COMPLEXITY GUARANTEES FOR HETEROGENEOUS FEDERATED BILEVEL OPTIMIZATION WITH LOWER-LEVEL CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated bilevel optimization (FBO) provides a framework for bilevel learning with decentralized data, yet existing methods largely focus on unconstrained lower-level problems and often rely on hypergradient computation. Lower-level constraints often induce nonsmoothness in the solution mapping, posing substantial challenges for standard implicit differentiation and hypergradient approaches. We propose an inexact zeroth-order framework for FBO that accommodates lower-level constraints and nonsmooth, nonconvex implicit upper-level objectives. To the best of our knowledge, this is the first zeroth-order method for constrained FBO that provides convergence and communication-complexity guarantees while accommodating client drift in both levels. Our contributions are threefold. (i) We design a SCAFFOLD-based zeroth-order scheme that uses control variates to correct client drift under heterogeneous data and partial client participation. To reduce repeated lower-level communication, we compute only two inexact lower-level solutions per outer round and reuse them across multiple local upper-level updates, while explicitly characterizing and controlling the resulting delay, bias, and lower-level inexactness. (ii) We propose Adaptive Penalty SCAFFOLD for constrained federated lower-level problems, using an increasing penalty parameter and round-dependent local stepsizes, and establish convergence guarantees for the resulting inexact solutions. (iii) We extend the framework to constrained personalized FBO with client-specific lower-level problems and a general bilevel structure, allowing distinct upper- and lower-level objectives and client-specific constraints. Our numerical experiments on hyperparameter optimization and personalized federated learning demonstrate the performance of the proposed methods.

1 INTRODUCTION

In modern large-scale learning systems, data are often distributed across multiple clients. Federated learning (FL) enables collaborative training in such settings while preserving privacy and reducing communication overhead. A key challenge in FL is data heterogeneity, where client data are not identically distributed Li et al. (2020), leading to client drift and slow or unstable convergence of standard methods such as FedAvg McMahan et al. (2017). Control-variate-based methods such as SCAFFOLD Karimireddy et al. (2020) mitigate this issue by correcting biased local updates. Many learning problems further exhibit a bilevel structure, such as hyperparameter optimization Ghadimi and Wang (2018); Ji et al. (2021). In these problems, the upper-level objective depends implicitly on the solution of a lower-level problem, making it difficult to evaluate. Furthermore, when the lower-level problem is constrained, the resulting implicit objective function may be nonsmooth and nonconvex. Bilevel optimization in FL introduces additional challenges due to the implicit objective and communication limitations. In this work, we study federated bilevel optimization problems with lower-level constraints. This includes the following problem class which we call global federated bilevel optimization (GFBO).

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_i\in D_i}[\tilde{f}_i(x,y(x),\xi_i)] \quad \text{s.t.} \quad y(x)\in \arg\min_{y\in Y(x)} \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\zeta_i\in \tilde{D}_i}[\tilde{h}_i(x,y,\zeta_i)]. \tag{GFBO}$$

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.

Here, there are m clients, each indexed by $i \in [m]$, holding possibly heterogeneous local datasets. The upper-level variable $x \in \mathbb{R}^n$ represents the global model parameter, while $y(x) \in \mathbb{R}^{\tilde{n}}$ denotes the solution to the lower-level problem associated with a given x. Each client i has a local validation dataset D_i and a local training dataset $\tilde{D}_i$. The function $\tilde{f}_i : \mathbb{R}^n \times \mathbb{R}^{\tilde{n}} \times D_i \to \mathbb{R}$ denotes a (possibly nonsmooth and nonconvex) local validation loss, and $\tilde{h}_i : \mathbb{R}^n \times \mathbb{R}^{\tilde{n}} \times \tilde{D}_i \to \mathbb{R}$ represents a smooth, strongly convex local training loss. The feasible set $Y(x) \subseteq \mathbb{R}^{\tilde{n}}$ is nonempty, convex, and closed for every x. The strong convexity of the lower-level objective ensures the existence and uniqueness of the solution $y(x)$. Let us define the local implicit function for client i as $f_i(x) \triangleq \mathbb{E}_{\xi_i \in D_i}[\tilde{f}_i(x, y(x), \xi_i)]$ and the global implicit objective as $f(x) \triangleq \frac{1}{m}\sum_{i=1}^m f_i(x)$.

The problem in GFBO assumes a shared lower-level response $y(x)$ across clients. In heterogeneous federated learning settings, clients may instead maintain client-specific responses $y_i(x)$, consistent with personalized federated learning, where clients share a global structure while learning models adapted to their local data Dinh et al. (2020); Fallah et al. (2020); McLaughlin and Su (2024); Allouah et al. (2024); M. Ghari and Shen (2024); Wang et al. (2024). To this end, we further study a variant of GFBO, in which each client has its own lower-level problem. Specifically, we study the following problem, referred to as personalized federated bilevel optimization (PFBO).

$$\min_{x \in \mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_i \in D_i}[\tilde{f}_i(x, y_i(x), \xi_i)] \text{ s.t. } y_i(x) \in \arg\min_{y_i \in Y_i(x)} \mathbb{E}_{\zeta_i \in \tilde{D}_i}[\tilde{h}_i(x, y_i, \zeta_i)],\ \forall i \in [m]. \tag{PFBO}$$

Problem PFBO captures personalized learning through client-specific models $y_i(x)$ coupled with a shared global model. Several standard personalized FL frameworks including pFedMe Dinh et al. (2020) and Ditto Li et al. (2021) can be viewed as special cases. For pFedMe, $\tilde{h}_i(x, y_i, \zeta_i) = \tilde{\ell}_i(y_i, \zeta_i) + \frac{\lambda}{2}\|y_i - x\|^2$ with $Y_i(x) = \mathbb{R}^n$, and the upper-level objective is the corresponding regularized local objective evaluated at $y_i(x)$. For Ditto, under the standard averaged global objective, $\tilde{f}_i(x, y_i(x), \xi_i) = \tilde{\ell}_i(x, \xi_i)$, while $\tilde{h}_i(x, y_i, \zeta_i) = \tilde{\ell}_i(y_i, \zeta_i) + \frac{\lambda}{2}\|y_i - x\|^2$, with $Y_i(x) = \mathbb{R}^n$. Thus, both formulations employ unconstrained personalized lower-level problems, while Ditto's upper-level objective is independent of the lower-level response.

Key challenges. *(i)* Lower-level constraints in both problems GFBO and PFBO may induce nonsmoothness in the solution mapping(s) and hence in the implicit upper-level objective, even when the explicit upper-level objective is smooth. This poses substantial challenges for standard hypergradient-based methods, which have been widely used for bilevel optimization, motivating the development of hypergradient-free and zeroth-order approaches. *(ii)* Communication efficiency and client drift, while well studied in standard FL, pose additional challenges in bilevel settings. For instance, in problem GFBO, constructing a zeroth-order gradient for f requires clients to inexactly compute $y(\bullet)$ at each local step, where the argument $\bullet$ is time-varying. However, $y(\bullet)$ itself solves a parametric lower-level FL problem, which requires additional communication. Developing a truly communication-efficient inexact zeroth-order framework therefore requires a careful algorithmic design. *(iii-1)* For problem GFBO, the lower-level problem is constrained and federated, with a feasible set that depends on the upper-level variable. Constructing inexact zeroth-order gradients for the upper level requires computing $y(\bullet)$ to a prescribed accuracy, which calls for novel FL methods that accommodate constraints and client drift while providing explicit complexity guarantees. *(iii-2)* For problem PFBO, the challenge in *(iii-1)* does not arise and the inexact $y_i(\bullet)$ can be computed locally without additional communication. However, integrating an upper-level inexact zeroth-order framework with local lower-level solvers subject to prescribed termination criteria, while addressing the nonsmoothness and nonconvexity of the implicit objective and client drift, remains challenging.

Our Approach and Contributions. To resolve these challenges, we propose a unified inexact zeroth-order method (see Algorithm 1) for a constrained FBO under partial client participation that avoids explicit hypergradient computation. In Algorithm 1, we combine randomized smoothing with control variates to construct zeroth-order updates for the implicit objective under heterogeneous client data. To resolve challenge (ii), we reuse inexact lower-level solutions across multiple local upper-level updates. This delayed-reuse mechanism improves communication efficiency but introduces an additional bias in the zeroth-order estimators, which we explicitly characterize and control in the convergence analysis. We establish convergence and communication-complexity guarantees for both problems GFBO and PFBO.

Table 1: Comparison of federated bilevel optimization methods.

	FedNest	AggITD	FedBiOAcc	FedMBO	SimFBO	FedRZO	Anarchic	MemFBO	SVRBGD	BR-FedBi	SCAFFOLD-RZO$_{\mathrm{bl}}$
ZO	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
FO	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗
SO	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓	✗
Constr. LL	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Smooth.	L-sm.	L-sm.	L-sm.	L-sm.	L-sm.	sm.	L-sm.	L-sm.	L-sm.	L-sm.	sm.
Hypergrad.	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗

Notation: ZO: zeroth-order; FO: first-order; SO: second-order; LL: lower-level; sm.: smooth; L-sm.: Lipschitz-smooth; Hypergrad.: hypergradient computation. *References:* FedNest (Tarzanagh et al., 2022); AggITD (Xiao and Ji, 2023); FedBiOAcc (Li et al., 2023); FedMBO (Huang et al., 2023); SimFBO (Yang et al., 2023); FedRZO (Qiu et al., 2023); Anarchic (Li et al., 2024a); MemFBO (Yang et al., 2025); SVRBGD (Zhang et al., 2025); BR-FedBi (Li et al., 2026).

To the best of our knowledge, this is the first zeroth-order framework for federated bilevel optimization with constrained lower-level problems that explicitly corrects client drift at both levels under heterogeneous data and partial client participation while providing convergence and communication-complexity guarantees. Next, we elaborate on our main contributions and position them relative to the most closely related recent work.

(i) Client-drift correction at both levels. We incorporate client-specific and global control variates into the upper-level zeroth-order updates to correct client drift under heterogeneous data and partial client participation. For problem GFBO, client drift is additionally controlled within the federated lower-level solver. Recently, inexact randomized zeroth-order schemes have been developed for addressing problem GFBO, including FedRZO$_{\mathrm{bl}}$ Qiu et al. (2023). However, FedRZO$_{\mathrm{bl}}$ does not correct client drift during local updates.

(ii) Adaptive Penalty SCAFFOLD for constrained federated lower-level problems. In addressing problem GFBO, to approximate $y(\bullet)$, we propose an Adaptive Penalty SCAFFOLD method (see Algorithm 2). This method uses an increasing penalty parameter and round-dependent local stepsizes, together with control variates to correct client drift under heterogeneous data and partial client participation. Existing single-level constrained FL methods, such as FedDualAvg Yuan et al. (2021) and FedDA Li et al. (2024b), do not directly provide the lower-level solution-error guarantees for such parameter-dependent feasible sets, and FedDA additionally assumes bounded gradient dissimilarity. We establish explicit bounds on the resulting inexact lower-level solutions, which are subsequently incorporated into the upper-level zeroth-order analysis.

(iii) Addressing personalized constrained FBO. We extend the proposed framework from problem GFBO to its personalized variant PFBO. Problem PFBO includes pFedMe Dinh et al. (2020) and Ditto Li et al. (2021) under their standard formulations as special cases, while allowing distinct upper- and lower-level objectives, nonsmooth implicit upper-level objectives, and client-specific lower-level constraints. Moreover, our analysis for problem PFBO does not require the bounded-diversity condition used in pFedMe or the bounded stochastic-gradient second-moment condition used in the convergence analysis of Ditto.

Related work and gaps in the literature Table 1 compares representative FBO methods. Most existing FBO methods use first- or second-order information to compute or approximate hypergradients and primarily consider smooth, unconstrained lower-level problems. Recent methods also develop communication-efficient, asynchronous, memory-based, and variance-reduced updates. Among zeroth-order approaches, FedRZO$_{\mathrm{bl}}$ Qiu et al. (2023) avoids explicit hypergradient computation through randomized smoothing, but relies on bounded set dissimilarity and an unconstrained lower-level problem. BR-FedBi (Li et al., 2026) considers heterogeneity at both levels under Byzantine attacks, but assumes an unconstrained lower-level problem and relies on gradient/Hessian dissimilarity assumptions, resulting in heterogeneity-dependent asymptotic error terms. The unconstrained assumption becomes restrictive when the feasible set depends on the upper-level variable. For example, $y(x) \in \arg\min_{y \le x} \frac{1}{2}(y-1)^2$ yields $y(x) = \min\{1, x\}$, which is nonsmooth at $x = 1$, whereas the corresponding unconstrained problem yields $y(x) = 1$. Hence, parameter-dependent lower-level constraints can alter the implicit solution mapping even when the lower-level objective is smooth and strongly convex. Recent works study constrained bilevel optimization in non-federated settings Yao et al. (2024); Kornowski et al. (2024); Yao et al. (2025); Shen et al. (2025), but consider non-federated problems. In single-level FL, FedDualAvg Yuan et al. (2021) addresses convex composite optimization, while FedDA Li et al. (2024b) studies federated constrained optimization. Their

existing guarantees do not directly provide the inexact lower-level solution bounds required for our constrained FBO problem with an upper-level-dependent feasible set.

Personalized FL methods use regularization, meta-learning, and model-personalization strategies to adapt models to client-specific data Dinh et al. (2020); Li et al. (2021); Fallah et al. (2020); Arivazhagan et al. (2019); Deng et al. (2020); Collins et al. (2021). More recent works employ bilevel formulations for personalized FL Xing et al. (2022); Hashemi et al. (2024); Zhang et al. (2025). In contrast with standard personalized FL frameworks, problem PFBO allows for distinct upper- and lower-level objectives and client-specific, parameter-dependent feasible sets. For example, in equation 2, considered in our numerical experiments, the upper level uses the nonsmooth, nonconvex unit-margin ramp loss Tsilivis et al. (2024); Tong et al. (2023), whereas the lower level uses a smooth regularized logistic loss subject to an explicit client-specific personalization constraint.

2 ALGORITHM OUTLINE

We propose a unified method in Algorithm 1 to address both GFBO and its personalized variant PFBO under data heterogeneity across clients. The proposed method extends the SCAFFOLD algorithm Karimireddy et al. (2020), originally developed for single-level smooth federated learning problems, to the bilevel setting, where we employ a zeroth-order scheme combined with inexact implicit evaluations of the lower-level problem. This enables the handling of implicit upper-level objectives that may be nonsmooth and nonconvex, as well as of constrained lower-level problems.

(I) Scheme for addressing the problem GFBO. At each communication round r, the server generates random perturbations $\{v^r\}$ and invokes a federated lower-level solver to obtain inexact solutions at two perturbed points $x = \hat{x}^{r-1} - v^r$ and $x = \hat{x}^{r-1} + v^r$. Specifically, the server uses Algorithm 2 to compute $y_{\epsilon_r}(\hat{x}^{r-1} - v^r)$ and $y_{\epsilon_r}(\hat{x}^{r-1} + v^r)$, which approximate the exact lower-level solutions and satisfy $\mathbb{E}[\|y_{\epsilon_r}(\bullet) - y(\bullet)\|^2] \leq \epsilon_r$. To elaborate on the design of the zeroth-order scheme, we note that a standard ε-inexact zeroth-order local gradient for an η-smoothed $\tilde{f}_i(\bullet_i, y(\bullet_i), \xi_i)$ takes the form $\frac{n}{2\eta^2}(\tilde{f}_i(\bullet_i + v, y_\varepsilon(\bullet_i + v), \xi_i) - \tilde{f}_i(\bullet_i - v, y_\varepsilon(\bullet_i - v), \xi_i))v^r$. However, computing $y_\varepsilon(\bullet_i + v)$ would require implementing a lower-level FL method and thereby incurring excessive communication during each local step in Algorithm 1. To resolve this, we use delayed information and instead use $y_\varepsilon(\hat{\bullet} + v)$ to construct the gradients, where $\hat{\bullet}$ denotes the most recent server's model. Through the analysis, we then carefully handle this delay-induced error. The server broadcasts y_r^- and y_r^+, the global control variate $\hat{c}^{r-1}$, and the inexact lower-level solutions to a randomly selected subset of clients S^r. Each client $i \in S^r$ initializes $z_{i,0}^r = \hat{x}^{r-1}$ and performs K local updates using the zeroth-order estimator $z_{i,k}^r = z_{i,k-1}^r - \gamma_l(\check{g}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1})$. After K steps, clients update their control variates and send $(z_{i,K}^r, c_i^r)$ to the server. The server aggregates updates to obtain $\hat{x}^r = \hat{x}^{r-1} + \frac{\gamma_g}{|S^r|}\sum_{i\in S^r}(z_{i,K}^r - \hat{x}^{r-1})$, and updates the global control variate by averaging across clients. The final output $\bar{x}^R$ is selected uniformly at random from $\{\hat{x}^0, \ldots, \hat{x}^R\}$, such that $\mathbb{P}(\bar{x}^R = \hat{x}^{r-1}) = \frac{1}{R+1}$ for $r = 1, \ldots, R+1$.

(I-1) Adaptive Penalty SCAFFOLD for the constrained federated lower-level problem. To solve the constrained lower-level problem in Algorithm 1, we use Adaptive Penalty SCAFFOLD in Algorithm 2. For a fixed outer query x, constraint violation is measured by $0.5\,\mathrm{dist}^2(y, Y(x))$, which has gradient $y - \Pi_{Y(x)}(y)$, where $\Pi_{Y(x)}(\bullet)$ denotes the Euclidean projection onto the set $Y(x)$. At each round, a subset of clients performs local SCAFFOLD updates augmented with the penalty gradient. The clients then update their control variates, and the server aggregates the local model and control-variate changes. The increasing penalty parameter ρ_r^l progressively enforces feasibility, while the SCAFFOLD correction controls client drift. We also use a round-dependent stepsize $\gamma_{l,r}^l$ to balance the tradeoff between infeasibility and suboptimality.

(II) Scheme for addressing the problem PFBO. In the personalized setting, each client i maintains its own lower-level variable $y_i(x)$, and the lower-level problems are solved locally without requiring federated communication. As a result, the inexact evaluations $y_{\epsilon_r}(\cdot)$ are computed directly by each client. In contrast to the global setting, the personalized formulation eliminates delayed inexact computation, since each client evaluates its own lower-level solution (e.g., employing projected stochastic gradient descent) locally without invoking a federated solver. This removes repeated

lower-level communication and significantly reduces overhead. The upper-level updates remain identical to those in addressing problem GFBO and follow the same SCAFFOLD-based structure.

Algorithm 1 Unified delay-enabled inexact randomized implicit zeroth-order SCAFFOLD (SCAFFOLD-RZO$_{\text{bl}}$). (Schemes for addressing problems GFBO and PFBO are highlighted.)

Server input: initial $\hat{x}^0 \in \mathbb{R}^n$ and $\hat{c}^0 = 0$, global step size $\gamma_g \geq 1$, sampling size s, smoothing parameter $\eta > 0$, and $K \in \mathbb{N}$.
Client i's input for $i \in \{1, 2, \ldots, m\}$: $z_{i,0}^0 = \hat{x}^0$, $c_i^0 = 0$, local step size $\gamma_l > 0$.
for each round $r = 1, 2, \ldots, R+1$ **do**
  Server samples $v^r \sim \text{Unif}(\eta\mathbb{S})$ and $\mathcal{S}^r \subseteq [m]$ with $|\mathcal{S}^r| = s$
  Server uses Algorithm 2 to compute $y_r^- = y_{\epsilon_r}(\hat{x}^{r-1} - v^r)$ and $y_r^+ = y_{\epsilon_r}(\hat{x}^{r-1} + v^r)$
  for $i \in \mathcal{S}^r$ **in parallel do**
    Initialize local model $z_{i,0}^r = \hat{x}^{r-1}$
    for $k = 1, 2, \ldots, K$ **do**
      Client i generates a random sample $\xi_{i,k-1}^r \in \mathcal{D}_i$
      Client i provides $y_r^- := y_{\epsilon_{i,r}}(z_{i,k-1}^r - v^r)$ and $y_r^+ = y_{\epsilon_{i,r}}(z_{i,k-1}^r + v^r)$
      $\check{g}_{i,k-1}^{\eta,r,z} := \frac{n}{2\eta^2}(\tilde{f}_i(z_{i,k-1}^r + v^r, y_r^+, \xi_{i,k-1}^r) - \tilde{f}_i(z_{i,k-1}^r - v^r, y_r^-, \xi_{i,k-1}^r))v^r$
      $z_{i,k}^r = z_{i,k-1}^r - \gamma_l(\check{g}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1})$
    end for
    Update $c_i^r = c_i^{r-1} - \hat{c}^{r-1} + \frac{1}{K\gamma_l}(\hat{x}^{r-1} - z_{i,K}^r)$, $\Delta z_i^r = z_{i,K}^r - \hat{x}^{r-1}$, $\Delta c_i^r = c_i^r - c_i^{r-1}$
  end for
  $\hat{x}^r = \hat{x}^{r-1} + \frac{\gamma_g}{|\mathcal{S}^r|}\sum_{i\in\mathcal{S}^r}\Delta z_i^r$ and $\hat{c}^r = \hat{c}^{r-1} + \frac{1}{m}\sum_{i\in\mathcal{S}^r}\Delta c_i^r$
end for
return $\bar{x}^R = \hat{x}^{\hat{r}-1}$, where $\hat{r} \sim \text{Unif}\{1, \ldots, R+1\}$

Assumption 2.1. Consider problems GFBO and PFBO. Let the following statements hold.
(i) Upper-level objective function. For all $i \in [m]$, $\tilde{f}_i(\bullet, y, \xi_i)$ is $L_{0,x}^f(\xi_i)$-Lipschitz continuous, where $L_{0,x}^f \triangleq \max_{i=1,\ldots,m}\sqrt{\mathbb{E}_{\xi_i\sim D_i}[(L_{0,x}^f(\xi_i))^2]} < \infty$. Also, for any $x \in \mathbb{R}^n$, $\tilde{f}_i(x, \bullet, \xi_i)$ is $L_{0,y}^f(\xi_i)$-Lipschitz continuous, where $L_{0,y}^f \triangleq \max_{i=1,\ldots,m}\sqrt{\mathbb{E}_{\xi_i\sim D_i}[(L_{0,y}^f(\xi_i))^2]} < \infty$.
(ii) Lower-level objective function. For all $i \in [m]$ and for any $x \in \mathbb{R}^n$, let us define $h_i(x, \bullet) \triangleq \mathbb{E}_{\zeta_i\sim\tilde{D}_i}[\tilde{h}_i(x, \bullet, \zeta_i)]$. Then, $h_i(x, \bullet)$ is $L_{1,y}^h$-smooth and μ_h-strongly convex function and for any $y \in \mathbb{R}^{\tilde{n}}$, the mapping $\nabla_y h_i(\bullet, y)$ is $L_{1,x}^h$-Lipschitz continuous.
(iii) Lower-level constraint set. For problem GFBO, the set $Y(x) \subseteq \mathbb{R}^{\tilde{n}}$ is nonempty, closed, and convex for all $x \in \mathbb{R}^n$, and satisfies $\|\Pi_{Y(x_1)}[\bullet] - \Pi_{Y(x_2)}[\bullet]\| \leq L_Y\|x_1 - x_2\|$ for all $x_1, x_2 \in \mathbb{R}^n$ and some $L_Y > 0$. The same assumption holds for $Y_i(x)$ in problem PFBO, for all $i \in [m]$.

Assumption 2.2. Consider problem GFBO. Assume that Assumption 2.1 (ii) holds, and the set $Y(x) \subseteq \mathbb{R}^{\tilde{n}}$ is nonempty, closed, and convex for all $x \in \mathbb{R}^n$. Suppose $\sup_x \left\|\nabla_y h\big(x, y(x)\big)\right\| < \infty$. Moreover, suppose for all $i \in [m]$ and for any $x \in \mathbb{R}^n$, $\mathbb{E}_{\zeta_i\sim\tilde{D}_i}[\|\nabla\tilde{h}_i(x, \bullet, \zeta_i) - \nabla h_i(x, \bullet)\|^2] \leq \sigma_h^2$.

Remark 2.3. For the hyperparameter optimization problem in equation 1, Lemma A.9 verifies Assumption 2.1(iii), while Lemma A.13 establishes $\sup_x \|\nabla_y h(x, y(x))\| < \infty$ in Assumption 2.2. Lemmas A.10 and A.11 further verify Assumption 2.1(iii) for additional feasible sets.

Remark 2.4. We clarify the main distinctions between the proposed framework and FedRZO$_{\text{bl}}$ Qiu et al. (2023), a zeroth-order method for FBO that avoids explicit hypergradient computation. In contrast, our method *(i)* addresses client drift at both levels through control variates, *(ii)* accommodates constrained lower-level problems through the proposed Algorithm 2, *(iii)* allows partial client participation at both levels, and *(iv)* explicitly characterizes and controls the additional bias in the zeroth-order estimator for $k \geq 2$ induced by delayed reuse of the lower-level solutions. Both methods compute only two lower-level solutions per outer round and reuse them across the K local upper-level updates. In our analysis, this additional bias is captured through a separate bias term and controlled

Algorithm 2 Adaptive Penalty SCAFFOLD (x, $\hat{y}^1$, $\hat{c}^1 = c_i^1 = 0$, $\{\gamma_{l,r}^l, \rho_r^l\}_{r=1}^{R^l}$, γ_g^l)

for $r = 1, \ldots, R^l$ **do**
  Sample $S^{l,r} \subseteq [m]$, $|S^{l,r}| = s^l$, and broadcast $(\hat{y}^r, \hat{c}^r)$
  for all $i \in S^{l,r}$ **in parallel do**
    Set $y_{i,0}^r = \hat{y}^r$ and, for $k = 0, \ldots, K^l - 1$, sample $\zeta_{i,k}^r \sim \mathcal{D}_i$ and update
    $y_{i,k+1}^r = y_{i,k}^r - \gamma_{l,r}^l(\nabla_y \tilde{h}_i(x, y_{i,k}^r, \zeta_{i,k}^r) + \rho_r^l(y_{i,k}^r - \Pi_{Y(x)}(y_{i,k}^r)) - c_i^r + \hat{c}^r)$
    $c_i^{r+1} = c_i^r - \hat{c}^r + (K^l \gamma_{l,r}^l)^{-1}(\hat{y}^r - y_{i,K^l}^r)$, $\Delta y_i^r = y_{i,K^l}^r - \hat{y}^r$, $\Delta c_i^r = c_i^{r+1} - c_i^r$
  end for
  $\hat{y}^{r+1} = \hat{y}^r + \frac{\gamma_g^l}{s^l} \sum_{i \in S^{l,r}} \Delta y_i^r$, $\quad \hat{c}^{r+1} = \hat{c}^r + \frac{1}{m} \sum_{i \in S^{l,r}} \Delta c_i^r$
end for
return $\hat{y}^{R^l+1}$

through the local drift, together with the error arising from inexact lower-level solutions. Such a bias is a distinct term and is not characterized in the analysis of FedRZO$_{\text{bl}}$.

