=\; 0. \tag{35}$$

Define the map

$$T(w,\theta) \;:=\; \nabla_\theta f(x,\theta) + z^\top \nabla_\theta g(x,\theta) + \frac{1}{\gamma_1}(\theta - y).$$

Because $\Phi(w,\cdot)$ is $\frac{1}{\gamma_1}$-strongly convex, the map $T(w,\cdot)$ is $\frac{1}{\gamma_1}$-strongly monotone, so that

$$\langle T(w,\theta) - T(w,\theta'),\, \theta - \theta' \rangle \;\geq\; \frac{1}{\gamma_1}\|\theta - \theta'\|^2, \qquad \forall\, \theta, \theta' \in \mathcal{Y}. \tag{36}$$

Combining equation 34 and equation 35 gives

$$T(w,\theta^*(w)) \;=\; -\nu, \qquad T(w',\theta^*(w')) \;=\; -\nu'.$$

Since the normal cone is monotone,

$$\langle \nu - \nu',\, \theta^*(w) - \theta^*(w') \rangle \;\geq\; 0,$$

which implies

$$\langle T(w,\theta^*(w)) - T(w',\theta^*(w')),\, \theta^*(w) - \theta^*(w') \rangle \;\leq\; 0. \tag{37}$$

Decomposing the left-hand side of equation 37 yields

$$\begin{aligned} &\langle T(w,\theta^*(w)) - T(w',\theta^*(w')),\, \theta^*(w) - \theta^*(w') \rangle \\ &\quad= \langle T(w,\theta^*(w)) - T(w,\theta^*(w')),\, \theta^*(w) - \theta^*(w') \rangle \\ &\qquad+ \langle T(w,\theta^*(w')) - T(w',\theta^*(w')),\, \theta^*(w) - \theta^*(w') \rangle. \end{aligned} \tag{38}$$

By the strong-monotonicity estimate equation 36, the first term on the right-hand side of equation 38 is bounded below by $\frac{1}{\gamma_1}\|\theta^*(w) - \theta^*(w')\|^2$. Substituting this bound into equation 37 gives

$$\frac{1}{\gamma_1}\|\theta^*(w) - \theta^*(w')\|^2 \;\leq\; \|T(w,\theta^*(w')) - T(w',\theta^*(w'))\| \cdot \|\theta^*(w) - \theta^*(w')\|. \tag{39}$$

If $\theta^*(w) = \theta^*(w')$, the claimed Lipschitz bound holds trivially. Otherwise, dividing both sides of equation 39 by $\|\theta^*(w) - \theta^*(w')\|$ gives

$$\|\theta^*(w) - \theta^*(w')\| \;\leq\; \gamma_1 \|T(w,\theta^*(w')) - T(w',\theta^*(w'))\|. \tag{40}$$

It remains to bound the right-hand side of equation 40. For any fixed $\theta \in \mathcal{Y}$,

$$\begin{aligned} T(w,\theta) - T(w',\theta) \;&=\; \big[\nabla_\theta f(x,\theta) - \nabla_\theta f(x',\theta)\big] \\ &\quad+\; \big[z^\top \nabla_\theta g(x,\theta) - (z')^\top \nabla_\theta g(x',\theta)\big] + \frac{1}{\gamma_1}(y' - y). \end{aligned} \tag{41}$$

By Assumption 4.2,

$$\|\nabla_\theta f(x,\theta) - \nabla_\theta f(x',\theta)\| \;\leq\; L_f \|x - x'\|.$$

For the second bracket in equation 41, adding and subtracting $(z')^\top \nabla_\theta g(x,\theta)$ gives

$$z^\top \nabla_\theta g(x,\theta) - (z')^\top \nabla_\theta g(x',\theta) \;=\; (z - z')^\top \nabla_\theta g(x,\theta) + (z')^\top \big[\nabla_\theta g(x,\theta) - \nabla_\theta g(x',\theta)\big].$$

Applying $\|\nabla_\theta g(x,\theta)\| \leq L_g$, $\|z'\| \leq r_2$, and the L_{gy}-Lipschitz continuity of $\nabla_\theta g$ in x (Assumption 4.3) bounds the two terms by

$$\|(z - z')^\top \nabla_\theta g(x,\theta)\| \;\leq\; L_g \|z - z'\|, \qquad \|(z')^\top (\nabla_\theta g(x,\theta) - \nabla_\theta g(x',\theta))\| \;\leq\; r_2 L_{gy} \|x - x'\|.$$

The remaining term in equation 41 satisfies $\left\| \frac{1}{\gamma_1}(y' - y) \right\| = \frac{1}{\gamma_1}\|y - y'\|$. Combining the preceding estimates gives

$$\|T(w,\theta) - T(w',\theta)\| \;\le\; (L_f + r_2 L_{gy})\|x - x'\| + \frac{1}{\gamma_1}\|y - y'\| + L_g\|z - z'\|. \tag{42}$$

Substituting equation 42 into equation 40 yields

$$\|\theta^*(w) - \theta^*(w')\| \;\le\; \gamma_1(L_f + r_2 L_{gy})\|x - x'\| + \|y - y'\| + \gamma_1 L_g\|z - z'\|.$$

Applying the Cauchy–Schwarz inequality $a_1u_1 + a_2u_2 + a_3u_3 \le \sqrt{3}\max\{a_1,a_2,a_3\}\sqrt{u_1^2+u_2^2+u_3^2}$ with u_i the three norm terms gives

$$\|\theta^*(w) - \theta^*(w')\| \;\le\; \sqrt{3}\max\{\gamma_1(L_f + r_2 L_{gy}),\, 1,\, \gamma_1 L_g\}\,\|w - w'\|,$$

which proves the bound for θ^*.

Bound for $\lambda^*(w)$. By definition,

$$\lambda^*(w) \;=\; \mathrm{Proj}_{\mathbb{R}^p_+}\big(z + \gamma_2 g(x,y)\big). \tag{43}$$

Since projection onto a closed convex set is nonexpansive,

$$\|\lambda^*(w) - \lambda^*(w')\| \;\le\; \big\|(z + \gamma_2 g(x,y)) - (z' + \gamma_2 g(x',y'))\big\| \;=\; \big\|(z - z') + \gamma_2\big(g(x,y) - g(x',y')\big)\big\|. \tag{44}$$

Squaring both sides of equation 44 and applying $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ yields

$$\|\lambda^*(w) - \lambda^*(w')\|^2 \;\le\; 2\|z - z'\|^2 + 2\gamma_2^2\|g(x,y) - g(x',y')\|^2.$$

By Assumption 4.3, the map g is L_g-Lipschitz, so

$$\|g(x,y) - g(x',y')\| \;\le\; L_g\|(x,y) - (x',y')\|.$$

Combining the previous two inequalities gives

$$\|\lambda^*(w) - \lambda^*(w')\|^2 \;\le\; 2\|z - z'\|^2 + 2\gamma_2^2 L_g^2\|(x,y) - (x',y')\|^2 \;\le\; 2\max\{1, \gamma_2^2 L_g^2\}\,\|w - w'\|^2.$$

Taking square roots yields

$$\|\lambda^*(w) - \lambda^*(w')\| \;\le\; \sqrt{2}\max\{1, \gamma_2 L_g\}\,\|w - w'\|,$$

which completes the proof. □

Lemma D.2 (Lipschitz continuity of $\nabla G_{\gamma,i}$)**.** *Under Assumptions 4.2 and 4.3, with compact sets $\mathcal{X}$, $\mathcal{Y}$ and $\gamma_1, \gamma_2 > 0$, for each client $i \in [N]$ the gap-function gradient satisfies*

$$\|\nabla G_{\gamma,i}(w) - \nabla G_{\gamma,i}(w')\| \;\le\; L_G\|w - w'\|, \qquad \forall\, w, w' \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}, \tag{45}$$

where $L_G > 0$ depends on $L_f, L_g, L_{gx}, L_{gy}, \gamma_1, \gamma_2, r$, and the Lipschitz constants L_θ, L_λ from Lemma D.1. Consequently, the gap function satisfies the descent inequality

$$G_{\gamma,i}(w') \;\le\; G_{\gamma,i}(w) + \langle \nabla G_{\gamma,i}(w),\, w' - w\rangle + \frac{L_G}{2}\|w - w'\|^2. \tag{46}$$

Proof. Recall from Lemma C.2 the gradient components

$$\nabla_x G_{\gamma,i}(w) \;=\; \nabla_x f_i(x,y) + (\lambda^*)^\top \nabla_x g_i(x,y) - \nabla_x f_i(x,\theta^*) - z^\top \nabla_x g_i(x,\theta^*), \tag{47}$$

$$\nabla_y G_{\gamma,i}(w) \;=\; \nabla_y f_i(x,y) + (\lambda^*)^\top \nabla_y g_i(x,y) - \frac{1}{\gamma_1}(y - \theta^*), \tag{48}$$

$$\nabla_z G_{\gamma,i}(w) \;=\; -\frac{1}{\gamma_2}(z - \lambda^*) - g_i(x,\theta^*), \tag{49}$$

where $\theta^* = \theta_i^*(w)$ and $\lambda^* = \lambda_i^*(w)$. Compactness of $\mathcal{X}$ and $\mathcal{Y}$ provides the finite bounds $M_g := \sup_{(x,y)}\|\nabla g_i(x,y)\|$ and $\widehat{M}_g := \sup_{(x,y)}\|g_i(x,y)\|$, and $M_\lambda := r_2 + \gamma_2\widehat{M}_g$ bounds $\|\lambda^*\|$, since $\lambda^* = \mathrm{Proj}_{\mathbb{R}^{p_i}_+}(z + \gamma_2 g_i(x,y))$.

We bound each block separately. Throughout, let $\theta^* = \theta_i^*(w)$, $\theta'^* = \theta_i^*(w')$, $\lambda^* = \lambda_i^*(w)$, and $\lambda'^* = \lambda_i^*(w')$.

Bound on the x-component. From equation 47,

$$\begin{aligned}\nabla_x G_{\gamma,i}(w) - \nabla_x G_{\gamma,i}(w') \;=\; & \underbrace{\big[\nabla_x f_i(x,y) - \nabla_x f_i(x',y')\big]}_{(a)} \\ & + \underbrace{\big[(\lambda^*)^\top \nabla_x g_i(x,y) - (\lambda'^*)^\top \nabla_x g_i(x',y')\big]}_{(b)} \\ & - \underbrace{\big[\nabla_x f_i(x,\theta^*) - \nabla_x f_i(x',\theta'^*)\big]}_{(c)} \\ & - \underbrace{\big[z^\top \nabla_x g_i(x,\theta^*) - (z')^\top \nabla_x g_i(x',\theta'^*)\big]}_{(d)}. \end{aligned} \tag{50}$$

We bound each term in turn.

Term (a). By Assumption 4.2,

$$\|\nabla_x f_i(x,y) - \nabla_x f_i(x',y')\| \;\le\; L_f \|(x,y) - (x',y')\| \;\le\; L_f \|w - w'\|.$$

Term (b). Adding and subtracting $(\lambda'^*)^\top \nabla_x g_i(x,y)$ gives

$$(\lambda^*)^\top \nabla_x g_i(x,y) - (\lambda'^*)^\top \nabla_x g_i(x',y') \;=\; (\lambda^* - \lambda'^*)^\top \nabla_x g_i(x,y) + (\lambda'^*)^\top \big[\nabla_x g_i(x,y) - \nabla_x g_i(x',y')\big].$$

Applying $\|\nabla_x g_i\| \le M_g$, Lemma D.1, and the L_{gx}-Lipschitz continuity of $\nabla_x g_i$ (Assumption 4.3) yields

$$\|\text{Term (b)}\| \;\le\; M_g L_\lambda \|w - w'\| + M_\lambda L_{gx} \|w - w'\|. \tag{51}$$

Term (c). Applying Assumption 4.2 and Lemma D.1 gives

$$\|\nabla_x f_i(x,\theta^*) - \nabla_x f_i(x',\theta'^*)\| \;\le\; L_f(\|x - x'\| + \|\theta^* - \theta'^*\|) \;\le\; L_f(1 + L_\theta)\|w - w'\|.$$

Term (d). A similar argument, together with the bound $\|\lambda^*\|, \|z\| \le r_2$, yields

$$\|\text{Term (d)}\| \;\le\; M_g \|z - z'\| + r_2 L_{gx}(\|x - x'\| + \|\theta^* - \theta'^*\|) \;\le\; \big(M_g + r_2 L_{gx}(1 + L_\theta)\big)\|w - w'\|.$$

Combining the four term bounds, we obtain

$$\|\nabla_x G_{\gamma,i}(w) - \nabla_x G_{\gamma,i}(w')\| \;\le\; C_x \|w - w'\|, \tag{52}$$

where $C_x := 2L_f + L_f L_\theta + M_g L_\lambda + M_\lambda L_{gx} + M_g + r_2 L_{gx}(1 + L_\theta)$.

Bound on the y-component. Applying the same style of decomposition to equation 48 yields

$$\|\nabla_y G_{\gamma,i}(w) - \nabla_y G_{\gamma,i}(w')\| \;\le\; C_y \|w - w'\|, \tag{53}$$

where $C_y := L_f + M_g L_\lambda + M_\lambda L_{gy} + \frac{1}{\gamma_1}(1 + L_\theta)$.

Bound on the z-component. Starting from equation 49,

$$\begin{aligned}\|\nabla_z G_{\gamma,i}(w) - \nabla_z G_{\gamma,i}(w')\| \;&\le\; \frac{1}{\gamma_2}\|(z - \lambda^*) - (z' - \lambda'^*)\| + \|g_i(x,\theta^*) - g_i(x',\theta'^*)\| \\ &\le\; \frac{1}{\gamma_2}\big(\|z - z'\| + \|\lambda^* - \lambda'^*\|\big) + L_g(\|x - x'\| + \|\theta^* - \theta'^*\|) \\ &\le\; \left(\frac{1 + L_\lambda}{\gamma_2} + L_g(1 + L_\theta)\right)\|w - w'\| \;=:\; C_z \|w - w'\|, \end{aligned} \tag{54}$$

where the second inequality applies Lemma D.1 and the L_g-Lipschitz continuity of g_i.

Combining the three block bounds equation 52, equation 53, and equation 54 yields

$$\|\nabla G_{\gamma,i}(w) - \nabla G_{\gamma,i}(w')\|^2 \;\le\; (C_x^2 + C_y^2 + C_z^2)\|w - w'\|^2, \tag{55}$$

so that setting $L_G := \sqrt{C_x^2 + C_y^2 + C_z^2}$ proves the Lipschitz bound equation 45. The descent inequality equation 46 then follows from the standard descent lemma for L_G-smooth functions. □

Definition D.3 (Approximation-error Lipschitz constant)**.** Under Assumptions 4.2 and 4.3, the gradient approximation error satisfies

$$\|\widetilde{\nabla} G_{\gamma,i}(w;\theta,\lambda) - \nabla G_{\gamma,i}(w)\| \;\le\; L_e \sqrt{\|\theta - \theta^*(w)\|^2 + \|\lambda - \lambda^*(w)\|^2}, \tag{56}$$

where

$$L_e \;:=\; \sqrt{(L_f + r_2 L_{gx})^2 + \frac{1}{\gamma_1^2} + 3L_g^2 + \frac{1}{\gamma_2^2}}. \tag{57}$$

Justification of equation 56. Define the tracking residuals

$$\delta_\theta \;:=\; \theta - \theta^*(w), \qquad \delta_\lambda \;:=\; \lambda - \lambda^*(w).$$

By Lemma C.2, the exact gradient $\nabla G_{\gamma,i}(w)$ is obtained by evaluating the θ^*- and λ^*-dependent terms at $(\theta^*(w), \lambda^*(w))$, whereas the approximate gradient $\widetilde{\nabla} G_{\gamma,i}(w;\theta,\lambda)$ substitutes the pair (θ,λ) in place of $(\theta^*(w), \lambda^*(w))$. We bound each block of the resulting gradient difference separately.

For the x-component,

$$\|\widetilde{\nabla}_x G_{\gamma,i}(w;\theta,\lambda) - \nabla_x G_{\gamma,i}(w)\| \;\le\; (L_f + r_2 L_{gx})\|\delta_\theta\| + L_g\|\delta_\lambda\|.$$

For the y-component,

$$\|\widetilde{\nabla}_y G_{\gamma,i}(w;\theta,\lambda) - \nabla_y G_{\gamma,i}(w)\| \;\le\; \frac{1}{\gamma_1}\|\delta_\theta\| + L_g\|\delta_\lambda\|.$$

For the z-component,

$$\|\widetilde{\nabla}_z G_{\gamma,i}(w;\theta,\lambda) - \nabla_z G_{\gamma,i}(w)\| \;\le\; L_g\|\delta_\theta\| + \frac{1}{\gamma_2}\|\delta_\lambda\|.$$

Each of these three estimates takes the form $\|v_k\| \le a_k\|\delta_\theta\| + b_k\|\delta_\lambda\|$ for nonnegative constants a_k, b_k depending on the block $k \in \{x,y,z\}$. Applying the Cauchy–Schwarz inequality $(a_k u + b_k v)^2 \le (a_k^2 + b_k^2)(u^2 + v^2)$ with $u = \|\delta_\theta\|$ and $v = \|\delta_\lambda\|$ gives

$$\|v_k\|^2 \;\le\; (a_k^2 + b_k^2)\big(\|\delta_\theta\|^2 + \|\delta_\lambda\|^2\big).$$

Summing this bound over the three blocks $k \in \{x,y,z\}$ yields

$$\begin{aligned}
\|\widetilde{\nabla} G_{\gamma,i}(w;\theta,\lambda) - \nabla G_{\gamma,i}(w)\|^2 \;&=\; \sum_{k\in\{x,y,z\}} \|v_k\|^2 \\
&\le\; \left(\sum_{k\in\{x,y,z\}} (a_k^2 + b_k^2)\right)\big(\|\delta_\theta\|^2 + \|\delta_\lambda\|^2\big) \\
&=\; \left((L_f + r_2 L_{gx})^2 + \frac{1}{\gamma_1^2} + 3L_g^2 + \frac{1}{\gamma_2^2}\right)\big(\|\delta_\theta\|^2 + \|\delta_\lambda\|^2\big),
\end{aligned}$$

where the final equality collects the block contributions $\sum_k a_k^2 = (L_f + r_2 L_{gx})^2 + 1/\gamma_1^2 + L_g^2$ and $\sum_k b_k^2 = L_g^2 + L_g^2 + 1/\gamma_2^2 = 2L_g^2 + 1/\gamma_2^2$. Taking square roots proves equation 56. □

Lemma D.4 (Smoothness of the normalized merit function)**.** *Under Assumptions 4.1–4.3, for any $c \ge c_0 > 0$, the normalized merit function $\phi_{c,i}$ defined in equation 8 satisfies*

$$\|\nabla\phi_{c,i}(w) - \nabla\phi_{c,i}(w')\| \;\le\; L_\phi\|w - w'\|, \qquad \forall\, w, w' \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}, \tag{58}$$

where

$$L_\phi \;:=\; \frac{L_F}{c_0} + L_G \tag{59}$$

is independent of c for all $c \ge c_0$. Consequently, for any $c \ge c_0$ the descent inequality

$$\phi_{c,i}(w') \;\le\; \phi_{c,i}(w) + \langle \nabla\phi_{c,i}(w),\, w' - w\rangle + \frac{L_\phi}{2}\|w - w'\|^2 \tag{60}$$

holds.