3 Convergence analysis

In this section, we first present Theorem 3.1, which establishes the error bound and communication complexity of Algorithm 2 for the constrained lower-level problem. We then present two theorems establishing the convergence and communication complexity guarantees of Algorithm 1 for problems GFBO and PFBO, respectively.

Theorem 3.1. Consider Algorithm 2 for a fixed outer query x, and suppose Assumption 2.2 holds. Let $y(x)$ denote the exact solution of the constrained lower-level problem. The quantity G^* and the choices of γ_g^l, $\gamma_{l,r}^l$, and ρ_r^l are specified in A.30. Then, after R^l communication rounds, Algorithm 2 returns an approximate lower-level solution $y_\varepsilon(x)$ satisfying $\mathbb{E}\big[\|y_\varepsilon(x) - y(x)\|^2\big] \le \varepsilon$, where ε satisfies $\varepsilon = \mathcal{O}(\max\{\frac{m(G^*)^2 \exp(2(L_{1,y}^h)^2/\mu_h)}{\mu_h^2 (R^l)^2 s^l}, \frac{\exp((L_{1,y}^h)^2/\mu_h)}{\mu_h^2 R^l}((G^*)^2 + (1 + \frac{1}{s^l})\frac{\sigma_h^2}{K^l})\})$.

Remark 3.2. Theorem 3.1 establishes an $\mathcal{O}(1/R^l)$ convergence rate despite the presence of constraints, matching the rate of unconstrained SCAFFOLD in the stochastic strongly convex setting (Karimireddy et al., 2020, Theorem VII). This rate is achieved through the joint design of the penalty parameter ρ_r^l and the effective stepsize γ_r^l. Our analysis yields a linearly increasing penalty and a reciprocally decreasing stepsize (cf. A.30).

Theorem 3.3. Consider Algorithm 1 for addressing the problem in GFBO. Suppose Assumption 2.1 holds. Let $L_0^{\text{imp}} := L_{0,y}^f L_{0,x}^y + L_{0,x}^f$. Let $\gamma_{\max} > 0$ satisfy $\gamma_{\max} \le N$ and $\frac{\eta}{\gamma_{\max}} > L_0^{\text{imp}} \sqrt{n}\, M$, where N and M are defined in Definition A.21. Assume that the lower-level problem is solved inexactly, and let ϵ_r be the corresponding error level at round $r \ge 1$, satisfying $\mathbb{E}\left[\|y_{\epsilon_r}(\bullet) - y(\bullet)\|^2\right] \le \epsilon_r$, where $\{\varepsilon_r\}$ is nonincreasing. Let $\partial_\eta f(x)$ denote the η-Clarke subdifferential of f at x. Then, there exists a constant effective stepsize $\tilde{\gamma} := K\gamma_g\gamma_l$, where $\gamma_g \ge 1$, satisfying $0 < \tilde{\gamma} \le \gamma_{\max}$, such that the following statements hold true for $K \ge 2$ and $R \ge 1$.

(i) [Error bound] The error metric satisfies

$$\mathbb{E}\big[\text{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] \le \mathcal{O}\Big(\frac{L_0^{\text{imp}}(n^{3/2} + n\sqrt{\frac{m}{s}})}{\eta R}\Big) + \mathcal{O}\Big(\frac{(L_0^{\text{imp}})^{3/2} n^{3/4}}{\sqrt{\eta R}}\Big) + \mathcal{O}\Big(\frac{n^2 (L_{0,y}^f)^2}{\eta^2} \frac{1}{R+1} \sum_{r=1}^{R+1} \epsilon_{r-1}\Big).$$

(ii) Suppose Assumption 2.2 holds. At each round r of Algorithm 1, Algorithm 2 is called twice for the lower-level problems, where R_r^l, K_r^l, and s_r^l denote the number of lower-level communication rounds, the number of local iterations per client per lower-level communication round, and the number of selected clients, respectively. We set $R_r^l := r + 1$, $K_r^l := \sqrt{K}$, and $s_r^l := \sqrt{s}$. Ignoring logarithmic factors, to ensure $\mathbb{E}\big[\text{dist}^2(0, \partial_\eta f(\bar{x}^R))\big] \le \epsilon$, we obtain the following results.

(ii.a) [Error bound] We obtain

$$\mathbb{E}\big[\text{dist}^2(0, \partial_\eta f(\bar{x}^R))\big] \le \mathcal{O}\Big(\frac{L_0^{\text{imp}}(n^{3/2} + n\sqrt{\frac{m}{s}})}{\eta R} + \frac{(L_0^{\text{imp}})^{3/2} n^{3/4}}{\sqrt{\eta R}}\Big) + \frac{n^2 (L_{0,y}^f)^2 \exp\big(\frac{2(L_{1,y}^h)^2}{\mu_h}\big)\big(\frac{m}{\sqrt{s}} + \frac{\sigma_h^2}{\sqrt{K}}\big)}{\eta^2 \mu_h^2 R}\Big).$$

(ii.b) [Upper-level communication complexity] Accordingly, we can ensure the required accuracy by taking $R_\varepsilon = \mathcal{O}(\max\{\frac{L_0^{\mathrm{imp}}(n^{3/2}+n\sqrt{\frac{m}{s}})}{\eta\varepsilon}, \frac{(L_0^{\mathrm{imp}})^3 n^{3/2}}{\eta\varepsilon^2}, \frac{n^2(L_{0,y}^f)^2\exp(2(L_{1,y}^h)^2/\mu_h)}{\eta^2\mu_h^2\varepsilon}(\frac{m}{\sqrt{s}}+\frac{\sigma_h^2}{\sqrt{K}})\})$.

(ii.c) [Total communication complexity] Accounting for the two lower-level calls at each outer round, we obtain $\bar{R}_\varepsilon = \mathcal{O}(\max\{\frac{(L_0^{\mathrm{imp}})^2(n^3+n^2\frac{m}{s})}{\eta^2\varepsilon^2}, \frac{(L_0^{\mathrm{imp}})^6 n^3}{\eta^2\varepsilon^4}, \frac{n^4(L_{0,y}^f)^4\exp(4(L_{1,y}^h)^2/\mu_h)(\frac{m}{\sqrt{s}}+\frac{\sigma_h^2}{\sqrt{K}})^2}{\eta^4\mu_h^4\varepsilon^2}\})$.

Proof sketch. We outline the main steps for establishing Theorem 3.3.

Step 1: Randomized smoothing and zeroth-order error decomposition. Since the implicit upper-level objective f may be nonsmooth and nonconvex, we apply randomized smoothing to obtain the differentiable surrogate f^η and use its two-point zeroth-order representation in Algorithm 1. For $k = 1$, the ideal estimator is conditionally unbiased at the round-start iterate $\hat{x}^{r-1}$. For $k \geq 2$, reusing the same perturbation v^r across local updates introduces the bias $b_{i,k-1}^r$. In addition, reusing lower-level solutions computed at the round-start perturbed points creates a delayed-reuse error, while their approximate computation introduces a lower-level inexactness error. The reuse bias and delayed-reuse error are controlled through the client-drift quantity D_r, whereas the lower-level inexactness is controlled through ϵ_r.

Step 2: Client drift, control-variate error, and Lyapunov descent. We next track two coupled quantities (i) D_r, which measures the within-round drift of the local iterates from the round-start global iterate $\hat{x}^{r-1}$, and (ii) C_r, which captures the lag/error associated with the control variates relative to the updated global iterate $\hat{x}^r$. Using the control-variate-corrected local updates, we derive coupled recursions for D_r and C_r under partial client participation, without introducing BGD-type heterogeneity parameters. We then combine these recursions with a one-round descent inequality for f^η through a weighted Lyapunov function involving $\mathbb{E}[f^\eta(\hat{x}^r)]$, D_r, and C_r. The stepsize conditions allow the drift and control-variate terms to be absorbed into the descent inequality. Summing over the outer rounds and using the uniformly selected output $\bar{x}^R$ yields a bound on $\mathbb{E}[\|\nabla f^\eta(\bar{x}^R)\|^2]$. Finally, using $\mathrm{dist}(0, \partial_\eta f(x)) \leq \|\nabla f^\eta(x)\|$, we obtain Theorem 3.3(i).

Step 3: Lower-level specialization and communication complexity. For problem GFBO, we apply Algorithm 2 to the two lower-level calls at each outer round with $R_r^l = r+1$, $K_r^l = \sqrt{K}$, and $s_r^l = \sqrt{s}$. Substituting the resulting lower-level error bound into part (i) gives part (ii.a). Solving the resulting bound for the required number of outer rounds R_ε yields part (ii.b). Finally, accounting for the communication required by the two lower-level calls across all outer rounds gives the total communication complexity in part (ii.c).

Remark 3.4. We summarize several observations based on Theorem 3.3. *(i) Robustness to data heterogeneity.* The convergence bounds do not depend on data-heterogeneity parameters and are established without imposing BGD assumptions, reflecting the effect of client-drift correction at both federated levels. *(ii) Effect of local steps and client participation.* With $K_r^l = \sqrt{K}$ and $s_r^l = \sqrt{s}$, the lower-level inexactness term decreases with K through $\sigma_h^2/\sqrt{K}$ and with s through $m/\sqrt{s}$. Increasing s additionally reduces the upper-level partial-participation term $\sqrt{m/s}$. *(iii) Lower-level communication tradeoff.* With $R_r^l = r+1$, the lower-level inexactness contributes $\mathcal{O}(\frac{1}{R}(\frac{m}{\sqrt{s}}+\frac{\sigma_h^2}{\sqrt{K}}))$ to the error bound, ignoring logarithmic and problem-dependent factors. Increasing the lower-level communication budget, e.g., $R_r^l = (r+1)^2$, removes the logarithmic factor in the exact bound, at the expense of higher total lower-level communication.

Theorem 3.5. Consider problem PFBO and Algorithm 1. Suppose Assumption 2.1 holds. Under the same parameter choices, error-sequence and output-selection conditions, and effective stepsize setting as in Theorem 3.3, suppose each client addresses its lower-level problem inexactly, such that at round $r \geq 1$, $\mathbb{E}\left[\|y_{\epsilon_r,i}(\bullet) - y_i(\bullet)\|^2\right] \leq \epsilon_r$, $\forall i \in [m]$. Let $\partial_\eta f(x)$ denote the η-Clarke subdifferential of f at x. Then, the following hold for $K \geq 2$ and $R \geq 1$.

(i) [Error bound] The error metric satisfies

$$\mathbb{E}\big[\mathrm{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] \leq \mathcal{O}\Big(\frac{L_0^{\mathrm{imp}}}{\eta R}(n^{3/2} + n\sqrt{\tfrac{m}{s}}) + \frac{(L_0^{\mathrm{imp}})^{3/2} n^{3/4}}{\sqrt{\eta R}}\Big) + \frac{n^2(L_{0,y}^f)^2}{\eta^2}\frac{1}{R+1}\sum_{r=1}^{R+1}\epsilon_{r-1}\Big).$$

(ii) Suppose $Y_i(x) \subseteq \mathbb{R}^{\tilde{n}}$ for every $i \in [m]$ in problem PFBO. Each client applies the projected SGD method in Algorithm 4 to its lower-level problem and performs $H_{i,r} = H_r$ local steps at round r of

Algorithm 1. This choice yields $\epsilon_r = \mathcal{O}(1/H_r)$. We set $H_r = r$ and $\epsilon_0 := \epsilon_1$. Ignoring logarithmic factors, to ensure $\mathbb{E}\big[\mathrm{dist}^2(0, \partial_\eta f(\bar{x}^R))\big] \le \varepsilon$, we obtain the following results.

(ii.a) [Error bound] $\mathbb{E}\big[\mathrm{dist}^2(0, \partial_\eta f(\bar{x}^R))\big] \le \mathcal{O}(\frac{L_0^{\mathrm{imp}}(n^{3/2}+n\sqrt{\frac{m}{s}})}{\eta R} + \frac{(L_0^{\mathrm{imp}})^{3/2}n^{3/4}}{\sqrt{\eta R}} + \frac{n^2(L_{0,y}^f)^2}{\eta^2 R})$.

(ii.b) [Upper-level communication complexity] Accordingly, we can ensure the required accuracy by taking $R_\varepsilon = \max\{\mathcal{O}(\frac{L_0^{\mathrm{imp}}(n^{3/2}+n\sqrt{\frac{m}{s}})}{\eta\varepsilon}), \mathcal{O}(\frac{(L_0^{\mathrm{imp}})^3 n^{3/2}}{\eta\varepsilon^2}), \mathcal{O}(\frac{n^2(L_{0,y}^f)^2}{\eta^2\varepsilon})\}$.

(ii.c) [Total communication complexity] The total number of communication rounds equals the number of upper-level communication rounds. Hence, $\bar{R}_\varepsilon = R_\varepsilon$, where R_ε appears in (ii.b).

Remark 3.6. For problem PFBO, each client solves its own lower-level problem locally without federated communication. Hence, the delay-related error vanishes, and Theorem 3.3(i) extends to Theorem 3.5(i) by setting the delay terms to zero. Theorems 3.3 and 3.5 therefore yield the same order of upper-level communication complexity. Moreover, since the personalized lower-level problems require no communication, the total communication complexity for problem PFBO equals its upper-level communication complexity.

4 Experiments

We evaluate Algorithm 1 on problems GFBO and PFBO by addressing problem1 on CIFAR-10 and problem 2 on MNIST, respectively. In these experiments, for generating non-IID data, we use a Dirichlet distribution $\mathrm{Dir}(\tau)$, where τ is the concentration parameter Li et al. (2022). When $\tau \le 1$, the resulting datasets are non-IID, while $\tau \to \infty$ approaches the IID case.

Constrained federated bilevel hyperparameter learning We study the following federated hyperparameter learning problem, where the lower-level problem is constrained to induce a sparse solution y^*. The hyperparameters are $\lambda \in \mathbb{R}^m$ characterize the regularization parameters, and $\alpha \in \mathbb{R}$, which defines a convex constraint set $Y(x) = \{y \in \mathbb{R}^{\tilde{n}} \mid \|y\|_1 \le \log(1+\exp(\alpha))\}$ for the lower-level problem. Hence, the full hyperparameter vector is $x \in \mathbb{R}^{m+1}$. Let $D_i = \{(U_{i,\ell}, v_{i,\ell}) \in \mathbb{R}^{\tilde{n}} \times \{-1,1\} \mid \ell \in I_i\}$ denote the local test dataset for client i, where $I_i = \{1, \dots, |D_i|\}$ is the corresponding index set. Similarly, let $\tilde{D}_i = \{(U_{i,\tilde{\ell}}, v_{i,\tilde{\ell}}) \in \mathbb{R}^{\tilde{n}} \times \{-1,1\} \mid \tilde{\ell} \in \tilde{I}_i\}$ denote the local training dataset for client i, with $\tilde{I}_i = \{1, \dots, |\tilde{D}_i|\}$. Specifically, we consider a binary classification problem of the form

$$\min_{x=[\lambda;\alpha]\in\mathbb{R}^{m+1}} \big\{f(x) := \tfrac{1}{m}\textstyle\sum_{i=1}^m \sum_{\ell\in I_i} \log\left(1+\exp\left(-v_{i,\ell}\langle U_{i,\ell}, y(x)\rangle\right)\right)\big\} \tag{1}$$

$$\text{s.t.} \quad y(x) \in \underset{y\in\mathbb{R}^{\tilde{n}}, \|y\|_1 \le \log(1+\exp(\alpha))}{\mathrm{argmin}} \tfrac{1}{m}\textstyle\sum_{i=1}^m (\sum_{\tilde{\ell}\in\tilde{I}_i} \log(1+\exp(-v_{i,\tilde{\ell}}\langle U_{i,\tilde{\ell}}, y\rangle)) + \frac{\log(1+\exp(\lambda_i))}{2}\|y\|^2).$$

Following Wen et al. (2016), we choose CIFAR-10 as a suitable dataset for inducing sparsity.

Results and insights. In Figure 1, we compare our method with FedRZO$_{\text{bl}}$ Qiu et al. (2023) under different levels of heterogeneity τ. We observe that our method consistently attains lower objective values across all settings, indicating improved convergence behavior. The performance gap becomes more evident as τ decreases (more heterogeneous), indicating that our method is more robust to data heterogeneity. In contrast, FedRZO$_{\text{bl}}$ shows slower convergence and higher objective values, particularly in more heterogeneous regimes. These findings align with the known advantages of SCAFFOLD in the FL framework and demonstrate that its benefits extend to the bilevel setting. Moreover, our proposed method attains slightly higher validation accuracy in most cases.

Constrained personalized federated learning. We consider a constrained personalized federated bilevel optimization problem in the form

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}_{(U_i,v_i)\sim\mathcal{D}_i}\left[\ell_{\mathrm{ramp}}(v_i\langle U_i, y_i(x)\rangle)\right] \tag{2}$$

$$\text{s.t. } y_i(x) \in \underset{y_i\in Y_i(x)}{\mathrm{argmin}}\{\mathbb{E}_{(\widetilde{U}_i,\widetilde{v}_i)\sim\widetilde{\mathcal{D}}_i}\left[\log(1+\exp(-\widetilde{v}_i\langle\widetilde{U}_i, y_i\rangle))\right] + \tfrac{a_i}{2}\|y_i\|^2\},\ i\in[m],$$

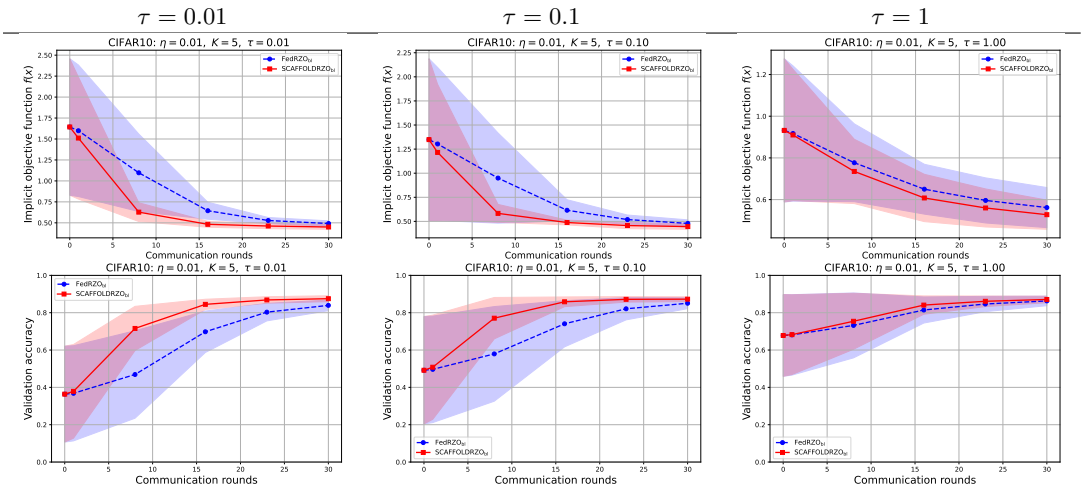

Figure 1: SCAFFOLD-RZO$_{\text{bl}}$ vs. FedRZO$_{\text{bl}}$ for problem 1 (95% confidence intervals, 6 runs).

where $Y_i(x) := \{y_i \in \mathbb{R}^n \mid \|y_i - x\| \leq b_i\}$. Here, $x \in \mathbb{R}^n$ denotes the shared global model, $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$, represents a given client-specific personalization budget. We evaluate binary MNIST 3-vs.-5 on ten clients with Dirichlet parameter $\tau = 0.5$, radius $b_i = 0.1$, 100 communication rounds, and five random seeds. Acc. and Loss denote client-mean test accuracy and unit-margin ramp loss. Max Vio. is $\max_i[\|y_i - x\| - b_i]_+$, and Feas. is the fraction of feasible clients. Ditto and pFedMe retain their original updates but add the soft penalty $\frac{\rho}{2}[\|y_i - x\| - b_i]_+^2$ with tuned penalty parameter ρ.

Table 2: Performance Metrics

	Ditto-Penalty	pFedMe-Penalty	Ours
Acc.↑	0.871 ± 0.039	0.803 ± 0.091	$\mathbf{0.879 \pm 0.024}$
Loss↓	0.796 ± 0.007	0.778 ± 0.020	$\mathbf{0.312 \pm 0.008}$
Max Vio.↓	0.0027 ± 0.0002	0.0050 ± 0.0004	$\mathbf{0.0000 \pm 0.0000}$
Feas.↑	0.54 ± 0.05	0.24 ± 0.09	$\mathbf{1.00 \pm 0.00}$

Results and insights. Our method obtains the highest accuracy and lowest ramp loss because its upper level directly optimizes the ramp loss of the personalized responses. Its projected lower-level updates enforce the hard constraint exactly, yielding zero violation and full feasibility. Ditto and pFedMe optimize different proximal objectives and enforce the constraint only through a finite penalty, which permits residual violations. The results therefore show that our method maintains exact feasibility while achieving competitive accuracy and substantially better classification margins.

5 Conclusion

We design an inexact zeroth-order framework for constrained global and personalized FBO with potentially nonsmooth and nonconvex implicit upper-level objectives. The proposed method avoids explicit hypergradient computation and uses control variates to correct client drift under heterogeneous data and partial client participation. To reduce repeated lower-level communication, we compute only two inexact lower-level solutions per outer round and reuse them across multiple local upper-level updates, while explicitly characterizing and controlling the resulting delay, bias, and lower-level inexactness. For constrained federated lower-level problems, we proposed Adaptive Penalty SCAFFOLD and established convergence guarantees for the resulting inexact solutions. We further extended the framework to constrained personalized FBO with client-specific lower-level problems.

Reproducibility Statement

To facilitate reproducibility, we provide complete descriptions of the proposed algorithms and their assumptions in the main text. Detailed proofs of the theoretical results, including the convergence and communication-complexity analyses, are provided in the appendix. The appendix also includes verification of the assumptions used in the numerical experiments. The datasets, parameter settings, and evaluation procedures used in the experiments are described in the corresponding experimental sections.

AI Use Statement

Generative AI tools were used to assist with language editing, including grammar correction, wording refinement, and improving the readability of the manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

References

Youssef Allouah, Abdellah El Mrini, Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Fine-tuning personalization in federated learning to mitigate adversarial clients. *Advances in Neural Information Processing Systems*, 37:100816–100844, 2024.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR, 2021.

Stella Dafermos. Sensitivity analysis in variational inequalities. *Mathematics of Operations Research*, 13(3):421–434, 1988.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

Mohammadjavad Ebrahimi, Uday V Shanbhag, and Farzad Yousefian. On the resolution of stochastic mpecs over networks: Distributed implicit zeroth-order gradient tracking methods. *arXiv preprint arXiv:2505.22916*, 2025.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

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

Diba Hashemi, Lie He, and Martin Jaggi. Cobo: Collaborative learning via bilevel optimization. *Advances in Neural Information Processing Systems*, 37:15550–15574, 2024.

Minhui Huang, Dewei Zhang, and Kaiyi Ji. Achieving linear speedup in non-iid federated bilevel learning. In *International conference on machine learning*, pages 14039–14059. PMLR, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.

Guy Kornowski, Swati Padmanabhan, Kai Wang, Jimmy Zhang, and Suvrit Sra. First-order methods for linearly constrained bilevel optimization. *Advances in neural information processing systems*, 37:141417–141460, 2024.

Dongsheng Li, Ye Zhu, Xiaowen Gong, Shiwen Mao, and Yang Zhou. Anarchic federated bilevel optimization. In *2024 22nd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 353–360. IEEE, 2024a.

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

Junyi Li, Feihu Huang, and Heng Huang. Fedda: Faster adaptive gradient methods for federated constrained optimization. In *International Conference on Learning Representations*, volume 2024, pages 53747–53759, 2024b.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.

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

Pouya M. Ghari and Yanning Shen. Personalized federated learning with mixture of models for adaptive prediction and model fine-tuning. *Advances in Neural Information Processing Systems*, 37:92155–92183, 2024.

Connor J McLaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. *Advances in Neural Information Processing Systems*, 37:77038–77059, 2024.

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*, pages 1273–1282. PMLR, 2017.

Yuyang Qiu, Uday Shanbhag, and Farzad Yousefian. Zeroth-order methods for nondifferentiable, nonconvex, and hierarchical federated optimization. *Advances in Neural Information Processing Systems*, 36:3425–3438, 2023.

Wei Shen, Jiawei Zhang, Minhui Huang, and Cong Shen. A single-loop first-order algorithm for linearly constrained bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 38, pages 57432–57479, 2025.

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

Peifeng Tong, Wu Su, He Li, Jialin Ding, Haoxiang Zhan, and Song Xi Chen. Distribution free domain generalization. In *International Conference on Machine Learning*, pages 34369–34378. PMLR, 2023.

Nikolaos Tsilivis, Natalie S Frank, Nathan Srebro, and Julia Kempe. The price of implicit bias in adversarially robust generalization. *Advances in Neural Information Processing Systems*, 37: 58023–58057, 2024.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jiaqi Wang, Qi Li, Lingjuan Lyu, and Fenglong Ma. pfedclub: Controllable heterogeneous model aggregation for personalized federated learning. *Advances in Neural Information Processing Systems*, 37:54692–54713, 2024.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Peiyao Xiao and Kaiyi Ji. Communication-efficient federated hypergradient computation via aggregated iterative differentiation. In *International Conference on Machine Learning*, pages 38059–38086. PMLR, 2023.

Pengwei Xing, Songtao Lu, Lingfei Wu, and Han Yu. Big-fed: Bilevel optimization enhanced graph-aided federated learning. *IEEE Transactions on Big Data*, 10(6):903–914, 2022.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. *Advances in Neural Information Processing Systems*, 36:33027–33040, 2023.

Yifan Yang, Peiyao Xiao, Shiqian Ma, and Kaiyi Ji. First-order federated bilevel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22029–22037, 2025.

Wei Yao, Chengming Yu, Shangzhi Zeng, and Jin Zhang. Constrained bi-level optimization: Proximal lagrangian value function approach and hessian-free algorithm. In *The Twelfth 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.

Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In *International conference on machine learning*, pages 12253–12266. PMLR, 2021.

Yihan Zhang, Rohit Dhaipule, Chiu C Tan, Haibin Ling, and Hongchang Gao. Federated stochastic bilevel optimization with fully first-order gradients. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7047–7055, 2025.

A Appendix

In this appendix, we provide a more detailed literature review and discussion of the gaps in the literature. In addition, the appendix mainly consists of the proofs of Theorem 3.3 and Theorem 3.5, together with the required preliminary lemmas and their proofs.

A.1 Related work and gaps in the literature

Federated bilevel optimization. Most existing methods for federated bilevel optimization rely on first- or second-order information together with hypergradient computation. Early works such as FedNest Tarzanagh et al. (2022) and FBO-AggITD Xiao and Ji (2023) use second-order information to estimate hypergradients. Later methods improve communication efficiency through more efficient gradient or hypergradient estimators. In particular, FedBiOAcc Li et al. (2023), FedMBO Huang et al. (2023), and SimFBO Yang et al. (2023) develop communication-efficient methods under smoothness assumptions. These methods generally require hypergradient computation and consider unconstrained lower-level problems. More recent works consider alternative federated designs. Anarchic FBO Li et al. (2024a) studies decentralized and asynchronous settings, FedSVRBGD-FO Zhang et al. (2025)

incorporates variance reduction, and MemFBO Yang et al. (2025) uses memory-based updates. These methods still rely on gradient-based or hypergradient-based information and primarily consider smooth, unconstrained formulations.

Among zeroth-order FBO methods, FedRZO$_{\text{bl}}$ Qiu et al. (2023) uses randomized smoothing to avoid explicit gradient and hypergradient computation. However, its analysis relies on BGD-type assumptions to control client heterogeneity and considers an unconstrained lower-level problem. In contrast, our method uses control variates to correct client drift without BGD-type assumptions and accommodates constrained lower-level problems.

BR-FedBi (Li et al., 2026) studies heterogeneous FBO under Byzantine attacks and considers heterogeneity in both the upper- and lower-level objectives. Its analysis relies on gradient/Hessian dissimilarity assumptions and assumes an unconstrained lower-level problem, resulting in heterogeneity-dependent asymptotic error terms. Our setting differs by considering constrained lower-level problems in the non-Byzantine setting and establishing convergence without BGD-type assumptions. Moreover, the lower-level feasible set in our formulation may depend on the upper-level variable, which can alter the implicit lower-level solution mapping even when the lower-level objective is smooth and strongly convex.

Constrained bilevel optimization. Recent works study bilevel problems with lower-level constraints, including linear, coupled, and more general constrained formulations Yao et al. (2024); Kornowski et al. (2024); Yao et al. (2025); Shen et al. (2025). These methods develop algorithmic and analytical tools for constrained bilevel optimization, but they consider non-federated problems. Consequently, they do not address the combined effects of lower-level constraints, heterogeneous client data, client drift, and partial participation arising in our federated setting.

Constrained federated optimization. Single-level federated methods also study constrained optimization. FedDualAvg Yuan et al. (2021) addresses convex composite optimization, where constraints can be represented through convex regularizers, while FedDA Li et al. (2024b) studies federated constrained optimization over a convex feasible set. Their existing guarantees do not directly provide the inexact lower-level solution bounds required for our constrained FBO problem with an upper-level-dependent feasible set. Moreover, FedDA assumes bounded gradient dissimilarity across clients.

Our outer bilevel analysis requires a controlled approximation to the unique solution of the strongly convex lower-level problem of the form $\mathbb{E}\left[\|y^{\epsilon_r}(x) - y(x)\|^2\right] \leq \epsilon_r$, where the error must be related to the lower-level communication rounds, local updates, and client participation. We therefore develop Adaptive Penalty SCAFFOLD, which combines an increasing penalty parameter, round-dependent local stepsizes, and SCAFFOLD control variates to obtain such inexact lower-level solutions under heterogeneous data and partial client participation without BGD-type assumptions.

Personalized federated optimization. Personalized federated optimization addresses statistical heterogeneity by allowing each client to learn a model adapted to its local data distribution. Regularization-based methods such as pFedMe Dinh et al. (2020) and Ditto Li et al. (2021) introduce personalized objectives to balance global and local models. Per-FedAvg Fallah et al. (2020) uses meta-learning to learn a shared initialization for fast client adaptation. Model-personalization methods, including FedPer Arivazhagan et al. (2019), APFL Deng et al. (2020), and FedRep Collins et al. (2021), separate shared and client-specific components. These methods improve client-level performance under non-IID data, but they are primarily formulated as single-level optimization problems.

More recent works study personalized federated learning through bilevel formulations. Graph-aided or collaboration-aware methods use bilevel optimization to learn client relationships and personalized models Xing et al. (2022). CoBo Hashemi et al. (2024) formulates collaborative learning and client selection through nested optimization. Other methods combine client clustering with personalized classifier adaptation through bilevel structures Zhang et al. (2025). These approaches explicitly model the dependence between global coordination and client-specific adaptation, but they typically rely on gradient-based or hypergradient-based updates and focus on smooth, unconstrained settings. Our personalized formulation instead accommodates client-specific constrained lower-level problems within a zeroth-order bilevel framework.

A.2 GAP IN THE LITERATURE

Despite these advances, several limitations remain in existing approaches.

(i) Dependence on hypergradient computation. Most existing FBO methods rely on first- or second-order information to estimate hypergradients, which increases computational requirements and limits applicability when such information is unavailable or expensive to compute. These methods also commonly rely on smoothness assumptions for the implicit upper-level objective.

(ii) Unconstrained lower-level formulations. Most existing FBO methods assume unconstrained lower-level problems. Although recent constrained bilevel methods address lower-level constraints, they primarily consider non-federated settings. Existing single-level constrained FL methods, in turn, do not directly provide the inexact lower-level solution guarantees required by our outer bilevel analysis.

(iii) Limitations of existing zeroth-order FBO methods. Existing zeroth-order FBO methods remain limited to unconstrained lower-level formulations and rely on BGD-type assumptions to control client heterogeneity. Our framework instead combines zeroth-order upper-level updates, constrained federated lower-level optimization, and client-drift correction without BGD-type assumptions.

A.3 NOTATION

Throughout the paper, $\|\cdot\|$ denotes the ℓ_2 norm, $\|\cdot\|_1$ denotes the ℓ_1 norm, and $\langle x, y\rangle$ denotes the Euclidean inner product for vectors $x, y \in \mathbb{R}^n$. We use $\Pi_Y(x)$ to denote the Euclidean projection of a point x onto a nonempty, closed, and convex set Y. The projection operator is nonexpansive, i.e., $\|\Pi_Y(x) - \Pi_Y(y)\| \leq \|x - y\|$, for all $x, y \in \mathbb{R}^n$. We denote by $\mathbb{B}$ and $\mathbb{S}$ the n-dimensional unit ball and the unit sphere, respectively, defined as $\mathbb{B} := \{u \in \mathbb{R}^n \mid \|u\| \leq 1\}$ and $\mathbb{S} := \{v \in \mathbb{R}^n \mid \|v\| = 1\}$. Throughout, $\mathbb{E}[\cdot]$ denotes the expectation operator, and $\mathcal{O}(\cdot)$ denotes the standard big-$\mathcal{O}$ notation. A function $f : Y \subseteq \mathbb{R}^n \to \mathbb{R}$ is said to be L_f-Lipschitz continuous on Y if $|f(x) - f(y)| \leq L_f\|x - y\|$ for all $x, y \in Y$. A continuously differentiable function f is μ-strongly convex on Y if $f(y) \geq f(x) + \langle \nabla f(x), (y - x)\rangle + \frac{\mu}{2}\|y - x\|^2$, for all $x, y \in Y$, and L-smooth if $f(y) \leq f(x) + \langle \nabla f(x), (y - x)\rangle + \frac{L}{2}\|y - x\|^2$, for all $x, y \in Y$. $\mathrm{conv}(\cdot)$ denotes the convex hull. For $\eta > 0$, $\partial_\eta f(x)$ denotes the η-Clarke subdifferential of f at x, and for $X \subseteq \mathbb{R}^n$, $\mathrm{dist}(x, X) := \inf_{y \in X} \|x - y\|$ denotes the distance from x to X. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\|\mathbf{A}\|_2$ denotes the spectral norm, defined as $\|\mathbf{A}\|_2 \triangleq \sigma_{\max}(\mathbf{A}) = \sqrt{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}$, where $\sigma_{\max}(\mathbf{A})$ denotes the largest singular value of $\mathbf{A}$. Let I denote the identity mapping on $\mathbb{R}^{\tilde{n}}$. For $\eta > 0$, we let $\partial_\eta f(x)$ denote the η-Clarke subdifferential of f at x, defined as $\partial_\eta f(x) := \mathrm{conv}\big(\bigcup_{y\in x+\eta\mathbb{B}} \partial f(y)\big)$, where $\partial f(y)$ denotes the Clarke subdifferential of f at y and $\mathbb{B}$ is the unit Euclidean ball. Let $\mathrm{Unif}(\mathcal{A})$ denote the uniform distribution over the set $\mathcal{A}$.

A.4 PRELIMINARIES

Lemma A.1. [Randomized spherical smoothing (Qiu et al., 2023, Lemma 1)] Let $h : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and define the smoothed function $h^\eta(x) := \mathbb{E}_{u\in\mathbb{B}}[h(x + \eta u)]$, where $\mathbb{B}$ is the n-dimensional unit ball. Then the following properties hold.

(i) The function h^η is continuously differentiable. Let $\mathbb{S}$ denote the unit sphere. For every $x \in \mathbb{R}^n$, its gradient is given by $\nabla h^\eta(x) = \frac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\frac{h(x+v)}{\|v\|}v\right] = \frac{n}{2\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\frac{h(x+v)-h(x-v)}{\|v\|}v\right]$.

(ii) Suppose h is Lipschitz continuous with parameter $L_0 > 0$. Then, the following statements hold.

(ii.1) h^η is Lipschitz with the same parameter L_0.

(ii.2) The smoothed function approximates h uniformly with an error of at most $L_0\eta$. That is $|h^\eta(x) - h(x)| \leq L_0\eta$, for all $x \in \mathbb{R}^n$.

(ii.3) The gradient of h^η is Lipschitz with constant $\frac{L_0\sqrt{n}}{\eta}$.

Lemma A.2. [Lévy concentration on sphere $\mathbb{S}$ (Wainwright, 2019, Proposition 3.11 and Example 3.12)] Let $h : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz continuous function and v be uniformly distributed on the sphere $\mathbb{S}$. Then, for all $\epsilon > 0$, we have $\mathbb{P}\left(|h(v) - \mathbb{E}[h(v)]| \geq \epsilon\right) \leq 2\sqrt{2\pi}e^{-\frac{n\epsilon^2}{8L^2}}$.

Lemma A.3. Let x be a random vector in $\mathbb{R}^n$ and $\mathbb{E}\left[\|x\|^2\right] \leq q$. Then, $\mathbb{E}\left[\|x - \mathbb{E}\left[x\right]\|^2\right] \leq q$.

Proof. We have $\|x - \mathbb{E}\left[x\right]\|^2 = \|x\|^2 + \|\mathbb{E}\left[x\right]\|^2 - 2\langle x, \mathbb{E}\left[x\right]\rangle$. Then, taking expectations on both sides, we obtain

$$\mathbb{E}\left[\|x - \mathbb{E}\left[x\right]\|^2\right] = \mathbb{E}\left[\|x\|^2\right] + \|\mathbb{E}\left[x\right]\|^2 - 2\langle \mathbb{E}\left[x\right], \mathbb{E}\left[x\right]\rangle = \mathbb{E}\left[\|x\|^2\right] - \|\mathbb{E}\left[x\right]\|^2 \leq \mathbb{E}\left[\|x\|^2\right] \leq q.$$

□

Lemma A.4. Let $x_1, x_2 \in \mathbb{R}^n$. Then, we have for any $t > 0$,

$$\|x_1 + x_2\|^2 \leq (1+t)\|x_1\|^2 + (1+\tfrac{1}{t})\|x_2\|^2.$$

Lemma A.5. Let $x_1, \ldots, x_T \in \mathbb{R}^n$. Then, we have

$$\|\sum_{i=1}^T x_i\|^2 \leq T \sum_{i=1}^T \|x_i\|^2.$$

Proof. By applying the triangle inequality, we have $\|\sum_{i=1}^T x_i\| \leq \sum_{i=1}^T \|x_i\|$. Now, let us consider $\sum_{i=1}^T \|x_i\| = \sum_{i=1}^T \|x_i\| \times 1 = \langle \tilde{x}, \mathbf{1}\rangle$, where $\tilde{x} = [\|x_1\|, \cdots, \|x_T\|] \in \mathbb{R}^T$ and $\mathbf{1} \in \mathbb{R}^T$. Then, by invoking the Cauchy–Schwarz inequality, we have $\langle \tilde{x}, \mathbf{1}\rangle \leq \|\tilde{x}\|\,\|\mathbf{1}\| = \|\tilde{x}\|\,\sqrt{T}$. Therefore, we have $\|\sum_{i=1}^T x_i\| \leq \sum_{i=1}^T \|x_i\| = \langle \tilde{x}, \mathbf{1}\rangle \leq \sqrt{T}\|\tilde{x}\|$. As a result, we obtain

$$\|\sum_{i=1}^T x_i\| \leq \sqrt{T}\|\tilde{x}\| = \sqrt{T}\sqrt{\sum_{i=1}^T \|x_i\|^2}.$$

By taking the square of both sides, we obtain $\|\sum_{i=1}^T x_i\|^2 \leq T\sum_{i=1}^T \|x_i\|^2$. □

Remark A.6. Note that in Algorithm 1, the variables $z_{i,k}^r$ and c_i^r are updated only for the clients selected in round r, i.e., for $i \in \mathcal{S}^r$. Therefore, for the purpose of analysis, we define for all $i \notin \mathcal{S}^r$ that $c_i^r = c_i^{r-1}$ and $z_{i,k}^r = z_{i,k}^{r-1}$ for all $k \in \{1, \ldots, K\}$ and $r \geq 1$.

Definition A.7. [Filtration] Consider Algorithm 1. We define a filtration $\{\mathcal{F}_{r,k}\}_{r,k\geq 0}$ such that $\mathcal{F}_{0,0} = \mathcal{F}_{0,K} = \emptyset$. For each $r \geq 1$, define $\mathcal{F}_{r,0} = \mathcal{F}_{r-1,K} \cup \{v^r\} \cup \mathcal{S}^r \cup \mathcal{F}_r^2$, and for $r \geq 1$ and $k \geq 1$, define $\mathcal{F}_{r,k} = \mathcal{F}_{r,k-1} \cup \left(\cup_{i\in\mathcal{S}^r} \cup_{j=1}^k \{\xi_{i,j-1}^r\}\right)$, where $\mathcal{S}^r$ defines the random set of selected clients at round r, $\xi_{i,k-1}^r$ denotes the random data set that client i uses at round r and iteration k, $\mathcal{F}_r^2$ denotes the sigma-field generated by all random variables arising from the two calls to the lower-level algorithm at round r, and v^r denotes the random direction sampled at round r and reused throughout all K local iterations.

Definition A.8. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function, and let $\partial f(x)$ denote the Clarke subdifferential of f at x Kornowski and Shamir (2024). A point $x \in \mathbb{R}^n$ is called an ϵ-Clarke stationary point of f if $\mathrm{dist}(0, \partial f(x)) \leq \epsilon$.

The three following lemmas provide some feasible sets that satisfies Assumption 2.1 (iii).

Lemma A.9. Let $x \in \mathbb{R}^n$, where α denotes one component of x, and consider $Y(x) := \left\{y \in \mathbb{R}^{\tilde{n}} : \|y\|_1 \leq \log(1+\exp(\alpha))\right\}$. Then $Y(x)$ is nonempty, closed, and convex for every $x \in \mathbb{R}^n$. Moreover, for any $x_1, x_2 \in \mathbb{R}^n$ and $z \in \mathbb{R}^{\tilde{n}}$, we have

$$\left\|\Pi_{Y(x_1)}[z] - \Pi_{Y(x_2)}[z]\right\| \leq \|x_1 - x_2\|.$$

Consequently, Assumption 2.1 (iii) is satisfied with $L_Y = 1$.

Proof. Since $\log(1+\exp(\alpha)) > 0$ for every $\alpha \in \mathbb{R}$, we have $0 \in Y(x)$, and hence $Y(x)$ is nonempty.

To show that $Y(x)$ is closed, let $\{y^j\}$ be a sequence in $Y(x)$ such that $y^j \to y$. Since the ℓ_1-norm is continuous, we have $\|y\|_1 = \lim_{j\to\infty} \|y^j\|_1 \leq \log(1+\exp(\alpha))$. Therefore, $y \in Y(x)$, and hence $Y(x)$ is closed.

To show that $Y(x)$ is convex, let $y_1, y_2 \in Y(x)$ and $\theta \in [0,1]$. Using the triangle inequality and the positive homogeneity of the ℓ_1-norm, we obtain

$$\begin{aligned}\|\theta y_1 + (1-\theta)y_2\|_1 &\leq \theta\|y_1\|_1 + (1-\theta)\|y_2\|_1 \leq \theta\log(1+\exp(\alpha)) + (1-\theta)\log(1+\exp(\alpha)) \\ &= \log(1+\exp(\alpha)).\end{aligned}$$

Therefore, $\theta y_1 + (1-\theta)y_2 \in Y(x)$, and hence $Y(x)$ is convex.

Next, let α_1 and α_2 denote the components of x_1 and x_2 corresponding to α, respectively. Without loss of generality, suppose that $\alpha_1 \le \alpha_2$. Since $\log(1+\exp(\alpha))$ is increasing in α, we have

$$\log(1+\exp(\alpha_1)) \le \log(1+\exp(\alpha_2)),$$

and consequently, $Y(x_1) \subseteq Y(x_2)$.

For $j \in \{1,2\}$, if $z \in Y(x_j)$, then $\Pi_{Y(x_j)}[z] = z$. Otherwise, the projection lies on the boundary of $Y(x_j)$. Therefore,

$$\left\|\Pi_{Y(x_j)}[z]\right\|_1 = \min\left\{\|z\|_1, \log(1+\exp(\alpha_j))\right\}. \tag{3}$$

Moreover, for Euclidean projections onto the nested ℓ_1-balls $Y(x_1) \subseteq Y(x_2)$ centered at the origin, the corresponding projected components have the same sign and satisfy

$$\left|\left[\Pi_{Y(x_1)}[z]\right]_\ell\right| \le \left|\left[\Pi_{Y(x_2)}[z]\right]_\ell\right|, \qquad \forall \ell \in [\tilde{n}].$$

Therefore, for every $\ell \in [\tilde{n}]$,

$$\left|\left[\Pi_{Y(x_2)}[z]\right]_\ell - \left[\Pi_{Y(x_1)}[z]\right]_\ell\right| = \left|\left[\Pi_{Y(x_2)}[z]\right]_\ell\right| - \left|\left[\Pi_{Y(x_1)}[z]\right]_\ell\right|.$$

Summing the preceding equality over $\ell \in [\tilde{n}]$, we obtain

$$\left\|\Pi_{Y(x_2)}[z] - \Pi_{Y(x_1)}[z]\right\|_1 = \left\|\Pi_{Y(x_2)}[z]\right\|_1 - \left\|\Pi_{Y(x_1)}[z]\right\|_1.$$

Then, using equation 3, we obtain

$$\begin{aligned}\left\|\Pi_{Y(x_2)}[z] - \Pi_{Y(x_1)}[z]\right\|_1 &= \min\left\{\|z\|_1, \log(1+\exp(\alpha_2))\right\} - \min\left\{\|z\|_1, \log(1+\exp(\alpha_1))\right\} \\ &\le \log(1+\exp(\alpha_2)) - \log(1+\exp(\alpha_1)),\end{aligned}$$

where the last inequality follows from $\min\{a, b_2\} - \min\{a, b_1\} \le b_2 - b_1$ for $b_1 \le b_2$.

Since $\|v\| \le \|v\|_1$ for every $v \in \mathbb{R}^{\tilde{n}}$, we further obtain

$$\left\|\Pi_{Y(x_2)}[z] - \Pi_{Y(x_1)}[z]\right\| \le \left\|\Pi_{Y(x_2)}[z] - \Pi_{Y(x_1)}[z]\right\|_1 \le \log(1+\exp(\alpha_2)) - \log(1+\exp(\alpha_1)). \tag{4}$$

Finally, since $\frac{d}{d\alpha}\log(1+\exp(\alpha)) = \frac{\exp(\alpha)}{1+\exp(\alpha)} \le 1$ for every $\alpha \in \mathbb{R}$, the mean value theorem yields

$$\log(1+\exp(\alpha_2)) - \log(1+\exp(\alpha_1)) \le |\alpha_2 - \alpha_1|.$$

Moreover, since α is one component of x, we have

$$|\alpha_2 - \alpha_1| \le \|x_2 - x_1\|.$$

Combining the preceding inequalities with equation 4, we obtain

$$\left\|\Pi_{Y(x_1)}[z] - \Pi_{Y(x_2)}[z]\right\| \le \|x_1 - x_2\|.$$

Thus, Assumption 2.1 (iii) holds with $L_Y = 1$. □

Lemma A.10. Consider the set $Y(x) = Y := \left\{y \in \mathbb{R}^C \mid y_c \ge 0,\ c \in [C],\ \sum_{c=1}^C y_c = 1\right\},\ x \in \mathbb{R}^n$. Then $Y(x)$ is nonempty, closed, and convex for every $x \in \mathbb{R}^n$. Moreover, for every $x_1, x_2 \in \mathbb{R}^n$ and $z \in \mathbb{R}^C$,

$$\left\|\Pi_{Y(x_1)}(z) - \Pi_{Y(x_2)}(z)\right\| = 0.$$

Consequently, Assumption 2.1 (iii) holds with $L_Y = 0$.

Proof. Since the vector $\frac{1}{C}\mathbf{1} = \left(\frac{1}{C}, \ldots, \frac{1}{C}\right)$ belongs to Y, the set $Y(x) = Y$ is nonempty for every $x \in \mathbb{R}^n$. To show closedness, let $\{y^j\}$ be a sequence in Y such that $y^j \to y$. Since $y_c^j \ge 0$ for every $c \in [C]$ and every j, taking the limit gives $y_c \ge 0,\ c \in [C]$. Moreover, by continuity, we have $\sum_{c=1}^C y_c = \lim_{j\to\infty}\sum_{c=1}^C y_c^j = 1$. Hence, $y \in Y$, and therefore Y is closed.

Next, let $y^1, y^2 \in Y$ and $\theta \in [0,1]$. For every $c \in [C]$, we have $\theta y_c^1 + (1-\theta) y_c^2 \geq 0$, and

$$\textstyle\sum_{c=1}^{C} \left(\theta y_c^1 + (1-\theta) y_c^2\right) = \theta \sum_{c=1}^{C} y_c^1 + (1-\theta) \sum_{c=1}^{C} y_c^2 = 1.$$

Thus, $\theta y^1 + (1-\theta) y^2 \in Y$, showing that Y is convex.

Finally, since $Y(x)$ is independent of x, for any $x_1, x_2 \in \mathbb{R}^n$, we have $Y(x_1) = Y(x_2) = Y$. Therefore, for every $z \in \mathbb{R}^C$,

$$\Pi_{Y(x_1)}(z) = \Pi_Y(z) = \Pi_{Y(x_2)}(z),$$

and hence

$$\left\|\Pi_{Y(x_1)}(z) - \Pi_{Y(x_2)}(z)\right\| = 0.$$

Thus, the projection mapping satisfies the required Lipschitz condition with $L_Y = 0$, and Assumption 2.1 (iii) holds. □

Lemma A.11. Let $Y_0 \subseteq \mathbb{R}^{\tilde{n}}$ be a nonempty, closed, and convex set, and consider the following feasible set $Y(x) \triangleq Y_0 + \mathbf{A}x + b$, where $\mathbf{A} \in \mathbb{R}^{\tilde{n} \times n}$ and $b \in \mathbb{R}^{\tilde{n}}$. Then, for any $x_1, x_2 \in \mathbb{R}^n$ and any $z \in \mathbb{R}^{\tilde{n}}$,

$$\|\Pi_{Y(x_1)}[z] - \Pi_{Y(x_2)}[z]\| \leq \|\mathbf{A}\|_2 \|x_1 - x_2\|,$$

where $\|\mathbf{A}\|_2$ denotes the spectral norm of $\mathbf{A}$. Consequently, Assumption 2.1 (iii) is satisfied with $L_Y = \|\mathbf{A}\|_2$.

Proof. Let $\mathbb{I}_{Y(x)}$ and $\mathbb{I}_{Y_0}$ denote the indicator functions of $Y(x)$ and Y_0, respectively. Since $Y(x) = Y_0 + \mathbf{A}x + b$, we have

$$\mathbb{I}_{Y(x)}(y) = \mathbb{I}_{Y_0}(y - \mathbf{A}x - b).$$

To apply (Beck, 2017, Theorem 6.11), for any fixed $x \in \mathbb{R}^n$, let $g(y) \triangleq \mathbb{I}_{Y_0}(y)$, $\lambda = 1$, and $a = -\mathbf{A}x - b$. Then, $f(y) = g(y - \mathbf{A}x - b) = \mathbb{I}_{Y_0}(y - \mathbf{A}x - b) = \mathbb{I}_{Y(x)}(y)$. Therefore, (Beck, 2017, Theorem 6.11) yields

$$\operatorname{prox}_{\mathbb{I}_{Y(x)}}(z) = \mathbf{A}x + b + \operatorname{prox}_{\mathbb{I}_{Y_0}}(z - \mathbf{A}x - b).$$

Moreover, the proximal mapping of the indicator function of a nonempty, closed, and convex set coincides with its Euclidean projection (Beck, 2017, Theorem 6.24). Hence,

$$\Pi_{Y(x)}[z] = \mathbf{A}x + b + \Pi_{Y_0}[z - \mathbf{A}x - b].$$

That is,

$$\Pi_{Y(x)}[z] = z - (I - \Pi_{Y_0})[z - \mathbf{A}x - b].$$

Therefore, applying the firm nonexpansiveness property in (Beck, 2017, Theorem 6.42 (a)), for any $u, v \in \mathbb{R}^{\tilde{n}}$, we have

$$\langle u - v, \Pi_{Y_0}[u] - \Pi_{Y_0}[v] \rangle \geq \|\Pi_{Y_0}[u] - \Pi_{Y_0}[v]\|^2.$$

Using the preceding inequality, we obtain

$$\begin{aligned}
\|(I - \Pi_{Y_0})[u] - (I - \Pi_{Y_0})[v]\|^2 &= \|(u - v) - (\Pi_{Y_0}[u] - \Pi_{Y_0}[v])\|^2 \\
&= \|u - v\|^2 + \|\Pi_{Y_0}[u] - \Pi_{Y_0}[v]\|^2 - 2\langle u - v, \Pi_{Y_0}[u] - \Pi_{Y_0}[v] \rangle \\
&\leq \|u - v\|^2 - \langle u - v, \Pi_{Y_0}[u] - \Pi_{Y_0}[v] \rangle \\
&= \langle (I - \Pi_{Y_0})[u] - (I - \Pi_{Y_0})[v], u - v \rangle.
\end{aligned}$$

Thus, $I - \Pi_{Y_0}$ is firmly nonexpansive and, consequently, nonexpansive (Beck, 2017, Theorem 6.42). Therefore, for any $x_1, x_2 \in \mathbb{R}^n$ and $z \in \mathbb{R}^{\tilde{n}}$,

$$\begin{aligned}
\|\Pi_{Y(x_1)}[z] - \Pi_{Y(x_2)}[z]\| &= \|(I - \Pi_{Y_0})[z - \mathbf{A}x_2 - b] - (I - \Pi_{Y_0})[z - \mathbf{A}x_1 - b]\| \\
&\leq \|(z - \mathbf{A}x_2 - b) - (z - \mathbf{A}x_1 - b)\| = \|\mathbf{A}(x_1 - x_2)\| \leq \|\mathbf{A}\|_2 \|x_1 - x_2\|,
\end{aligned}$$

where the last inequality follows from the spectral norm property in (Beck, 2017, Section 1.8.2). Thus, Assumption 2.1 (iii) is satisfied with $L_Y = \|\mathbf{A}\|_2$. □

Remark A.12. In Table 3, we show several commonly used constraint sets that can be represented in the form $Y(x) = Y_0 + \mathrm{A}x + b$ and are therefore covered by Lemma A.11.