Proof. The gradient of the normalized merit function is

$$\nabla\phi_{c,i}(w) \;=\; \frac{1}{c}\nabla F_i(w) + \nabla G_{\gamma,i}(w).$$

Applying the triangle inequality to the difference $\nabla\phi_{c,i}(w) - \nabla\phi_{c,i}(w')$ yields

$$\|\nabla\phi_{c,i}(w) - \nabla\phi_{c,i}(w')\| \;=\; \left\|\frac{1}{c}\big[\nabla F_i(w) - \nabla F_i(w')\big] + \big[\nabla G_{\gamma,i}(w) - \nabla G_{\gamma,i}(w')\big]\right\|$$
$$\leq\; \frac{1}{c}\|\nabla F_i(w) - \nabla F_i(w')\| + \|\nabla G_{\gamma,i}(w) - \nabla G_{\gamma,i}(w')\|.$$

By Assumption 4.1, the function F_i is L_F-smooth, which gives

$$\|\nabla F_i(w) - \nabla F_i(w')\| \;\leq\; L_F\|w - w'\|.$$

By Lemma D.2, the gap-function gradient satisfies

$$\|\nabla G_{\gamma,i}(w) - \nabla G_{\gamma,i}(w')\| \;\leq\; L_G\|w - w'\|.$$

Substituting these two bounds into the triangle-inequality estimate yields

$$\|\nabla\phi_{c,i}(w) - \nabla\phi_{c,i}(w')\| \;\leq\; \frac{L_F}{c}\|w - w'\| + L_G\|w - w'\| \;=\; \left(\frac{L_F}{c} + L_G\right)\|w - w'\|. \quad (61)$$

Since $c \geq c_0$ implies $\frac{1}{c} \leq \frac{1}{c_0}$, we conclude from equation 61 that

$$\|\nabla\phi_{c,i}(w) - \nabla\phi_{c,i}(w')\| \;\leq\; \left(\frac{L_F}{c_0} + L_G\right)\|w - w'\| \;=\; L_\phi\|w - w'\|, \quad (62)$$

which establishes equation 58 with L_ϕ independent of c (and hence of the iteration index k). The descent inequality equation 60 follows from the standard descent lemma for L_ϕ-smooth functions. □

Lemma D.4 is the cornerstone of our analysis. Although the unnormalized penalized objective $\psi_{c_k} = F + c_k G_\gamma$ has a smoothness constant $L_F + c_k L_G$ that grows with k, the normalized objective ϕ_{c_k} has the bounded smoothness constant $L_\phi = L_F/c_0 + L_G$. This uniform bound enables a single descent analysis that applies across all iterations.

Lemma D.5 (Auxiliary variable contraction)**.** *Consider the inner loop of Algorithm 1 for client i at outer iteration k, and let $\theta_i^{k,e}$ denote the auxiliary primal variable at inner iteration $e \in \{0, 1, \ldots, E\}$ with initial value $\theta_i^{k,0} = \theta_i^k$.*

Throughout the inner loop of Algorithm 1, the quantities x^k, $\tilde{y}_i$, and $\tilde{z}_i$ remain fixed: line 4 initializes $(\tilde{y}_i, \tilde{z}_i) = (y_i^k, z_i^k)$, and lines 5–8 update only the auxiliary variables. Consequently, the exact target is stationary and given by

$$\theta_i^\star \;:=\; \theta_i^*(x^k, \tilde{y}_i, \tilde{z}_i) \;=\; \theta_i^*(x^k, y_i^k, z_i^k) \;=\; \theta_i^*(w_i^k).$$

Under the step-size condition

$$\eta \;\leq\; \bar{\eta} := \min\big\{(L_f + r_2 L_{gy} + 1/\gamma_1)^{-1},\, \gamma_2\big\},$$

where the second entry is inherited from Lemma D.6; note $(L_f + r_2 L_{gy} + 1/\gamma_1)^{-1} \leq \gamma_1$ already, so $\eta \leq \bar{\eta}$ implies $\eta \leq \min\{\gamma_1, \gamma_2\}$ and the condition of Lemma D.8 holds. Under this condition, the auxiliary tracking error contracts geometrically,

$$\|\theta_i^{k,E} - \theta_i^\star\|^2 \;\leq\; \kappa_\theta^E\, \|\theta_i^{k,0} - \theta_i^\star\|^2, \quad (63)$$

with contraction rate $\kappa_\theta := 1 - \eta/\gamma_1 \in [0, 1)$.

Proof. At inner iteration e, the auxiliary update is

$$\theta_i^{k,e+1} \;=\; \mathrm{Proj}_{\mathcal{Y}}\Big(\theta_i^{k,e} - \eta\, \nabla_\theta \Phi_i(x^k, \tilde{y}_i, \tilde{z}_i, \theta_i^{k,e})\Big),$$

where the inner objective is

$$\Phi_i(x^k, \tilde{y}_i, \tilde{z}_i, \theta) \;=\; f_i(x^k, \theta) + \tilde{z}_i^\top g_i(x^k, \theta) + \frac{1}{2\gamma_1}\|\theta - \tilde{y}_i\|^2.$$

Since x^k, $\tilde{y}_i$, and $\tilde{z}_i$ remain fixed throughout the inner loop, the target $\theta_i^\star$ is stationary for all inner iterations. By Assumptions 4.2–4.3, the mapping $\Phi_i(x^k, \tilde{y}_i, \tilde{z}_i, \cdot)$ is μ-strongly convex on $\mathcal{Y}$ with $\mu = 1/\gamma_1$, since $f_i + \tilde{z}_i^\top g_i$ is convex in θ and the proximal term contributes $\|\theta - \tilde{y}_i\|^2/(2\gamma_1)$; it is L-smooth with $L \leq L_f + r_2 L_{gy} + 1/\gamma_1$, using $\|\tilde{z}_i\| \leq r_2$ on $\mathcal{Z}_i$. The step size satisfies $\eta \leq \bar{\eta} \leq 1/L$. Since $\theta_i^\star$ minimises Φ_i over $\mathcal{Y}$ it is a fixed point, $\theta_i^\star = \mathrm{Proj}_{\mathcal{Y}}\big(\theta_i^\star - \eta\nabla_\theta\Phi_i(\theta_i^\star)\big)$, and $\mathrm{Proj}_{\mathcal{Y}}$ is nonexpansive, so, abbreviating $\Delta\theta := \theta_i^{k,e} - \theta_i^\star$ and $\Delta\nabla := \nabla_\theta\Phi_i(\theta_i^{k,e}) - \nabla_\theta\Phi_i(\theta_i^\star)$,

$$\|\theta_i^{k,e+1} - \theta_i^\star\|^2 \;\leq\; \|\Delta\theta - \eta\,\Delta\nabla\|^2 \;=\; \|\Delta\theta\|^2 - 2\eta\langle\Delta\nabla, \Delta\theta\rangle + \eta^2\|\Delta\nabla\|^2.$$

Since $\Phi_i(x^k, \tilde{y}_i, \tilde{z}_i, \cdot)$ is L-smooth, its gradient is $1/L$-cocoercive, so $\|\Delta\nabla\|^2 \le L\langle\Delta\nabla, \Delta\theta\rangle$; with $\eta \le 1/L$ this gives $\eta^2\|\Delta\nabla\|^2 \le \eta\langle\Delta\nabla, \Delta\theta\rangle$ and hence

$$\|\theta_i^{k,e+1} - \theta_i^\star\|^2 \;\le\; \|\Delta\theta\|^2 - \eta\langle\Delta\nabla, \Delta\theta\rangle.$$

By μ-strong convexity the gradient is μ-strongly monotone, $\langle\Delta\nabla, \Delta\theta\rangle \ge \mu\|\Delta\theta\|^2$, which with $\mu = 1/\gamma_1$ gives

$$\|\theta_i^{k,e+1} - \theta_i^\star\|^2 \;\le\; \kappa_\theta\, \|\theta_i^{k,e} - \theta_i^\star\|^2, \qquad \kappa_\theta = 1 - \eta/\gamma_1.$$

Iterating this inequality from $e = 0$ to $e = E - 1$ yields

$$\|\theta_i^{k,E} - \theta_i^\star\|^2 \;\le\; \kappa_\theta^E\, \|\theta_i^{k,0} - \theta_i^\star\|^2,$$

which is precisely equation 63. □

Lemma D.6 (Dual auxiliary variable contraction)**.** *Consider the inner loop of Algorithm 1 for client i at outer iteration k, and let $\bar{\lambda}_i^{k,e}$ denote the auxiliary dual variable at inner iteration $e \in \{0, 1, \dots, E\}$ with initial value $\bar{\lambda}_i^{k,0} = \lambda_i^k$.*

Throughout the inner loop of Algorithm 1, the quantities x^k, $\tilde{y}_i$, and $\tilde{z}_i$ remain fixed: line 4 initializes $(\tilde{y}_i, \tilde{z}_i) = (y_i^k, z_i^k)$, and lines 5–8 update only the auxiliary variables. Consequently, the exact dual target is stationary and given by

$$\lambda_i^\star \;:=\; \mathrm{Proj}_{\mathbb{R}_+^{p_i}}\big(\tilde{z}_i + \gamma_2\, g_i(x^k, \tilde{y}_i)\big) \;=\; \mathrm{Proj}_{\mathbb{R}_+^{p_i}}\big(z_i^k + \gamma_2\, g_i(x^k, y_i^k)\big) \;=\; \lambda_i^*(w_i^k).$$

Under the step-size condition $\eta \le \gamma_2$, the dual tracking error contracts geometrically,

$$\|\bar{\lambda}_i^{k,E} - \lambda_i^\star\|^2 \;\le\; \kappa_\lambda^E\, \|\bar{\lambda}_i^{k,0} - \lambda_i^\star\|^2, \tag{64}$$

with contraction rate $\kappa_\lambda := 1 - \eta/\gamma_2 \in [0, 1)$.

Proof. For fixed $(x^k, \tilde{y}_i, \tilde{z}_i)$, the λ-subproblem in the gap function is

$$\max_{\lambda_i \in \mathbb{R}_+^{p_i}} \left\{ \lambda_i^\top g_i(x^k, \tilde{y}_i) - \frac{\|\lambda_i - \tilde{z}_i\|^2}{2\gamma_2} \right\}.$$

This objective is $1/\gamma_2$-strongly concave in λ_i, with unique maximizer $\lambda_i^\star$. The projected gradient ascent update is

$$\bar{\lambda}_i^{k,e+1} \;=\; \mathrm{Proj}_{\mathbb{R}_+^{p_i}}\left(\bar{\lambda}_i^{k,e} + \eta\left[g_i(x^k, \tilde{y}_i) - \frac{1}{\gamma_2}\big(\bar{\lambda}_i^{k,e} - \tilde{z}_i\big)\right]\right).$$

Since x^k, $\tilde{y}_i$, and $\tilde{z}_i$ remain fixed throughout the inner loop, the target $\lambda_i^\star$ is stationary for all inner iterations. Writing $t := \eta/\gamma_2 \in (0, 1]$ and $v_i := \tilde{z}_i + \gamma_2\, g_i(x^k, \tilde{y}_i)$, the update above is exactly an averaging step followed by a projection,

$$\bar{\lambda}_i^{k,e+1} \;=\; \mathrm{Proj}_{\mathbb{R}_+^{p_i}}\big((1-t)\,\bar{\lambda}_i^{k,e} + t\, v_i\big), \qquad \lambda_i^\star = \mathrm{Proj}_{\mathbb{R}_+^{p_i}}(v_i).$$

The target is a fixed point of the same map: componentwise, if $v_{i,j} \ge 0$ then $\lambda_{i,j}^\star = v_{i,j}$ and $(1-t)v_{i,j} + t\, v_{i,j} = v_{i,j} \ge 0$, while if $v_{i,j} < 0$ then $\lambda_{i,j}^\star = 0$ and $(1-t)\cdot 0 + t\, v_{i,j} < 0$ projects to 0; in both cases $\mathrm{Proj}_{\mathbb{R}_+^{p_i}}((1-t)\lambda_i^\star + t\, v_i) = \lambda_i^\star$. Since $\mathrm{Proj}_{\mathbb{R}_+^{p_i}}$ is nonexpansive,

$$\|\bar{\lambda}_i^{k,e+1} - \lambda_i^\star\| \;\le\; (1-t)\,\|\bar{\lambda}_i^{k,e} - \lambda_i^\star\|,$$

and squaring, with $(1-t)^2 \le 1 - t$ for $t \in [0, 1]$,

$$\|\bar{\lambda}_i^{k,e+1} - \lambda_i^\star\|^2 \;\le\; \kappa_\lambda\, \|\bar{\lambda}_i^{k,e} - \lambda_i^\star\|^2, \qquad \kappa_\lambda = 1 - \eta/\gamma_2.$$

Iterating this inequality from $e = 0$ to $e = E - 1$ yields

$$\|\bar{\lambda}_i^{k,E} - \lambda_i^\star\|^2 \;\le\; \kappa_\lambda^E\, \|\bar{\lambda}_i^{k,0} - \lambda_i^\star\|^2,$$

which proves equation 64. □

At the endpoint $\eta = \gamma_2$, admissible whenever $\gamma_2 \le (L_f + r_2 L_{gy} + 1/\gamma_1)^{-1}$, the factor κ_λ vanishes and the update collapses, for every e, to $\bar{\lambda}_i^{k,e+1} = \mathrm{Proj}_{\mathbb{R}_+^{p_i}}(\tilde{z}_i + \gamma_2 g_i(x^k, \tilde{y}_i)) = \lambda_i^\star$. A single step is then exact, and the loop over e is needed only when the primal block $\bar{\theta}_i$ forces $\eta < \gamma_2$.

Definition D.7 and Lemma D.8 hold for any participation rate; Theorem D.16 specializes them to *full participation* ($p = 1$), where every stored gradient is refreshed each round, so no staleness from skipped rounds arises.

Definition D.7 (Minimum number of inner iterations)**.** Let $\alpha > 0$ and $\eta \in (0, \bar{\eta}]$, and define

$$\kappa_\theta := 1 - \eta/\gamma_1, \qquad \kappa_\lambda := 1 - \eta/\gamma_2.$$

Both contraction factors belong to $[0, 1)$. Define

$$E_0 := \min \left\{ e \in \{1, 2, \ldots\} : \begin{array}{l} \kappa_\theta^e \le \min\left\{\frac{1}{4}, \frac{C_\theta \eta}{8\gamma_1 L_e^2 \alpha}\right\}, \\ \kappa_\lambda^e \le \min\left\{\frac{1}{4}, \frac{C_\lambda \eta}{8\gamma_2 L_e^2 \alpha}\right\} \end{array} \right\}.$$

This integer is finite because both contraction factors are strictly less than one. If either contraction factor is zero, its corresponding condition holds for every $e \ge 1$. For positive contraction factors, the definition is equivalently evaluated using the logarithmic ceiling formulas.

Lemma D.8 (Single-client descent)**.** *Under Assumptions 4.1– 4.3, consider a participating client $i \in \mathcal{A}_k$ with step sizes*

$$\alpha_k \;=\; \frac{\alpha}{c_k}, \qquad \beta_k \;=\; \frac{\beta}{c_k}, \qquad \eta_k \;=\; \eta,$$

where $\beta = \alpha$ and the base parameters $\alpha, \eta > 0$ satisfy

$$\alpha \;\le\; \bar{\alpha} \;:=\; \min\left\{ \frac{1}{4L_\phi},\, \frac{1}{64 C_\theta L_\theta^2},\, \frac{1}{64 C_\lambda L_\lambda^2} \right\}, \qquad \eta \;\le\; \min\{\bar{\eta},\, \gamma_1,\, \gamma_2\}, \qquad E \;\ge\; E_0.$$

Writing $u_i := (y_i, z_i)$ for the local (y, z) block and $\mathcal{U}_i := \mathcal{Y} \times \mathcal{Z}_i$ for its feasible set, the local merit function satisfies

$$\begin{aligned} V_{k+1,i} - V_{k,i} \;\le\; & -\frac{1}{16\alpha}\|u_i^{k+1} - u_i^k\|^2 + \left\langle \nabla_x \phi_{c_k,i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k \right\rangle \\ & + \left(\frac{L_\phi}{2} + 2C_\theta L_\theta^2 + 2C_\lambda L_\lambda^2 \right) \|x^{k+1} - x^k\|^2 \\ & - \frac{C_\theta\, \eta}{4\gamma_1}\|\theta_i^k - \theta_i^*(w_i^k)\|^2 - \frac{C_\lambda\, \eta}{4\gamma_2}\|\lambda_i^k - \lambda_i^*(w_i^k)\|^2. \end{aligned} \tag{65}$$

Here $V_{k+1,i}$ carries the penalty c_{k+1} while the right-hand side is evaluated at c_k; the omitted difference $(1/c_{k+1} - 1/c_k)(F_i(w_i^{k+1}) - \underline{F}_i)$ is nonpositive because $c_{k+1} \ge c_k$, so the inequality holds as stated.

Proof. The local update modifies only the tuple $(y_i, z_i, \theta_i, \lambda_i)$, whereas the broadcast variable x^k is shared across all clients. Since $\beta = \alpha$, lines 9–10 of Algorithm 1 constitute a single projected step on the product set $\mathcal{U}_i$,

$$u_i^{k+1} \;=\; \mathrm{Proj}_{\mathcal{U}_i}\big(u_i^k - \alpha\, d_i^k\big), \qquad d_i^k := \widetilde{\nabla}_{(y,z)} \phi_{c_k,i}(w_i^k), \tag{66}$$

while $x^{k+1} - x^k$ is determined by the server-side aggregation step and is identical across clients.

Define the local approximation error

$$e_i^k \;:=\; \widetilde{\nabla}_{(y,z)} \phi_{c_k,i}(w_i^k) - \nabla_{(y,z)} \phi_{c_k,i}(w_i^k).$$

By Definition D.3 together with Lemmas D.5 and D.6, after $E \ge E_0$ inner iterations the approximation error satisfies

$$\|e_i^k\|^2 \;\le\; L_e^2 \left(\kappa_\theta^E \|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \kappa_\lambda^E \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2\right) \;\le\; \frac{L_e^2}{4}\left(\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2\right), \tag{67}$$

where the second inequality applies the bounds $\kappa_\theta^E \le 1/4$ and $\kappa_\lambda^E \le 1/4$ from Definition D.7.