Table 3: Examples of constraint sets covered by Lemma A.11.

Constraint	Constraint sets	Condition
Coupled linear	$Y(x) = \{y : Cx + By \le c\}$ $Y_0 = \{z : Bz \le c\}$	Here, $b = 0$, and there exists A such that $B\mathsf{A} = -C$, equivalently, $\mathrm{Range}(C) \subseteq \mathrm{Range}(B)$. If B has full row rank, one may take $\mathsf{A} = -B^{\dagger}C$.
Ball	$Y(x) = \{y : \|y - (\mathsf{A}x + b)\| \le r\}$ $Y_0 = \{z : \|z\| \le r\}$	The radius r is fixed, while the center $\mathsf{A}x + b$ depends affinely on x.
Box	$Y(x) = \{y : \ell + \mathsf{A}x + b \le y \le u + \mathsf{A}x + b\}$ $Y_0 = \{z : \ell \le z \le u\}$	The lower and upper bounds undergo the same affine shift $\mathsf{A}x + b$, while the size of the box remains fixed.
Ellipsoid	$Y(x) = \{y : (y - \mathsf{A}x - b)^{\top} Q (y - \mathsf{A}x - b) \le r^2\}$ $Y_0 = \{z : z^{\top} Q z \le r^2\}$	The matrix $Q \succ 0$ and radius $r > 0$ are fixed, while the center $\mathsf{A}x + b$ depends affinely on x.

In the following lemma, we show that $\sup_x \left\|\nabla_y h\big(x, y(x)\big)\right\| < \infty$, and hence Assumption 2.2 holds for our first numerical experiment in problem equation 1.

Lemma A.13. For the problem equation 1, Assumption 2.2 holds. In particular, for every $x = [\lambda; \alpha] \in \mathbb{R}^{m+1}$ and $y(x)$ solving the lower-level problem,

$$\|\nabla_y h(x, y(x))\| \le \tfrac{2}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} \|\tilde{U}_{i,\ell}\| < \infty.$$

Proof. From the definition of the lower-level objective, we have

$$\nabla_y h(x, y(x)) = \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} -\frac{\tilde{v}_{i,\ell} \tilde{U}_{i,\ell}}{1 + \exp(\tilde{v}_{i,\ell} \langle \tilde{U}_{i,\ell}, y(x) \rangle)} + \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log(1 + \exp(\lambda_i)) y(x).$$

Therefore,

$$\|\nabla_y h(x, y(x))\| \le \| \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} -\frac{\tilde{v}_{i,\ell} \tilde{U}_{i,\ell}}{1 + \exp(\tilde{v}_{i,\ell} \langle \tilde{U}_{i,\ell}, y(x) \rangle)} \| + \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log(1 + \exp(\lambda_i)) \|y(x)\|. \tag{5}$$

Since $\tilde{v}_{i,\ell} \in \{-1, 1\}$ and $0 < \frac{1}{1 + \exp(\tilde{v}_{i,\ell} \langle \tilde{U}_{i,\ell}, y(x) \rangle)} < 1$, the first term in equation 5 satisfies

$$\| \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} -\frac{\tilde{v}_{i,\ell} \tilde{U}_{i,\ell}}{1 + \exp(\tilde{v}_{i,\ell} \langle \tilde{U}_{i,\ell}, y(x) \rangle)} \| \le \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} \|\tilde{U}_{i,\ell}\|. \tag{6}$$

Moreover, since $Y(x)$ is convex and $0 \in Y(x)$, the first-order optimality condition for $y(x)$, evaluated at 0, gives

$$\langle \nabla_y h(x, y(x)), 0 - y(x) \rangle = -\langle \nabla_y h(x, y(x)), y(x) \rangle \ge 0. \tag{7}$$

Substituting the expression for $\nabla_y h(x, y(x))$ into equation 7, we obtain

$$0 \le -\langle \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} -\frac{\tilde{v}_{i,\ell} \tilde{U}_{i,\ell}}{1 + \exp(\tilde{v}_{i,\ell} \langle \tilde{U}_{i,\ell}, y(x) \rangle)}, y(x) \rangle - \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log(1 + \exp(\lambda_i)) \|y(x)\|^2.$$

Rearranging gives

$$\tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log(1 + \exp(\lambda_i)) \|y(x)\|^2 \le -\langle \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} -\frac{\tilde{v}_{i,\ell} \tilde{U}_{i,\ell}}{1 + \exp(\tilde{v}_{i,\ell} \langle \tilde{U}_{i,\ell}, y(x) \rangle)}, y(x) \rangle.$$

By the Cauchy–Schwarz inequality and equation 6, we obtain

$$\tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log(1 + \exp(\lambda_i)) \|y(x)\|^2 \le \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} \|\tilde{U}_{i,\ell}\| \|y(x)\|.$$

If $y(x) \neq 0$, dividing the preceding inequality by $\|y(x)\|$ gives

$$\tfrac{1}{m} \textstyle\sum_{i=1}^{m} \log(1 + \exp(\lambda_i)) \|y(x)\| \le \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} \|\tilde{U}_{i,\ell}\|. \tag{8}$$

If $y(x) = 0$, equation 8 holds immediately since its left-hand side is zero. Hence, equation 8 holds for every $y(x)$.

Finally, substituting equation 6 and equation 8 into equation 5 gives

$$\|\nabla_y h(x, y(x))\| \le \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} \|\tilde{U}_{i,\ell}\| + \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} \|\tilde{U}_{i,\ell}\| = \tfrac{2}{m} \textstyle\sum_{i=1}^{m} \sum_{\ell \in \tilde{I}_i} \|\tilde{U}_{i,\ell}\| < \infty.$$

□

Lemma A.14. Let Assumption 2.1 hold. Then, the following statements hold.

(i) The mapping $y(x)$ is $L^y_{0,x}$-Lipschitz continuous.

(ii) The random implicit function $\tilde{f}_i(x, y(x), \xi_i)$ is Lipschitz continuous with parameter $L_0^{\text{imp}}(\xi_i) := L^f_{0,y}(\xi_i)L^y_{0,x} + L^f_{0,x}(\xi_i)$, almost surely.

(iii) The implicit function $f(x)$ is Lipschitz continuous with parameter $L_0^{\text{imp}} := L^f_{0,y}L^y_{0,x} + L^f_{0,x}$.

Proof. (i) Let us define the mapping $G(x, y) = \Pi_{Y(x)}[y - \rho\nabla_y h(x, y)]$, where $y \in Y(x)$ and $x \in \mathbb{R}^n$. By (Dafermos, 1988, Lemma 2.2), for all $y_1, y_2 \in Y(x)$, there exists $\beta < 1$ such that

$$\|G(x, y_1) - G(x, y_2)\| \leq \beta\|y_1 - y_2\|.$$

Since $y(x)$ is the unique solution of the lower-level problem (Dafermos, 1988, Lemma 2.1), we have

$$y(x) = G(x, y(x)) = \Pi_{Y(x)}[y(x) - \rho\nabla_y h(x, y(x))].$$

Then, for any $x_1, x_2 \in \mathbb{R}^n$, we have

$$\begin{aligned}\|y(x_1) - y(x_2)\| &= \|G(x_1, y(x_1)) - G(x_2, y(x_2))\| \\ &\leq \underbrace{\|G(x_1, y(x_1)) - G(x_1, y(x_2))\|}_{\leq\beta\|y(x_1)-y(x_2)\|} + \underbrace{\|G(x_1, y(x_2)) - G(x_2, y(x_2))\|}_{\text{Term 1}}.\end{aligned}$$

Now consider Term 1 in the preceding inequality. Using the nonexpansiveness of projection, we have

$$\begin{aligned}&\|G(x_1, y(x_2)) - G(x_2, y(x_2))\| \\ &= \|\Pi_{Y(x_1)}[y(x_2) - \rho\nabla_y h(x_1, y(x_2))] - \Pi_{Y(x_2)}[y(x_2) - \rho\nabla_y h(x_2, y(x_2))]\| \\ &= \|\Pi_{Y(x_1)}[y(x_2) - \rho\nabla_y h(x_1, y(x_2))] - \Pi_{Y(x_1)}[y(x_2) - \rho\nabla_y h(x_2, y(x_2))] \\ &+ \Pi_{Y(x_1)}[y(x_2) - \rho\nabla_y h(x_2, y(x_2))] - \Pi_{Y(x_2)}[y(x_2) - \rho\nabla_y h(x_2, y(x_2))]\| \\ &\leq \underbrace{\|\Pi_{Y(x_1)}[y(x_2) - \rho\nabla_y h(x_1, y(x_2))] - \Pi_{Y(x_1)}[y(x_2) - \rho\nabla_y h(x_2, y(x_2))]\|}_{\leq\rho\underbrace{\|\nabla_y h(x_1, y(x_2)) - \nabla_y h(x_2, y(x_2))\|}_{\leq L^h_{1,x}\|x_1 - x_2\|}} \\ &+ \|\Pi_{Y(x_1)}[y(x_2) - \rho\nabla_y h(x_2, y(x_2))] - \Pi_{Y(x_2)}[y(x_2) - \rho\nabla_y h(x_2, y(x_2))]\|.\end{aligned}$$

Therefore, we obtain

$$\|y(x_1) - y(x_2)\| \leq \tfrac{\rho}{1-\beta}L^h_{1,x}\|x_1 - x_2\| + \tfrac{1}{1-\beta}\|\Pi_{Y(x_1)}[p] - \Pi_{Y(x_2)}[p]\|,$$

where $p = y(x_2) - \rho\nabla_y h(x_2, y(x_2))$. Using Assumption 2.1 (iii), we have $\|\Pi_{Y(x_1)}[\bullet] - \Pi_{Y(x_2)}[\bullet]\| \leq L_Y\|x_1 - x_2\|$. Then, we obtain

$$\begin{aligned}\|y(x_1) - y(x_2)\| &\leq \tfrac{\rho}{1-\beta}L^h_{1,x}\|x_1 - x_2\| + \tfrac{1}{1-\beta}\underbrace{\|\Pi_{Y(x_1)}[p] - \Pi_{Y(x_2)}[p]\|}_{\leq L_Y\|x_1 - x_2\|} \\ &\leq \underbrace{\left(\tfrac{\rho}{1-\beta}L^h_{1,x} + \tfrac{1}{1-\beta}L_Y\right)}_{=L^y_{0,x}}\|x_1 - x_2\|.\end{aligned}$$

(ii) We have

$$\begin{aligned}&|\tilde{f}_i(x_1, y(x_1), \xi_i) - \tilde{f}_i(x_2, y(x_2), \xi_i)| \\ &= |\tilde{f}_i(x_1, y(x_1), \xi_i) - \tilde{f}_i(x_1, y(x_2), \xi_i) + \tilde{f}_i(x_1, y(x_2), \xi_i) - \tilde{f}_i(x_2, y(x_2), \xi_i)| \\ &\leq |\tilde{f}_i(x_1, y(x_1), \xi_i) - \tilde{f}_i(x_1, y(x_2), \xi_i)| + |\tilde{f}_i(x_1, y(x_2), \xi_i) - \tilde{f}_i(x_2, y(x_2), \xi_i)| \\ &\leq L^f_{0,y}(\xi_i)\|y(x_1) - y(x_2)\| + L^f_{0,x}(\xi_i)\|x_1 - x_2\| \leq \underbrace{\left(L^f_{0,y}(\xi_i)L_y + L^f_{0,x}(\xi_i)\right)}_{=L_0^{\text{imp}}(\xi_i)}\|x_1 - x_2\|.\end{aligned}$$

(iii) First, we show that for any $x_1, x_2 \in \mathbb{R}^n$, and $y \in \mathbb{R}^{\tilde{n}}$, we have $|f_i(x_1, y) - f_i(x_2, y)| \leq L^f_{0,x}\|x_1 - x_2\|$. From Assumption 2.1 (i), we have

$$|f_i(x_1,y) - f_i(x_2,y)| = |\mathbb{E}_{\xi_i}\left[\tilde{f}_i(x_1,y,\xi_i) - \tilde{f}_i(x_2,y,\xi_i)\right]| \leq \mathbb{E}_{\xi_i}\left[|\tilde{f}_i(x_1,y,\xi_i) - \tilde{f}_i(x_2,y,\xi_i)|\right]$$
$$\leq \mathbb{E}_{\xi_i}\left[L^f_{0,x}(\xi_i)\right]\|x_1 - x_2\|.$$

Also, we have $L^f_{0,x} \triangleq \max_{i=1,\ldots,m}\sqrt{\mathbb{E}_{\xi_i\sim D_i}[(L^f_{0,x}(\xi_i))^2]} < \infty$ and from Jensen's inequality, we have $\left(\mathbb{E}_{\xi_i}\left[L^f_{0,x}(\xi_i)\right]\right)^2 \leq \mathbb{E}_{\xi_i}\left[(L^f_{0,x}(\xi_i))^2\right]$. Therefore, we have

$$\mathbb{E}_{\xi_i}\left[L^f_{0,x}(\xi_i)\right] \leq \sqrt{\mathbb{E}_{\xi_i}\left[(L^f_{0,x}(\xi_i))^2\right]} \leq L^f_{0,x}.$$

As a result, we obtain $|f_i(x_1, y) - f_i(x_2, y)| \leq L^f_{0,x}\|x_1 - x_2\|$. Similarly, we obtain

$$|f_i(x, y_1) - f_i(x, y_2)| \leq L^f_{0,y}\|y_1 - y_2\|.$$

Then, from the last two inequalities and Lemma A.14 (i), we obtain

$$\begin{aligned}
|f(x_1) - f(x_2)| &= |f(x_1, y(x_1)) - f(x_2, y(x_2))| \\
&= |f(x_1, y(x_1)) - f(x_1, y(x_2)) + f(x_1, y(x_2)) - f(x_2, y(x_2))| \\
&\leq |f(x_1, y(x_1)) - f(x_1, y(x_2))| + |f(x_1, y(x_2)) - f(x_2, y(x_2))| \\
&\leq \tfrac{1}{m}\textstyle\sum_{i=1}^m |f_i(x_1, y(x_1)) - f_i(x_1, y(x_2))| \\
&+ \tfrac{1}{m}\textstyle\sum_{i=1}^m |f_i(x_1, y(x_2)) - f_i(x_2, y(x_2))| \\
&\leq L^f_{0,y}\underbrace{\|y(x_1) - y(x_2)\|}_{\leq L^y_{0,x}\|x_1 - x_2\|} + L^f_{0,x}\|x_1 - x_2\| \leq (L^f_{0,y}L^y_{0,x} + L^f_{0,x})\|x_1 - x_2\|.
\end{aligned}$$

□

We introduce the following definitions used in the convergence rate analysis of Algorithm 1.

Definition A.15. Let us define the following terms for all $i \in [m]$, $r \geq 1$, and $k \in \{1, \ldots, K\}$.

(i) Zeroth-order stochastic gradient of the local implicit objective function of client i at round r and iteration k, is given by

$$g^{\eta,r,z}_{i,k-1} := \tfrac{n}{2\eta^2}\Big(\tilde{f}_i\big(z^r_{i,k-1} + v^r,\, y(z^r_{i,k-1} + v^r),\, \xi^r_{i,k-1}\big) - \tilde{f}_i\big(z^r_{i,k-1} - v^r,\, y(z^r_{i,k-1} - v^r),\, \xi^r_{i,k-1}\big)\Big)v^r.$$

(ii) A variant of $g^{\eta,r,z}_{i,k-1}$ with $y(\bullet)$ given at a delayed update, is given by

$$\hat{g}^{\eta,r,z}_{i,k-1} := \tfrac{n}{2\eta^2}\Big(\tilde{f}_i\big(z^r_{i,k-1} + v^r,\, y(\hat{x}^{r-1} + v^r),\, \xi^r_{i,k-1}\big) - \tilde{f}_i\big(z^r_{i,k-1} - v^r,\, y(\hat{x}^{r-1} - v^r),\, \xi^r_{i,k-1}\big)\Big)v^r.$$

(iii) Inexact variant of $g^{\eta,r,z}_{i,k-1}$ with a given inexact delayed update with an accuracy ϵ_r, is given by

$$\check{g}^{\eta,r,z}_{i,k-1} := \tfrac{n}{2\eta^2}\Big(\tilde{f}_i\big(z^r_{i,k-1} + v^r,\, y_{\epsilon_r}(\hat{x}^{r-1} + v^r),\, \xi^r_{i,k-1}\big) - \tilde{f}_i\big(z^r_{i,k-1} - v^r,\, y_{\epsilon_r}(\hat{x}^{r-1} - v^r),\, \xi^r_{i,k-1}\big)\Big)v^r.$$

(iv) Delayed error is given by $\hat{w}^{\eta,r,z}_{i,k-1} = \hat{g}^{\eta,r,z}_{i,k-1} - g^{\eta,r,z}_{i,k-1}$.

(v) Inexact delayed error is given by $\check{w}^{\eta,r,z}_{i,k-1} = \check{g}^{\eta,r,z}_{i,k-1} - \hat{g}^{\eta,r,z}_{i,k-1}$.

(vi) $\check{g}^{\eta,r,z}_{i,k-1} = g^{\eta,r,z}_{i,k-1} + \hat{w}^{\eta,r,z}_{i,k-1} + \check{w}^{\eta,r,z}_{i,k-1}$.

Definition A.16 (Error in control variate)**.** We define the following $\mathcal{C}_r$ to be the error in the control variate at round $r \geq 0$ as $\mathcal{C}_r := \frac{1}{Km}\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|z^r_{i,k-1} - \hat{x}^r\|^2\right]$. Noting that $z^0_{i,k-1} = \hat{x}^0$, we obtain $\mathcal{C}_0 = 0$.

Definition A.17 (Client drift)**.** Let us define $\mathcal{D}_r$ to measure the client drift caused by local updates for $r \geq 1$ as $\mathcal{D}_r := \frac{1}{Km}\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right]$. Noting that $z_{i,0}^r = \hat{x}^{r-1}$, therefore $\|z_{i,0}^r - \hat{x}^{r-1}\|^2 = 0$. As a result, we have $\mathcal{D}_r = \frac{1}{Km}\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] = \frac{1}{Km}\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\|z_{i,k}^r - \hat{x}^{r-1}\|^2\right]$. For notational convenience, let $\mathcal{D}_0 := 0$.

Lemma A.18. Consider Algorithm 1. We have $c_i^r = \frac{1}{K}\sum\limits_{k=1}^{K}\check{g}_{i,k-1}^{\eta,r,z}$ for $r \geq 1$, $k \geq 1$, and $i \in [m]$.

Proof. Regarding Algorithm 1, for $i \in \mathcal{S}^r$, we have $z_{i,k}^r - z_{i,k-1}^r = -\gamma_l(\check{g}_{i,k-1}^{\eta,r,z} + \hat{c}^{r-1} - c_i^{r-1})$. Taking the sum from $k = 1$ to K, we obtain

$$\begin{aligned} z_{i,K}^r - \underbrace{z_{i,0}^r}_{=\hat{x}^{r-1}} &= -\gamma_l\textstyle\sum_{k=1}^{K}\check{g}_{i,k-1}^{\eta,r,z} - \gamma_l\sum_{k=1}^{K}(\hat{c}^{r-1} - c_i^{r-1}) \\ &\Rightarrow \hat{x}^{r-1} - z_{i,K}^r = \gamma_l\textstyle\sum_{k=1}^{K}\check{g}_{i,k-1}^{\eta,r,z} + \gamma_l K(\hat{c}^{r-1} - c_i^{r-1}). \end{aligned}$$

Now, multiply both sides by $\frac{1}{K\gamma_l}$. Then, we obtain

$$\begin{aligned} \tfrac{1}{K\gamma_l}(\hat{x}^{r-1} - z_{i,K}^r) &= \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\check{g}_{i,k-1}^{\eta,r,z} + (\hat{c}^{r-1} - c_i^{r-1}) \\ &\Rightarrow \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\check{g}_{i,k-1}^{\eta,r,z} = \underbrace{c_i^{r-1} - \hat{c}^{r-1} + \tfrac{1}{K\gamma_l}(\hat{x}^{r-1} - z_{i,K}^r)}_{=c_i^r} \\ &\Rightarrow c_i^r = \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\check{g}_{i,k-1}^{\eta,r,z}. \end{aligned}$$

By Remark A.6, we have $c_i^r = c_i^{r-1}$ for $i \notin \mathcal{S}^r$ and $c_i^0 \in \mathbb{R}^n$. As a result, we conclude $c_i^r = \frac{1}{K}\sum\limits_{k=1}^{K}\check{g}_{i,k-1}^{\eta,r,z}$ for $r \geq 1$, $k \geq 1$, and $i \in [m]$. □

In the next lemma, we define an upper bound for the terms defined in Definition A.15.

Lemma A.19. Consider Definition A.15 and Definition A.7. Let Assumption 2.1 hold. Then, for all $i \in [m]$ and $r \in \{1, \ldots, R\}$, the following statements hold.
(i) For $k = 1$, we have

$$\mathbb{E}\left[g_{i,0}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right] = \nabla f_i^{\eta}(\hat{x}^{r-1}).$$

For $k \in \{2, \ldots, K\}$, we have

$$\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r,k-1}\right] = \nabla f_i^{\eta}(z_{i,k-1}^r) + b_{i,k-1}^r,$$

where the bias term is defined as

$$b_{i,k-1}^r := \tfrac{n}{2\eta^2}\left(f_i\left(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r)\right) - f_i\left(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r)\right)\right)v^r - \nabla f_i^{\eta}(z_{i,k-1}^r). \tag{9}$$

Moreover, for $k \in \{2, \ldots, K\}$,

$$\mathbb{E}\left[\|\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \tfrac{2n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right]. \tag{10}$$

(ii) For $k \in \{1, \ldots, K\}$, we have

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z}\|^2\right] \leq \tfrac{2n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] + 32\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

In particular, for $k = 1$, we have

$$\mathbb{E}\left[\|g_{i,0}^{\eta,r,z}\|^2\right] \leq 16\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

(iii) For $k \in \{1, \ldots, K\}$, we have

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^{\eta}(z_{i,k-1}^r)\|^2\right] \leq \tfrac{3n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] + 48\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

In particular, for $k = 1$, we have

$$\mathbb{E}\left[\|g_{i,0}^{\eta,r,z} - \nabla f_i^{\eta}(\hat{x}^{r-1})\|^2\right] \leq 16\sqrt{2\pi}\, n(L_0^{\text{imp}})^2.$$

(iv) For $k \in \{1, \ldots, K\}$, we have

$$\mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right] \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 \mathbb{E}\left[\|\hat{x}^{r-1} - z_{i,k-1}^r\|^2\right].$$

(v) For $k \in \{1, \ldots, K\}$, we have

$$\mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r,z}\|^2\right] \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f)^2 \epsilon_r.$$

(vi) For $k \in \{2, \ldots, K\}$, we have

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^{\eta}(\hat{x}^{r-1})\|^2\right] \leq \tfrac{2n^2(L_0^{\text{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] + 32\sqrt{2\pi}\, n(L_0^{\text{imp}})^2. \tag{11}$$

Proof. (i) First, consider $k = 1$. We have $z_{i,0}^r = \hat{x}^{r-1}$. Moreover, $\hat{x}^{r-1}$ is measurable with respect to $\mathcal{F}_{r-1,K}$, v^r is independent of $\mathcal{F}_{r-1,K}$, and $\xi_{i,0}^r$ is independent of $\mathcal{F}_{r,0}$, which contains v^r. Therefore,

$$\begin{aligned}
\mathbb{E}\left[g_{i,0}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right] &= \mathbb{E}_{v^r,\xi_{i,0}^r}\Big[\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,0}^r) \\
&\quad - \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,0}^r))v^r \mid \mathcal{F}_{r-1,K}\Big] \\
&= \mathbb{E}_{v^r}\left[\tfrac{n}{2\eta^2}(f_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r)) - f_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r)))v^r \mid \mathcal{F}_{r-1,K}\right] \\
&= \nabla f_i^{\eta}(\hat{x}^{r-1}),
\end{aligned}$$

where the first equality follows from the definition of $g_{i,0}^{\eta,r,z}$ and $z_{i,0}^r = \hat{x}^{r-1}$. The second equality follows by taking the expectation with respect to $\xi_{i,0}^r$ and using $f_i(x, y) = \mathbb{E}_{\xi_i}[\tilde{f}_i(x, y, \xi_i)]$. The last equality follows from Lemma A.1 (i), since $\hat{x}^{r-1}$ is measurable with respect to $\mathcal{F}_{r-1,K}$ and v^r is independent of $\mathcal{F}_{r-1,K}$.

Next, consider $k \in \{2, \ldots, K\}$. In this case, $z_{i,k-1}^r$ and v^r are measurable with respect to $\mathcal{F}_{r,k-1}$, while $\xi_{i,k-1}^r$ is independent of $\mathcal{F}_{r,k-1}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r,k-1}\right] &= \mathbb{E}_{\xi_{i,k-1}^r}\Big[\tfrac{n}{2\eta^2}(\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) \\
&\quad - \tilde{f}_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r), \xi_{i,k-1}^r))v^r \mid \mathcal{F}_{r,k-1}\Big] \\
&= \tfrac{n}{2\eta^2}(f_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r)) - f_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r)))v^r \\
&= \nabla f_i^{\eta}(z_{i,k-1}^r) + b_{i,k-1}^r,
\end{aligned}$$

where the first equality follows from the definition of $g_{i,k-1}^{\eta,r,z}$, the second equality follows by taking the expectation with respect to $\xi_{i,k-1}^r$ and using $f_i(x, y) = \mathbb{E}_{\xi_i}[\tilde{f}_i(x, y, \xi_i)]$, and the last equality follows from the definition of $b_{i,k-1}^r$ given in equation 9.

Next, consider equation 9 and then by adding and subtracting $\frac{n}{2\eta^2}(f_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r)))v^r + \nabla f_i^{\eta}(\hat{x}^{r-1})$, we obtain

$$\begin{aligned}
b_{i,k-1}^r =& \tfrac{n}{2\eta^2}(f_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r)) - f_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r)) \\
&- f_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r)) + f_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r)))v^r \\
&+ \tfrac{n}{2\eta^2}(f_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r)) - f_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r)))v^r \\
&- \nabla f_i^{\eta}(\hat{x}^{r-1}) + \nabla f_i^{\eta}(\hat{x}^{r-1}) - \nabla f_i^{\eta}(z_{i,k-1}^r).
\end{aligned} \tag{12}$$

Since $\hat{x}^{r-1}$ is measurable with respect to $\mathcal{F}_{r-1,K}$ and v^r is independent of $\mathcal{F}_{r-1,K}$, Lemma A.1 (i) gives

$$\mathbb{E}\left[\tfrac{n}{2\eta^2}(f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r)) - f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r)))v^r - \nabla f_i^\eta(\hat{x}^{r-1}) \mid \mathcal{F}_{r-1,K}\right] \tag{13}$$

$$= \mathbb{E}\left[\tfrac{n}{2\eta^2}(f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r)) - f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r)))v^r \mid \mathcal{F}_{r-1,K}\right] - \nabla f_i^\eta(\hat{x}^{r-1})$$
$$= \nabla f_i^\eta(\hat{x}^{r-1}) - \nabla f_i^\eta(\hat{x}^{r-1}) = 0.$$

Therefore, taking the conditional expectations on both sides of equation 12 and using equation 13, we obtain

$$\begin{aligned}\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right] = \mathbb{E}\Big[&\tfrac{n}{2\eta^2}(f_i(z_{i,k-1}^r+v^r, y(z_{i,k-1}^r+v^r)) - f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r))\\ &-f_i(z_{i,k-1}^r-v^r, y(z_{i,k-1}^r-v^r)) + f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r)))v^r \mid \mathcal{F}_{r-1,K}\Big]\\ &+\mathbb{E}\left[\nabla f_i^\eta(\hat{x}^{r-1}) - \nabla f_i^\eta(z_{i,k-1}^r) \mid \mathcal{F}_{r-1,K}\right],\end{aligned}$$

Then, using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, conditional Jensen's inequality $\|\mathbb{E}[X \mid \mathcal{F}]\|^2 \le \mathbb{E}\left[\|X\|^2 \mid \mathcal{F}\right]$, and the tower property, we obtain

$$\begin{aligned}&\mathbb{E}\left[\|\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &\le 2\mathbb{E}\Big[\|\mathbb{E}\Big[\tfrac{n}{2\eta^2}(f_i(z_{i,k-1}^r+v^r, y(z_{i,k-1}^r+v^r)) - f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r))\\ &-f_i(z_{i,k-1}^r-v^r, y(z_{i,k-1}^r-v^r)) + f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r)))v^r \mid \mathcal{F}_{r-1,K}\Big]\|^2\Big]\\ &+ 2\mathbb{E}\left[\|\mathbb{E}\left[\nabla f_i^\eta(\hat{x}^{r-1}) - \nabla f_i^\eta(z_{i,k-1}^r) \mid \mathcal{F}_{r-1,K}\right]\|^2\right]\\ &\le 2\mathbb{E}\Big[\mathbb{E}\Big[\|\tfrac{n}{2\eta^2}(f_i(z_{i,k-1}^r+v^r, y(z_{i,k-1}^r+v^r)) - f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r))\\ &-f_i(z_{i,k-1}^r-v^r, y(z_{i,k-1}^r-v^r)) + f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r)))v^r\|^2 \mid \mathcal{F}_{r-1,K}\Big]\Big]\\ &+ 2\mathbb{E}\left[\mathbb{E}\left[\|\nabla f_i^\eta(\hat{x}^{r-1}) - \nabla f_i^\eta(z_{i,k-1}^r)\|^2 \mid \mathcal{F}_{r-1,K}\right]\right]\\ &= 2\mathbb{E}\Big[\|\tfrac{n}{2\eta^2}(f_i(z_{i,k-1}^r+v^r, y(z_{i,k-1}^r+v^r)) - f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r))\\ &-f_i(z_{i,k-1}^r-v^r, y(z_{i,k-1}^r-v^r)) + f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r)))v^r\|^2\Big]\\ &+ 2\mathbb{E}\left[\|\nabla f_i^\eta(\hat{x}^{r-1}) - \nabla f_i^\eta(z_{i,k-1}^r)\|^2\right].\end{aligned} \tag{14}$$

Since $f_i(x, y(x))$ is L_0^{imp}-Lipschitz continuous and $\|v^r\| = \eta$, we obtain

$$\begin{aligned}&\|\tfrac{n}{2\eta^2}(f_i(z_{i,k-1}^r+v^r, y(z_{i,k-1}^r+v^r)) - f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r))\\ &- f_i(z_{i,k-1}^r-v^r, y(z_{i,k-1}^r-v^r)) + f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r)))v^r\|^2\\ &= \tfrac{n^2}{4\eta^4}\|v^r\|^2\Big(f_i(z_{i,k-1}^r+v^r, y(z_{i,k-1}^r+v^r)) - f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r))\\ &- f_i(z_{i,k-1}^r-v^r, y(z_{i,k-1}^r-v^r)) + f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r))\Big)^2\\ &\le \tfrac{n^2}{2\eta^2}\left|f_i(z_{i,k-1}^r+v^r, y(z_{i,k-1}^r+v^r)) - f_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r))\right|^2\\ &+ \tfrac{n^2}{2\eta^2}\left|f_i(z_{i,k-1}^r-v^r, y(z_{i,k-1}^r-v^r)) - f_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r))\right|^2\\ &\le \tfrac{n^2(L_0^{\mathrm{imp}})^2}{2\eta^2}\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2 + \tfrac{n^2(L_0^{\mathrm{imp}})^2}{2\eta^2}\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\\ &= \tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2.\end{aligned} \tag{15}$$

Moreover, Lemma A.14 together with Lemma A.1 (ii.3) gives

$$\mathbb{E}\left[\|\nabla f_i^\eta(\hat{x}^{r-1}) - \nabla f_i^\eta(z_{i,k-1}^r)\|^2\right] \le \tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|\hat{x}^{r-1} - z_{i,k-1}^r\|^2\right]. \tag{16}$$

Substituting equation 15 and equation 16 and equation 14 yields equation 10 as follows.

$$\mathbb{E}\left[\|\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \frac{2n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right].$$

(ii) First, in the expression defining $g_{i,k-1}^{\eta,r,z}$ given in Definition A.15 (i), we add and subtract $\frac{n}{2\eta^2}\tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r)v^r$. Then, taking squared norms and expectations and using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z}\|^2\right] \qquad (17)\\
&\leq 2\mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r)\right.\\
&\left.-\tilde{f}_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r), \xi_{i,k-1}^r) + \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r))v^r\|^2\right]\\
&+ 2\mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r))v^r\|^2\right].
\end{aligned}$$

By Lemma A.14 (ii), the Cauchy–Schwarz inequality, and Assumption 2.1 (i), we obtain

$$\begin{aligned}
\mathbb{E}\left[(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2\right] &= (L_{0,x}^y)^2\mathbb{E}\left[(L_{0,y}^f(\xi_{i,k-1}^r))^2\right] + 2L_{0,x}^y\mathbb{E}\left[L_{0,y}^f(\xi_{i,k-1}^r)L_{0,x}^f(\xi_{i,k-1}^r)\right] \qquad (18)\\
&\quad + \mathbb{E}\left[(L_{0,x}^f(\xi_{i,k-1}^r))^2\right]\\
&\leq (L_{0,x}^y)^2\mathbb{E}\left[(L_{0,y}^f(\xi_{i,k-1}^r))^2\right] + 2L_{0,x}^y\left(\mathbb{E}\left[(L_{0,y}^f(\xi_{i,k-1}^r))^2\right]\right)^{1/2}\left(\mathbb{E}\left[(L_{0,x}^f(\xi_{i,k-1}^r))^2\right]\right)^{1/2}\\
&\quad + \mathbb{E}\left[(L_{0,x}^f(\xi_{i,k-1}^r))^2\right] \leq (L_{0,x}^y)^2(L_{0,y}^f)^2 + 2L_{0,x}^yL_{0,y}^fL_{0,x}^f + (L_{0,x}^f)^2\\
&= \left(L_{0,y}^fL_{0,x}^y + L_{0,x}^f\right)^2 = (L_0^{\mathrm{imp}})^2,
\end{aligned}$$

where the first equality follows from Lemma A.14 (ii), the first inequality follows from the Cauchy–Schwarz inequality, the second inequality follows from Assumption 2.1 (i), and the last equality follows from Lemma A.14 (iii).

By Lemma A.14 (ii), $\tilde{f}_i(x, y(x), \xi_{i,k-1}^r)$ is $L_0^{\mathrm{imp}}(\xi_{i,k-1}^r)$-Lipschitz continuous. Since $\|v^r\| = \eta$, the first expectation on the right-hand side of equation 17 satisfies

$$\begin{aligned}
&\mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) \qquad (19)\right.\\
&\left.-\tilde{f}_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r), \xi_{i,k-1}^r) + \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r))v^r\|^2\right]\\
&\leq \tfrac{n^2}{\eta^2}\mathbb{E}\left[(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] = \tfrac{n^2}{\eta^2}\mathbb{E}\left[(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2\right]\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right]\\
&\leq \tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right],
\end{aligned}$$

where the equality follows because $\xi_{i,k-1}^r$ is independent of $z_{i,k-1}^r$ and $\hat{x}^{r-1}$, and the last inequality uses equation 18.

We next consider the second expectation in equation 17. Since $\hat{x}^{r-1}$ is measurable with respect to $\mathcal{F}_{r-1,K}$ and v^r is independent of $\mathcal{F}_{r-1,K}$ and $\xi_{i,k-1}^r$, we condition on $\mathcal{F}_{r-1,K} \cup \{\xi_{i,k-1}^r\}$. Let $\bar{v}^r := \frac{v^r}{\eta}$, which is uniformly distributed on the unit sphere $\mathbb{S}$, and define $h_i(\bar{v}^r) := \tilde{f}_i(\hat{x}^{r-1} + \eta\bar{v}^r, y(\hat{x}^{r-1} + \eta\bar{v}^r), \xi_{i,k-1}^r)$. Moreover, $\mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right] = \mathbb{E}_{\hat{v}}\left[\tilde{f}_i(\hat{x}^{r-1} + \hat{v}, y(\hat{x}^{r-1} + \hat{v}), \xi_{i,k-1}^r)\right]$, where $\hat{v}$ is an independent random vector uniformly distributed on $\eta\mathbb{S}$. By Lemma A.14 (ii), for any $\bar{v}_1^r, \bar{v}_2^r \in \mathbb{S}$,

$$|h_i(\bar{v}_1^r) - h_i(\bar{v}_2^r)| \leq L_0^{\mathrm{imp}}(\xi_{i,k-1}^r)\|\eta\bar{v}_1^r - \eta\bar{v}_2^r\| = \eta L_0^{\mathrm{imp}}(\xi_{i,k-1}^r)\|\bar{v}_1^r - \bar{v}_2^r\|.$$

Thus, h_i is $\eta L_0^{\mathrm{imp}}(\xi_{i,k-1}^r)$-Lipschitz continuous. Therefore, by Lemma A.2, for every $t \geq 0$,

$$\mathbb{P}_{\bar{v}^r}\left[|h_i(\bar{v}^r) - \mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right]| \geq t \mid \mathcal{F}_{r-1,K} \cup \{\xi_{i,k-1}^r\}\right] \leq 2\sqrt{2\pi}\exp\left(-\frac{nt^2}{8\eta^2(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2}\right).$$

Using the tail-integral identity, we obtain

$$\begin{aligned}
&\mathbb{E}_{\bar{v}^r}\left[\left|h_i(\bar{v}^r)-\mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right]\right|^2 \mid \mathcal{F}_{r-1,K}\cup\{\xi_{i,k-1}^r\}\right]\\
&=\int_0^\infty \mathbb{P}_{\bar{v}^r}\left[\left|h_i(\bar{v}^r)-\mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right]\right|^2\geq\beta \mid \mathcal{F}_{r-1,K}\cup\{\xi_{i,k-1}^r\}\right]\,d\beta\\
&=\int_0^\infty \mathbb{P}_{\bar{v}^r}\left[\left|h_i(\bar{v}^r)-\mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right]\right|\geq\sqrt{\beta} \mid \mathcal{F}_{r-1,K}\cup\{\xi_{i,k-1}^r\}\right]\,d\beta\\
&\leq 2\sqrt{2\pi}\int_0^\infty \exp\left(-\tfrac{n\beta}{8\eta^2(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2}\right)\,d\beta\\
&=\tfrac{16\sqrt{2\pi}\eta^2(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2}{n}.
\end{aligned}$$

Next, using $\|v^r\|=\eta$, $\|a-b\|^2\leq 2\|a\|^2+2\|b\|^2$, and the symmetry of $\bar{v}^r$ and $-\bar{v}^r$, we obtain

$$\begin{aligned}
&\mathbb{E}_{v^r}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r,y(\hat{x}^{r-1}+v^r),\xi_{i,k-1}^r)-\tilde{f}_i(\hat{x}^{r-1}-v^r,y(\hat{x}^{r-1}-v^r),\xi_{i,k-1}^r))v^r\|^2\right.\\
&\left.\mid \mathcal{F}_{r-1,K}\cup\{\xi_{i,k-1}^r\}\right]\\
&=\tfrac{n^2}{4\eta^2}\mathbb{E}_{\bar{v}^r}\left[|h_i(\bar{v}^r)-h_i(-\bar{v}^r)|^2 \mid \mathcal{F}_{r-1,K}\cup\{\xi_{i,k-1}^r\}\right]\leq\tfrac{n^2}{2\eta^2}\mathbb{E}_{\bar{v}^r}\left[|h_i(\bar{v}^r)-\mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right]|^2\right.\\
&\left.+|h_i(-\bar{v}^r)-\mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right]|^2 \mid \mathcal{F}_{r-1,K}\cup\{\xi_{i,k-1}^r\}\right]=\tfrac{n^2}{\eta^2}\mathbb{E}_{\bar{v}^r}\left[|h_i(\bar{v}^r)-\mathbb{E}_{\bar{v}^r}\left[h_i(\bar{v}^r)\right]|^2 \mid \mathcal{F}_{r-1,K}\cup\{\xi_{i,k-1}^r\}\right]\\
&\leq 16\sqrt{2\pi}\,n(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2.
\end{aligned}$$

Taking expectations in the preceding conditional bound, using the tower property, and invoking equation 18, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r,y(\hat{x}^{r-1}+v^r),\xi_{i,k-1}^r)-\tilde{f}_i(\hat{x}^{r-1}-v^r,y(\hat{x}^{r-1}-v^r),\xi_{i,k-1}^r))v^r\|^2\right] \quad (20)\\
&\leq 16\sqrt{2\pi}\,n\mathbb{E}\left[(L_0^{\mathrm{imp}}(\xi_{i,k-1}^r))^2\right]\leq 16\sqrt{2\pi}\,n(L_0^{\mathrm{imp}})^2.
\end{aligned}$$

Combining this bound with the bound for the first expectation in equation 17, we obtain

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z}\|^2\right]\leq\tfrac{2n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r-\hat{x}^{r-1}\|^2\right]+32\sqrt{2\pi}\,n(L_0^{\mathrm{imp}})^2.$$

For $k=1$, since $z_{i,0}^r=\hat{x}^{r-1}$, the first difference term in equation 17 vanishes. Therefore, applying the bound for the second expectation directly gives $\mathbb{E}\left[\|g_{i,0}^{\eta,r,z}\|^2\right]\leq 16\sqrt{2\pi}\,n(L_0^{\mathrm{imp}})^2$. (iii) First, adding and subtracting $\frac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r,y(\hat{x}^{r-1}+v^r),\xi_{i,k-1}^r)-\tilde{f}_i(\hat{x}^{r-1}-v^r,y(\hat{x}^{r-1}-v^r),\xi_{i,k-1}^r))v^r+\nabla f_i^\eta(\hat{x}^{r-1})$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z}-\nabla f_i^\eta(z_{i,k-1}^r)\|^2\right]\\
&=\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z}-\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r,y(\hat{x}^{r-1}+v^r),\xi_{i,k-1}^r)-\tilde{f}_i(\hat{x}^{r-1}-v^r,y(\hat{x}^{r-1}-v^r),\xi_{i,k-1}^r))v^r\right.\\
&+\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r,y(\hat{x}^{r-1}+v^r),\xi_{i,k-1}^r)-\tilde{f}_i(\hat{x}^{r-1}-v^r,y(\hat{x}^{r-1}-v^r),\xi_{i,k-1}^r))v^r-\nabla f_i^\eta(\hat{x}^{r-1})\\
&\left.+\nabla f_i^\eta(\hat{x}^{r-1})-\nabla f_i^\eta(z_{i,k-1}^r)\|^2\right].
\end{aligned}$$

Applying Lemma A.5, equation 19, and Lemma A.1 (ii.3), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z}-\nabla f_i^\eta(z_{i,k-1}^r)\|^2\right]\leq\tfrac{3n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r-\hat{x}^{r-1}\|^2\right]\\
&+3\mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r,y(\hat{x}^{r-1}+v^r),\xi_{i,k-1}^r)-\tilde{f}_i(\hat{x}^{r-1}-v^r,y(\hat{x}^{r-1}-v^r),\xi_{i,k-1}^r))v^r-\nabla f_i^\eta(\hat{x}^{r-1})\|^2\right]\\
&+\tfrac{3n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|\hat{x}^{r-1}-z_{i,k-1}^r\|^2\right].
\end{aligned}$$

Since $\hat{x}^{r-1}$ is measurable with respect to $\mathcal{F}_{r-1,K}$, v^r is independent of $\mathcal{F}_{r-1,K}$, and $\xi_{i,k-1}^r$ is independent of $\mathcal{F}_{r,k-1}$, which contains v^r, Lemma A.1 (i) gives

$$\begin{aligned}
&\mathbb{E}\left[\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r,y(\hat{x}^{r-1}+v^r),\xi_{i,k-1}^r)-\tilde{f}_i(\hat{x}^{r-1}-v^r,y(\hat{x}^{r-1}-v^r),\xi_{i,k-1}^r))v^r \mid \mathcal{F}_{r-1,K}\right] \quad (21)\\
&=\nabla f_i^\eta(\hat{x}^{r-1}).
\end{aligned}$$

By equation 21, the conditional expectation appearing in Lemma A.3 is equal to $\nabla f_i^\eta(\hat{x}^{r-1})$. Therefore, applying Lemma A.3 conditionally on $\mathcal{F}_{r-1,K}$, taking expectations, using the tower property, and invoking equation 20, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r), \xi_{i,k-1}^r))v^r - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\
&= \mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r), \xi_{i,k-1}^r))v^r\|^2\right] \\
&- \mathbb{E}\left[\|\nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\
&\leq \mathbb{E}\left[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r), \xi_{i,k-1}^r))v^r\|^2\right] \\
&\leq 16\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2,
\end{aligned}$$

Hence,

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(z_{i,k-1}^r)\|^2\right] \leq \tfrac{3n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] + 48\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

For $k = 1$, since $z_{i,0}^r = \hat{x}^{r-1}$, the first and third terms in the preceding decomposition vanish. Therefore, the bound for the remaining term gives

$$\mathbb{E}\left[\|g_{i,0}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \leq 16\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

(iv) Regarding Definition A.15, using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and $\|v^r\| = \eta$, we obtain

$$\begin{aligned}
\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2 \leq\ & \tfrac{n^2}{2\eta^4}|\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(z_{i,k-1}^r + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r)|^2\|v^r\|^2 \\
&+ \tfrac{n^2}{2\eta^4}|\tilde{f}_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r), \xi_{i,k-1}^r) - \tilde{f}_i(z_{i,k-1}^r - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r)|^2\|v^r\|^2.
\end{aligned}$$

By Assumption 2.1 (i) and Lemma A.14 (i), we have

$$\begin{aligned}
&|\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(z_{i,k-1}^r + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r)| \\
&\leq L_{0,y}^f(\xi_{i,k-1}^r) L_{0,x}^y \|z_{i,k-1}^r - \hat{x}^{r-1}\|,
\end{aligned}$$

and analogously for the term involving $-v^r$. Therefore,

$$\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2 \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f(\xi_{i,k-1}^r) L_{0,x}^y)^2 \|z_{i,k-1}^r - \hat{x}^{r-1}\|^2.$$

Taking conditional expectations in the preceding inequality with respect to $\mathcal{F}_{r,k-1}$, using the $\mathcal{F}_{r,k-1}$-measurability of $\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2$ and the independence of $\xi_{i,k-1}^r$ from $\mathcal{F}_{r,k-1}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2 \mid \mathcal{F}_{r,k-1}\right] &\leq \tfrac{n^2}{\eta^2}(L_{0,x}^y)^2 \mathbb{E}\left[(L_{0,y}^f(\xi_{i,k-1}^r))^2 \|z_{i,k-1}^r - \hat{x}^{r-1}\|^2 \mid \mathcal{F}_{r,k-1}\right] \\
&= \tfrac{n^2}{\eta^2}(L_{0,x}^y)^2 \|z_{i,k-1}^r - \hat{x}^{r-1}\|^2 \mathbb{E}\left[(L_{0,y}^f(\xi_{i,k-1}^r))^2 \mid \mathcal{F}_{r,k-1}\right] \\
&\leq \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 \|z_{i,k-1}^r - \hat{x}^{r-1}\|^2.
\end{aligned}$$

Taking expectations on both sides and using the law of total expectation, we obtain

$$\mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right] \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 \mathbb{E}\left[\|\hat{x}^{r-1} - z_{i,k-1}^r\|^2\right].$$

(v) Regarding Definition A.15, using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and $\|v^r\| = \eta$, we obtain

$$\begin{aligned}
\|\check{w}_{i,k-1}^{\eta,r,z}\|^2 \leq\ & \tfrac{n^2}{2\eta^2}(L_{0,y}^f(\xi_{i,k-1}^r))^2 \|y_{\epsilon_r}(\hat{x}^{r-1} + v^r) - y(\hat{x}^{r-1} + v^r)\|^2 \\
&+ \tfrac{n^2}{2\eta^2}(L_{0,y}^f(\xi_{i,k-1}^r))^2 \|y_{\epsilon_r}(\hat{x}^{r-1} - v^r) - y(\hat{x}^{r-1} - v^r)\|^2.
\end{aligned}$$

Taking conditional expectations in the preceding inequality with respect to $\mathcal{F}_{r,k-1}$, using the $\mathcal{F}_{r,k-1}$-measurability of the two lower-level error terms and the independence of $\xi_{i,k-1}^r$ from $\mathcal{F}_{r,k-1}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r,z}\|^2 \mid \mathcal{F}_{r,k-1}\right] \leq\ & \tfrac{n^2}{2\eta^2}(L_{0,y}^f)^2 \|y_{\epsilon_r}(\hat{x}^{r-1} + v^r) - y(\hat{x}^{r-1} + v^r)\|^2 \\
&+ \tfrac{n^2}{2\eta^2}(L_{0,y}^f)^2 \|y_{\epsilon_r}(\hat{x}^{r-1} - v^r) - y(\hat{x}^{r-1} - v^r)\|^2.
\end{aligned}$$

Taking expectations on both sides, using the law of total expectation and $\mathbb{E}\left[\|y_{\epsilon_r}(\bullet) - y(\bullet)\|^2\right] \leq \epsilon_r$, we obtain $\mathbb{E}\left[\|\breve{w}_{i,k-1}^{\eta,r,z}\|^2\right] \leq \frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r$.

(vi) For $k = 1$, since $z_{i,0}^r = \hat{x}^{r-1}$ and the estimator is conditionally unbiased at $\hat{x}^{r-1}$, we have

$$\mathbb{E}\left[\|g_{i,0}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \leq 16\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

Moreover, for $k \in \{2, \ldots, K\}$, by adding and subtracting $\frac{n}{2\eta^2}\big(\tilde{f}_i(\hat{x}^{r-1}+v^r, y(\hat{x}^{r-1}+v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1}-v^r, y(\hat{x}^{r-1}-v^r), \xi_{i,k-1}^r)\big)v^r$, we obtain

$$\begin{aligned}
&g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1}) \\
&= \tfrac{n}{2\eta^2}\Big(\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) \\
&-\tilde{f}_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r), \xi_{i,k-1}^r) + \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r)\Big)\, v^r \\
&+ \tfrac{n}{2\eta^2}\Big(\tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r)\Big)\, v^r \\
&- \nabla f_i^\eta(\hat{x}^{r-1}).
\end{aligned}$$

Using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\
&\leq 2\mathbb{E}\Big[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) \\
&-\tilde{f}_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r), \xi_{i,k-1}^r) + \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r))v^r\|^2\Big] \\
&+ 2\mathbb{E}\Big[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r))v^r - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\Big].
\end{aligned}$$

By the bound derived in part (ii),

$$\begin{aligned}
&\mathbb{E}\Big[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(z_{i,k-1}^r + v^r, y(z_{i,k-1}^r + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) \\
&-\tilde{f}_i(z_{i,k-1}^r - v^r, y(z_{i,k-1}^r - v^r), \xi_{i,k-1}^r) + \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r))v^r\|^2\Big] \\
&\leq \tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right].
\end{aligned}$$

Moreover, as established above,

$$\begin{aligned}
&\mathbb{E}\Big[\|\tfrac{n}{2\eta^2}(\tilde{f}_i(\hat{x}^{r-1} + v^r, y(\hat{x}^{r-1} + v^r), \xi_{i,k-1}^r) - \tilde{f}_i(\hat{x}^{r-1} - v^r, y(\hat{x}^{r-1} - v^r), \xi_{i,k-1}^r))v^r - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\Big] \\
&\leq 16\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.
\end{aligned}$$

Therefore, for $k \in \{2, \ldots, K\}$, we obtain the result in equation 11. □

Lemma A.20. Consider Algorithm 1. Suppose Assumption 2.1 holds, and suppose that $\mathcal{S}^r$ is sampled uniformly from all subsets of $[m]$ of cardinality s. Then, the following statements hold true.