We decompose the change in the normalized merit function by splitting the move $(x^k, u_i^k) \mapsto (x^{k+1}, u_i^{k+1})$ into a (y, z)-update at fixed x^k followed by an x-update at fixed u_i^{k+1}:

$$\begin{aligned} \phi_{c_k,i}(w_i^{k+1}) - \phi_{c_k,i}(w_i^k) \;=\; & \underbrace{\left[\phi_{c_k,i}(x^k, u_i^{k+1}) - \phi_{c_k,i}(x^k, u_i^k)\right]}_{=:\ \text{(A)}} \\ & + \underbrace{\left[\phi_{c_k,i}(x^{k+1}, u_i^{k+1}) - \phi_{c_k,i}(x^k, u_i^{k+1})\right]}_{=:\ \text{(B)}}. \end{aligned}$$

We bound (A) and (B) in turn.

Step 1: descent in the (y, z) block. Write $\Delta u_i := u_i^{k+1} - u_i^k$. Since $u_i^k \in \mathcal{U}_i$, the variational characterisation of the projection in equation 66 with test point u_i^k gives $\langle u_i^k - \alpha d_i^k - u_i^{k+1},\, u_i^k - u_i^{k+1}\rangle \le 0$, that is,

$$\big\langle d_i^k,\, \Delta u_i\big\rangle \;\le\; -\frac{1}{\alpha}\|\Delta u_i\|^2. \tag{68}$$

No unconstrained identity for Δu_i is used; this is the projected counterpart of the sufficient-decrease step for projected gradient methods (Beck, 2017, Lemma 5.7), as used in the centralized analysis (Yao et al., 2025).

By Lemma D.4, the map $(y, z) \mapsto \phi_{c_k,i}(x^k, y, z)$ is L_ϕ-smooth, so

$$\text{(A)} \;\le\; \big\langle \nabla_{(y,z)}\phi_{c_k,i}(w_i^k),\, \Delta u_i\big\rangle + \frac{L_\phi}{2}\|\Delta u_i\|^2.$$

Writing $\nabla_{(y,z)}\phi_{c_k,i}(w_i^k) = d_i^k - e_i^k$, applying equation 68, and using Young's inequality with parameter α on the error term,

$$\big\langle \nabla_{(y,z)}\phi_{c_k,i}(w_i^k),\, \Delta u_i\big\rangle \;\le\; -\frac{1}{\alpha}\|\Delta u_i\|^2 + \frac{\alpha}{2}\|e_i^k\|^2 + \frac{1}{2\alpha}\|\Delta u_i\|^2.$$

Combining the two displays and using $\alpha \le 1/(4L_\phi)$, so that $\frac{1}{2\alpha} - \frac{L_\phi}{2} \ge \frac{1}{4\alpha}$, yields

$$\text{(A)} \;\le\; -\frac{1}{4\alpha}\|u_i^{k+1} - u_i^k\|^2 + \frac{\alpha}{2}\|e_i^k\|^2. \tag{69}$$

Step 2: variation in the x block. Applying the L_ϕ-smoothness of the map $x \mapsto \phi_{c_k,i}(x, u_i^{k+1})$ gives

$$\text{(B)} \;\le\; \big\langle \nabla_x\phi_{c_k,i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k\big\rangle + \frac{L_\phi}{2}\|x^{k+1} - x^k\|^2. \tag{70}$$

Step 3: auxiliary-variable tracking. Since $\theta_i^{k+1} = \theta_i^{k,E}$ is produced by the inner tracking loop with target $\theta_i^*(w_i^k)$, Lemma D.5 gives

$$\|\theta_i^{k+1} - \theta_i^*(w_i^k)\|^2 \;\le\; \kappa_\theta^E\|\theta_i^k - \theta_i^*(w_i^k)\|^2.$$

Combining this contraction bound with Lemma D.1 and the identity

$$\|w_i^{k+1} - w_i^k\|^2 \;=\; \|x^{k+1} - x^k\|^2 + \|u_i^{k+1} - u_i^k\|^2$$

yields

$$\|\theta_i^{k+1} - \theta_i^*(w_i^{k+1})\|^2 \;\le\; 2\kappa_\theta^E\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + 2L_\theta^2\big(\|x^{k+1} - x^k\|^2 + \|u_i^{k+1} - u_i^k\|^2\big). \tag{71}$$

Applying $\kappa_\theta^E \le 1/4$ to equation 71 and rearranging yields

$$C_\theta\|\theta_i^{k+1} - \theta_i^*(w_i^{k+1})\|^2 - C_\theta\|\theta_i^k - \theta_i^*(w_i^k)\|^2 \;\le\; -\frac{C_\theta}{2}\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + 2C_\theta L_\theta^2\big(\|x^{k+1} - x^k\|^2 + \|u_i^{k+1} - u_i^k\|^2\big).$$

The same argument applied to λ_i gives

$$C_\lambda\|\lambda_i^{k+1} - \lambda_i^*(w_i^{k+1})\|^2 - C_\lambda\|\lambda_i^k - \lambda_i^*(w_i^k)\|^2 \;\le\; -\frac{C_\lambda}{2}\|\lambda_i^k - \lambda_i^*(w_i^k)\|^2 + 2C_\lambda L_\lambda^2\big(\|x^{k+1} - x^k\|^2 + \|u_i^{k+1} - u_i^k\|^2\big).$$

Step 4: conclusion. Combining equation 69, equation 70, and the two tracking bounds from Step 3 yields

$$\begin{aligned} V_{k+1,i} - V_{k,i} \;\le\; &-\left(\frac{1}{4\alpha} - 2C_\theta L_\theta^2 - 2C_\lambda L_\lambda^2\right)\|u_i^{k+1} - u_i^k\|^2 \\ &+ \big\langle \nabla_x\phi_{c_k,i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k\big\rangle \\ &+ \left(\frac{L_\phi}{2} + 2C_\theta L_\theta^2 + 2C_\lambda L_\lambda^2\right)\|x^{k+1} - x^k\|^2 \\ &- \left(\frac{C_\theta}{2} - \frac{\alpha L_e^2\kappa_\theta^E}{2}\right)\|\theta_i^k - \theta_i^*(w_i^k)\|^2 \\ &- \left(\frac{C_\lambda}{2} - \frac{\alpha L_e^2\kappa_\lambda^E}{2}\right)\|\lambda_i^k - \lambda_i^*(w_i^k)\|^2. \end{aligned}$$

The bound $\alpha \le \min\{1/(64C_\theta L_\theta^2),\, 1/(64C_\lambda L_\lambda^2)\}$ gives $2C_\theta L_\theta^2 + 2C_\lambda L_\lambda^2 \le 1/(16\alpha)$ and hence

$$\frac{1}{4\alpha} - 2C_\theta L_\theta^2 - 2C_\lambda L_\lambda^2 \;\ge\; \frac{3}{16\alpha} \;\ge\; \frac{1}{16\alpha}.$$

Moreover, Definition D.7 together with the conditions $\eta \le \gamma_1$ and $\eta \le \gamma_2$ implies that the coefficients of the tracking-error terms are bounded below by $C_\theta\eta/(4\gamma_1)$ and $C_\lambda\eta/(4\gamma_2)$, respectively. Substituting these three lower bounds into the preceding display proves equation 65. □

Lemma D.9 (Non-participating client drift)**.** *Let $\tau \in (0, 1]$ be a free parameter and let $i \notin \mathcal{A}_k$ be a non-participating client, so that*

$$y_i^{k+1} = y_i^k, \qquad z_i^{k+1} = z_i^k, \qquad \theta_i^{k+1} = \theta_i^k, \qquad \lambda_i^{k+1} = \lambda_i^k.$$

Then the local merit function satisfies

$$\begin{aligned} V_{k+1,i} - V_{k,i} \;\le\; &\big\langle \nabla_x \phi_{c_{k+1},i}(w_i^k),\, x^{k+1} - x^k \big\rangle + L^{(1)}_{\mathrm{drift}}(\tau)\, \|x^{k+1} - x^k\|^2 \\ &+ \tau\, C_\theta \|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \tau\, C_\lambda \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2 + \Delta_c^{(i)}, \end{aligned} \tag{72}$$

where

$$L^{(1)}_{\mathrm{drift}}(\tau) \;:=\; \frac{L_\phi}{2} + \Big(1 + \frac{1}{\tau}\Big)\big(C_\theta L_\theta^2 + C_\lambda L_\lambda^2\big), \qquad \Delta_c^{(i)} \;:=\; \Big(\frac{1}{c_{k+1}} - \frac{1}{c_k}\Big)\big(F_i(w_i^k) - \underline{F}_i\big) \;\le\; 0.$$

Since $u_i^{k+1} = u_i^k$ for $i \notin \mathcal{A}_k$, the inner product in equation 72 may equivalently be written $\big\langle \nabla_x \phi_{c_{k+1},i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k \big\rangle$, matching the form appearing in Lemma D.8.

Proof. For a non-participating client, the only change in the local state comes from the global variable x. Accordingly, define

$$w_i^{k+1} \;:=\; (x^{k+1}, y_i^k, z_i^k), \qquad w_i^k \;:=\; (x^k, y_i^k, z_i^k).$$

We decompose the merit-function increment as

$$\begin{aligned} V_{k+1,i} - V_{k,i} \;=\; &\underbrace{\big[\phi_{c_{k+1},i}(w_i^{k+1}) - \phi_{c_{k+1},i}(w_i^k)\big]}_{\text{(I-a): spatial variation}} + \underbrace{\big[\phi_{c_{k+1},i}(w_i^k) - \phi_{c_k,i}(w_i^k)\big]}_{\text{(I-b): penalty variation}} \\ &+ C_\theta \underbrace{\big[\|\theta_i^k - \theta_i^*(w_i^{k+1})\|^2 - \|\theta_i^k - \theta_i^*(w_i^k)\|^2\big]}_{\text{(II): primal auxiliary variation}} \\ &+ C_\lambda \underbrace{\big[\|\lambda_i^k - \lambda_i^*(w_i^{k+1})\|^2 - \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2\big]}_{\text{(III): dual auxiliary variation}}. \end{aligned} \tag{73}$$

We bound each of the four terms in turn.

Step 1: spatial variation. By Lemma D.4, the map $x \mapsto \phi_{c_{k+1},i}(x, y_i^k, z_i^k)$ is L_ϕ-smooth, which gives

$$\text{(I-a)} \;\le\; \big\langle \nabla_x \phi_{c_{k+1},i}(w_i^k),\, x^{k+1} - x^k \big\rangle + \frac{L_\phi}{2}\|x^{k+1} - x^k\|^2. \tag{74}$$

We retain the inner product rather than bounding it by Young's inequality. The server update aggregates the contributions of all N clients (Algorithm 1, line 17), so the corresponding inner products combine across the whole client population and cancel against the descent in the shared variable; see Theorem D.14. Bounding them client by client would discard this cancellation and leave a term that is not proportional to the step size.

Step 2: penalty variation. Since $c_{k+1} \ge c_k$ and $F_i(w_i^k) - \underline{F}_i \ge 0$,

$$\text{(I-b)} \;=\; \Big(\frac{1}{c_{k+1}} - \frac{1}{c_k}\Big)\big(F_i(w_i^k) - \underline{F}_i\big) \;=\; \Delta_c^{(i)} \;\le\; 0. \tag{75}$$

Step 3: auxiliary variation in θ_i. Define the shift $\delta\theta_i^* := \theta_i^*(w_i^{k+1}) - \theta_i^*(w_i^k)$, which satisfies $\|\delta\theta_i^*\| \le L_\theta \|x^{k+1} - x^k\|$ by Lemma D.1. Applying the identity $\|a-b\|^2 - \|a\|^2 = -2\langle a, b\rangle + \|b\|^2$ with $a = \theta_i^k - \theta_i^*(w_i^k)$ and $b = \delta\theta_i^*$ yields

$$\text{(II)} \;=\; -2\langle \theta_i^k - \theta_i^*(w_i^k),\, \delta\theta_i^* \rangle + \|\delta\theta_i^*\|^2.$$

Young's inequality with parameter τ bounds the inner product by $\tau\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \frac{1}{\tau}\|\delta\theta_i^*\|^2$, so that, with the Lipschitz estimate,

$$\text{(II)} \;\le\; \tau\, \|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \Big(1 + \frac{1}{\tau}\Big) L_\theta^2\, \|x^{k+1} - x^k\|^2. \tag{76}$$

The parameter τ trades the drift charged to the tracking error against the drift charged to the motion of the shared variable. It is an analysis parameter only: it does not appear in Algorithm 1, and the algorithm is unchanged. Taking $\tau = 1$ recovers the elementary bound.

Step 4: auxiliary variation in λ_i. Define $\delta\lambda_i^* := \lambda_i^*(w_i^{k+1}) - \lambda_i^*(w_i^k)$, which satisfies $\|\delta\lambda_i^*\| \le L_\lambda \|x^{k+1} - x^k\|$ by Lemma D.1. Repeating the argument of Step 3 with the same parameter τ yields

$$\text{(III)} \;\le\; \tau\, \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2 + \Big(1 + \frac{1}{\tau}\Big) L_\lambda^2\, \|x^{k+1} - x^k\|^2. \tag{77}$$

Step 5: conclusion. Substituting equation 74, equation 75, equation 76, and equation 77 into equation 73 gives equation 72. □

The quantity $\Delta_c^{(i)}$ is non-positive under the normalized formulation: a useful structural distinction from the unnormalized penalty method, for which the corresponding penalty increment is generally nonnegative.

Since $\tau \le 1$ gives $1 + 1/\tau \ge 2$, the coefficient $L_{\text{drift}}^{(1)}(\tau)$ dominates the coefficient of $\|x^{k+1} - x^k\|^2$ in Lemma D.8, so both populations may be bounded using the same constant. Together with the retained inner product of Step 1, this is what allows active and inactive clients to be treated by a single aggregate argument in Theorem D.14. Under a fixed-penalty schedule $c_k \equiv c$ the penalty-variation term vanishes identically, $\Delta_c^{(i)} = 0$, and the penalty indices in Lemmas D.8 and D.9 coincide. Under the growing-penalty schedule the two indices differ, but Theorem D.16 assumes full participation, so no client is inactive and Lemma D.9 is never invoked.

Lemma D.10 (Global variable update bound)**.** *Under Assumption 2.1, with step size $\alpha_x > 0$, the global variable update satisfies*

$$\|x^{k+1} - x^k\|^2 \;\le\; D_{\mathcal{X}}^2, \tag{78}$$

where $D_{\mathcal{X}} := \sup_{a,b\in\mathcal{X}} \|a - b\|$ denotes the diameter of $\mathcal{X}$, finite by Assumption 2.1.

Proof. Since $\mathcal{X}$ is compact by Assumption 2.1, the update rule $x^{k+1} = \text{Proj}_{\mathcal{X}}(x^k - \alpha_x d_x^k)$ together with $x^k \in \mathcal{X}$ implies

$$\|x^{k+1} - x^k\| \;=\; \|\text{Proj}_{\mathcal{X}}(x^k - \alpha_x d_x^k) - x^k\| \;\le\; \sup_{a,b\in\mathcal{X}} \|a - b\| \;=\; D_{\mathcal{X}}.$$

Squaring both sides yields $\|x^{k+1} - x^k\|^2 \le D_{\mathcal{X}}^2$. □

The next lemma is not used in the proofs of Theorems D.14, D.16 or D.21. It records the sampling term that the server-side gradient table removes: because the shared-variable update reads one stored entry from *every* client rather than averaging the participating clients' gradients, no client-sampling deviation $\frac{1}{A}\sum_{i\in\mathcal{A}_k} \nabla\phi_{c,i} - \frac{1}{N}\sum_{i=1}^N \nabla\phi_{c,i}$ ever appears, and the heterogeneity constants of Definition 4.4 do not enter the rates.

Lemma D.11 (Heterogeneity and sampling variance bounds)**.** *With the client heterogeneity constants for the objective F and the gap function G_γ as in Definition 4.4, the heterogeneity of the normalized gradients satisfies*

$$\zeta_\phi^2 := \sup_w \frac{1}{N}\sum_{i=1}^N \|\nabla\phi_{c,i}(w) - \nabla\phi_c(w)\|^2 \le \frac{2\zeta_F^2}{c_0^2} + 2\zeta_G^2. \tag{79}$$

Moreover, if $\mathcal{A}_k$ is sampled uniformly without replacement with $|\mathcal{A}_k| = A$, then for any vectors $\{v_i\}_{i=1}^N$ and $\bar{v} := N^{-1}\sum_{i=1}^N v_i$,

$$\mathbb{E}\left[\left\|\frac{1}{A}\sum_{i\in\mathcal{A}_k} v_i - \bar{v}\right\|^2\right] = \frac{N-A}{A(N-1)}\cdot\frac{1}{N}\sum_{i=1}^N \|v_i-\bar{v}\|^2. \tag{80}$$

In particular,

$$\mathbb{E}\left[\left\|\frac{1}{A}\sum_{i\in\mathcal{A}_k} v_i - \bar{v}\right\|^2\right] \le \frac{1}{A}\cdot\frac{1}{N}\sum_{i=1}^N \|v_i-\bar{v}\|^2. \tag{81}$$

Proof. We first prove the heterogeneity bound for the normalized gradient. By definition,

$$\phi_{c,i}(w) = \frac{1}{c}\big(F_i(w) - \underline{F}_i\big) + G_{\gamma,i}(w),$$

and hence

$$\nabla\phi_{c,i}(w) = \frac{1}{c}\nabla F_i(w) + \nabla G_{\gamma,i}(w).$$

Averaging over clients gives

$$\nabla\phi_c(w) = \frac{1}{c}\nabla F(w) + \nabla G_\gamma(w),$$

where

$$\nabla F(w) = \frac{1}{N}\sum_{i=1}^N \nabla F_i(w), \qquad \nabla G_\gamma(w) = \frac{1}{N}\sum_{i=1}^N \nabla G_{\gamma,i}(w).$$

Therefore,

$$\nabla\phi_{c,i}(w) - \nabla\phi_c(w) = \frac{1}{c}\big(\nabla F_i(w) - \nabla F(w)\big) + \big(\nabla G_{\gamma,i}(w) - \nabla G_\gamma(w)\big).$$

Using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, averaging over i, and using $c \ge c_0$, we obtain

$$\zeta_\phi^2 \le \frac{2\zeta_F^2}{c^2} + 2\zeta_G^2 \le \frac{2\zeta_F^2}{c_0^2} + 2\zeta_G^2.$$

This proves equation 79.