(i) For $r \geq 1$ and $K \geq 1$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbb{E}\left[\hat{x}^r \mid \mathcal{F}_{r-1,K}\right] - \hat{x}^{r-1}\|^2\right] \leq &\left(9\tilde{\gamma}^2\tfrac{n(2n+3)(L_0^{\mathrm{imp}})^2}{\eta^2} + 3\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_r \\
&+ 9\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] + 3\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r.
\end{aligned}$$

(ii) For $r \geq 1$ and $K \geq 1$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right] \leq &\left(6\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 12\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\mathcal{D}_r \\
&+ \left(216\tilde{\gamma}^2\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 36\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\
&+ 72\tilde{\gamma}^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1} + 6\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\
&+ 36\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} + 6\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&+ \left(3648 - \tfrac{2400}{K}\right)\sqrt{2\pi}\,\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.
\end{aligned}$$

Proof. (i) Considering Algorithm 1, we have

$$\begin{aligned}\hat{x}^r - \hat{x}^{r-1} &= -\tfrac{\tilde{\gamma}}{Ks}\textstyle\sum_{k=1}^{K}\sum_{i\in\mathcal{S}^r}(\check{g}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}) \\ &= -\tfrac{\tilde{\gamma}}{Ks}\textstyle\sum_{k=1}^{K}\sum_{i\in\mathcal{S}^r}(g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}).\end{aligned}$$

Since $\mathcal{S}^r$ is sampled uniformly, $\mathbb{P}[i \in \mathcal{S}^r] = \frac{s}{m}$ for every $i \in [m]$. Moreover, c_i^{r-1} and $\hat{c}^{r-1}$ are measurable with respect to $\mathcal{F}_{r-1,K}$. Thus,

$$\mathbb{E}\left[\tfrac{1}{s}\textstyle\sum_{i\in\mathcal{S}^r}(-c_i^{r-1} + \hat{c}^{r-1}) \mid \mathcal{F}_{r-1,K}\right] = \tfrac{1}{m}\textstyle\sum_{i=1}^{m}(-c_i^{r-1} + \hat{c}^{r-1}) = -\tfrac{1}{m}\textstyle\sum_{i=1}^{m} c_i^{r-1} + \hat{c}^{r-1} = 0,$$

where the last equality follows from $\hat{c}^{r-1} = \frac{1}{m}\sum_{i=1}^{m} c_i^{r-1}$. Therefore,

$$\mathbb{E}[\hat{x}^r \mid \mathcal{F}_{r-1,K}] - \hat{x}^{r-1} = -\tfrac{\tilde{\gamma}}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right].$$

Taking the squared norm, using $\|a+b+c\|^2 \le 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, and then taking expectations, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{E}[\hat{x}^r \mid \mathcal{F}_{r-1,K}] - \hat{x}^{r-1}\|^2\right] &\le 3\tilde{\gamma}^2\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &\quad + 3\tilde{\gamma}^2\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\hat{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &\quad + 3\tilde{\gamma}^2\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\check{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right].\end{aligned} \tag{22}$$

We first bound the term containing the conditional mean of $g_{i,k-1}^{\eta,r,z}$. For $k = 1$, Lemma A.19(i) gives

$$\mathbb{E}\left[g_{i,0}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right] = \nabla f_i^{\eta}(\hat{x}^{r-1}).$$

For $k \in \{2, \dots, K\}$, since $\mathcal{F}_{r-1,K} \subseteq \mathcal{F}_{r,k-1}$, the tower property and Lemma A.19(i) give

$$\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right] = \mathbb{E}\left[\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r,k-1}\right] \mid \mathcal{F}_{r-1,K}\right] = \mathbb{E}\left[\nabla f_i^{\eta}(z_{i,k-1}^r) + b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right].$$

Using $\nabla f^{\eta}(\hat{x}^{r-1}) = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i^{\eta}(\hat{x}^{r-1})$, we therefore obtain

$$\begin{aligned}\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right] &= \nabla f^{\eta}(\hat{x}^{r-1}) \\ &\quad + \tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\nabla f_i^{\eta}(z_{i,k-1}^r) - \nabla f_i^{\eta}(\hat{x}^{r-1}) \mid \mathcal{F}_{r-1,K}\right] \\ &\quad + \tfrac{1}{Km}\textstyle\sum_{k=2}^{K}\sum_{i=1}^{m}\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right].\end{aligned}$$

Using $\|a+b+c\|^2 \le 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$ gives

$$\begin{aligned}&\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &\le 3\mathbb{E}\left[\|\nabla f^{\eta}(\hat{x}^{r-1})\|^2\right] \\ &+ 3\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\nabla f_i^{\eta}(z_{i,k-1}^r) - \nabla f_i^{\eta}(\hat{x}^{r-1}) \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &+ 3\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=2}^{K}\sum_{i=1}^{m}\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right].\end{aligned} \tag{23}$$

Finite-sum Jensen's inequality, conditional Jensen's inequality, and the tower property give

$$\begin{aligned}&\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\nabla f_i^{\eta}(z_{i,k-1}^r) - \nabla f_i^{\eta}(\hat{x}^{r-1}) \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &\le \tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\|\nabla f_i^{\eta}(z_{i,k-1}^r) - \nabla f_i^{\eta}(\hat{x}^{r-1})\|^2\right].\end{aligned}$$

By Lemma A.1(ii.3), ∇f_i^{η} is $\frac{\sqrt{n}L_0^{\mathrm{imp}}}{\eta}$-Lipschitz continuous. Hence,

$$\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\nabla f_i^{\eta}(z_{i,k-1}^r) - \nabla f_i^{\eta}(\hat{x}^{r-1}) \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \le \tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_r.$$

For $k \in \{2, \ldots, K\}$, equation 10 gives

$$\mathbb{E}\left[\|\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \tfrac{2n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right].$$

Applying Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=2}^K\sum_{i=1}^m \mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right] &\leq \tfrac{1}{Km}\textstyle\sum_{k=2}^K\sum_{i=1}^m \mathbb{E}\left[\|\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &\leq \tfrac{2n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\tfrac{1}{Km}\textstyle\sum_{k=2}^K\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] \\ &= \tfrac{2n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_r,\end{aligned}$$

where the last equality follows from $z_{i,0}^r = \hat{x}^{r-1}$. Substituting the preceding bounds in equation 23 yields

$$\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq 3\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] + \tfrac{3n(2n+3)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_r. \tag{24}$$

First, Jensen's inequality for the average over i and k gives

$$\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\hat{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|\mathbb{E}\left[\hat{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right].$$

For every $i \in [m]$ and $k \in \{1, \ldots, K\}$, conditional Jensen's inequality gives

$$\|\mathbb{E}\left[\hat{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2 \leq \mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2 \mid \mathcal{F}_{r-1,K}\right].$$

Taking expectations on both sides and using the tower property, we obtain

$$\mathbb{E}\left[\|\mathbb{E}\left[\hat{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2 \mid \mathcal{F}_{r-1,K}\right]\right] = \mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right].$$

Therefore,

$$\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\hat{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right].$$

Applying Lemma A.19(iv) to each term on the right-hand side gives

$$\begin{aligned}\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right] &\leq \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] \\ &= \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\mathcal{D}_r,\end{aligned}$$

where the last equality follows from the definition of $\mathcal{D}_r$. Combining the preceding two bounds, we conclude that

$$\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\hat{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\mathcal{D}_r. \tag{25}$$

Similarly, Lemma A.19(v) gives

$$\mathbb{E}\left[\|\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\check{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \leq \tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r,z}\|^2\right] \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r. \tag{26}$$

Substituting the bounds obtained in equation 24, equation 25, and equation 26 into equation 22 gives the result in part (i).

(ii) Considering Algorithm 1, we have

$$\hat{x}^r - \hat{x}^{r-1} = -\tfrac{\tilde{\gamma}}{Ks}\textstyle\sum_{k=1}^K\sum_{i\in\mathcal{S}^r}(g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}).$$

Therefore, we obtain

$$\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right] = \tfrac{\tilde{\gamma}^2}{K^2s^2}\mathbb{E}\left[\|\textstyle\sum_{k=1}^K\sum_{i\in\mathcal{S}^r}(g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1})\|^2\right].$$

Applying Lemma A.5 to the Ks terms in the preceding sum gives

$$\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right] \leq \frac{\tilde{\gamma}^2}{Ks}\sum_{k=1}^{K}\mathbb{E}\left[\sum_{i\in\mathcal{S}^r}\|g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}\|^2\right].$$

Uniform sampling and $\mathbb{P}\left[i \in \mathcal{S}^r\right] = \frac{s}{m}$ give

$$\begin{aligned}&\mathbb{E}\left[\sum_{i\in\mathcal{S}^r}\|g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}\|^2\right]\\ &= \frac{s}{m}\sum_{i=1}^{m}\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}\|^2\right].\end{aligned}$$

Consequently,

$$\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right] \leq \frac{\tilde{\gamma}^2}{Km}\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}\|^2\right]. \quad (27)$$

We first consider $r = 1$. By the initialization of Algorithm 1, we have $c_i^0 = \hat{c}^0 = 0,\ i \in [m]$. Therefore,

$$g_{i,k-1}^{\eta,1,z} + \hat{w}_{i,k-1}^{\eta,1,z} + \check{w}_{i,k-1}^{\eta,1,z} - c_i^0 + \hat{c}^0 = \hat{w}_{i,k-1}^{\eta,1,z} + \check{w}_{i,k-1}^{\eta,1,z} + \left(g_{i,k-1}^{\eta,1,z} - \nabla f_i^\eta(\hat{x}^0)\right) + \nabla f_i^\eta(\hat{x}^0).$$

By Lemma A.14(ii), Jensen's inequality, and the definition of L_0^{imp}, for all $x_1, x_2 \in \mathbb{R}^n$,

$$|f_i(x_1) - f_i(x_2)| \leq \mathbb{E}_{\xi_i}\left[L_0^{\mathrm{imp}}(\xi_i)\right]\|x_1 - x_2\| \leq L_0^{\mathrm{imp}}\|x_1 - x_2\|.$$

Thus, Lemma A.1(ii.1) implies that f_i^η is L_0^{imp}-Lipschitz. Since f_i^η is continuously differentiable,

$$\|\nabla f_i^\eta(\hat{x}^0)\| \leq L_0^{\mathrm{imp}}. \quad (28)$$

Using $\|a+b+c+d\|^2 \leq 4\|a\|^2 + 4\|b\|^2 + 4\|c\|^2 + 4\|d\|^2$ in equation 27, invoking Lemma A.19(iii)–(v), using equation 28, and averaging over i and k, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}^1 - \hat{x}^0\|^2\right] \leq{}& \left(4\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 8\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\mathcal{D}_1 + 4\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1 + 4\tilde{\gamma}^2(L_0^{\mathrm{imp}})^2\\ &+ 4\tilde{\gamma}^2\left(32 - \tfrac{16}{K}\right)\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.\end{aligned}$$

For $K \geq 1$ and $n \geq 1$, we have

$$4 + 4\left(32 - \tfrac{16}{K}\right)\sqrt{2\pi}n \leq \left(132 - \tfrac{64}{K}\right)\sqrt{2\pi}n \leq \left(3648 - \tfrac{2400}{K}\right)\sqrt{2\pi}n.$$

Moreover, $4 \leq 6$,, $8 \leq 12$, and $36\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_0 + 6\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^0)\|^2\right] \geq 0$. Using $\mathcal{C}_0 = \mathcal{D}_0 = 0$, the preceding inequalities give the result in part (ii) for $r = 1$.

We next consider $r \geq 2$. For every $i \in [m]$ and $k \in \{1, \ldots, K\}$, adding and subtracting $\nabla f_i^\eta(\hat{x}^{r-1})$ and $\nabla f^\eta(\hat{x}^{r-1})$, and then using $\|\sum_{j=1}^6 a_j\|^2 \leq 6\sum_{j=1}^6\|a_j\|^2$, give

$$\begin{aligned}\|g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} - c_i^{r-1} + \hat{c}^{r-1}\|^2 \leq{}& 6\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2 + 6\|\check{w}_{i,k-1}^{\eta,r,z}\|^2\\ &+ 6\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\\ &+ 6\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\\ &+ 6\|\hat{c}^{r-1} - \nabla f^\eta(\hat{x}^{r-1})\|^2 + 6\|\nabla f^\eta(\hat{x}^{r-1})\|^2.\end{aligned} \quad (29)$$

Lemma A.19(iv) gives

$$\frac{6\tilde{\gamma}^2}{Km}\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right] \leq 6\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\mathcal{D}_r. \quad (30)$$

Lemma A.19(v) gives

$$\frac{6\tilde{\gamma}^2}{Km}\sum_{k=1}^{K}\sum_{i=1}^{m}\mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r,z}\|^2\right] \leq 6\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r. \quad (31)$$

For $k = 1$, equation 11 gives

$$\mathbb{E}\left[\|g_{i,0}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \leq 16\sqrt{2\pi}\,n(L_0^{\mathrm{imp}})^2.$$

For $k \in \{2, \dots, K\}$, equation 11 gives

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \le \tfrac{2n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] + 32\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

Averaging these bounds over i and k, using $z_{i,0}^r = \hat{x}^{r-1}$ and the definition of $\mathcal{D}_r$, gives

$$\begin{aligned}\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] &\le \tfrac{2n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_r + \tfrac{16+32(K-1)}{K}\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2 \\ &= \tfrac{2n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_r + \left(32 - \tfrac{16}{K}\right)\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.\end{aligned}$$

Multiplying by $6\tilde{\gamma}^2$ gives

$$\tfrac{6\tilde{\gamma}^2}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \le 12\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_r + 6\tilde{\gamma}^2\left(32 - \tfrac{16}{K}\right)\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2. \tag{32}$$

We next bound the two terms containing c_i^{r-1} and $\hat{c}^{r-1}$. Since $\hat{c}^{r-1} - \nabla f^\eta(\hat{x}^{r-1}) = \frac{1}{m}\sum_{i=1}^m (c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1}))$, Jensen's inequality gives

$$\mathbb{E}\left[\|\hat{c}^{r-1} - \nabla f^\eta(\hat{x}^{r-1})\|^2\right] \le \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right].$$

Thus, the combined contribution of the two control-variate terms is bounded by

$$\begin{aligned}&6\tilde{\gamma}^2\mathbb{E}\left[\|\hat{c}^{r-1} - \nabla f^\eta(\hat{x}^{r-1})\|^2\right] + \tfrac{6\tilde{\gamma}^2}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\ &\le \tfrac{12\tilde{\gamma}^2}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right].\end{aligned} \tag{33}$$

By Lemma A.18, for every $i \in [m]$, we have

$$c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1}) = \tfrac{1}{K}\textstyle\sum_{k=1}^K (\check{g}_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})).$$

Applying Jensen's inequality to this average and then averaging over i, we obtain

$$\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \le \tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|\check{g}_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right]. \tag{34}$$

By Definition A.15(vi), we have

$$\check{g}_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1}) = g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1}) + \hat{w}_{i,k-1}^{\eta,r-1,z} + \check{w}_{i,k-1}^{\eta,r-1,z}.$$

Taking the squared Euclidean norm on both sides and using $\|a+b+c\|^2 \le 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, we obtain

$$\|\check{g}_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2 \le 3\|g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2 + 3\|\hat{w}_{i,k-1}^{\eta,r-1,z}\|^2 + 3\|\check{w}_{i,k-1}^{\eta,r-1,z}\|^2.$$

Taking expectations on both sides yields

$$\begin{aligned}\mathbb{E}\left[\|\check{g}_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] &\le 3\mathbb{E}\left[\|g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\ &\quad + 3\mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r-1,z}\|^2\right] + 3\mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r-1,z}\|^2\right].\end{aligned} \tag{35}$$

Adding and subtracting $\nabla f_i^\eta(z_{i,k-1}^{r-1})$ gives

$$g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1}) = g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(z_{i,k-1}^{r-1}) + \nabla f_i^\eta(z_{i,k-1}^{r-1}) - \nabla f_i^\eta(\hat{x}^{r-1}).$$

Thus,

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \le 2\mathbb{E}\left[\|g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(z_{i,k-1}^{r-1})\|^2\right] + 2\mathbb{E}\left[\|\nabla f_i^\eta(z_{i,k-1}^{r-1}) - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right]. \tag{36}$$

For $k = 1$, Lemma A.19(iii) gives

$$\mathbb{E}\left[\|g_{i,0}^{\eta,r-1,z} - \nabla f_i^\eta(z_{i,0}^{r-1})\|^2\right] \le 16\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

For $k \in \{2, \ldots, K\}$, Lemma A.19(iii) gives

$$\mathbb{E}\left[\|g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(z_{i,k-1}^{r-1})\|^2\right] \leq 3\frac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathbb{E}\left[\|z_{i,k-1}^{r-1} - \hat{x}^{r-2}\|^2\right] + 48\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.$$

Averaging over i and k, using $z_{i,0}^{r-1} = \hat{x}^{r-2}$ and the definition of $\mathcal{D}_{r-1}$, gives

$$\begin{aligned}
&\tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(z_{i,k-1}^{r-1})\|^2\right] \\
&\leq 3\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_{r-1} + \tfrac{16+48(K-1)}{K}\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2 \\
&= 3\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_{r-1} + \left(48 - \tfrac{32}{K}\right)\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.
\end{aligned} \tag{37}$$

By Lemma A.1(ii.3), ∇f_i^η is $\frac{\sqrt{n}L_0^{\mathrm{imp}}}{\eta}$-Lipschitz continuous. Hence,

$$\begin{aligned}
&\tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|\nabla f_i^\eta(z_{i,k-1}^{r-1}) - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\
&\leq \tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|z_{i,k-1}^{r-1} - \hat{x}^{r-1}\|^2\right] \\
&= \tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}.
\end{aligned} \tag{38}$$

Averaging equation 36 over i and k, and invoking equation 37 and equation 38, we obtain

$$\begin{aligned}
\tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|g_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] &\leq 6\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_{r-1} + 2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1} \\
&\quad + \left(96 - \tfrac{64}{K}\right)\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.
\end{aligned} \tag{39}$$

Lemma A.19(iv) gives

$$\tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r-1,z}\|^2\right] \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\mathcal{D}_{r-1}. \tag{40}$$

Lemma A.19(v) gives

$$\tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r-1,z}\|^2\right] \leq \tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1}. \tag{41}$$

Averaging equation 35 over i and k, and invoking equation 39, equation 40, and equation 41, we obtain

$$\begin{aligned}
\tfrac{1}{Km}\textstyle\sum_{i=1}^m\sum_{k=1}^K \mathbb{E}\left[\|\check{g}_{i,k-1}^{\eta,r-1,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] &\leq \left(18\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 3\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\
&\quad + 6\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1} + 3\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} \\
&\quad + 3\left(96 - \tfrac{64}{K}\right)\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.
\end{aligned} \tag{42}$$

Combining equation 33, equation 34, and equation 42, we obtain

$$\begin{aligned}
&6\tilde{\gamma}^2\mathbb{E}\left[\|\hat{c}^{r-1} - \nabla f^\eta(\hat{x}^{r-1})\|^2\right] + \tfrac{6\tilde{\gamma}^2}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\
&\leq \left(216\tilde{\gamma}^2\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 36\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\
&+ 72\tilde{\gamma}^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1} + 36\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} + 36\tilde{\gamma}^2\left(96 - \tfrac{64}{K}\right)\sqrt{2\pi}\, n(L_0^{\mathrm{imp}})^2.
\end{aligned} \tag{43}$$

The contribution of the term $6\|\nabla f^\eta(\hat{x}^{r-1})\|^2$ is

$$\tfrac{6\tilde{\gamma}^2}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] = 6\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right]. \tag{44}$$

Finally, applying equation 29 in equation 27, and then invoking equation 30, equation 31, equation 32, equation 43, and equation 44, together with $6\left(32 - \frac{16}{K}\right) + 36\left(96 - \frac{64}{K}\right) = 3648 - \frac{2400}{K}$, we obtain the result in part (ii). □

In the following definition, we introduce the notations used in the convergence analysis.

Definition A.21. Regarding Assumption 2.1, let $L_0^{\mathrm{imp}} := L_{0,y}^f L_{0,x}^y + L_{0,x}^f$. Moreover, let $\Theta := 2\frac{m}{s} + \frac{1}{24}$, $A_0 := 21(n+1)$, $A_1 := 126n + 108$, and $B := 6A_1 + 1296\Theta$. Define $N := \min\{\frac{\eta}{30\sqrt{3}\,nL_0^{\mathrm{imp}}}, \frac{\eta}{\sqrt{288}\,nL_0^{\mathrm{imp}}}\sqrt{\frac{s}{3m}}\}$ and $M := \max\{4A_0,\, 2\sqrt{B},\, \left(20736A_1\frac{m}{s}\right)^{1/4},\, \left(62208\frac{m}{s}\right)^{1/3}\}$.

In the next lemma, we compute an upper bound for the error in the control variate defined in Definition A.16.

Lemma A.22. Consider Algorithm 1 for addressing problem GFBO, and suppose that Assumption 2.1 holds. Suppose that $\mathcal{S}^r$ is sampled uniformly from all subsets of $[m]$ of cardinality s, and let $\mathcal{C}_r$ be defined as in Definition A.16. Then, for $K \geq 2$ and $r \geq 1$, we obtain

$$\begin{aligned}\mathcal{C}_r \leq{}& \left(1 - \tfrac{s}{2m} + 144\tilde{\gamma}^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\mathcal{C}_{r-1} \\ &+ \left(432\tilde{\gamma}^2\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\ &+ \left(12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 24\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} + 2\right)\mathcal{D}_r \\ &+ 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} + 12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\ &+ 12\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ &+ (7296 - \tfrac{4800}{K})\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{45}$$

Proof. For every $i \in [m]$ and $k \in \{1, \ldots, K\}$, we have

$$z_{i,k-1}^r - \hat{x}^r = z_{i,k-1}^r - \hat{x}^{r-1} + \hat{x}^{r-1} - \hat{x}^r.$$

Thus, using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^r\|^2\right] \leq 2\mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] + 2\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right].$$

Summing this inequality over $i \in [m]$ and $k \in \{1, \ldots, K\}$ and dividing both sides by Km gives

$$\begin{aligned}\mathcal{C}_r = \tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^r\|^2\right] \leq \tfrac{2}{Km}\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] \\ + \tfrac{2}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right] = 2\mathcal{D}_r + 2\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right],\end{aligned}$$

where the last equality follows from the definition of $\mathcal{D}_r$ and

$$\tfrac{1}{Km}\textstyle\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right] = \mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right].$$

Therefore,

$$\mathcal{C}_r \leq 2\mathcal{D}_r + 2\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right]. \tag{46}$$

Since $1 \leq s \leq m$ and $\mathcal{C}_{r-1} \geq 0$, we have $1 - \frac{s}{2m} \geq \frac{1}{2}$ and $(1 - \frac{s}{2m})\mathcal{C}_{r-1} \geq 0$. Adding this nonnegative term to the right-hand side of equation 46 gives

$$\mathcal{C}_r \leq (1 - \tfrac{s}{2m})\mathcal{C}_{r-1} + 2\mathcal{D}_r + 2\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right]. \tag{47}$$

Multiplying the inequality in Lemma A.20(ii) by 2 gives

$$\begin{aligned}2\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2\right] \leq{}& \left(12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 24\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\mathcal{D}_r \\ &+ \left(432\tilde{\gamma}^2\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\ &+ 144\tilde{\gamma}^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1} \\ &+ 12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r + 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} \\ &+ 12\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ &+ (7296 - \tfrac{4800}{K})\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{48}$$

Substituting equation 48 into equation 47 gives the desired result. □

Lemma A.23. Consider Algorithm 1 for addressing problem GFBO, and suppose that Assumption 2.1 holds. Suppose that $\mathcal{S}^r$ is sampled uniformly from all subsets of $[m]$ of cardinality s. Let $\tilde{\gamma} = K\gamma_l\gamma_g$, where $\gamma_g \geq 1$, and suppose that $\tilde{\gamma} \leq N$, where N is given in Definition A.21. Assume that the lower-level problem is solved inexactly and that ϵ_r satisfies $\mathbb{E}\left[\|y_{\epsilon_r}(\bullet) - y(\bullet)\|^2\right] \leq \epsilon_r$. Then, for $K \geq 2$ and $r \geq 1$, the following statements hold.

$$\begin{aligned}
\text{(i)}\quad \mathcal{C}_r &\leq (1 - \tfrac{s}{6m})\mathcal{C}_{r-1} + \left(432\tilde{\gamma}^2\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} + \left(2 + \tfrac{s}{24m}\right)\mathcal{D}_r \\
&\quad + 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} + 12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r + 12\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&\quad + (7296 - \tfrac{4800}{K})\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\text{(ii)}\quad \mathcal{D}_r &\leq 216\tilde{\gamma}^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1} + \left(648\tilde{\gamma}^2\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 108\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\
&\quad + 18\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r + 108\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} + 18\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&\quad + (10944 - \tfrac{6912}{K})\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.
\end{aligned}$$

Proof. (i) By Lemma A.14(iii), we have $L_{0,y}^f L_{0,x}^y \leq L_0^{\mathrm{imp}}$. Moreover, Lemma A.22 gives

$$\begin{aligned}
\mathcal{C}_r &\leq \left(1 - \tfrac{s}{2m} + 144\tilde{\gamma}^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\mathcal{C}_{r-1} \\
&\quad + \left(432\tilde{\gamma}^2\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2} + 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\
&\quad + \left(12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 24\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} + 2\right)\mathcal{D}_r \\
&\quad + 72\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} + 12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\
&\quad + 12\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&\quad + (7296 - \tfrac{4800}{K})\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.
\end{aligned} \tag{49}$$

The second stepsize condition in Definition A.21 gives $\tilde{\gamma} \leq \frac{\eta}{\sqrt{288}\,nL_0^{\mathrm{imp}}}\sqrt{\frac{s}{3m}}$. Squaring both sides and rearranging gives

$$\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} \leq \tfrac{s}{864m}. \tag{50}$$

Since $n \geq 1$, equation 50 gives $144\tilde{\gamma}^2\frac{n(L_0^{\mathrm{imp}})^2}{\eta^2} \leq 144\tilde{\gamma}^2\frac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} \leq \frac{s}{6m}$. Therefore,

$$1 - \tfrac{s}{2m} + 144\tilde{\gamma}^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2} \leq 1 - \tfrac{s}{3m} \leq 1 - \tfrac{s}{6m}. \tag{51}$$

Similarly, using $L_{0,y}^f L_{0,x}^y \leq L_0^{\mathrm{imp}}$ and equation 50, we obtain

$$12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 24\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} + 2 \leq 36\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} + 2 \leq 2 + \tfrac{s}{24m}. \tag{52}$$

Substituting equation 51 and equation 52 into equation 49 gives the result in part (i).

(ii) For every $k \in \{1, \ldots, K\}$, Lemma A.4 with $t = K - 1$ gives

$$\|z_{i,k}^r - \hat{x}^{r-1}\|^2 \leq \left(1 + \tfrac{1}{K-1}\right)\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2 + K\gamma_l^2\|g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} + \hat{c}^{r-1} - c_i^{r-1}\|^2. \tag{53}$$

We first consider $r = 1$. By the initialization of Algorithm 1, we have $c_i^0 = \hat{c}^0 = 0,\ i \in [m]$. Therefore, using $\|a + b + c + d\|^2 \leq 4\|a\|^2 + 4\|b\|^2 + 4\|c\|^2 + 4\|d\|^2$, Lemma A.19(iii)–(v), and equation 28, we obtain

$$\begin{aligned}
&\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|g_{i,k-1}^{\eta,1,z} + \hat{w}_{i,k-1}^{\eta,1,z} + \check{w}_{i,k-1}^{\eta,1,z}\|^2\right] \\
&\leq \left(4\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 8\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^1 - \hat{x}^0\|^2\right] \\
&\quad + 4\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1 + 4(L_0^{\mathrm{imp}})^2 + 128\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.
\end{aligned}$$

Using $L_{0,y}^f L_{0,x}^y \le L_0^{\mathrm{imp}}$, we obtain $4\frac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + 8\frac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} \le 12\frac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}$. Moreover, the first stepsize condition in Definition A.21 gives $\tilde{\gamma}^2\frac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} \le \frac{1}{2700}$. Since $\tilde{\gamma} = K\gamma_l\gamma_g$ and $\gamma_g \ge 1$, we have

$$12K\gamma_l^2\frac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} = \frac{12}{K\gamma_g^2}\tilde{\gamma}^2\frac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2} \le \frac{1}{225K} \le \frac{1}{225(K-1)}.$$

Substituting the preceding inequalities into equation 53 yields

$$\begin{aligned}\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k}^1 - \hat{x}^0\|^2\right] &\le \left(1 + \tfrac{1}{K-1} + \tfrac{1}{225(K-1)}\right)\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^1 - \hat{x}^0\|^2\right] \\ &\quad + K\gamma_l^2\left(4\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1 + 4(L_0^{\mathrm{imp}})^2 + 128\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2\right).\end{aligned}$$

For every $k \in \{1,\dots,K\}$,

$$\textstyle\sum_{j=0}^{k-1}\left(1 + \tfrac{1}{K-1} + \tfrac{1}{225(K-1)}\right)^j \le K\left(1 + \tfrac{226}{225(K-1)}\right)^{K-1} \le K\exp\left(\tfrac{226}{225}\right) < 3K.$$

Since $z_{i,0}^1 = \hat{x}^0$, recursively applying the preceding inequality gives

$$\begin{aligned}\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k}^1 - \hat{x}^0\|^2\right] &\le 3K^2\gamma_l^2\left(4\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1 + 4(L_0^{\mathrm{imp}})^2 + 128\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2\right) \\ &\le 3\tilde{\gamma}^2\left(4\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1 + 4(L_0^{\mathrm{imp}})^2 + 128\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2\right).\end{aligned}$$

Using Definition A.17, we obtain

$$\mathcal{D}_1 \le 12\tilde{\gamma}^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1 + 12\tilde{\gamma}^2(L_0^{\mathrm{imp}})^2 + 384\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.$$

For $K \ge 2$ and $n \ge 1$, we have $12 + 384\sqrt{2\pi}n \le 396\sqrt{2\pi}n \le \left(10944 - \frac{6912}{K}\right)\sqrt{2\pi}n$. Moreover, $12\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1 \le 18\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_1$, and $108\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_0 + 18\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^0)\|^2\right] \ge 0$. Using $\mathcal{C}_0 = \mathcal{D}_0 = 0$, the preceding inequalities give the result in part (ii) for $r = 1$.

We next consider $r \ge 2$. Adding and subtracting $\nabla f_i^\eta(\hat{x}^{r-1})$ and $\nabla f^\eta(\hat{x}^{r-1})$ into equation 53, and invoking Lemma A.5, we obtain

$$\begin{aligned}&\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} + \hat{c}^{r-1} - c_i^{r-1}\|^2\right] \\ &\le \tfrac{6}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right] + \tfrac{6}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r,z}\|^2\right] \\ &+ \tfrac{6}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\ &+ \tfrac{6}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\ &+ 6\mathbb{E}\left[\|\hat{c}^{r-1} - \nabla f^\eta(\hat{x}^{r-1})\|^2\right] + 6\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right].\end{aligned} \tag{54}$$

Lemma A.19(iv) gives

$$\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|\hat{w}_{i,k-1}^{\eta,r,z}\|^2\right] \le \tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right]. \tag{55}$$

Lemma A.19(v) gives

$$\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|\check{w}_{i,k-1}^{\eta,r,z}\|^2\right] \le \tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r. \tag{56}$$

Moreover, equation 11 gives

$$\begin{aligned}&\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \\ &\le 2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r - \hat{x}^{r-1}\|^2\right] + 32\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{57}$$

By Jensen's inequality,

$$\mathbb{E}\left[\|\hat{c}^{r-1} - \nabla f^\eta(\hat{x}^{r-1})\|^2\right] \le \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1} - \nabla f_i^\eta(\hat{x}^{r-1})\|^2\right]. \tag{58}$$

Using Lemma A.18, Definition A.15, Lemma A.19(iii)–(v), and Lemma A.1(ii.3), we obtain

$$\begin{aligned}\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|c_i^{r-1}-\nabla f_i^\eta(\hat{x}^{r-1})\|^2\right] \le \left(18\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+3\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\ +6\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+3\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} \\ +3(96-\tfrac{64}{K})\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{59}$$

Substituting equation 55, equation 56, equation 57, equation 58, and equation 59 into equation 54 gives

$$\begin{aligned}&\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|g_{i,k-1}^{\eta,r,z}+\hat{w}_{i,k-1}^{\eta,r,z}+\check{w}_{i,k-1}^{\eta,r,z}+\hat{c}^{r-1}-c_i^{r-1}\|^2\right] \\ &\le \left(6\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2+12\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r-\hat{x}^{r-1}\|^2\right] \\ &+\left(216\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+36\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\ &+72\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+6\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\ &+36\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1}+6\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ &+(3648-\tfrac{2304}{K})\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{60}$$

Averaging equation 53 over $i\in[m]$ and taking expectations, and then invoking equation 60, gives

$$\begin{aligned}\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k}^r-\hat{x}^{r-1}\|^2\right] \le \left(1+\tfrac{1}{K-1}+K\gamma_l^2\left(6\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2+12\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\right) \\ \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r-\hat{x}^{r-1}\|^2\right] \\ +K\gamma_l^2\left(216\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+36\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\ +72K\gamma_l^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+6K\gamma_l^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\ +36K\gamma_l^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1}+6K\gamma_l^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ +K\gamma_l^2(3648-\tfrac{2304}{K})\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{61}$$

Using $L_{0,y}^f L_{0,x}^y \le L_0^{\mathrm{imp}}$, we obtain $K\gamma_l^2\left(6\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2+12\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\right) \le 18K\gamma_l^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}$. The first stepsize condition in Definition A.21 gives $\tilde{\gamma}\le\tfrac{\eta}{30\sqrt{3}\,nL_0^{\mathrm{imp}}}$, and therefore $\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\le\tfrac{1}{2700}$. Therefore, using $\tilde{\gamma}=K\gamma_l\gamma_g$, $\gamma_g\ge 1$, and $\tilde{\gamma}^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\le\tfrac{1}{2700}$., we obtain

$$18K\gamma_l^2\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}=18\tfrac{\tilde{\gamma}^2}{K\gamma_g^2}\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\le 18\tfrac{\tilde{\gamma}^2}{K}\tfrac{n^2(L_0^{\mathrm{imp}})^2}{\eta^2}\le\tfrac{1}{150K}\le\tfrac{1}{150(K-1)},$$

where the last inequality follows from $K\ge 2$. Substituting this bound into equation 61 gives

$$\begin{aligned}\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k}^r-\hat{x}^{r-1}\|^2\right] \le \left(1+\tfrac{1}{K-1}+\tfrac{1}{150(K-1)}\right)\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k-1}^r-\hat{x}^{r-1}\|^2\right] \\ +K\gamma_l^2\left(216\tfrac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+36\tfrac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\ +72K\gamma_l^2\tfrac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+6K\gamma_l^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\ +36K\gamma_l^2\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1}+6K\gamma_l^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ +K\gamma_l^2(3648-\tfrac{2304}{K})\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{62}$$

Since $z_{i,0}^r = \hat{x}^{r-1}$, recursively applying equation 62 gives, for every $k \in \{1, \dots, K\}$,

$$\begin{aligned}
&\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k}^r - \hat{x}^{r-1}\|^2\right] \\
&\le K\gamma_l^2 \textstyle\sum_{j=0}^{k-1}\left(1+\frac{1}{K-1}+\frac{1}{150(K-1)}\right)^j \left(\left(216\frac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+36\frac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1}\right. \\
&+72\frac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+6\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r+36\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1} \\
&\left.+6\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right]+(3648-\tfrac{2304}{K})\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2\right).
\end{aligned} \tag{63}$$

For $K \ge 2$ and $k \le K$, we have

$$\textstyle\sum_{j=0}^{k-1}\left(1+\frac{1}{K-1}+\frac{1}{150(K-1)}\right)^j \le \sum_{j=0}^{K-1}\left(1+\frac{1}{K-1}+\frac{1}{150(K-1)}\right)^j \le 3K. \tag{64}$$

Combining equation 63 and equation 64 gives, for every $k \in \{1, \dots, K\}$,

$$\begin{aligned}
\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k}^r - \hat{x}^{r-1}\|^2\right] &\le 3K^2\gamma_l^2\left(216\frac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+36\frac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\
&+216K^2\gamma_l^2\frac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+18K^2\gamma_l^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\
&+108K^2\gamma_l^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1}+18K^2\gamma_l^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&+3K^2\gamma_l^2(3648-\tfrac{2304}{K})\sqrt{2\pi}n(L_0^{\mathrm{imp}})^2.
\end{aligned}$$

Using $K^2\gamma_l^2 = \frac{\tilde{\gamma}^2}{\gamma_g^2} \le \tilde{\gamma}^2$, we obtain, for every $k \in \{1, \dots, K\}$,

$$\begin{aligned}
\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\|z_{i,k}^r - \hat{x}^{r-1}\|^2\right] &\le \left(648\tilde{\gamma}^2\frac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+108\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1} \\
&+216\tilde{\gamma}^2\frac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+18\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r \\
&+108\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1}+18\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&+(10944-\tfrac{6912}{K})\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2.
\end{aligned} \tag{65}$$

Since $z_{i,0}^r = \hat{x}^{r-1}$, the definition of $\mathcal{D}_r$ gives

$$\mathcal{D}_r = \tfrac{1}{K}\textstyle\sum_{j=0}^{K-1}\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\|z_{i,j}^r - \hat{x}^{r-1}\|^2\right]\right) = \frac{1}{K}\sum_{j=1}^{K-1}\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\|z_{i,j}^r - \hat{x}^{r-1}\|^2\right]\right).$$

Applying equation 65 to every $j \in \{1, \dots, K-1\}$ gives

$$\begin{aligned}
\mathcal{D}_r \le \tfrac{K-1}{K}&\left(216\tilde{\gamma}^2\frac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{C}_{r-1}+\left(648\tilde{\gamma}^2\frac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}+108\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2\right)\mathcal{D}_{r-1}\right. \\
&+18\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r+108\tilde{\gamma}^2\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_{r-1}+18\tilde{\gamma}^2\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&\left.+(10944-\tfrac{6912}{K})\sqrt{2\pi}\tilde{\gamma}^2 n(L_0^{\mathrm{imp}})^2\right).
\end{aligned}$$

Since $\frac{K-1}{K} \le 1$, the result in part (ii) follows. □

In the next lemma, we show the progress at each round r.

Lemma A.24. Consider Algorithm 1 for addressing problem GFBO, and suppose that Assumption 2.1 holds. Let $\tilde{\gamma} := K\gamma_g\gamma_l$, where $\gamma_g \ge 1$, and suppose that $0 < \tilde{\gamma} \le N$ and $\frac{\eta}{\tilde{\gamma}} > L_0^{\mathrm{imp}}\sqrt{n}\,M$, where N and M are given in Definition A.21. Suppose that $\epsilon_r \le \epsilon_{r-1}$. Then, for $r \ge 1$, we have

$$\begin{aligned}
&\left(\mathbb{E}\left[f^\eta(\hat{x}^r)\right]+y\tilde{\gamma}^2\left(\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\frac{m}{s}\mathcal{C}_r+u\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\mathcal{D}_r\right) \\
&\le\left(\mathbb{E}\left[f^\eta(\hat{x}^{r-1})\right]+y\tilde{\gamma}^2\left(\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\frac{m}{s}\mathcal{C}_{r-1}+u\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\mathcal{D}_{r-1}\right) \\
&+(c+g+o)\epsilon_{r-1}+\underbrace{(w-d+p)}_{<0}\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\
&+\sqrt{2\pi}\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2\left(1824-\tfrac{1200}{K}+y\tfrac{m}{s}\rho^2\left(7296-\tfrac{4800}{K}\right)+\tfrac{y-216}{1296}\left(10944-\tfrac{6912}{K}\right)\right),
\end{aligned} \tag{66}$$

where $\rho := \tilde{\gamma}\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}$, $y := \frac{216\left[1+A_0\rho+54n\rho^2\right]}{1-B\rho^2-5184A_1\frac{m}{s}\rho^4}$, $u := \frac{A_1 y\rho^2}{216}\left(1+864\frac{m}{s}\rho^2\right)$, $c := 21\tilde{\gamma}^2\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\frac{n^2}{\eta^2}(L_{0,y}^f)^2 + \frac{3\tilde{\gamma}}{2}\frac{n^2}{\eta^2}(L_{0,y}^f)^2$, $g := 84y\tilde{\gamma}^4\left(\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\frac{m}{s}\frac{n^2}{\eta^2}(L_{0,y}^f)^2$, $o := \frac{7(y-216)}{72}\tilde{\gamma}^2\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\frac{n^2}{\eta^2}(L_{0,y}^f)^2$, $w := 12y\tilde{\gamma}^4\left(\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\frac{m}{s}$, $d := \frac{\tilde{\gamma}(K+1)}{4K} - 3\tilde{\gamma}^2\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}$, and $p := \frac{y-216}{72}\tilde{\gamma}^2\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}$.

Proof. By Lemma A.14(iii) and Lemma A.1(ii.3), f^η is $\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}$-smooth. Therefore,

$$f^\eta(\hat{x}^r) \le f^\eta(\hat{x}^{r-1}) + \langle \nabla f^\eta(\hat{x}^{r-1}), \hat{x}^r - \hat{x}^{r-1}\rangle + \frac{L_0^{\mathrm{imp}}\sqrt{n}}{2\eta}\|\hat{x}^r - \hat{x}^{r-1}\|^2.$$

Taking the conditional expectation given $\mathcal{F}_{r-1,K}$ gives

$$\begin{aligned}\mathbb{E}\left[f^\eta(\hat{x}^r) \mid \mathcal{F}_{r-1,K}\right] \le f^\eta(\hat{x}^{r-1}) + \langle \nabla f^\eta(\hat{x}^{r-1}), \mathbb{E}\left[\hat{x}^r \mid \mathcal{F}_{r-1,K}\right] - \hat{x}^{r-1}\rangle \\ + \frac{L_0^{\mathrm{imp}}\sqrt{n}}{2\eta}\mathbb{E}\left[\|\hat{x}^r - \hat{x}^{r-1}\|^2 \mid \mathcal{F}_{r-1,K}\right].\end{aligned} \tag{67}$$

Regarding the update of $\hat{x}^r$, the control-variate terms cancel after taking the conditional expectation over the sampled clients. Hence,

$$\mathbb{E}\left[\hat{x}^r \mid \mathcal{F}_{r-1,K}\right] - \hat{x}^{r-1} = -\frac{\tilde{\gamma}}{Km}\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[g_{i,k-1}^{\eta,r,z} + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right].$$

By the tower property and Lemma A.19(i), for $k = 1$, we have

$$\mathbb{E}\left[g_{i,0}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right] = \nabla f_i^\eta(\hat{x}^{r-1}),$$

whereas, for $k \in \{2,\dots,K\}$,

$$\mathbb{E}\left[g_{i,k-1}^{\eta,r,z} \mid \mathcal{F}_{r-1,K}\right] = \mathbb{E}\left[\nabla f_i^\eta(z_{i,k-1}^r) + b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right].$$

Adding and subtracting $\nabla f_i^\eta(\hat{x}^{r-1})$, using $\frac{1}{m}\sum_{i=1}^m \nabla f_i^\eta(\hat{x}^{r-1}) = \nabla f^\eta(\hat{x}^{r-1})$, and invoking $-\langle a,c\rangle \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|c\|^2$, $\|a+b+c\|^2 \le 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, Jensen's inequality, Lemma A.1(ii.3), and Lemma A.19(iv)–(v), we obtain

$$\begin{aligned}&-\frac{\tilde{\gamma}}{Km}\sum_{k=1}^K\sum_{i=1}^m \mathbb{E}\left[\langle \nabla f^\eta(\hat{x}^{r-1}), \nabla f_i^\eta(z_{i,k-1}^r) + \hat{w}_{i,k-1}^{\eta,r,z} + \check{w}_{i,k-1}^{\eta,r,z}\rangle\right] \\ &\le \frac{3\tilde{\gamma}}{2}\left(\frac{n^2}{\eta^2}(L_{0,y}^f L_{0,x}^y)^2 + \frac{n(L_0^{\mathrm{imp}})^2}{\eta^2}\right)\mathcal{D}_r + \frac{3\tilde{\gamma}}{2}\frac{n^2}{\eta^2}(L_{0,y}^f)^2\epsilon_r - \frac{\tilde{\gamma}}{2}\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right].\end{aligned}$$

We next consider the additional bias term. Since $\nabla f^\eta(\hat{x}^{r-1})$ is $\mathcal{F}_{r-1,K}$-measurable, we have

$$\mathbb{E}\left[\langle \nabla f^\eta(\hat{x}^{r-1}), b_{i,k-1}^r\rangle\right] = \mathbb{E}\left[\langle \nabla f^\eta(\hat{x}^{r-1}), \mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\rangle\right].$$

Using $-\langle a,c\rangle \le \frac{1}{4}\|a\|^2 + \|c\|^2$ and Lemma A.19(i), we obtain

$$\begin{aligned}&-\frac{\tilde{\gamma}}{Km}\sum_{k=2}^K\sum_{i=1}^m \mathbb{E}\left[\langle \nabla f^\eta(\hat{x}^{r-1}), b_{i,k-1}^r\rangle\right] \\ &= -\frac{\tilde{\gamma}}{Km}\sum_{k=2}^K\sum_{i=1}^m \mathbb{E}\left[\langle \nabla f^\eta(\hat{x}^{r-1}), \mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\rangle\right] \\ &\le \frac{\tilde{\gamma}(K-1)}{4K}\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ &+ \frac{\tilde{\gamma}}{Km}\sum_{k=2}^K\sum_{i=1}^m \mathbb{E}\left[\|\mathbb{E}\left[b_{i,k-1}^r \mid \mathcal{F}_{r-1,K}\right]\|^2\right] \\ &\le \frac{\tilde{\gamma}(K-1)}{4K}\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] + 2\tilde{\gamma}\frac{n(n+1)(L_0^{\mathrm{imp}})^2}{\eta^2}\mathcal{D}_r.\end{aligned}$$

Substituting the preceding two inequalities into equation 67, taking total expectation, invoking Lemma A.20(ii), using $L_{0,y}^f L_{0,x}^y \le L_0^{\mathrm{imp}}$ and $\epsilon_r \le \epsilon_{r-1}$, we obtain

$$\begin{aligned}\mathbb{E}\left[f^\eta(\hat{x}^r)\right] \le \mathbb{E}\left[f^\eta(\hat{x}^{r-1})\right] + \frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\left(\frac{7}{2}(n+1)\rho + 9n\rho^2\right)\mathcal{D}_r \\ + A_1\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\rho^2\mathcal{D}_{r-1} + 36\tilde{\gamma}^2\left(\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\mathcal{C}_{r-1} \\ + c\epsilon_{r-1} - d\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ + \left(1824 - \frac{1200}{K}\right)\sqrt{2\pi}\tilde{\gamma}^2\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{68}$$

We next verify that y and u are well defined and positive. By the stepsize condition and the definition of ρ, we have

$$\rho < \min\left\{\tfrac{1}{4A_0}, \tfrac{1}{2\sqrt{B}}, \left(20736A_1\tfrac{m}{s}\right)^{-1/4}, \left(62208\tfrac{m}{s}\right)^{-1/3}\right\}. \tag{69}$$

Consequently, $A_0\rho < \frac{1}{4}$, $B\rho^2 < \frac{1}{4}$, $5184A_1\frac{m}{s}\rho^4 < \frac{1}{4}$. Thus, $1-B\rho^2-5184A_1\frac{m}{s}\rho^4 > 1-\frac{1}{4}-\frac{1}{4} = \frac{1}{2} > 0$, and hence y is well defined. Moreover, $1-B\rho^2-5184A_1\frac{m}{s}\rho^4 < 1$ and $1+A_0\rho+54n\rho^2 > 1$, which give $y = \frac{216\left[1+A_0\rho+54n\rho^2\right]}{1-B\rho^2-5184A_1\frac{m}{s}\rho^4} > 216$. Since $A_0 = 21(n+1)$ and $\rho < \frac{1}{4A_0}$, we also have $54n\rho^2 < \frac{54n}{16A_0^2} = \frac{54n}{16\cdot 21^2(n+1)^2} < \frac{1}{4}$. Therefore,

$$216 < y = \frac{216\left(1+A_0\rho+54n\rho^2\right)}{1-B\rho^2-5184A_1\frac{m}{s}\rho^4} < \frac{216\left(1+\frac{1}{4}+\frac{1}{4}\right)}{\frac{1}{2}} = 648. \tag{70}$$

Furthermore, $u = \frac{A_1 y\rho^2}{216}\left(1+864\frac{m}{s}\rho^2\right) > 0$, since $A_1 > 0$, $y > 0$, and $\rho > 0$.

Multiplying the inequality in Lemma A.23(i) by $y\tilde{\gamma}^2\left(\frac{L_0^{\rm imp}\sqrt{n}}{\eta}\right)^3\frac{m}{s} > 0$, using $\frac{m}{s}\left(2+\frac{s}{24m}\right) = \frac{2m}{s}+\frac{1}{24} = \Theta$ and $\epsilon_r \le \epsilon_{r-1}$, we obtain

$$\begin{aligned} y\tilde{\gamma}^2\left(\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\right)^3\tfrac{m}{s}\mathcal{C}_r \le{}& y\tilde{\gamma}^2\left(\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\right)^3\tfrac{m}{s}\mathcal{C}_{r-1} - \tfrac{y}{6}\tilde{\gamma}^2\left(\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\right)^3\mathcal{C}_{r-1} \\ &+ y\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\Theta\rho^2\mathcal{D}_r + 4A_1 y\tfrac{m}{s}\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\rho^4\mathcal{D}_{r-1} \\ &+ g\epsilon_{r-1} + w\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ &+ y\tfrac{m}{s}\rho^2\left(7296-\tfrac{4800}{K}\right)\sqrt{2\pi}\tilde{\gamma}^2\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}n(L_0^{\rm imp})^2. \end{aligned} \tag{71}$$

Regarding the definitions of u and y, direct calculation gives

$$u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2 = \tfrac{y-216}{1296} \tag{72}$$

and

$$u = A_1\rho^2\left(1+4y\tfrac{m}{s}\rho^2+6\left(u+\tfrac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right)\right). \tag{73}$$

By equation 72 and $y > 216$, we obtain $\frac{L_0^{\rm imp}\sqrt{n}}{\eta}\left(u+\frac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right) = \frac{L_0^{\rm imp}\sqrt{n}}{\eta}\frac{y-216}{1296} > 0$. Therefore, multiplying the inequality in Lemma A.23(ii) by this positive quantity and using $\epsilon_r \le \epsilon_{r-1}$ give

$$\begin{aligned} &\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\left(u+\tfrac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right)\mathcal{D}_r \\ &\le 216\tilde{\gamma}^2\left(\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\right)^3\left(u+\tfrac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right)\mathcal{C}_{r-1} \\ &+ 6A_1\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\rho^2\left(u+\tfrac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right)\mathcal{D}_{r-1} \\ &+ 126\tilde{\gamma}^2\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\left(u+\tfrac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right)\epsilon_{r-1} \\ &+ 18\tilde{\gamma}^2\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}\left(u+\tfrac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right)\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \\ &+ \left(u+\tfrac{7}{2}(n+1)\rho+(9n+y\Theta)\rho^2\right)\left(10944-\tfrac{6912}{K}\right)\sqrt{2\pi}\tilde{\gamma}^2\tfrac{L_0^{\rm imp}\sqrt{n}}{\eta}n(L_0^{\rm imp})^2. \end{aligned} \tag{74}$$

Adding equation 68 and equation 71, adding $u\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\mathcal{D}_r$ to both sides, applying equation 74, and adding and subtracting $u\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\mathcal{D}_{r-1}$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[f^{\eta}(\hat{x}^r)\right] + y\tilde{\gamma}^2\left(\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\tfrac{m}{s}\mathcal{C}_r + u\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\mathcal{D}_r \qquad (75)\\
&\leq \mathbb{E}\left[f^{\eta}(\hat{x}^{r-1})\right] + y\tilde{\gamma}^2\left(\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\tfrac{m}{s}\mathcal{C}_{r-1} + u\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\mathcal{D}_{r-1}\\
&+ \tilde{\gamma}^2\left(\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\right)^3\left(36 - \tfrac{y}{6} + 216\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\right)\mathcal{C}_{r-1}\\
&+ \tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\left(A_1\rho^2 + 4A_1 y\tfrac{m}{s}\rho^4 - u + 6A_1\rho^2\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\right)\mathcal{D}_{r-1}\\
&+ \left(c + g + 126\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\right)\epsilon_{r-1}\\
&+ \left(w - d + 18\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\right)\mathbb{E}\left[\|\nabla f^{\eta}(\hat{x}^{r-1})\|^2\right]\\
&+ \sqrt{2\pi}\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2\left(1824 - \tfrac{1200}{K} + y\tfrac{m}{s}\rho^2\left(7296 - \tfrac{4800}{K}\right)\right.\\
&\left.+ \left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\left(10944 - \tfrac{6912}{K}\right)\right).
\end{aligned}
$$

By equation 72, we have $36 - \frac{y}{6} + 216\left(u + \frac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right) = 36 - \frac{y}{6} + 216\frac{y-216}{1296} = 0$. Moreover, equation 73 gives

$$
\begin{aligned}
&A_1\rho^2 + 4A_1 y\tfrac{m}{s}\rho^4 - u + 6A_1\rho^2\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\\
&= A_1\rho^2\left(1 + 4y\tfrac{m}{s}\rho^2 + 6\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\right) - u = 0.
\end{aligned}
$$

Also, using equation 72, we have

$$
\begin{aligned}
&126\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\\
&= 126\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2\tfrac{y-216}{1296} = \tfrac{7(y-216)}{72}\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\tfrac{n^2}{\eta^2}(L_{0,y}^f)^2 = o
\end{aligned}
$$

and

$$
18\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right) = 18\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\tfrac{y-216}{1296} = \tfrac{y-216}{72}\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta} = p.
$$

Finally,

$$
\left(u + \tfrac{7}{2}(n+1)\rho + (9n+y\Theta)\rho^2\right)\left(10944 - \tfrac{6912}{K}\right) = \tfrac{y-216}{1296}\left(10944 - \tfrac{6912}{K}\right).
$$

We next show that $w - d + p < 0$. By the definitions of w, d, and p, we obtain

$$
\tfrac{w-d+p}{\tilde{\gamma}} = -\tfrac{K+1}{4K} + 3\rho + \tfrac{y-216}{72}\rho + 12y\tfrac{m}{s}\rho^3 = -\tfrac{K+1}{4K} + \tfrac{y}{72}\rho + 12y\tfrac{m}{s}\rho^3. \qquad (76)
$$

By equation 70, equation 69, and $A_0 = 21(n+1) \geq 42$, we have $\frac{y}{72}\rho < 9\rho < \frac{9}{4A_0} \leq \frac{9}{168} < \frac{1}{8}$. Furthermore, $62208\frac{m}{s}\rho^3 < 1$ implies $648 \cdot 12\frac{m}{s}\rho^3 < \frac{648\cdot 12}{62208} = \frac{1}{8}$. Since $y < 648$, we have $12y\frac{m}{s}\rho^3 < 648 \cdot 12\frac{m}{s}\rho^3 < \frac{1}{8}$. Therefore, $\frac{y}{72}\rho + 12y\frac{m}{s}\rho^3 < \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$. Since $K \geq 2$, $\frac{K+1}{4K} = \frac{1}{4} + \frac{1}{4K} > \frac{1}{4}$. Thus, equation 76 gives $\frac{w-d+p}{\tilde{\gamma}} < -\frac{K+1}{4K} + \frac{1}{4} < 0$. Since $\tilde{\gamma} > 0$, we obtain $w - d + p < 0$.