We now prove the sampling variance identity. Let

$$\hat{v} := \frac{1}{A}\sum_{i\in\mathcal{A}_k} v_i.$$

Then

$$\|\hat{v}-\bar{v}\|^2 = \frac{1}{A^2}\sum_{i\in\mathcal{A}_k}\sum_{j\in\mathcal{A}_k}\langle v_i-\bar{v},\, v_j-\bar{v}\rangle.$$

For uniform sampling without replacement,

$$\mathbb{P}(i\in\mathcal{A}_k) = \frac{A}{N}, \qquad \mathbb{P}(i,j\in\mathcal{A}_k) = \frac{A(A-1)}{N(N-1)} \quad (i\ne j).$$

Since $\sum_{i=1}^N (v_i-\bar{v}) = 0$, we also have

$$\sum_{i\ne j}\langle v_i-\bar{v},\, v_j-\bar{v}\rangle = -\sum_{i=1}^N \|v_i-\bar{v}\|^2.$$

Taking expectation and substituting the inclusion probabilities gives

$$\begin{aligned}\mathbb{E}\|\hat{v}-\bar{v}\|^2 &= \frac{1}{A^2}\left[\frac{A}{N} - \frac{A(A-1)}{N(N-1)}\right]\sum_{i=1}^N \|v_i-\bar{v}\|^2 \\ &= \frac{N-A}{A(N-1)}\cdot\frac{1}{N}\sum_{i=1}^N \|v_i-\bar{v}\|^2.\end{aligned}$$

This proves equation 80. Since $(N-A)/(N-1) \le 1$, the bound equation 81 follows. □

Remark D.12. The factor $(N-A)/(N-1)$ in equation 80 is the finite-population correction. It equals zero under full participation ($A = N$) and approaches one when $A \ll N$. Consequently, equation 81 shows that the sampling variance scales as $O(1/A)$, which matches the standard mini-batch averaging behavior. In particular, applying equation 81 with $v_i = \nabla\phi_{c,i}$ gives

$$\mathbb{E}\left[\left\|\frac{1}{A}\sum_{i\in\mathcal{A}_k}\nabla\phi_{c,i} - \nabla\phi_c\right\|^2\right] \le \frac{\zeta_\phi^2}{A}.$$

Lemma D.13 (Stationarity measure bound)**.** *Consider the full-participation regime used in this part of the analysis, with the aggregate normalized objective ϕ_{c_k}, the block scaling D and the gradient mapping equation 12 as defined in Section 4. Then for every $k \ge 0$,*

$$\mathbb{E}\big[\|\mathcal{G}_\alpha(w^k)\|^2\big] \le \frac{C_1}{\alpha^2}\mathbb{E}\big[\|w^{k+1}-w^k\|^2\big] + C_2\,\mathbb{E}\big[\Theta^k+\Lambda^k\big] + C_2'\,\mathbb{E}\Big[\tfrac{1}{N}\textstyle\sum_{i=1}^N\|u_i^{k+1}-u_i^k\|^2\Big], \tag{82}$$

where $C_1 = 2$, $C_2 = 2(N+4)L_e^2\kappa_{\max}^E$ with $\kappa_{\max} := \max\{\kappa_\theta, \kappa_\lambda\}$, and $C_2' = 8L_e^2(L_\theta^2+L_\lambda^2)+4L_\phi^2$; all constants are independent of c_k. Moreover, the stationarity measure evaluated at the updated iterate satisfies

$$R_k \;\le\; c_k\,(1+\alpha L_\phi)\,\|\mathcal{G}_\alpha(w^k)\| \;+\; c_k\,(2+\alpha L_\phi)\,\|e_{\mathrm{aux}}^k\|. \tag{83}$$

Proof. We first relate the penalized objective to the normalized merit function. Since

$$\psi_{c_k} = F + c_k G_\gamma, \qquad \phi_{c_k} = \frac{1}{c_k}(F-\underline{F}) + G_\gamma,$$

we have

$$\nabla\psi_{c_k} = \nabla F + c_k\nabla G_\gamma = c_k\nabla\phi_{c_k}. \tag{84}$$

The residual is evaluated at the concatenated iterate $w^k = (x^k, u_1^k, \dots, u_N^k)$. The local blocks of $D\nabla\phi_{c_k}(w^k)$ are the exact per-client gradients $\nabla_{(y,z)}\phi_{c_k,i}(w_i^k)$, so no averaging across clients enters them and consequently no heterogeneity term arises; the client average appears only in the shared block $\nabla_x\phi_{c_k}(w^k)$, which is the quantity the server actually forms.

Step 1: comparing the realised step with the exact one. In the full-participation regime, Algorithm 1 performs the projected step

$$w^{k+1} = \mathrm{Proj}_{\mathcal{W}}\big(w^k - \alpha\widetilde{\nabla}^k\big), \qquad \widetilde{\nabla}^k := D\nabla\phi_{c_k}(w^k) + e_{\mathrm{aux}}^k,$$

blockwise. Indeed, with $\alpha_x = \alpha/c_k$ line 17 of Algorithm 1 gives $x^{k+1} = \mathrm{Proj}_{\mathcal{X}}(x^k - \alpha\widetilde{h}_x^k)$, and with $\alpha_k = \beta_k = \alpha/c_k$ lines 9–10 give $u_i^{k+1} = \mathrm{Proj}_{\mathcal{U}_i}(u_i^k - \alpha d_i^k)$ with d_i^k as in equation 66; these are the blocks of $D\nabla\phi_{c_k}(w^k)$ perturbed by the finite-inner-loop approximation error $e_{\mathrm{aux}}^k = (e_x^k, e_1^k, \dots, e_N^k)$. Since the projection onto the closed convex set $\mathcal{W}$ is nonexpansive,

$$\Big\|\mathcal{G}_\alpha(w^k) - \frac{1}{\alpha}\big(w^k - w^{k+1}\big)\Big\| = \frac{1}{\alpha}\Big\|\mathrm{Proj}_{\mathcal{W}}\big(w^k - \alpha D\nabla\phi_{c_k}(w^k)\big) - \mathrm{Proj}_{\mathcal{W}}\big(w^k - \alpha\widetilde{\nabla}^k\big)\Big\| \le \|e_{\mathrm{aux}}^k\|. \tag{85}$$

No unconstrained identity for $w^{k+1} - w^k$ is used.

Step 2: bounding the approximation error. The local blocks e_i^k are the errors bounded in equation 67, so $\sum_{i=1}^N\|e_i^k\|^2 \le NL_e^2\kappa_{\max}^E(\Theta^k+\Lambda^k)$. The shared block e_x^k is evaluated at (x^k, u_i^{k+1}) with plug-ins $(\theta_i^{k+1}, \lambda_i^{k+1})$, whereas the inner loop tracks the target at w_i^k; by Jensen's inequality, Definition D.3 and the splitting

$$\|\theta_i^{k+1} - \theta_i^*(x^k, u_i^{k+1})\|^2 \le 2\kappa_\theta^E\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + 2L_\theta^2\|u_i^{k+1} - u_i^k\|^2$$

(Lemmas D.5 and D.1, and analogously for λ_i via Lemma D.6), the plug-in inexactness relative to $h_x^k := \frac{1}{N}\sum_{i=1}^N\nabla_x\phi_{c_k,i}(x^k, u_i^{k+1})$ obeys

$$\|\widetilde{h}_x^k - h_x^k\|^2 \le 2L_e^2\kappa_{\max}^E(\Theta^k+\Lambda^k) + 2L_e^2(L_\theta^2+L_\lambda^2)\,\tfrac{1}{N}\textstyle\sum_{i=1}^N\|u_i^{k+1}-u_i^k\|^2.$$

The shared block of e_{aux}^k is measured against $\nabla_x\phi_{c_k}(w^k)$, which evaluates the local blocks at u_i^k rather than u_i^{k+1}; by Jensen's inequality and the L_ϕ-smoothness of $\phi_{c_k,i}$ (Lemma D.4), $\|h_x^k - \nabla_x\phi_{c_k}(w^k)\|^2 \le L_\phi^2\,\frac{1}{N}\sum_{i=1}^N\|u_i^{k+1}-u_i^k\|^2$. Combining the two with $\|a+b\|^2 \le 2(\|a\|^2+\|b\|^2)$,

$$\|e_x^k\|^2 \le 4L_e^2\kappa_{\max}^E(\Theta^k+\Lambda^k) + \big(4L_e^2(L_\theta^2+L_\lambda^2) + 2L_\phi^2\big)\,\tfrac{1}{N}\textstyle\sum_{i=1}^N\|u_i^{k+1}-u_i^k\|^2.$$

Adding the two contributions,

$$\mathbb{E}\big[\|e^k_{\mathrm{aux}}\|^2\big] \le (N+4)L_e^2\kappa^E_{\max}\,\mathbb{E}\big[\Theta^k+\Lambda^k\big] + \big(4L_e^2(L_\theta^2+L_\lambda^2)+2L_\phi^2\big)\,\mathbb{E}\Big[\tfrac{1}{N}\textstyle\sum_{i=1}^N \|u_i^{k+1}-u_i^k\|^2\Big]. \tag{86}$$

Squaring equation 85, using $\|a+b\|^2 \le 2(\|a\|^2+\|b\|^2)$, and substituting equation 86 proves equation 82.

Step 3: from the gradient mapping to the residual. Write $\hat{w} := \mathrm{Proj}_{\mathcal{W}}(w^k - \alpha D\nabla\phi_{c_k}(w^k))$, so that $\mathcal{G}_\alpha(w^k) = (w^k-\hat{w})/\alpha$. The variational characterisation of the projection gives $w^k - \alpha D\nabla\phi_{c_k}(w^k) - \hat{w} \in \mathcal{N}_{\mathcal{W}}(\hat{w})$, that is,

$$\mathcal{G}_\alpha(w^k) - D\nabla\phi_{c_k}(w^k) \in \mathcal{N}_{\mathcal{W}}(\hat{w}).$$

Because $\mathcal{W} = \mathcal{X}\times\prod_i \mathcal{U}_i$ is a product, $\mathcal{N}_{\mathcal{W}}(\hat{w})$ is the corresponding product of normal cones, and D^{-1} rescales each block by a positive scalar; the cone is therefore invariant under D^{-1}, and

$$D^{-1}\mathcal{G}_\alpha(w^k) - \nabla\phi_{c_k}(w^k) + \nabla\phi_{c_k}(\hat{w}) \in \nabla\phi_{c_k}(\hat{w}) + \mathcal{N}_{\mathcal{W}}(\hat{w}).$$

By the L_ϕ-smoothness of ϕ_{c_k} (Lemma D.4), $\|\hat{w}-w^k\| = \alpha\|\mathcal{G}_\alpha(w^k)\|$ and $\|D^{-1}\|\le 1$,

$$\mathrm{dist}\big(0, \nabla\phi_{c_k}(\hat{w}) + \mathcal{N}_{\mathcal{W}}(\hat{w})\big) \le (1+\alpha L_\phi)\,\|\mathcal{G}_\alpha(w^k)\|. \tag{87}$$

Repeating the argument at $w^{k+1} = \mathrm{Proj}_{\mathcal{W}}(w^k - \alpha\widetilde{\nabla}^k)$ in place of $\hat{w}$, with $\widetilde{\nabla}^k$ in place of $D\nabla\phi_{c_k}(w^k)$ and $\widetilde{\mathcal{G}} := (w^k - w^{k+1})/\alpha$, gives

$$\mathrm{dist}\big(0, \nabla\phi_{c_k}(w^{k+1}) + \mathcal{N}_{\mathcal{W}}(w^{k+1})\big) \le (1+\alpha L_\phi)\,\|\widetilde{\mathcal{G}}\| + \|e^k_{\mathrm{aux}}\|,$$

and $\|\widetilde{\mathcal{G}}\| \le \|\mathcal{G}_\alpha(w^k)\| + \|e^k_{\mathrm{aux}}\|$ by equation 85. Multiplying by c_k and applying equation 84 proves equation 83. □

The residual is measured at the updated iterate w^{k+1} rather than at w^k. This is the standard convention for projected methods: the gradient mapping certifies stationarity at the point the projection produces, and $\mathcal{G}_\alpha(w^k) = 0$ holds precisely when w^k is stationary for ϕ_{c_k} over $\mathcal{W}$. When $\mathcal{W}$ has empty boundary the gradient mapping reduces to $D\nabla\phi_{c_k}(w^k)$ and the two conventions coincide; on the boundary, only the gradient mapping vanishes at a solution.

Lemma D.13 is used only in the full-participation analysis. The partial-participation case is handled separately in Theorem D.21, where the sampling and staleness effects are controlled through the augmented Lyapunov function.

Theorem D.14 (Global merit function descent)**.** *Under Assumptions 4.1–4.3, let $\tau \in (0,1]$ be the parameter of Lemma D.9 and suppose that the participation rate $p = A/N$ satisfies*

$$p \;\ge\; p_{\min}(\tau) + \delta, \qquad p_{\min}(\tau) \;:=\; \max\left\{\frac{4\gamma_1\tau}{4\gamma_1\tau+\eta},\, \frac{4\gamma_2\tau}{4\gamma_2\tau+\eta}\right\}, \tag{88}$$

for some constant $\delta \in (0, 1-p_{\min}(\tau))$. Then the global merit function satisfies

$$\begin{aligned}\mathbb{E}[V^G_{k+1}] \;\le\; &\mathbb{E}[V^G_k] - \frac{1}{16\alpha}\,\mathbb{E}\big[\|\Delta u^k\|^2\big] - \frac{p_{\mathrm{eff},\theta}\,C_\theta\,\eta}{4\gamma_1}\,\mathbb{E}\big[\Theta^k\big] - \frac{p_{\mathrm{eff},\lambda}\,C_\lambda\,\eta}{4\gamma_2}\,\mathbb{E}\big[\Lambda^k\big] \\ &+ \mathbb{E}\big[\langle H_k,\, x^{k+1}-x^k\rangle\big] + L^{(1)}_{\mathrm{drift}}(\tau)\,\mathbb{E}\big[\|x^{k+1}-x^k\|^2\big] + \mathbb{E}\big[\Delta^G_c\big],\end{aligned} \tag{89}$$

where the aggregate shared-variable gradient is

$$H_k \;:=\; \frac{1}{N}\sum_{i=1}^N \nabla_x\phi_{c,i}\big(x^k,\, u_i^{k+1}\big), \tag{90}$$

the averaged local displacement is

$$\|\Delta u^k\|^2 \;:=\; \frac{1}{N}\sum_{i=1}^N \|u_i^{k+1}-u_i^k\|^2, \qquad u_i \;:=\; (y_i, z_i),$$

the tracking errors Θ^k and Λ^k are as defined in Section 4, and the effective participation rates are

$$p_{\mathrm{eff},\theta} \;:=\; p - (1-p)\frac{4\gamma_1\tau}{\eta}, \qquad p_{\mathrm{eff},\lambda} \;:=\; p - (1-p)\frac{4\gamma_2\tau}{\eta}.$$

Moreover, the condition equation 88 ensures

$$p_{\mathrm{eff},\theta} \;\ge\; \delta\cdot\frac{4\gamma_1\tau+\eta}{\eta}, \qquad p_{\mathrm{eff},\lambda} \;\ge\; \delta\cdot\frac{4\gamma_2\tau+\eta}{\eta}. \tag{91}$$

The averaged penalty-variation term $\Delta_c^G := \frac{1}{N}\sum_{i=1}^N \Delta_c^{(i)} \le 0$ *vanishes under a fixed penalty schedule.*

Proof. For each client i, conditioning on the participation event yields

$$\mathbb{E}[V_{k+1,i} - V_{k,i}] \;=\; p\,\mathbb{E}[V_{k+1,i} - V_{k,i} \mid i \in \mathcal{A}_k] + (1-p)\,\mathbb{E}[V_{k+1,i} - V_{k,i} \mid i \notin \mathcal{A}_k].$$

If $i \in \mathcal{A}_k$, Lemma D.8 gives

$$\begin{aligned} V_{k+1,i} - V_{k,i} \;\le\; & -\frac{1}{16\alpha}\|u_i^{k+1} - u_i^k\|^2 + \big\langle \nabla_x \phi_{c,i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k \big\rangle + L^{(1)}_{\mathrm{drift}}(\tau)\|x^{k+1} - x^k\|^2 \\ & - \frac{C_\theta\,\eta}{4\gamma_1}\|\theta_i^k - \theta_i^*(w_i^k)\|^2 - \frac{C_\lambda\,\eta}{4\gamma_2}\|\lambda_i^k - \lambda_i^*(w_i^k)\|^2, \end{aligned}$$

where we used that $\tau \le 1$ implies $\frac{L_\phi}{2} + 2C_\theta L_\theta^2 + 2C_\lambda L_\lambda^2 \le L^{(1)}_{\mathrm{drift}}(\tau)$, so the same coefficient bounds both populations. If $i \notin \mathcal{A}_k$, Lemma D.9 yields

$$\begin{aligned} V_{k+1,i} - V_{k,i} \;\le\; & \big\langle \nabla_x \phi_{c,i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k \big\rangle + L^{(1)}_{\mathrm{drift}}(\tau)\|x^{k+1} - x^k\|^2 \\ & + \tau\, C_\theta \|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \tau\, C_\lambda \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2 + \Delta_c^{(i)}. \end{aligned}$$

Both conditional bounds contain the *same* inner product $\langle \nabla_x \phi_{c,i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k \rangle$ — in Lemma D.8 at the updated local state, and in Lemma D.9 at the retained state $u_i^{k+1} = u_i^k$ — together with the same coefficient $L^{(1)}_{\mathrm{drift}}(\tau)$ on $\|x^{k+1} - x^k\|^2$. Since u_i^{k+1} is defined for every client, these two contributions combine without a case split: averaging over clients and taking total expectation gives

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\langle \nabla_x \phi_{c,i}(x^k, u_i^{k+1}),\, x^{k+1} - x^k \rangle\big] \;=\; \mathbb{E}\big[\langle H_k,\, x^{k+1} - x^k \rangle\big],$$

with H_k as in equation 90. In particular no sampled average of client gradients appears, and no term proportional to $(1-p)$ but independent of the step size is generated. The displacement term combines in the same way: for $i \notin \mathcal{A}_k$ one has $u_i^{k+1} = u_i^k$, so $-\frac{1}{16\alpha}\|u_i^{k+1} - u_i^k\|^2 = 0$ and the bound of Lemma D.8 holds verbatim for both populations. Participation is therefore already encoded in the realised displacement, and this term too carries coefficient 1 rather than p. Averaging the remaining terms with weights p and $1-p$ and taking total expectation gives

$$\begin{aligned} \mathbb{E}[V_{k+1}^G - V_k^G] \;\le\; & -\frac{1}{16\alpha}\,\mathbb{E}\big[\|\Delta u^k\|^2\big] - \Big(p\frac{C_\theta \eta}{4\gamma_1} - (1-p)\,\tau\, C_\theta\Big)\mathbb{E}\big[\Theta^k\big] \\ & - \Big(p\frac{C_\lambda \eta}{4\gamma_2} - (1-p)\,\tau\, C_\lambda\Big)\mathbb{E}\big[\Lambda^k\big] + \mathbb{E}\big[\langle H_k, x^{k+1} - x^k\rangle\big] \\ & + L^{(1)}_{\mathrm{drift}}(\tau)\,\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] + \mathbb{E}\big[\Delta_c^G\big]. \end{aligned}$$

Substituting

$$p\frac{C_\theta\,\eta}{4\gamma_1} - (1-p)\,\tau\, C_\theta \;=\; \frac{p_{\mathrm{eff},\theta}\, C_\theta\,\eta}{4\gamma_1}, \qquad p\frac{C_\lambda\,\eta}{4\gamma_2} - (1-p)\,\tau\, C_\lambda \;=\; \frac{p_{\mathrm{eff},\lambda}\, C_\lambda\,\eta}{4\gamma_2}$$

proves equation 89.

It remains to establish equation 91. Define

$$\varphi(p) \;:=\; p_{\mathrm{eff},\theta} \;=\; p\left(1 + \frac{4\gamma_1\tau}{\eta}\right) - \frac{4\gamma_1\tau}{\eta},$$

which is linear and strictly increasing in p, with $\varphi(p_{\min}(\tau)) = 0$ at $p_{\min}(\tau) = 4\gamma_1\tau/(4\gamma_1\tau + \eta)$. Hence for $p \ge p_{\min}(\tau) + \delta$,

$$p_{\mathrm{eff},\theta} \;\ge\; \varphi(p_{\min}(\tau) + \delta) - \varphi(p_{\min}(\tau)) \;=\; \delta\left(1 + \frac{4\gamma_1\tau}{\eta}\right) \;=\; \delta \cdot \frac{4\gamma_1\tau + \eta}{\eta},$$

and the same argument with γ_2 in place of γ_1 gives the bound on $p_{\mathrm{eff},\lambda}$. □

Remark D.15. The threshold $p_{\min}(\tau)$ arises from balancing the contraction obtained from updated lower-level problems against the drift of those left unvisited. The parameter τ governs how that balance is struck: a smaller τ charges less of the inactive-client drift to the tracking error and more to the motion of the shared variable, so that $p_{\min}(\tau) \to 0$ as $\tau \to 0$. Every participation rate

$p > 0$ is therefore admissible, at the price of a smaller step size: since $L^{(1)}_{\mathrm{drift}}(\tau) = \Theta(1/\tau)$, the admissible range of α scales as $O(\tau)$ and the constants in the rate degrade accordingly. Choosing $\tau = \Theta(p)$ makes any fixed participation rate admissible while keeping the dependence on p explicit. The threshold is a sufficient condition for the present Lyapunov argument, not a necessary one for convergence.

Theorem D.16 (Convergence rate under full participation)**.** *Under Assumptions 4.1–4.3, suppose that full participation holds ($p = 1$), the step sizes satisfy*

$$\alpha \in (0, \bar{\alpha}_{\mathrm{full}}], \qquad \alpha_x = \frac{\alpha}{c_k}, \qquad \eta \in (0, \bar{\eta}],$$

and the penalty parameter follows the schedule $c_k = c_0(k+1)^\rho$ with $\rho \in (0, 1/2)$, where

$$\bar{\alpha}_{\mathrm{full}} := \min\left\{\bar{\alpha},\ \frac{1}{4L^{(1)}_{\mathrm{drift}}(1)},\ \frac{C_\theta \eta}{16\gamma_1 L_e^2},\ \frac{C_\lambda \eta}{16\gamma_2 L_e^2},\ \frac{1}{\sqrt{32L_e^2(L_\theta^2 + L_\lambda^2)}}\right\}.$$

Since $p = 1 > p_{\min}(\tau)$ for every $\tau \in (0, 1]$, we take $\tau = 1$ throughout this theorem, which minimises $L^{(1)}_{\mathrm{drift}}(\tau)$; the growing-penalty result is unaffected by the reparameterisation of Lemma D.9. Then the stationarity measure satisfies

$$\min_{0 \le k \le K-1} \mathbb{E}[R_k] = O\Big(K^{-(1-2\rho)/2}\Big). \tag{92}$$

Moreover, define the client-level penalized objective

$$\Psi_{c_k}(w^k) := \frac{1}{N}\sum_{i=1}^N \left[F_i(x^k, y_i^k) + c_k G_{\gamma,i}(x^k, y_i^k, z_i^k)\right].$$

If $\Psi_{c_k}(w^k) \le M$ for all k and some constant $M > 0$, then the constraint violation satisfies

$$G_{\gamma,i}(w_i^K) = O(K^{-\rho}), \qquad \forall i \in [N], \tag{93}$$

where $w_i^K := (x^K, y_i^K, z_i^K)$.

Proof. Under full participation the iterates of Algorithm 1 are deterministic — no client sampling occurs and all oracles are exact — so the expectations below are vacuous; they are retained for uniformity with the partial-participation analysis of Section D.1.

Let $B_k := \Theta^k + \Lambda^k$, with Θ^k and Λ^k as defined in Section 4. The proof proceeds in four steps.

Step 1: Averaged x-descent. Under full participation, the server update uses only fresh client information. Define the exact and approximate averaged x-gradients by

$$h_x^k := \frac{1}{N}\sum_{i=1}^N \nabla_x \phi_{c_k,i}(x^k, u_i^{k+1}), \qquad \widetilde{h}_x^k := \frac{1}{N}\sum_{i=1}^N \widetilde{\nabla}_x \phi_{c_k,i}(x^k, u_i^{k+1}),$$

so that the server step is

$$x^{k+1} = \mathrm{Proj}_{\mathcal{X}}\big(x^k - \alpha \widetilde{h}_x^k\big). \tag{94}$$

Let $e_x^k := \widetilde{h}_x^k - h_x^k$. This error is evaluated at (x^k, u_i^{k+1}) while the inner loop tracks its target at w_i^k, so by the bound on the shared block established in Step 2 of Lemma D.13, and using $\kappa^E_{\max} \le 1/4$,

$$\mathbb{E}\big[\|e_x^k\|^2\big] \le L_e^2 \mathbb{E}[B_k] + 2L_e^2(L_\theta^2 + L_\lambda^2)\,\mathbb{E}\big[\|\Delta u^k\|^2\big]. \tag{95}$$

Applying Theorem D.14 with $p = 1$, in which case $H_k = h_x^k$, gives

$$\begin{aligned}\mathbb{E}[V_{k+1}^G] \le \mathbb{E}[V_k^G] &- \frac{1}{16\alpha}\mathbb{E}\big[\|\Delta u^k\|^2\big] - \frac{C_\theta \eta}{4\gamma_1}\mathbb{E}[\Theta^k] - \frac{C_\lambda \eta}{4\gamma_2}\mathbb{E}[\Lambda^k] \\ &+ \mathbb{E}\big[\langle h_x^k, x^{k+1} - x^k\rangle\big] + L^{(1)}_{\mathrm{drift}}(1)\,\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] + \mathbb{E}[\Delta_c^G].\end{aligned} \tag{96}$$

Since $\Delta_c^G \le 0$, this term can be dropped.

Since $x^k \in \mathcal{X}$, the variational characterisation of the projection in equation 94 with test point x^k gives

$$\big\langle \widetilde{h}_x^k,\ x^{k+1} - x^k\big\rangle \ \le\ -\frac{1}{\alpha}\|x^{k+1} - x^k\|^2. \tag{97}$$

Writing $h_x^k = \widetilde{h}_x^k - e_x^k$ and applying Young's inequality with parameter α,

$$\left\langle h_x^k,\, x^{k+1} - x^k \right\rangle \;\le\; -\frac{1}{2\alpha}\|x^{k+1} - x^k\|^2 + \frac{\alpha}{2}\|e_x^k\|^2.$$

Adding $L_{\mathrm{drift}}^{(1)}(1)\|x^{k+1} - x^k\|^2$ and using $\alpha \le 1/(4L_{\mathrm{drift}}^{(1)}(1))$, so that $\frac{1}{2\alpha} - L_{\mathrm{drift}}^{(1)}(1) \ge \frac{1}{4\alpha}$, together with equation 95,

$$\begin{aligned}
&\mathbb{E}\big[\langle h_x^k, x^{k+1} - x^k\rangle\big] + L_{\mathrm{drift}}^{(1)}(1)\,\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] \\
&\quad\le -\frac{1}{4\alpha}\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] + \frac{\alpha}{2}L_e^2\,\mathbb{E}[B_k] + \alpha L_e^2(L_\theta^2 + L_\lambda^2)\,\mathbb{E}\big[\|\Delta u^k\|^2\big]. \qquad (98)
\end{aligned}$$

No unconstrained identity for $x^{k+1} - x^k$ is used.

Substituting equation 98 into equation 96, and using

$$\alpha \le \min\left\{\frac{C_\theta \eta}{16\gamma_1 L_e^2}, \frac{C_\lambda \eta}{16\gamma_2 L_e^2}\right\},$$

to absorb $\frac{\alpha}{2}L_e^2 B_k$ into the tracking-error descent, and $\alpha^2 \le 1/\big(32L_e^2(L_\theta^2 + L_\lambda^2)\big)$ to absorb $\alpha L_e^2(L_\theta^2 + L_\lambda^2)\|\Delta u^k\|^2$ into $-\frac{1}{16\alpha}\|\Delta u^k\|^2$, gives

$$\begin{aligned}
\mathbb{E}[V_{k+1}^G] \le \mathbb{E}[V_k^G] &- \frac{1}{32\alpha}\mathbb{E}\big[\|\Delta u^k\|^2\big] - \frac{1}{4\alpha}\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] \\
&- \frac{C_\theta \eta}{8\gamma_1}\mathbb{E}[\Theta^k] - \frac{C_\lambda \eta}{8\gamma_2}\mathbb{E}[\Lambda^k]. \qquad (99)
\end{aligned}$$

Step 2: Telescoping and basic summability. Summing equation 99 over $k = 0, \dots, K-1$ and using the non-negativity of V_k^G yields

$$\sum_{k=0}^{K-1} \mathbb{E}\big[\|\Delta u^k\|^2\big] = O(\alpha), \qquad (100)$$

$$\sum_{k=0}^{K-1} \mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] = O(\alpha), \qquad (101)$$

$$\sum_{k=0}^{K-1} \mathbb{E}[\Theta^k + \Lambda^k] = O(1), \qquad (102)$$

with all constants independent of K.

Step 3: Cumulative bound on the gradient mapping. Applying Lemma D.13 in the full-participation regime gives

$$\begin{aligned}
\mathbb{E}\big[\|\mathcal{G}_\alpha(w^k)\|^2\big] \le \frac{2}{\alpha^2}\mathbb{E}\big[\|w^{k+1} - w^k\|^2\big] &+ 2(N+4)L_e^2\kappa_{\max}^E\,\mathbb{E}[B_k] \\
&+ \big(8L_e^2(L_\theta^2 + L_\lambda^2) + 4L_\phi^2\big)\,\mathbb{E}\big[\|\Delta u^k\|^2\big]. \qquad (103)
\end{aligned}$$

Since the concatenated iterate collects the shared and local blocks,

$$\|w^{k+1} - w^k\|^2 = \|x^{k+1} - x^k\|^2 + \sum_{i=1}^{N}\|u_i^{k+1} - u_i^k\|^2 = \|x^{k+1} - x^k\|^2 + N\,\|\Delta u^k\|^2,$$

and the factor N is absorbed into the constants; N is fixed and independent of K, so this affects the constants only. Summing equation 103 over k and applying equation 100, equation 101, and equation 102 — the last two terms contributing $O(\alpha)$ and $O(1)$ respectively yields

$$\sum_{k=0}^{K-1} \mathbb{E}\big[\|\mathcal{G}_\alpha(w^k)\|^2\big] = O(1/\alpha). \qquad (104)$$

Step 4: Weighted pigeonhole argument. By Lemma D.13 and the schedule $c_k = c_0(k+1)^\rho$,

$$c_k^{-2} R_k^2 \le 2(1+\alpha L_\phi)^2 \, \|\mathcal{G}_\alpha(w^k)\|^2 + 2(2+\alpha L_\phi)^2 \, \|e_{\mathrm{aux}}^k\|^2.$$

Let $\omega_k := c_k^{-2} = c_0^{-2}(k+1)^{-2\rho}$. Since $\sum_k \mathbb{E}\|e_{\mathrm{aux}}^k\|^2 = O(1)$ by equation 86, equation 100 and equation 102, and the weights ω_k are bounded, equation 104 gives the uniformly bounded weighted partial sums

$$\sup_{K \ge 1} \sum_{k=0}^{K-1} \omega_k \, \mathbb{E}[R_k^2] \;=\; O(1/\alpha), \qquad \omega_k := c_k^{-2}, \tag{105}$$

the bound being independent of the horizon K. Since $\rho < 1/2$,

$$\sum_{k=0}^{K-1} \omega_k = c_0^{-2} \sum_{k=1}^{K} k^{-2\rho} = \Theta(K^{1-2\rho}).$$

Hence

$$\min_{0 \le k \le K-1} \mathbb{E}[R_k^2] \le \frac{\sum_{k=0}^{K-1} \omega_k \mathbb{E}[R_k^2]}{\sum_{k=0}^{K-1} \omega_k} = O\Big(\frac{1}{\alpha K^{1-2\rho}}\Big),$$

and by Jensen's inequality

$$\min_{0 \le k \le K-1} \mathbb{E}[R_k] \le \min_{0 \le k \le K-1} \sqrt{\mathbb{E}[R_k^2]} = O\Big(K^{-(1-2\rho)/2}\Big),$$

where the factor $\alpha^{-1/2}$ is absorbed into the constant because α is fixed independently of K. This proves equation 92.

Constraint violation. By definition,

$$\Psi_{c_k}(w^k) = \frac{1}{N} \sum_{i=1}^{N} F_i(x^k, y_i^k) + \frac{c_k}{N} \sum_{i=1}^{N} G_{\gamma,i}(x^k, y_i^k, z_i^k).$$

Since each F_i is bounded below on $\mathcal{X} \times \mathcal{Y}$, define

$$F^* := \frac{1}{N} \sum_{i=1}^{N} \underline{F}_i, \qquad \underline{F}_i := \inf_{(x,y) \in \mathcal{X} \times \mathcal{Y}} F_i(x, y).$$

The uniform upper bound $\Psi_{c_k}(w^k) \le M$ implies

$$\frac{c_k}{N} \sum_{i=1}^{N} G_{\gamma,i}(x^k, y_i^k, z_i^k) \le M - F^*, \qquad \forall k,$$

and therefore

$$\frac{1}{N} \sum_{i=1}^{N} G_{\gamma,i}(x^K, y_i^K, z_i^K) \le \frac{M - F^*}{c_K} = \frac{M - F^*}{c_0(K+1)^\rho} = O(K^{-\rho}).$$

Since each $G_{\gamma,i}$ is nonnegative by Lemma C.1, the same rate holds componentwise, which proves equation 93. □

When $N = 1$ the server-side gradient table reduces to the identity map, and under full participation the single entry is refreshed every round, so no staleness from skipped rounds arises. In this case FedBiC with $\alpha_k = \alpha/c_k$ reduces to the centralized BiC-GAFFA method of Yao et al. (2025). Therefore, under the growing-penalty schedule in Theorem D.16, FedBiC recovers the same asymptotic guarantees: an $O(K^{-(1-2\rho)/2})$ stationarity rate and an $O(K^{-\rho})$ feasibility rate.

D.1 Convergence under Partial Participation

We now establish that Algorithm 1 with participation rate $p < 1$, fixed penalty $c_k \equiv c > 0$, and a server-side gradient table achieves $O(1/\sqrt{pK})$ convergence of the stationarity measure

$$\|\mathcal{G}_\alpha(w^k)\|, \qquad w^k := (x^k, u_1^k, \ldots, u_N^k),$$

with $\mathcal{G}_\alpha$ the block-scaled gradient mapping equation 12 with the standard linear-speedup dependence on p. The proof combines the descent inequality of Lemma D.19 with the main statement of Theorem D.21. The analysis uses one auxiliary quantity,

$$S^k := \frac{1}{N}\sum_{i=1}^{N}\big\|g_{x,i}^k - \nabla_x\phi_{c,i}(w_i^k)\big\|^2 \qquad \text{(gradient-table staleness).} \tag{106}$$

Throughout this subsection, fix $\tau \in (0,1]$ and constants

$$p_0 \in (p_{\min}(\tau), 1), \qquad \beta_0 := \frac{p_0}{2(1-p_0)},$$

with $p_{\min}(\tau)$ as in Theorem 4.5; the special case $\tau = 1$ recovers $\max\{4\gamma_1/(4\gamma_1+\eta),\ 4\gamma_2/(4\gamma_2+\eta)\}$. and assume that the realized participation rate satisfies $p \in [p_0, 1]$.

Lemma D.17 (Gradient-table staleness recursion)**.** *Under Algorithm 1 with fixed penalty $c_k \equiv c > 0$ and participation rate $p \in [p_0, 1]$, there exists a constant $C_S^{(x)} > 0$, depending only on the problem constants and on p_0, such that*

$$\mathbb{E}\big[S^{k+1} \mid \mathcal{F}_k\big] \le \Big(1 - \tfrac{p_0}{2}\Big)S^k + C_S^{(x)}\,\mathbb{E}\big[\|x^{k+1}-x^k\|^2 + \|\Delta u^k\|^2 \,\big|\, \mathcal{F}_k\big] + 4L_e^2\,\kappa_{\max}^E\,B_k, \tag{107}$$

where S^k is as in equation 106, $\kappa_{\max} := \max\{\kappa_\theta, \kappa_\lambda\}$, and

$$\mathcal{G}_\alpha^{u_i}(w_i^k) := \frac{1}{\alpha}\Big(u_i^k - \mathrm{Proj}_{\mathcal{U}_i}\big(u_i^k - \alpha\nabla_{(y,z)}\phi_{c,i}(w_i^k)\big)\Big), \qquad G_k := \frac{1}{N}\sum_{i=1}^{N}\big\|\mathcal{G}_\alpha^{u_i}(w_i^k)\big\|^2,$$

$$B_k := \frac{1}{N}\sum_{i=1}^{N}\Big(\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2\Big).$$

Proof. Fix a client i and write $T_i^k := \nabla_x\phi_{c,i}(x^k, u_i^{k+1})$ for its exact contribution at the post-update state, and $e_i^k := g_{x,i}^{k+1} - T_i^k$ for the entry error after round k.

Active clients. If $i \in \mathcal{A}_k$, the entry is refreshed, so e_i^k is the finite-inner-loop error at (x^k, u_i^{k+1}), whereas the reference point of S^{k+1} is (x^{k+1}, u_i^{k+1}). Splitting the two with $\|a+b\|^2 \le 2\|a\|^2+2\|b\|^2$ and using the L_ϕ-smoothness of $\phi_{c,i}$ (Lemma D.4),

$$\big\|g_{x,i}^{k+1} - \nabla_x\phi_{c,i}(x^{k+1}, u_i^{k+1})\big\|^2 \le 2\|e_i^k\|^2 + 2L_\phi^2\|x^{k+1}-x^k\|^2,$$

the second term being absorbed into $C_S^{(x)}$ below. As in the proof of Lemma D.19, Definition D.3 together with Lemmas D.5, D.6 and D.1 gives

$$\|e_i^k\|^2 \le 2L_e^2\kappa_{\max}^E\big(\|\theta_i^k - \theta_i^*(w_i^k)\|^2 + \|\lambda_i^k - \lambda_i^*(w_i^k)\|^2\big) + 2L_e^2(L_\theta^2 + L_\lambda^2)\,\|u_i^{k+1} - u_i^k\|^2. \tag{108}$$

Inactive clients. If $i \notin \mathcal{A}_k$, the entry is retained, $g_{x,i}^{k+1} = g_{x,i}^k$, while the reference point moves from w_i^k to $(x^{k+1}, u_i^{k+1}) = (x^{k+1}, u_i^k)$. Adding and subtracting $\nabla_x\phi_{c,i}(w_i^k)$ and applying Young's inequality with parameter $\beta_0 := p_0/(2(1-p_0))$ together with the L_ϕ-smoothness of $\phi_{c,i}$ (Lemma D.4),

$$\big\|g_{x,i}^{k+1} - \nabla_x\phi_{c,i}(x^{k+1}, u_i^{k+1})\big\|^2 \le (1+\beta_0)\,\big\|g_{x,i}^k - \nabla_x\phi_{c,i}(w_i^k)\big\|^2 + \Big(1 + \tfrac{1}{\beta_0}\Big)L_\phi^2\,\|x^{k+1}-x^k\|^2. \tag{109}$$

Combining. Averaging over clients and taking the conditional expectation, each client is inactive with probability $1 - p \le 1 - p_0$, and the retained-entry term of equation 109 is the only one carrying S^k. Since S^k is $\mathcal{F}_k$-measurable and the participation indicator is conditionally independent of it,

$$(1-p)(1+\beta_0) \ \le\ (1-p_0)\Big(1 + \tfrac{p_0}{2(1-p_0)}\Big) \ =\ 1 - p_0 + \tfrac{p_0}{2} \ =\ 1 - \tfrac{p_0}{2}.$$

The quantity $\|x^{k+1}-x^k\|^2$ is $\mathcal{A}_k$-measurable, so its coefficient is bounded by the number of clients rather than by $1-p$; this contributes $C_S^{(x)} \ge (1 + 1/\beta_0)L_\phi^2 + 2L_\phi^2$. Adding twice the active-client bound equation 108, whose $\|u_i^{k+1} - u_i^k\|^2$ terms average to $\|\Delta u^k\|^2$ and whose tracking terms average to $4L_e^2\kappa_{\max}^E B_k$, and enlarging $C_S^{(x)}$ to also cover $4L_e^2(L_\theta^2 + L_\lambda^2)$, gives equation 107. The unconditional forms used in Theorem D.21 follow by the tower property. □

Remark D.18. Because the local states (y_i, z_i) are never averaged, the staleness of an inactive client's table entry is driven only by the motion of the shared variable, which is $O(\alpha^2)$ per round. This is what makes the entry usable for $\Theta(1/p)$ rounds and is the mechanism behind the participation threshold $p_{\min}(\tau)$.

Lemma D.19 (Lyapunov descent)**.** *Under Assumptions 4.1–4.3, suppose that $c_k \equiv c > 0$, $p \in [p_0, 1]$ for some $p_0 \in (p_{\min}(\tau), 1)$, $\alpha_x = \alpha/c$, $\eta \in (0, \bar{\eta}]$, and $E \geq E_0$. Then there exist constants $c_g, c_B > 0$, depending only on the problem constants and on p_0, such that every $\alpha \in (0, \bar{\alpha}_{\mathrm{pp}}]$ satisfies*

$$\mathbb{E}[V_{k+1}^G] \leq V_k^G - \frac{1}{4\alpha}\,\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] - c_g\, p\alpha\, \mathbb{E}[G_k] - c_B\, \mathbb{E}[B_k] + \alpha\, \mathbb{E}[S^k], \tag{110}$$

where G_k and B_k are as in Lemma D.17, and

$$\bar{\alpha}_{\mathrm{pp}} := \min\Big\{\bar{\alpha},\ \tfrac{1}{4L^{(1)}_{\mathrm{drift}}(\tau)},\ \tfrac{a_B}{4A_B},\ \sqrt{\tfrac{a_g}{4A_u}},\ \big(\tfrac{a_B}{4A_u L_e^2}\big)^{1/3}\Big\}.$$

Since $L^{(1)}_{\mathrm{drift}}(\tau) = \Theta(1/\tau)$, the admissible range of α scales as $O(\tau)$.

Proof. Local blocks. By Theorem D.14 with $\Delta_c^G = 0$ under the fixed penalty, and using $p \geq p_0 > p_{\min}(\tau)$ so that the effective participation rates in equation 91 are positive, the descent inequality equation 89 holds. It remains to convert its displacement term into the gradient mapping. For $i \in \mathcal{A}_k$ the update $u_i^{k+1} = \mathrm{Proj}_{\mathcal{U}_i}(u_i^k - \alpha d_i^k)$ is a projected step along $d_i^k = \nabla_{(y,z)}\phi_{c,i}(w_i^k) + e_i^k$, so $\frac{1}{\alpha}\|u_i^{k+1} - u_i^k\|$ is the gradient mapping of $\phi_{c,i}$ at u_i^k taken along d_i^k; since the projection is nonexpansive it differs from $\|\mathcal{G}_\alpha^{u_i}(w_i^k)\|$ by at most $\|e_i^k\|$, whence

$$\tfrac{1}{\alpha^2}\|u_i^{k+1} - u_i^k\|^2 \ \geq\ \tfrac{1}{2}\big\|\mathcal{G}_\alpha^{u_i}(w_i^k)\big\|^2 - \|e_i^k\|^2.$$

Averaging over clients, taking the conditional expectation over $\mathcal{A}_k$, and using equation 67 gives

$$-\frac{1}{16\alpha}\,\mathbb{E}\big[\|\Delta u^k\|^2\big] \ \leq\ -\frac{p\alpha}{32}\,\mathbb{E}[G_k] + \frac{p\alpha}{16} L_e^2 \kappa^E_{\max}\,\mathbb{E}[B_k],$$

in which the positive term is absorbed into the tracking-error descent of equation 89 for α small. Hence there exist constants $a_g, a_B > 0$ depending only on the problem constants and on p_0 with

$$\mathbb{E}\big[V_{k+1}^G\big] \leq V_k^G - a_g p\alpha\, \mathbb{E}[G_k] - a_B\, \mathbb{E}[B_k] + \mathbb{E}\big[\langle H_k, x^{k+1} - x^k\rangle\big] + L^{(1)}_{\mathrm{drift}}(\tau)\,\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big]. \tag{111}$$

Server direction. The server aggregates $d_x^k = \frac{c}{N}\sum_{i=1}^N g_{x,i}^{k+1}$ and applies $\alpha_x = \alpha/c$, so that

$$x^{k+1} = \mathrm{Proj}_{\mathcal{X}}\big(x^k - \alpha\, \bar{g}^k\big), \qquad \bar{g}^k := \frac{1}{N}\sum_{i=1}^N g_{x,i}^{k+1}.$$

Writing $T_i^k := \nabla_x \phi_{c,i}(x^k, u_i^{k+1})$, so that $H_k = \frac{1}{N}\sum_i T_i^k$, the identity $g_{x,i}^{k+1} = T_i^k + (g_{x,i}^{k+1} - T_i^k)$ summed over the two populations, together with $g_{x,i}^{k+1} = g_{x,i}^k$ for $i \notin \mathcal{A}_k$, gives

$$\bar{g}^k = H_k + E^k_{\mathrm{aux}} + E^k_{\mathrm{stale}}, \quad E^k_{\mathrm{aux}} := \frac{1}{N}\sum_{i\in\mathcal{A}_k}\big(g_{x,i}^{k+1} - T_i^k\big), \quad E^k_{\mathrm{stale}} := \frac{1}{N}\sum_{i\notin\mathcal{A}_k}\big(g_{x,i}^k - T_i^k\big). \tag{112}$$

For $i \notin \mathcal{A}_k$ we have $u_i^{k+1} = u_i^k$, hence $T_i^k = \nabla_x \phi_{c,i}(w_i^k)$, and the i-th summand of E^k_{stale} is the i-th summand of equation 106; Cauchy–Schwarz therefore gives

$$\|E^k_{\mathrm{stale}}\|^2 \ \leq\ S^k. \tag{113}$$

For $i \in \mathcal{A}_k$ the table entry is the inexact gradient at (x^k, u_i^{k+1}) with plug-ins $(\theta_i^{k+1}, \lambda_i^{k+1})$, so Definition D.3 applied at that point gives

$$\|g_{x,i}^{k+1} - T_i^k\|^2 \leq L_e^2\Big(\|\theta_i^{k+1} - \theta_i^*(x^k, u_i^{k+1})\|^2 + \|\lambda_i^{k+1} - \lambda_i^*(x^k, u_i^{k+1})\|^2\Big).$$

The inner loop tracks the target $\theta_i^*(w_i^k)$, not $\theta_i^*(x^k, u_i^{k+1})$. Splitting the residual and applying Lemma D.5 together with Lemma D.1,

$$\|\theta_i^{k+1} - \theta_i^*(x^k, u_i^{k+1})\|^2 \leq 2\kappa_\theta^E \|\theta_i^k - \theta_i^*(w_i^k)\|^2 + 2L_\theta^2 \|u_i^{k+1} - u_i^k\|^2,$$

and analogously for λ_i via Lemma D.6. Summing the two, applying Cauchy–Schwarz to E^k_{aux}, averaging over clients, and taking the conditional expectation over the uniformly sampled $\mathcal{A}_k$ yields

$$\mathbb{E}\big[\|E^k_{\mathrm{aux}}\|^2 \,\big|\, \mathcal{F}_k\big] \ \leq\ A_B\, p\, B_k \ +\ A_u\, \mathbb{E}\big[\|\Delta u^k\|^2 \,\big|\, \mathcal{F}_k\big], \tag{114}$$

with $A_B := 2L_e^2\kappa_{\max}^E$ and $A_u := 2L_e^2(L_\theta^2 + L_\lambda^2)$. The factor p multiplies B_k because only sampled clients contribute, whereas $\|\Delta u^k\|^2$ already sums over active clients only.

Shared block. Since $x^k \in \mathcal{X}$, the variational characterisation of the projection with test point x^k gives $\langle x^k - \alpha\bar{g}^k - x^{k+1},\, x^k - x^{k+1}\rangle \le 0$, that is,

$$\langle \bar{g}^k,\, x^{k+1} - x^k\rangle \;\le\; -\frac{1}{\alpha}\|x^{k+1} - x^k\|^2. \tag{115}$$

No unconstrained identity for $x^{k+1} - x^k$ is used. Writing $H_k = \bar{g}^k - E_{\text{aux}}^k - E_{\text{stale}}^k$ from equation 112 and applying Young's inequality with parameter α together with $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$,

$$\langle H_k, x^{k+1} - x^k\rangle \le -\frac{1}{2\alpha}\|x^{k+1} - x^k\|^2 + \alpha\|E_{\text{aux}}^k\|^2 + \alpha\|E_{\text{stale}}^k\|^2.$$

Adding $L_{\text{drift}}^{(1)}(\tau)\|x^{k+1} - x^k\|^2$ and using $\alpha \le 1/(4L_{\text{drift}}^{(1)}(\tau))$, so that $\frac{1}{2\alpha} - L_{\text{drift}}^{(1)}(\tau) \ge \frac{1}{4\alpha}$, and applying equation 113,

$$\langle H_k, x^{k+1} - x^k\rangle + L_{\text{drift}}^{(1)}(\tau)\|x^{k+1} - x^k\|^2 \le -\frac{1}{4\alpha}\|x^{k+1} - x^k\|^2 + \alpha\|E_{\text{aux}}^k\|^2 + \alpha S^k. \tag{116}$$

Absorption. Substituting equation 116 into equation 111 and applying equation 114 leaves the two positive terms $\alpha A_B p\,\mathbb{E}[B_k]$ and $\alpha A_u\,\mathbb{E}[\|\Delta u^k\|^2]$. For the latter, the same gradient-mapping comparison gives $\|u_i^{k+1} - u_i^k\| \le \alpha\big(\|\mathcal{G}_\alpha^{u_i}(w_i^k)\| + \|e_i^k\|\big)$ with e_i^k as in equation 67, whence $\mathbb{E}[\|\Delta u^k\|^2 \mid \mathcal{F}_k] \le 2p\alpha^2(G_k + L_e^2 B_k)$, so

$$\alpha A_u\,\mathbb{E}\big[\|\Delta u^k\|^2\big] \le 2A_u\alpha^3 p\,\mathbb{E}[G_k] + 2A_u L_e^2\alpha^3 p\,\mathbb{E}[B_k].$$

The conditions $\alpha^2 \le a_g/(4A_u)$ and $\alpha^3 \le a_B/(4A_u L_e^2)$ absorb these into $-a_g p\alpha\, G_k$ and $-a_B B_k$ respectively, and $\alpha \le a_B/(4A_B)$ absorbs $\alpha A_B p\, B_k$ (using $p \le 1$). The same conditions cover the $\|\Delta u^k\|^2$ contribution of Lemma D.17 after enlarging A_u to $A_u + \lambda_S C_S^{(x)}$. Renaming the surviving constants c_g, c_B gives equation 110. □

Remark D.20. The shared-variable descent above uses the projection inequality equation 115 rather than an unconstrained update identity.

Theorem D.21 (Fixed-penalty convergence under partial participation)**.** *Under Assumptions 4.1–4.3, suppose that $c_k \equiv c > 0$ and $p_0 \in (p_{\min}(\tau), 1)$. Assume that $p \in [p_0, 1]$, $\alpha_x = \alpha/c$, $\eta \in (0, \bar{\eta}]$, and $E \ge E_0$. Then there exists $\bar{\alpha}_{\text{pp}} > 0$, depending only on the problem constants and on p_0, such that for every $\alpha \in (0, \bar{\alpha}_{\text{pp}}]$, Algorithm 1 satisfies*

$$\min_{0\le k\le K-1} \mathbb{E}\big[\|\mathcal{G}_\alpha(w^k)\|\big] = O\!\left(\frac{1}{\sqrt{pK}}\right),$$

equivalently $K = O(1/(p\varepsilon^2))$ rounds to reach an ε-stationary point.

Proof. Throughout, constants may depend on the problem constants and on p_0, but not on K or on the realized participation rate $p \in [p_0, 1]$. By Lemma D.19,

$$\mathbb{E}[V_{k+1}^G] \le V_k^G - \frac{1}{4\alpha}\,\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] - c_g p\alpha\,\mathbb{E}[G_k] - c_B\,\mathbb{E}[B_k] + \alpha\,\mathbb{E}[S^k]. \tag{117}$$

Define the augmented Lyapunov function

$$\mathcal{L}^k := V_k^G + \lambda_S S^k, \qquad \lambda_S := \frac{4\alpha}{p_0}.$$

Staleness. By Lemma D.17,

$$\lambda_S\big(\mathbb{E}[S^{k+1}] - \mathbb{E}[S^k]\big) \le -2\alpha\,\mathbb{E}[S^k] + \lambda_S C_S^{(x)}\,\mathbb{E}\big[\|x^{k+1} - x^k\|^2 + \|\Delta u^k\|^2\big] + 4\lambda_S L_e^2\kappa_{\max}^E\,\mathbb{E}[B_k].$$

The first term cancels the $+\alpha\,\mathbb{E}[S^k]$ of equation 117 and leaves a residual $-\alpha\,\mathbb{E}[S^k]$. The $\|x^{k+1} - x^k\|^2$ term is absorbed by $-\frac{1}{4\alpha}\mathbb{E}\|x^{k+1} - x^k\|^2$ provided

$$\alpha^2 \;\le\; \frac{p_0}{32\,C_S^{(x)}}, \tag{118}$$

leaving at least $-\frac{1}{8\alpha}\mathbb{E}\|x^{k+1} - x^k\|^2$. For the remaining two terms, $\lambda_S = 4\alpha/p_0$ and the local-update bound $\mathbb{E}[\|\Delta u^k\|^2 \mid \mathcal{F}_k] \le 2p\alpha^2(G_k + L_e^2 B_k)$ give contributions of order $\alpha^3\mathbb{E}[G_k]$ and $\alpha\,\mathbb{E}[B_k]$, which are absorbed into $-c_g p\alpha\, G_k$ and $-c_B B_k$ after decreasing those constants.

Collecting, and choosing $\alpha \le \bar{\alpha}_{\mathrm{pp}}$ small enough that equation 118 also holds, we obtain positive constants $\hat{c}_x, \hat{c}_g, \hat{c}_B, \hat{c}_S$ with

$$\mathbb{E}[\mathcal{L}^{k+1}] \le \mathcal{L}^k - \frac{\hat{c}_x}{\alpha}\,\mathbb{E}\big[\|x^{k+1}-x^k\|^2\big] - \hat{c}_g p\alpha\,\mathbb{E}[G_k] - \hat{c}_B\,\mathbb{E}[B_k] - \hat{c}_S\alpha\,\mathbb{E}[S^k]. \tag{119}$$

Summing over $k = 0, \dots, K-1$ and using the lower boundedness of $\mathcal{L}^k$ yields

$$\sum_{k=0}^{K-1}\mathbb{E}\big[\|x^{k+1}-x^k\|^2\big] = O(\alpha), \qquad \sum_{k=0}^{K-1}\mathbb{E}[G_k] = O\big(1/(p\alpha)\big),$$
$$\sum_{k=0}^{K-1}\mathbb{E}[B_k] = O(1), \qquad \sum_{k=0}^{K-1}\mathbb{E}[S^k] = O(1/\alpha).$$

It remains to convert these bounds into a rate for the stationarity measure. We bound $\mathcal{G}_\alpha(w^k)$ blockwise; Lemma D.13 is not used, since it applies only under full participation.

Local blocks. By construction the u_i-block of $\mathcal{G}_\alpha(w^k)$ in equation 12 is exactly $\mathcal{G}_\alpha^{u_i}(w_i^k)$, so $\sum_{i=1}^N \|\mathcal{G}_\alpha^{u_i}(w_i^k)\|^2 = N\,G_k$, with no inequality incurred.

Shared block. Write $\mathcal{G}_\alpha^x(w^k)$ for the x-block of $\mathcal{G}_\alpha(w^k)$. The realised step is $x^{k+1} = \mathrm{Proj}_{\mathcal{X}}(x^k - \alpha\bar{g}^k)$, so by nonexpansiveness

$$\Big\|\mathcal{G}_\alpha^x(w^k) - \tfrac{1}{\alpha}(x^k - x^{k+1})\Big\| \;\le\; \big\|\nabla_x\phi_c(w^k) - \bar{g}^k\big\|.$$

By equation 112, $\bar{g}^k = H_k + E^k_{\mathrm{aux}} + E^k_{\mathrm{stale}}$, and $\|\nabla_x\phi_c(w^k) - H_k\| \le L_\phi\|\Delta u^k\|$ by Jensen's inequality and the L_ϕ-smoothness of $\phi_{c,i}$ (Lemma D.4), since the two differ only in evaluating the local blocks at u_i^k rather than u_i^{k+1}. Combining with $\|E^k_{\mathrm{stale}}\|^2 \le S^k$ from equation 113 and equation 114,

$$\mathbb{E}\big\|\mathcal{G}_\alpha(w^k)\big\|^2 \le N\,\mathbb{E}[G_k] + \frac{4}{\alpha^2}\mathbb{E}\|x^{k+1}-x^k\|^2 + 4(L_\phi^2 + A_u)\,\mathbb{E}\|\Delta u^k\|^2 + 4A_B p\,\mathbb{E}[B_k] + 4\,\mathbb{E}[S^k].$$

Every term is summable by the four displays above, together with $\sum_k \mathbb{E}\|\Delta u^k\|^2 \le 2p\alpha^2\sum_k(\mathbb{E}[G_k] + L_e^2\mathbb{E}[B_k]) = O(\alpha)$, and N is fixed; using $p \le 1$ to absorb the $O(1/\alpha)$ contributions into $O(1/(p\alpha))$,

$$\sum_{k=0}^{K-1}\mathbb{E}\big\|\mathcal{G}_\alpha(w^k)\big\|^2 = O\big(1/(p\alpha)\big).$$

Averaging over k and applying Jensen's inequality,

$$\min_{0\le k\le K-1}\mathbb{E}\big[\|\mathcal{G}_\alpha(w^k)\|\big] \le \Big(\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big\|\mathcal{G}_\alpha(w^k)\big\|^2\Big)^{1/2} = O\left(\frac{1}{\sqrt{p\alpha K}}\right),$$

and $\alpha^{-1/2}$ is absorbed into the constant since α is fixed independently of K and p. □

Remark D.22. Both results are stated at the concatenated iterate $w^k = (x^k, u_1^k, \dots, u_N^k)$. Under full participation every block is updated by a projected step, so the residual $R_k = \mathrm{dist}(0, \nabla\psi_{c_k}(w^{k+1}) + \mathcal{N}_{\mathcal{W}}(w^{k+1}))$ is available and is bounded through $\mathcal{G}_\alpha(w^k)$ (Lemma D.13). Under partial participation the local blocks of an inactive client are not updated, so no projection identity holds at w^{k+1} for those blocks and we report $\|\mathcal{G}_\alpha(w^k)\|$ directly; by Lemma D.13 this is the stronger of the two whenever both apply. Under full participation all stored gradients are fresh and no staleness from skipped rounds arises. The server-side gradient table adds $O(Nd)$ memory, and each round transmits one d-dimensional vector per participating client in each direction; no lower-level state is communicated. Each client stores only its own (y_i, z_i) and the transient $(\bar{\theta}_i, \bar{\lambda}_i)$, i.e. $O(m + p_i)$ beyond the shared x, so per-client memory is independent of N.

E Correspondence with the Original Bilevel Problem

Throughout this section we work under the hypotheses of Theorem D.16 (full participation, growing penalty $c_k = c_0(k+1)^\rho$ with $\rho \in (0, \frac{1}{2})$, and $\psi_{c_k}(w^k) \le M$), and we add regularity used nowhere else in the paper.

Assumption E.1 (Second-order regularity)**.** For each $i \in [N]$, F_i, f_i and g_i are twice continuously differentiable with locally Lipschitz second derivatives on a neighbourhood of $\mathcal{W}$.

Assumption E.1 is what makes the first-order expansions below carry a remainder that is $O(\|\cdot\|^2)$ *uniformly* on a compact neighbourhood of a limit point. It strengthens Assumption 2.2 and is required only here.

Notation. Write $L_i(x, y, \mu) := f_i(x, y) + \mu^\top g_i(x, y)$. Second derivatives are ordered so that $\nabla^2_{yy} L_i \in \mathbb{R}^{m\times m}$ and $\nabla^2_{xy} L_i \in \mathbb{R}^{d\times m}$ with $(\nabla^2_{xy} L_i)_{st} = \partial^2 L_i/\partial x_s \partial y_t$, so $\nabla^2_{xy} L_i\, v \in \mathbb{R}^d$ for $v \in \mathbb{R}^m$; likewise $\nabla_y g_i \in \mathbb{R}^{p_i\times m}$, $\nabla_x g_i \in \mathbb{R}^{p_i\times d}$.

The point of evaluation and the residual vector. By equation 11 the residual R_k is that of the *unnormalized* penalized objective ψ_{c_k}, evaluated at the post-update iterate. We therefore write

$$\widehat{w}^k := w^{k+1} = \big(\widehat{x}^k, \widehat{u}^k_1, \dots, \widehat{u}^k_N\big), \qquad \widehat{u}^k_i = (\widehat{y}^k_i, \widehat{z}^k_i), \tag{120}$$

and define the *rescaled proximal displacements* at that same point:

$$A^k_i := c_k\big(\theta^*_i(\widehat{w}^k_i) - \widehat{y}^k_i\big), \qquad B^k_i := c_k\big(\widehat{z}^k_i - \lambda^*_i(\widehat{w}^k_i)\big). \tag{121}$$

The set $\nabla\psi_{c_k}(\widehat{w}^k) + \mathcal{N}_{\mathcal{W}}(\widehat{w}^k)$ is a translate of the closed convex cone $\mathcal{N}_{\mathcal{W}}(\widehat{w}^k)$, hence closed, so the distance defining R_k is attained. Choose $v^k \in \mathcal{N}_{\mathcal{W}}(\widehat{w}^k)$ attaining

$$R_k = \big\|\nabla\psi_{c_k}(\widehat{w}^k) + v^k\big\|, \qquad \text{and set} \qquad e^k := \nabla\psi_{c_k}(\widehat{w}^k) + v^k. \tag{122}$$

Since $\mathcal{W} = \mathcal{X} \times \prod_i (\mathcal{Y} \times \mathcal{Z}_i)$ is a product, $\mathcal{N}_{\mathcal{W}}$ splits blockwise: $v^k = (v^k_x, v^k_{y_1}, v^k_{z_1}, \dots)$ with $v^k_x \in \mathcal{N}_{\mathcal{X}}(\widehat{x}^k)$, $v^k_{y_i} \in \mathcal{N}_{\mathcal{Y}}(\widehat{y}^k_i)$, $v^k_{z_i} \in \mathcal{N}_{\mathcal{Z}_i}(\widehat{z}^k_i)$, and every block of e^k satisfies $\|e^k_\bullet\| \le \|e^k\| = R_k$. All statements of the form "$\cdot = O(R_k)$" below refer to these blocks.

Lemma E.2 (Gap lower bound)**.** *For every i and every $w_i \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}_i$,*

$$G_{\gamma,i}(w_i) \;\ge\; \frac{1}{2\gamma_1}\|\theta^*_i(w_i) - y_i\|^2 + \frac{1}{2\gamma_2}\|\lambda^*_i(w_i) - z_i\|^2, \tag{123}$$

and consequently $\big\|(A^k_i, B^k_i)\big\| \le c_k\sqrt{2\max\{\gamma_1,\gamma_2\}\, G_{\gamma,i}(\widehat{w}^k_i)}$.

Proof. Write $G_{\gamma,i} = \Phi_1 - \Phi_2$ with $\Phi_1 := \max_{\lambda\ge 0}\{L_i(x, y_i, \lambda) - \|\lambda - z_i\|^2/(2\gamma_2)\}$ and $\Phi_2 := \min_{\theta\in\mathcal{Y}}\{L_i(x, \theta, z_i) + \|\theta - y_i\|^2/(2\gamma_1)\}$. The objective defining Φ_2 is $1/\gamma_1$-strongly convex with minimizer θ^*_i and $\theta = y_i$ feasible, so $L_i(x, y_i, z_i) - \Phi_2 \ge \frac{1}{2\gamma_1}\|y_i - \theta^*_i\|^2$; the objective defining Φ_1 is $1/\gamma_2$-strongly concave with maximizer λ^*_i and $\lambda = z_i$ feasible, so $\Phi_1 - L_i(x, y_i, z_i) \ge \frac{1}{2\gamma_2}\|\lambda^*_i - z_i\|^2$. Adding gives equation 123. Substituting equation 121 at $\widehat{w}^k_i$ gives $\|A^k_i\|^2/(2\gamma_1 c_k^2) + \|B^k_i\|^2/(2\gamma_2 c_k^2) \le G_{\gamma,i}(\widehat{w}^k_i)$, and bounding both denominators below by $2\max\{\gamma_1,\gamma_2\}c_k^2$ yields the claim. □

E.1 A SYNCHRONIZED SUBSEQUENCE

Remark E.3 (Determinism under full participation). Under $p = 1$, Algorithm 1 selects $\mathcal{A}_k = [N]$ at every round, so no client sampling occurs; lines 5–10 and 17 evaluate f_i, g_i, F_i and their gradients exactly. The iterates are therefore a deterministic function of the initialization and every expectation in Theorem D.16 is vacuous; we write R_k and $G_{\gamma,i}$ without expectations below. With a stochastic oracle an almost-sure subsequence argument would replace Lemma E.4.

Lemma E.4 (Synchronized feasible-stationary subsequence)**.** *Under the hypotheses of Theorem D.16 and Remark E.3, there exist $k_j \to \infty$ and $\bar{w} \in \mathcal{W}$ with*

$$\widehat{w}^{k_j} \to \bar{w}, \qquad R_{k_j} \to 0, \qquad G_{\gamma,i}(\widehat{w}^{k_j}_i) \to 0 \quad \textit{for every } i \in [N].$$

Proof. By equation 105, $\sup_{K\ge 1}\sum_{k=0}^{K-1}\omega_k R_k^2 < \infty$ with $\omega_k := c_k^{-2} = c_0^{-2}(k+1)^{-2\rho}$, so $\sum_{k\ge 0}\omega_k R_k^2 < \infty$. Since $\rho < \frac{1}{2}$, $\sum_{k\ge 0}\omega_k = \infty$. If $R_k^2 \ge \varepsilon > 0$ for all $k \ge K_0$ then $\sum_{k\ge K_0}\omega_k R_k^2 \ge \varepsilon\sum_{k\ge K_0}\omega_k = \infty$, a contradiction; hence $\liminf_k R_k = 0$ and there are $k_j \to \infty$ with $R_{k_j} \to 0$.

The hypothesis $\psi_{c_k}(w^k) \le M$ gives $\frac{1}{N}\sum_i G_{\gamma,i}(w_i^k) \le (M - \underline{F})/c_k$ for every k, where $\underline{F} := \frac{1}{N}\sum_i \underline{F}_i$ is the lower bound of Assumption 4.1; hence $G_{\gamma,i}(w_i^k) \to 0$ for every i. By equation 101 and equation 100, $\sum_k \|w^{k+1} - w^k\|^2 < \infty$, so $\widehat{w}^k - w^k \to 0$; $G_{\gamma,i}$ is continuous on the compact set $\mathcal{W}$, hence uniformly continuous there, so $G_{\gamma,i}(\widehat{w}_i^k) \to 0$. Finally $\mathcal{W}$ is compact, so a subsequence of $\{\widehat{w}^{k_j}\}$ converges; relabel. □

By Lemma C.1, the limit satisfies $\bar{y}_i \in \mathcal{S}_i(\bar{x})$ and $\bar{z}_i \in \mathcal{M}_i(\bar{x}, \bar{y}_i)$ for every i. Throughout the rest of this section we set

$$\bar{\mu}_i := \bar{z}_i \ \in \ \mathcal{M}_i(\bar{x}, \bar{y}_i), \tag{124}$$

so that no multiplier other than the limit of the algorithm's dual iterates is ever invoked.

E.2 Regularity at the limit

Assumption E.5 (Lower-level regularity at the limit)**.** At the limit point $\bar{w}$ of Lemma E.4, for every $i \in [N]$, with $\bar{\mu}_i$ as in equation 124 and $\mathcal{A}_i := \{\ell : g_{i\ell}(\bar{x}, \bar{y}_i) = 0\}$:

(i) $\bar{y}_i \in \operatorname{int} \mathcal{Y}$;

(ii) $\{\nabla_y g_{i\ell}(\bar{x}, \bar{y}_i)\}_{\ell \in \mathcal{A}_i}$ are linearly independent (LICQ);

(iii) $\bar{\mu}_{i\ell} > 0$ for $\ell \in \mathcal{A}_i$ (strict complementarity), and $\bar{\mu}_{i\ell} < r$;

(iv) $d^\top \nabla^2_{yy} L_i(\bar{x}, \bar{y}_i, \bar{\mu}_i)\, d > 0$ for every nonzero d with $\nabla_y g_{i,\mathcal{A}_i} d = 0$ (SSOSC).

Conditions (ii)–(iv) are the classical sufficient conditions for Robinson's strong regularity of the lower-level KKT generalized equation (Robinson, 1980; Dontchev & Rockafellar, 2009). Two consequences are used below. First, strict complementarity fixes the active set locally: for (x, y, μ) near $(\bar{x}, \bar{y}_i, \bar{\mu}_i)$ the constraints active at $(\bar{x}, \bar{y}_i)$ remain active with positive multiplier and the others remain strictly slack, so the KKT system reduces locally to the smooth square system

$$\nabla_y L_i(x, y, \mu_{\mathcal{A}_i}) = 0, \qquad g_{i,\mathcal{A}_i}(x, y) = 0, \qquad \mu_{i\ell} = 0 \ \ (\ell \notin \mathcal{A}_i).$$

Second, the Jacobian of that reduced system in $(y, \mu_{\mathcal{A}_i})$ is

$$J_i := \begin{bmatrix} \nabla^2_{yy} L_i(\bar{x}, \bar{y}_i, \bar{\mu}_i) & \nabla_y g_{i,\mathcal{A}_i}^\top \\ \nabla_y g_{i,\mathcal{A}_i} & 0 \end{bmatrix}, \tag{125}$$

which is nonsingular under (ii) and (iv). The classical implicit function theorem, applicable by Assumption E.1, therefore yields $\varepsilon, \delta > 0$ and a *continuously differentiable* map $x \mapsto (y_i^*(x), \mu_i^*(x))$ on $\mathbb{B}(\bar{x}, \delta)$ such that $(y_i^*(x), \mu_i^*(x))$ is the unique KKT pair in $\mathbb{B}((\bar{y}_i, \bar{\mu}_i), \varepsilon)$, with $(y_i^*(\bar{x}), \mu_i^*(\bar{x})) = (\bar{y}_i, \bar{\mu}_i)$.

Lemma E.6 (The local branch is the lower-level solution map)**.** *Under Assumptions 2.1 and E.5, $\mathcal{S}_i(\bar{x}) = \{\bar{y}_i\}$ for every i, and there is a neighbourhood $\mathcal{U} \subseteq \mathbb{B}(\bar{x}, \delta)$ of $\bar{x}$ on which $\mathcal{S}_i(x) = \{y_i^*(x)\}$. Consequently $\varphi(x) := \frac{1}{N}\sum_i F_i(x, y_i^*(x))$ is well defined and continuously differentiable on $\mathcal{U}$ and coincides there with the objective of equation 1.*

Proof. Step 1: uniqueness at $\bar{x}$. The lower-level problem is convex (Assumption 2.1(ii)), so $\mathcal{S}_i(\bar{x})$ is convex and every KKT point is a global minimizer. By (iv), $\bar{y}_i$ is a strict local minimizer, hence isolated in $\mathcal{S}_i(\bar{x})$; a convex set with an isolated point is a singleton, so $\mathcal{S}_i(\bar{x}) = \{\bar{y}_i\}$.

Step 2: recovering feasibility along $x^\nu \to \bar{x}$. We claim: for every $\hat{y} \in \mathcal{Y}$ with $g_i(\bar{x}, \hat{y}) \le 0$ and every $x^\nu \to \bar{x}$ in $\mathcal{X}$, there are $\hat{y}^\nu \in \mathcal{Y}$ with $g_i(x^\nu, \hat{y}^\nu) \le 0$ and $\hat{y}^\nu \to \hat{y}$. Let $s^\nu := \hat{y}_i(x^\nu)$ be the Slater point of Assumption 2.1(iv), so $g_i(x^\nu, s^\nu) \le -\sigma_i \mathbf{1}$, put $\epsilon_\nu := \|g_i(x^\nu, \hat{y}) - g_i(\bar{x}, \hat{y})\|_\infty$, and set

$$\tau_\nu := \begin{cases} 0, & \epsilon_\nu = 0, \\ \min\{1, \ \sqrt{\epsilon_\nu}\}, & \epsilon_\nu > 0, \end{cases} \qquad \hat{y}^\nu := (1 - \tau_\nu)\hat{y} + \tau_\nu s^\nu \in \mathcal{Y},$$

which lies in $\mathcal{Y}$ by convexity. If $\epsilon_\nu = 0$ then $\hat{y}^\nu = \hat{y}$ is already feasible at x^ν. Otherwise convexity of $g_i(x^\nu, \cdot)$ gives

$$g_i(x^\nu, \hat{y}^\nu) \le (1 - \tau_\nu) g_i(x^\nu, \hat{y}) + \tau_\nu g_i(x^\nu, s^\nu) \le \big(\epsilon_\nu - \tau_\nu \sigma_i\big)\mathbf{1},$$

using $g_i(x^\nu, \hat{y}) \le g_i(\bar{x}, \hat{y}) + \epsilon_\nu \mathbf{1} \le \epsilon_\nu \mathbf{1}$. By continuity $\epsilon_\nu \to 0$, so $\tau_\nu \sigma_i = \sigma_i \sqrt{\epsilon_\nu} \ge \epsilon_\nu$ once $\sqrt{\epsilon_\nu} \le \sigma_i$, i.e. for all large ν; hence $g_i(x^\nu, \hat{y}^\nu) \le 0$. Finally $\|\hat{y}^\nu - \hat{y}\| \le \tau_\nu \operatorname{diam}(\mathcal{Y}) \to 0$.

Step 3: uniqueness nearby. Suppose not: there are $x^\nu \to \bar{x}$ and $\widetilde{y}_i^\nu \in \mathcal{S}_i(x^\nu)$ with $\widetilde{y}_i^\nu \neq y_i^*(x^\nu)$. By compactness of $\mathcal{Y}$, pass to a subsequence with $\widetilde{y}_i^\nu \to \widetilde{y}_i$; continuity gives $g_i(\bar{x}, \widetilde{y}_i) \leq 0$. For optimality, let $\hat{y}$ satisfy $g_i(\bar{x}, \hat{y}) \leq 0$ and take $\hat{y}^\nu$ from Step 2; then $f_i(x^\nu, \widetilde{y}_i^\nu) \leq f_i(x^\nu, \hat{y}^\nu)$, and in the limit $f_i(\bar{x}, \widetilde{y}_i) \leq f_i(\bar{x}, \hat{y})$. Hence $\widetilde{y}_i \in \mathcal{S}_i(\bar{x}) = \{\bar{y}_i\}$, so $\widetilde{y}_i^\nu \to \bar{y}_i$.

Let $\widetilde{\mu}_i^\nu \in \mathcal{M}_i(x^\nu, \widetilde{y}_i^\nu)$, which exists by Assumption 2.3. By Lemma C.4, $\|\widetilde{\mu}_i^\nu\|_1 \leq M_{f,i}/\sigma_i$ uniformly in ν, so a further subsequence gives $\widetilde{\mu}_i^\nu \to \widetilde{\mu}_i$. Passing to the limit in the KKT conditions, legitimate by Assumption 2.2, gives $\widetilde{\mu}_i \in \mathcal{M}_i(\bar{x}, \bar{y}_i)$, which LICQ makes a singleton, so $\widetilde{\mu}_i = \bar{\mu}_i$.

Therefore both $(\widetilde{y}_i^\nu, \widetilde{\mu}_i^\nu)$ and $(y_i^*(x^\nu), \mu_i^*(x^\nu))$ converge to $(\bar{y}_i, \bar{\mu}_i)$, so for large ν both lie in $\mathbb{B}((\bar{y}_i, \bar{\mu}_i), \varepsilon)$ and both are KKT pairs at $x^\nu \in \mathbb{B}(\bar{x}, \delta)$. Uniqueness of the KKT pair there forces $\widetilde{y}_i^\nu = y_i^*(x^\nu)$, a contradiction. Hence $\mathcal{S}_i(x) = \{y_i^*(x)\}$ on a neighbourhood $\mathcal{U}$, and φ inherits continuous differentiability from y_i^* and Assumption E.1. □

Lemma E.7 (Active-set identification)**.** *Under Assumptions E.1 and E.5, for all j sufficiently large and every i, with all quantities at $\widehat{w}_i^{k_j}$:*

(a) Both $\theta_i^*(\widehat{w}_i^{k_j})$ and $\widehat{y}_i^{k_j}$ lie in $\operatorname{int}\mathcal{Y}$. Consequently the normal-cone term in the optimality condition for θ_i^* vanishes, and $v_{y_i}^{k_j} = 0$;

(b) for $\ell \in \mathcal{A}_i$: $\widehat{z}_{i\ell}^{k_j} \in (0, r)$, $v_{z_i\ell}^{k_j} = 0$, and $\widehat{z}_{i\ell}^{k_j} + \gamma_2 g_{i\ell}(\widehat{x}^{k_j}, \widehat{y}_i^{k_j}) > 0$, so the projection defining $\lambda_{i\ell}^*$ acts as the identity;

(c) for $\ell \notin \mathcal{A}_i$: $\widehat{z}_{i\ell}^{k_j} = 0$, $\lambda_{i\ell}^* = 0$, and hence $B_{i\ell}^{k_j} = 0$.

Proof. Since $G_{\gamma,i}(\widehat{w}_i^{k_j}) \to 0$, Lemma C.1 and LICQ give $\widehat{z}_i^{k_j} \to \bar{\mu}_i$, and Lemma C.1 also gives $\widehat{y}_i^{k_j} \to \bar{y}_i$. By Lemma E.2, $\|\theta_i^* - \widehat{y}_i^{k_j}\| \to 0$, so $\theta_i^* \to \bar{y}_i$ as well. Both limits are interior by (i), so both θ_i^* and $\widehat{y}_i^{k_j}$ lie in $\operatorname{int}\mathcal{Y}$ for large j; the first makes the normal-cone term in the θ_i^* subproblem vanish, the second gives $v_{y_i}^{k_j} = 0$. This is (a).

For (b), let $\ell \in \mathcal{A}_i$. By (iii), $0 < \bar{\mu}_{i\ell} < r$, so $\widehat{z}_{i\ell}^{k_j} \in (0, r)$ for large j, whence $v_{z_i\ell}^{k_j} = 0$. Since $g_{i\ell}(\bar{x}, \bar{y}_i) = 0$, $\widehat{z}_{i\ell}^{k_j} + \gamma_2 g_{i\ell} \to \bar{\mu}_{i\ell} > 0$ and the projection onto $\mathbb{R}_+^{p_i}$ acts as the identity in that coordinate.

For (c), let $\ell \notin \mathcal{A}_i$, so $g_{i\ell}(\bar{x}, \bar{y}_i) =: -2\sigma < 0$ and $\bar{\mu}_{i\ell} = 0$. For large j, $g_{i\ell} \leq -\sigma$ and $\widehat{z}_{i\ell}^{k_j} \to 0$, hence $\widehat{z}_{i\ell}^{k_j} + \gamma_2 g_{i\ell} < 0$ and $\lambda_{i\ell}^* = 0$. Since $\widehat{z}_{i\ell}^{k_j} \in [0, r]$ and $\widehat{z}_{i\ell}^{k_j} \to 0 < r$, the upper boundary is excluded for large j, leaving $\widehat{z}_{i\ell}^{k_j} = 0$ or $\widehat{z}_{i\ell}^{k_j} \in (0, r)$. Suppose the latter; then $v_{z_i\ell}^{k_j} = 0$ and the z_i-block of equation 122 reads

$$\frac{c_{k_j}}{N}\left[\frac{\lambda_{i\ell}^* - \widehat{z}_{i\ell}^{k_j}}{\gamma_2} - g_{i\ell}\big(\widehat{x}^{k_j}, \theta_i^*\big)\right] = e_{z_i\ell}^{k_j}, \qquad \|e_{z_i\ell}^{k_j}\| \leq R_{k_j},$$

so, multiplying by $N\gamma_2/c_{k_j}$ and using $\lambda_{i\ell}^* = 0$,

$$\widehat{z}_{i\ell}^{k_j} = -\gamma_2 g_{i\ell}\big(\widehat{x}^{k_j}, \theta_i^*\big) - \frac{N\gamma_2}{c_{k_j}} e_{z_i\ell}^{k_j} \;\geq\; \frac{\gamma_2\sigma}{2} \;>\; 0$$

for large j, contradicting $\widehat{z}_{i\ell}^{k_j} \to 0$. Hence $\widehat{z}_{i\ell}^{k_j} = 0$ and $B_{i\ell}^{k_j} = 0$. □

E.3 The rescaled system

For brevity write, at the j-th index of the subsequence,

$$H_i^j := \nabla_{yy}^2 L_i\big(\widehat{x}^{k_j}, \theta_i^*(\widehat{w}_i^{k_j}), \widehat{z}_i^{k_j}\big), \qquad K_i^j := \nabla_y g_{i,\mathcal{A}_i}\big(\widehat{x}^{k_j}, \theta_i^*(\widehat{w}_i^{k_j})\big),$$
$$M_i^j := \nabla_{xy}^2 L_i\big(\widehat{x}^{k_j}, \theta_i^*(\widehat{w}_i^{k_j}), \widehat{z}_i^{k_j}\big), \qquad D_i^j := \nabla_x g_{i,\mathcal{A}_i}\big(\widehat{x}^{k_j}, \theta_i^*(\widehat{w}_i^{k_j})\big), \tag{126}$$

so that derivatives of L_i and g_i are evaluated at $\theta_i^*(\widehat{w}_i^{k_j})$, whereas derivatives of F_i are evaluated at the iterate $(\widehat{x}^{k_j}, \widehat{y}_i^{k_j})$. By Assumption E.1 and $\widehat{w}^{k_j} \to \bar{w}$, all four converge to their values at $(\bar{x}, \bar{y}_i, \bar{\mu}_i)$, denoted $\bar{H}_i, \bar{K}_i, \bar{M}_i, \bar{D}_i$.

Proposition E.8. *Under Assumptions E.1 and E.5, along the subsequence of Lemma E.4 and for every $i \in [N]$, with $E_j := \|(A_i^{k_j}, B_i^{k_j})\|^2/c_{k_j}$:*

$$H_i^j A_i^{k_j} + (K_i^j)^\top B_{i,\mathcal{A}_i}^{k_j} = \nabla_y F_i\big(\widehat{x}^{k_j}, \widehat{y}_i^{k_j}\big) + O(E_j) + O(NR_{k_j}), \tag{127}$$

$$K_i^j A_i^{k_j} = O(E_j) + O(NR_{k_j}), \tag{128}$$

$$c_{k_j} \nabla_x G_{\gamma,i}(\widehat{w}_i^{k_j}) = -M_i^j A_i^{k_j} - (D_i^j)^\top B_{i,\mathcal{A}_i}^{k_j} + O(E_j). \tag{129}$$

Proof. All quantities are at $\widehat{w}^{k_j}$; write $\delta_i := (\widehat{y}_i - \theta_i^*, \lambda_i^* - \widehat{z}_i)$, so $\|\delta_i\| = \|(A_i, B_i)\|/c_{k_j}$ by equation 121.

Equation equation 127. By Lemma E.7(a), θ_i^* is an interior minimizer of $L_i(x, \theta, \widehat{z}_i) + \|\theta - \widehat{y}_i\|^2/(2\gamma_1)$, so

$$\nabla_y L_i(x, \theta_i^*, \widehat{z}_i) + \tfrac{1}{\gamma_1}(\theta_i^* - \widehat{y}_i) = 0. \tag{130}$$

Expanding $\nabla_y L_i(x, \widehat{y}_i, \lambda_i^*)$ about $(\theta_i^*, \widehat{z}_i)$ with uniform quadratic remainder (Assumption E.1),

$$\nabla_y L_i(x, \widehat{y}_i, \lambda_i^*) = \nabla_y L_i(x, \theta_i^*, \widehat{z}_i) + H_i^j(\widehat{y}_i - \theta_i^*) + \nabla_y g_i^\top(\lambda_i^* - \widehat{z}_i) + O(\|\delta_i\|^2).$$

Substituting into $\nabla_y G_{\gamma,i} = \nabla_y L_i(x, \widehat{y}_i, \lambda_i^*) + \frac{1}{\gamma_1}(\theta_i^* - \widehat{y}_i)$ and using equation 130, the proximal terms cancel; multiplying by c_{k_j} and using $c_{k_j}(\widehat{y}_i - \theta_i^*) = -A_i$, $c_{k_j}(\lambda_i^* - \widehat{z}_i) = -B_i$,

$$c_{k_j} \nabla_y G_{\gamma,i} = -H_i^j A_i - \nabla_y g_i^\top B_i + O(E_j),$$

and $B_{i\ell} = 0$ off $\mathcal{A}_i$ by Lemma E.7(c) turns $\nabla_y g_i^\top B_i$ into $(K_i^j)^\top B_{i,\mathcal{A}_i}$. The y_i-block of equation 122, with $v_{y_i}^{k_j} = 0$ by Lemma E.7(a), reads $\frac{1}{N}[\nabla_y F_i(\widehat{x}^{k_j}, \widehat{y}_i^{k_j}) + c_{k_j}\nabla_y G_{\gamma,i}] = e_{y_i}^{k_j}$ with $\|e_{y_i}^{k_j}\| \le R_{k_j}$; multiplying by N gives equation 127.

Equation equation 128. Fix $\ell \in \mathcal{A}_i$. By Lemma E.7(b), $v_{z_i\ell}^{k_j} = 0$, so multiplying the z_i-block by $N\gamma_2/c_{k_j}$,

$$-\frac{B_{i\ell}}{\gamma_2} = c_{k_j}\, g_{i\ell}(x, \theta_i^*) + O(NR_{k_j}),$$

while the same lemma gives $\lambda_{i\ell}^* = \widehat{z}_{i\ell} + \gamma_2 g_{i\ell}(x, \widehat{y}_i)$, hence $-B_{i\ell}/\gamma_2 = c_{k_j} g_{i\ell}(x, \widehat{y}_i)$. Subtracting,

$$c_{k_j}\big[g_{i\ell}(x, \widehat{y}_i) - g_{i\ell}(x, \theta_i^*)\big] = O(NR_{k_j}),$$

and since $\widehat{y}_i - \theta_i^* = -A_i/c_{k_j}$, expanding with uniform quadratic remainder gives $-\nabla_y g_{i\ell}^\top A_i = O(E_j) + O(NR_{k_j})$; multiplying by -1 and collecting over $\ell \in \mathcal{A}_i$ yields equation 128.

Equation equation 129. Expanding the first two terms of $\nabla_x G_{\gamma,i}$ (Lemma C.2) about $(\theta_i^*, \widehat{z}_i)$ in the same way and multiplying by c_{k_j}, using $B_{i\ell} = 0$ off $\mathcal{A}_i$, gives equation 129. □

E.4 Boundedness of the rescaled displacements

Lemma E.9. *Under Assumptions E.1 and E.5, the sequences $\{A_i^{k_j}\}$ and $\{B_i^{k_j}\}$ are bounded.*

Proof. Fix i and suppose $t_j := \|(A_i^{k_j}, B_{i,\mathcal{A}_i}^{k_j})\| \to \infty$ along a further subsequence. By Lemma E.7(c) the inactive coordinates of $B_i^{k_j}$ vanish, so $E_j = t_j^2/c_{k_j}$ and $E_j/t_j = t_j/c_{k_j}$. Dividing equation 127 and equation 128 by t_j and writing $(\widetilde{A}_j, \widetilde{B}_j) := (A_i^{k_j}, B_{i,\mathcal{A}_i}^{k_j})/t_j$, so $\|(\widetilde{A}_j, \widetilde{B}_j)\| = 1$,

$$H_i^j \widetilde{A}_j + (K_i^j)^\top \widetilde{B}_j = \frac{\nabla_y F_i}{t_j} + O\!\left(\frac{t_j}{c_{k_j}}\right) + O\!\left(\frac{NR_{k_j}}{t_j}\right), \qquad K_i^j \widetilde{A}_j = O\!\left(\frac{t_j}{c_{k_j}}\right) + O\!\left(\frac{NR_{k_j}}{t_j}\right).$$

By Lemma E.2, $t_j/c_{k_j} \le \sqrt{2\max\{\gamma_1, \gamma_2\}\, G_{\gamma,i}(\widehat{w}_i^{k_j})} \to 0$; and $R_{k_j} \to 0$ while $t_j \to \infty$, so every right-hand side tends to zero. Passing to a limit along a subsequence on which $(\widetilde{A}_j, \widetilde{B}_j)$ converges — possible on the unit sphere — and using $H_i^j \to \bar{H}_i$, $K_i^j \to \bar{K}_i$, which holds by Assumption E.1 and $\widehat{w}^{k_j} \to \bar{w}$ with $\mathcal{A}_i$ the fixed limiting index set and $B_{i\ell}^{k_j} = 0$ off it for large j, the limit is a unit vector in the kernel of J_i, contradicting nonsingularity of equation 125. Hence t_j is bounded. □

E.5 THE CORRESPONDENCE THEOREM

Theorem E.10. *Let $\{w^k\}$ be generated by Algorithm 1 under the hypotheses of Theorem D.16, and let $\bar{w}$ be the limit point of Lemma E.4. If Assumptions E.1 and E.5 hold at $\bar{w}$, then:*

(i) $\bar{w}$ is bilevel feasible, and each $\bar{y}_i$ is the unique global solution of client i's lower-level problem at $\bar{x}$;

(ii) $\big(A_i^{k_j}, B_{i,\mathcal{A}_i}^{k_j}\big) \to (a_i, b_i)$, the unique solution of the lower-level adjoint system

$$\bar{H}_i\, a_i + \bar{K}_i^\top b_i = \nabla_y F_i(\bar{x}, \bar{y}_i), \qquad \bar{K}_i\, a_i = 0;$$

(iii) $\bar{x}$ is a stationary point of the reduced upper-level problem: $0 \in \nabla\varphi(\bar{x}) + \mathcal{N}_{\mathcal{X}}(\bar{x})$, with φ as in Lemma E.6, which coincides with the objective of equation 1 near $\bar{x}$.

Proof. Part (i) is Lemma E.6 together with the remark following Lemma E.4.

Part (ii). By Lemma E.9 the sequence is bounded, so $E_j \to 0$ and the remainders in equation 127–equation 128 vanish; the matrices converge to $\bar{H}_i, \bar{K}_i$ and $\nabla_y F_i(\widehat{x}^{k_j}, \widehat{y}_i^{k_j}) \to \nabla_y F_i(\bar{x}, \bar{y}_i)$. Any accumulation point therefore satisfies $J_i(A_i^\infty, B_{i,\mathcal{A}_i}^\infty)^\top = (\nabla_y F_i(\bar{x}, \bar{y}_i), 0)^\top$, which by nonsingularity of J_i has the unique solution (a_i, b_i); since every accumulation point of a bounded sequence equals this value, the whole sequence converges to it. By Lemma E.7(c), $B_{i\ell}^{k_j} = 0$ off $\mathcal{A}_i$, matching $b_{i\ell} = 0$ there.

Part (iii). By Lemma E.6, φ is continuously differentiable near $\bar{x}$. Differentiating the reduced lower-level KKT system along the branch and eliminating $\partial\mu_i/\partial x$ using $\bar{K}_i a_i = 0$ gives

$$\nabla\varphi_i(\bar{x}) = \nabla_x F_i(\bar{x}, \bar{y}_i) - \bar{M}_i\, a_i - \bar{D}_i^\top b_i. \tag{131}$$

By equation 129 and part (ii), $c_{k_j} \nabla_x G_{\gamma,i}(\widehat{w}_i^{k_j}) \to -\bar{M}_i a_i - \bar{D}_i^\top b_i$, so, averaging over i,

$$\nabla_x \psi_{c_{k_j}}(\widehat{w}^{k_j}) = \frac{1}{N}\sum_{i=1}^N \big[\nabla_x F_i + c_{k_j}\nabla_x G_{\gamma,i}\big] \;\longrightarrow\; \frac{1}{N}\sum_{i=1}^N \nabla\varphi_i(\bar{x}) = \nabla\varphi(\bar{x}).$$

The x-block of equation 122 gives $v_x^{k_j} = e_x^{k_j} - \nabla_x \psi_{c_{k_j}}(\widehat{w}^{k_j})$ with $v_x^{k_j} \in \mathcal{N}_{\mathcal{X}}(\widehat{x}^{k_j})$ and $\|e_x^{k_j}\| \le R_{k_j} \to 0$; hence $v_x^{k_j} \to -\nabla\varphi(\bar{x})$. Since $\widehat{x}^{k_j} \to \bar{x}$ and the normal-cone map of the closed convex set $\mathcal{X}$ has closed graph, $-\nabla\varphi(\bar{x}) \in \mathcal{N}_{\mathcal{X}}(\bar{x})$. □

Remark E.11. Theorem E.10 is conditional on regularity *at the limit point*, the standard form of results in constrained sensitivity analysis; it says nothing at degenerate limit points, where J_i is singular and the lower-level solution map need not be differentiable.Full participation and exact oracles are conditions of Theorem D.16 itself rather than additional restrictions imposed here: the correspondence is a statement about the iterates that theorem produces, and cannot be broader than it. Under a fixed penalty the question does not arise, since $G_{\gamma,i}$ is not driven to zero and no limit point is bilevel feasible to begin with. Under the stated conditions, exact oracles and full participation, the boundedness hypothesis of Theorem D.16, twice-differentiable data, and LICQ, strict complementarity and SSOSC at $\bar{w}$ — the scaled gap gradients $c_k \nabla_x G_{\gamma,i}$ converge along the selected subsequence to the sensitivity terms of the reduced hypergradient equation 131. Stationarity is not transferred through the constraint system, and cannot be: $G_{\gamma,i} \ge 0$ vanishes on the entire feasible set, so every feasible point minimizes it over $\mathcal{W}$ and no MFCQ-type direction exists at any of them. Making a lower-level multiplier an explicit optimization variable is known to create stationary points with no counterpart in the original problem, even in the convex case (Dempe & Dutta, 2012), and z_i is such a variable in equation 6. The centralized analysis of Yao et al. (2025) relates the reformulation to the original problem at the level of global minimizers (their Theorems 2.3 and 3.3), and its accumulation-point result presumes exact global minimization of each penalized subproblem; its algorithmic guarantees, like ours, are stated for the residual of the penalized objective. The step taken here is therefore the one that argument does not take: it works along the sequence the algorithm generates, at which exact minimization is unavailable. Theorem E.10 instead identifies conditions on the limit point under which the correspondence holds for the iterates FedBiC produces.