Substituting the preceding equalities and $w - d + p < 0$ into equation 75 gives equation 66. □

Lemma A.25. [Sub-linear convergence rate (Karimireddy et al., 2020, Lemma 2)] For every non-negative sequence $\{u_{r-1}\}_{r\geq 1}$ and any parameters $\gamma_{\max} \geq 0$, $c \geq 0$, and $R \geq 0$, there exists a constant stepsize $0 \leq \gamma \leq \gamma_{\max}$ such that

$$
\phi_R := \tfrac{1}{R+1}\textstyle\sum_{r=1}^{R+1}\left(\tfrac{u_{r-1}}{\gamma} - \tfrac{u_r}{\gamma} + \lambda_1\gamma + \lambda_2\gamma^2\right) \leq \tfrac{u_0}{\gamma_{\max}(R+1)} + \tfrac{2\sqrt{\lambda_1 u_0}}{\sqrt{R+1}} + 2\left(\tfrac{u_0}{R+1}\right)^{\frac{2}{3}}\lambda_2^{\frac{1}{3}}. \qquad (77)
$$

Lemma A.26. For any $R \geq 1$, the following statements hold.

$$\text{(i)} \quad \sum_{r=1}^{R+1} \frac{1}{r} \leq 1 + \log(R+1),$$
$$\text{(ii)} \quad \sum_{r=1}^{R+1} \frac{1}{\sqrt{r}} \leq 2\sqrt{R+1} - 1 \leq 2\sqrt{R+1}.$$

Proof. (i) Since $g(x) = 1/x$ is decreasing for $0 < x$, we obtain

$$\sum_{r=1}^{R+1} \frac{1}{r} = 1 + \sum_{r=2}^{R+1} \frac{1}{r} \leq 1 + \int_1^{R+1} \frac{1}{x}\, dx = 1 + \log(R+1).$$

(ii) Similarly, since $g(x) = 1/\sqrt{x}$ is decreasing for $0 < x$, we obtain

$$\sum_{r=1}^{R+1} \frac{1}{\sqrt{r}} = 1 + \sum_{r=2}^{R+1} \frac{1}{\sqrt{r}} \leq 1 + \int_1^{R+1} \frac{1}{\sqrt{x}}\, dx = 2\sqrt{R+1} - 1 \leq 2\sqrt{R+1}.$$

□

A.5 Constrained lower-level problem

In this section, we propose a novel Algorithm 3, which is called Adaptive Penalty SCAFFOLD, and its convergence analysis to address a constrained FL problem with a heterogeneous dataset.

A.5.1 Algorithm outline

We present the compact version of Algorithm 3 in Algorithm 2. Algorithm 3 summarizes Adaptive Penalty SCAFFOLD for a fixed outer query x. At each communication round, the server samples a subset of clients and broadcasts the current global iterate and control variate. Each selected client initializes from the global iterate and performs K^l local stochastic updates using the penalized objective together with the SCAFFOLD correction. The penalty parameter ρ_r^l progressively enforces feasibility, while the round-dependent local stepsize $\gamma_{l,r}^l$ adapts to the increasing penalty and the control variates mitigate client drift. After the local updates, selected clients update their control variates and return the model and control-variate changes to the server. The server aggregates these changes to update the global iterate and global control variate, while unselected clients keep their previous control variates.

Algorithm 3 Adaptive Penalty SCAFFOLD

Require: Fixed outer query x; server initialization $\widehat{y}^1$; controls $\widehat{c}^1 = 0$ and $c_i^1 = 0$ for $i \in [m]$; local stepsizes $\{\gamma_{l,r}^l\}_{r=1}^{R^l}$; global stepsize γ_g^l; penalties $\{\rho_r^l\}_{r=1}^{R^l}$.

for $r = 1, \ldots, R^l$ **do**
  Server samples $S^{l,r} \subseteq [m]$, $|S^{l,r}| = s^l$, and broadcasts $\widehat{y}^r, \widehat{c}^r$.
  for all $i \in S^{l,r}$ **in parallel do**
    Initialize $y_{i,0}^r = \widehat{y}^r$.
    for $k = 0, \ldots, K^l - 1$ **do**
      Sample $\zeta_{i,k}^r \sim \widetilde{D}_i$ and compute
      $y_{i,k+1}^r = y_{i,k}^r - \gamma_{l,r}^l(\nabla_y \widetilde{h}_i(x, y_{i,k}^r, \zeta_{i,k}^r) + \rho_r^l\big(y_{i,k}^r - \Pi_{Y(x)}(y_{i,k}^r)\big) - c_i^r + \widehat{c}^r)$.
    end for
    $c_i^{r+1} = c_i^r - \widehat{c}^r + (K^l \gamma_{l,r}^l)^{-1}(\widehat{y}^r - y_{i,K^l}^r)$.
    $\Delta y_i^r = y_{i,K^l}^r - \widehat{y}^r$ and $\Delta c_i^r = c_i^{r+1} - c_i^r$.
  end for
  Server aggregates $\widehat{y}^{r+1} = \widehat{y}^r + \frac{\gamma_g^l}{s^l} \sum_{i \in S^{l,r}} \Delta y_i^r$ and $\widehat{c}^{r+1} = \widehat{c}^r + \frac{1}{m} \sum_{i \in S^{l,r}} \Delta c_i^r$.
  Unselected clients retain $c_i^{r+1} = c_i^r$.
end for
return $\hat{y}^{R^l+1}$

A.5.2 Convergence analysis

In this section, we establish the results needed to prove Theorem A.30. We present a compact version of Theorem A.30 in Theorem 3.1.

Lemma A.27. For every $y \in \mathbb{R}^{\tilde{n}}$,

$$h(x,y) - h(x,y(x)) \geq \tfrac{\mu_h}{2}\|y - y(x)\|^2 - G^*\sqrt{2g_x(y)}. \tag{78}$$

Consequently, for every $\rho > 0$,

$$2[h(x,y) - h(x,y(x))] + \rho g_x(y) \geq \mu_h\|y - y(x)\|^2 - \tfrac{2(G^*)^2}{\rho}. \tag{79}$$

Proof. Let $z = \Pi_{Y(x)}(y)$. The first-order optimality condition over the convex set $Y(x)$ gives $\langle \nabla_y h(x,y(x)), z - y(x)\rangle \geq 0$. Thus

$$\langle \nabla_y h(x,y(x)), y - y(x)\rangle \geq -G^*\|y - z\| = -G^*\sqrt{2g_x(y)}.$$

Combining this inequality with strong convexity proves equation 78. To obtain equation 79, use

$$\rho g_x(y) - 2G^*\sqrt{2g_x(y)} = \rho\left(\sqrt{g_x(y)} - \tfrac{\sqrt{2}G^*}{\rho}\right)^2 - \tfrac{2(G^*)^2}{\rho}.$$

□

For the analysis, define

$$e_r := \mathbb{E}[\|\widehat{y}^r - y(x)\|^2], \qquad p_r := \mathbb{E}[g_x(\widehat{y}^r)],$$
$$D_r^l := \tfrac{1}{mK^l}\textstyle\sum_{i=1}^m \sum_{k=0}^{K^l-1} \mathbb{E}[\|y_{i,k}^r - \widehat{y}^r\|^2], \qquad C_r^l := \tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}[\|c_i^r - \nabla_y h_i(x,y(x))\|^2].$$

and $\widetilde{g}_{i,k}^r = \widetilde{\nabla_y h_i}(x, y_{i,k}^r, \zeta_{i,k}^r) + \rho_r^l\big(y_{i,k}^r - \Pi_{Y(x)}(y_{i,k}^r)\big)$.

Lemma A.28. Define

$$\Phi_r := e_r + \tfrac{24m(\gamma_r^l)^2}{s^l} C_r^l$$

and

$$\mathcal{R}_r := (\gamma_r^l)^2\left(\left(8 + \tfrac{12A_r^l}{s^l}\right)(G^*)^2 + \left(48 + \tfrac{8}{s^l} + \tfrac{4A_r^l}{s^l}\right)\tfrac{\sigma^2}{K^l}\right).$$

Then

$$\Phi_{r+1} \leq \left(1 - \tfrac{\mu_h\gamma_r^l}{2} + B_r^l\right)\Phi_r - 2\gamma_r^l\mathbb{E}[h(x,\widehat{y}^r) - h(x,y(x))] - \gamma_r^l\rho_r^l p_r + \mathcal{R}_r. \tag{80}$$

Proof. Let the following conditions hold, and γ_r^l is nonincreasing.

$$\begin{aligned} K^l\gamma_{l,r}^l L_r^l \leq 2, \qquad & 0 < \mu_h\gamma_r^l < 2, \\ A_r^l + \mu_h m\gamma_r^l \leq s^l, \quad & \gamma_r^l\rho_r^l\left(400 + \tfrac{48A_r^l}{s^l}\right) \leq 1. \end{aligned} \tag{81}$$

We first record the gradient bound

$$\tfrac{1}{m}\textstyle\sum_{i=1}^m \mathbb{E}\left[\left\|\nabla_y F_{i,r}^l(x,\widehat{y}^r) - \nabla_y h_i(x,y(x))\right\|^2\right] \leq 2L_{h1,y}^2 e_r + 4(\rho_r^l)^2 p_r. \tag{82}$$

By the definition of $F_{i,r}^l$,

$$\nabla_y F_{i,r}^l(x,\widehat{y}^r) - \nabla_y h_i(x,y(x))$$

is the sum of the lower-level gradient displacement

$$\nabla_y h_i(x,\widehat{y}^r) - \nabla_y h_i(x,y(x))$$

and the penalty-gradient term. Applying $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the $L_{h1,y}$-smoothness of h_i, and the standard projection properties of the penalty function gives equation 82.

We now establish three inequalities that will be used in the one-round recursion.

Step 1

Let

$$\epsilon_{i,k}^r := \widetilde{g}_{i,k}^r - \nabla_y F_{i,r}^l(x, y_{i,k}^r).$$

By the unbiased stochastic-gradient assumption,

$$\mathbb{E}[\epsilon_{i,k}^r \mid \mathcal{F}_{r,k}] = 0, \qquad \mathbb{E}[\|\epsilon_{i,k}^r\|^2 \mid \mathcal{F}_{r,k}] \le \sigma^2.$$

Moreover, stochastic-gradient noises are conditionally independent across clients, and noise terms generated at different local iterations form martingale differences. Hence all cross terms vanish, and therefore

$$\mathbb{E}\left[\left\|\tfrac{1}{s^l K^l}\sum_{i\in S^{l,r}}\sum_{k=0}^{K^l-1}\epsilon_{i,k}^r\right\|^2\right] \le \tfrac{\sigma^2}{s^l K^l}. \tag{83}$$

From the server update,

$$\widehat{y}^{r+1} - \widehat{y}^r = -\tfrac{\gamma_r^l}{s^l K^l}\sum_{i\in S^{l,r}}\sum_{k=0}^{K^l-1}\left(\widetilde{g}_{i,k}^r - c_i^r + \widehat{c}^r\right).$$

Adding and subtracting $\nabla_y F_{i,r}^l(x,\widehat{y}^r)$ and $\nabla_y h_i(x,y(x))$, we obtain

$$\mathbb{E}\left[\|\widehat{y}^{r+1} - \widehat{y}^r\|^2\right] \le T_1 + T_2 + T_3 + T_4,$$

where

$$\begin{aligned}
T_1 &:= 4(\gamma_r^l)^2\mathbb{E}\left[\left\|\tfrac{1}{s^l K^l}\sum_{i\in S^{l,r}}\sum_{k=0}^{K^l-1}\left(\widetilde{g}_{i,k}^r - \nabla_y F_{i,r}^l(x,\widehat{y}^r)\right)\right\|^2\right],\\
T_2 &:= 4(\gamma_r^l)^2\mathbb{E}[\|\widehat{c}^r\|^2],\\
T_3 &:= 4(\gamma_r^l)^2\mathbb{E}\left[\left\|\tfrac{1}{s^l}\sum_{i\in S^{l,r}}\left(\nabla_y h_i(x,y(x)) - c_i^r\right)\right\|^2\right],\\
T_4 &:= 4(\gamma_r^l)^2\mathbb{E}\left[\left\|\tfrac{1}{s^l}\sum_{i\in S^{l,r}}\left(\nabla_y F_{i,r}^l(x,\widehat{y}^r) - \nabla_y h_i(x,y(x))\right)\right\|^2\right].
\end{aligned}$$

For T_1, add and subtract $\nabla_y F_{i,r}^l(x,y_{i,k}^r)$. By $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, Jensen's inequality, uniform client sampling, smoothness, and equation 83,

$$\begin{aligned}
T_1 &\le 8(\gamma_r^l)^2\mathbb{E}\left[\tfrac{1}{s^l K^l}\sum_{i\in S^{l,r}}\sum_{k=0}^{K^l-1}\left\|\nabla_y F_{i,r}^l(x,y_{i,k}^r) - \nabla_y F_{i,r}^l(x,\widehat{y}^r)\right\|^2\right]\\
&\quad + 8(\gamma_r^l)^2\mathbb{E}\left[\left\|\tfrac{1}{s^l K^l}\sum_{i\in S^{l,r}}\sum_{k=0}^{K^l-1}\epsilon_{i,k}^r\right\|^2\right]\\
&\le 8(\gamma_r^l)^2(L_r^l)^2 D_r^l + \tfrac{8(\gamma_r^l)^2\sigma^2}{s^l K^l}.
\end{aligned}$$

For T_2, write

$$\widehat{c}^r = \tfrac{1}{m}\sum_{i=1}^m\left(c_i^r - \nabla_y h_i(x,y(x))\right) + \tfrac{1}{m}\sum_{i=1}^m \nabla_y h_i(x,y(x)).$$

Thus, by Jensen's inequality and the definition of C_r^l,

$$T_2 \le 8(\gamma_r^l)^2 C_r^l + 8(\gamma_r^l)^2(G^*)^2.$$

Similarly, uniform sampling and Jensen's inequality give

$$T_3 \le 4(\gamma_r^l)^2 C_r^l.$$

Finally, by uniform sampling, Jensen's inequality, and equation 82,

$$\begin{aligned}
T_4 &\le \tfrac{4(\gamma_r^l)^2}{m}\sum_{i=1}^m\mathbb{E}\left[\left\|\nabla_y F_{i,r}^l(x,\widehat{y}^r) - \nabla_y h_i(x,y(x))\right\|^2\right]\\
&\le 8(\gamma_r^l)^2 L_{h1,y}^2 e_r + 16(\gamma_r^l)^2(\rho_r^l)^2 p_r.
\end{aligned}$$

Combining the four estimates yields

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{y}^{r+1} - \widehat{y}^r\|^2\right] &\le 8(\gamma_r^l)^2(L_r^l)^2 D_r^l + 12(\gamma_r^l)^2 C_r^l + 8(\gamma_r^l)^2(G^*)^2\\
&\quad + \tfrac{8(\gamma_r^l)^2\sigma^2}{s^l K^l} + 8(\gamma_r^l)^2 L_{h1,y}^2 e_r + 16(\gamma_r^l)^2(\rho_r^l)^2 p_r.
\end{aligned} \tag{84}$$

Step 2:

Let

$$I_{i,r} = \begin{cases} 1, & i \in S^{l,r}, \\ 0, & i \notin S^{l,r}. \end{cases}$$

For a selected client,

$$c_i^{r+1} = \frac{1}{K^l} \sum_{k=0}^{K^l-1} \widetilde{g}_{i,k}^r,$$

whereas for an unselected client, $c_i^{r+1} = c_i^r$. Hence

$$c_i^{r+1} = (1 - I_{i,r}) c_i^r + I_{i,r} \frac{1}{K^l} \sum_{k=0}^{K^l-1} \widetilde{g}_{i,k}^r.$$

Since $I_{i,r}(1 - I_{i,r}) = 0$,

$$\begin{aligned} \left\| c_i^{r+1} - \nabla_y h_i(x, y(x)) \right\|^2 &= (1 - I_{i,r}) \left\| c_i^r - \nabla_y h_i(x, y(x)) \right\|^2 \\ &\quad + I_{i,r} \left\| \frac{1}{K^l} \sum_{k=0}^{K^l-1} \widetilde{g}_{i,k}^r - \nabla_y h_i(x, y(x)) \right\|^2. \end{aligned}$$

Taking expectations and averaging over clients gives

$$C_{r+1}^l = \left(1 - \frac{s^l}{m}\right) C_r^l + \frac{s^l}{m^2} \sum_{i=1}^m \mathbb{E}\left[\left\| \frac{1}{K^l} \sum_{k=0}^{K^l-1} \widetilde{g}_{i,k}^r - \nabla_y h_i(x, y(x)) \right\|^2 \right]. \tag{85}$$

For the second term, add and subtract $\nabla_y F_{i,r}^l(x, y_{i,k}^r)$. Using the same martingale-difference argument as in Step 1,

$$\begin{aligned} &\mathbb{E}\left[\left\| \frac{1}{K^l} \sum_{k=0}^{K^l-1} \widetilde{g}_{i,k}^r - \nabla_y h_i(x, y(x)) \right\|^2 \right] \\ &\leq \frac{2\sigma^2}{K^l} + 2\mathbb{E}\left[\left\| \frac{1}{K^l} \sum_{k=0}^{K^l-1} \left(\nabla_y F_{i,r}^l(x, y_{i,k}^r) - \nabla_y h_i(x, y(x)) \right) \right\|^2 \right]. \end{aligned}$$

Add and subtract $\nabla_y F_{i,r}^l(x, \widehat{y}^r)$ in the second term. By Jensen's inequality,

$$\begin{aligned} &\mathbb{E}\left[\left\| \frac{1}{K^l} \sum_{k=0}^{K^l-1} \left(\nabla_y F_{i,r}^l(x, y_{i,k}^r) - \nabla_y h_i(x, y(x)) \right) \right\|^2 \right] \\ &\leq \frac{2(L_r^l)^2}{K^l} \sum_{k=0}^{K^l-1} \mathbb{E}\left[\| y_{i,k}^r - \widehat{y}^r \|^2 \right] \\ &\quad + 2\mathbb{E}\left[\left\| \nabla_y F_{i,r}^l(x, \widehat{y}^r) - \nabla_y h_i(x, y(x)) \right\|^2 \right]. \end{aligned}$$

Substituting this bound into equation 85, averaging over i, and applying equation 82, we obtain

$$C_{r+1}^l \leq \left(1 - \frac{s^l}{m}\right) C_r^l + \frac{2s^l\sigma^2}{mK^l} + \frac{4s^l(L_r^l)^2}{m} D_r^l + \frac{8s^l L_{h1,y}^2}{m} e_r + \frac{16s^l(\rho_r^l)^2}{m} p_r. \tag{86}$$

Step 3:

Define, for $k = 0, \ldots, K^l$,

$$M_k := \frac{1}{m} \sum_{i=1}^m \mathbb{E}\left[\| y_{i,k}^r - \widehat{y}^r \|^2 \right].$$

Since all local iterates are initialized from the server iterate, $M_0 = 0$.

We first bound

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E}\left[\left\| \widehat{c}^r - c_i^r + \nabla_y F_{i,r}^l(x, \widehat{y}^r) \right\|^2 \right].$$

Add and subtract $\nabla_y h_i(x, y(x))$ and its average. Using $\|a + b + c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, together with the centered-variance identity

$$\frac{1}{m} \sum_{i=1}^m \|z_i - \bar{z}\|^2 = \frac{1}{m} \sum_{i=1}^m \|z_i\|^2 - \|\bar{z}\|^2,$$

gives

$$\begin{aligned}&\tfrac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\left\|\widehat{c}^r - c_i^r + \nabla_y F_{i,r}^l(x,\widehat{y}^r)\right\|^2\right]\\&\le 3C_r^l + 3(G^*)^2 + \tfrac{3}{m}\sum_{i=1}^m \mathbb{E}\left[\left\|\nabla_y F_{i,r}^l(x,\widehat{y}^r) - \nabla_y h_i(x,y(x))\right\|^2\right]\\&\le 3C_r^l + 3(G^*)^2 + 6L_{h1,y}^2 e_r + 12(\rho_r^l)^2 p_r. \qquad (87)\end{aligned}$$

If $K^l = 1$, then $D_r^l = 0$, and equation 88 below holds trivially. Suppose therefore that $K^l > 1$.

Using the local update and the zero conditional mean of the oracle noise,

$$M_{k+1} \le \tfrac{1}{m}\sum_{i=1}^m \mathbb{E}\Big[\|y_{i,k}^r - \gamma_{l,r}^l \nabla_y F_{i,r}^l(x,y_{i,k}^r) - \gamma_{l,r}^l(\widehat{c}^r - c_i^r) - \widehat{y}^r\|^2\Big] + (\gamma_{l,r}^l)^2\sigma^2.$$

Add and subtract $\gamma_{l,r}^l \nabla_y F_{i,r}^l(x,\widehat{y}^r)$ and apply

$$\|a+b\|^2 \le (1+\alpha)\|a\|^2 + \left(1+\tfrac{1}{\alpha}\right)\|b\|^2$$

with

$$\alpha = \tfrac{1}{K^l - 1}.$$

Because $K^l\gamma_{l,r}^l L_r^l \le 2$ implies in particular $\gamma_{l,r}^l L_r^l \le 2$, the gradient mapping

$$y \mapsto y - \gamma_{l,r}^l \nabla_y F_{i,r}^l(x,y)$$

is nonexpansive. Hence, using equation 87,

$$\begin{aligned}M_{k+1} \le{}& \left(1 + \tfrac{1}{K^l-1}\right) M_k + (\gamma_{l,r}^l)^2\sigma^2\\&+ K^l(\gamma_{l,r}^l)^2\left(3C_r^l + 3(G^*)^2 + 6L_{h1,y}^2 e_r + 12(\rho_r^l)^2 p_r\right).\end{aligned}$$

Unrolling this recursion from $M_0 = 0$ gives

$$M_k \le \left((\gamma_{l,r}^l)^2\sigma^2 + K^l(\gamma_{l,r}^l)^2\left(3C_r^l + 3(G^*)^2 + 6L_{h1,y}^2 e_r + 12(\rho_r^l)^2 p_r\right)\right) \times \sum_{\tau=0}^{k-1}\left(1 + \tfrac{1}{K^l-1}\right)^\tau.$$

For $1 \le k \le K^l$,

$$\sum_{\tau=0}^{k-1}\left(1 + \tfrac{1}{K^l-1}\right)^\tau \le 4K^l.$$

Averaging over the K^l local iterations and using

$$(K^l)^2(\gamma_{l,r}^l)^2 = \tfrac{(\gamma_r^l)^2}{s^l},$$

we obtain

$$D_r^l \le \tfrac{12(\gamma_r^l)^2}{s^l}C_r^l + \tfrac{12(\gamma_r^l)^2}{s^l}(G^*)^2 + \tfrac{24(\gamma_r^l)^2 L_{h1,y}^2}{s^l}e_r + \tfrac{48(\gamma_r^l)^2(\rho_r^l)^2}{s^l}p_r + \tfrac{4(\gamma_r^l)^2\sigma^2}{s^l K^l}. \qquad (88)$$

Step 4:

We first bound the cross term appearing in the expansion of e_{r+1}. For every client and local iteration, strong convexity of $F_{i,r}^l(x,\cdot)$ gives

$$F_{i,r}^l(x,y(x)) \ge F_{i,r}^l(x,y_{i,k}^r) + \left\langle \nabla_y F_{i,r}^l(x,y_{i,k}^r), y(x) - y_{i,k}^r\right\rangle + \tfrac{\mu_h}{2}\|y_{i,k}^r - y(x)\|^2,$$

while smoothness gives

$$F_{i,r}^l(x,\widehat{y}^r) \le F_{i,r}^l(x,y_{i,k}^r) + \left\langle \nabla_y F_{i,r}^l(x,y_{i,k}^r), \widehat{y}^r - y_{i,k}^r\right\rangle + \tfrac{L_r^l}{2}\|y_{i,k}^r - \widehat{y}^r\|^2.$$

Adding the two inequalities yields

$$\begin{aligned}\left\langle \nabla_y F_{i,r}^l(x,y_{i,k}^r), \widehat{y}^r - y(x)\right\rangle \ge{}& F_{i,r}^l(x,\widehat{y}^r) - F_{i,r}^l(x,y(x))\\&+ \tfrac{\mu_h}{2}\|y_{i,k}^r - y(x)\|^2 - \tfrac{L_r^l}{2}\|y_{i,k}^r - \widehat{y}^r\|^2.\end{aligned}$$

Using

$$\|y_{i,k}^r - y(x)\|^2 \geq \tfrac{1}{2}\|\widehat{y}^r - y(x)\|^2 - \|y_{i,k}^r - \widehat{y}^r\|^2$$

and $\mu_h \leq L_r^l$, we obtain

$$\begin{aligned}\left\langle \nabla_y F_{i,r}^l(x, y_{i,k}^r), \widehat{y}^r - y(x)\right\rangle \geq F_{i,r}^l(x, \widehat{y}^r) - F_{i,r}^l(x, y(x)) \\ + \tfrac{\mu_h}{4}\|\widehat{y}^r - y(x)\|^2 - L_r^l\|y_{i,k}^r - \widehat{y}^r\|^2.\end{aligned}$$

The control terms cancel after taking conditional expectation under uniform client sampling. Therefore,

$$2\mathbb{E}\left[\left\langle \widehat{y}^{r+1} - \widehat{y}^r, \widehat{y}^r - y(x)\right\rangle\right] \leq -\tfrac{\mu_h\gamma_r^l}{2}e_r - 2\gamma_r^l\mathbb{E}\left[h(x,\widehat{y}^r) - h(x,y(x))\right] - 2\gamma_r^l\rho_r^l p_r + 2\gamma_r^l L_r^l D_r^l. \tag{89}$$

Expanding

$$e_{r+1} = e_r + 2\mathbb{E}\left[\left\langle \widehat{y}^{r+1} - \widehat{y}^r, \widehat{y}^r - y(x)\right\rangle\right] + \mathbb{E}\left[\|\widehat{y}^{r+1} - \widehat{y}^r\|^2\right]$$

and applying equation 84 and equation 89 gives

$$\begin{aligned}e_{r+1} \leq \left(1 - \tfrac{\mu_h\gamma_r^l}{2} + 8(\gamma_r^l)^2 L_{h1,y}^2\right)e_r - 2\gamma_r^l\mathbb{E}\left(h(x,\widehat{y}^r) - h(x,y(x))\right) \\ - \left(2\gamma_r^l\rho_r^l - 16(\gamma_r^l)^2(\rho_r^l)^2\right)p_r + \left(2\gamma_r^l L_r^l + 8(\gamma_r^l)^2(L_r^l)^2\right)D_r^l \\ + 12(\gamma_r^l)^2 C_r^l + 8(\gamma_r^l)^2(G^*)^2 + \tfrac{8(\gamma_r^l)^2\sigma^2}{s^l K^l}.\end{aligned} \tag{90}$$

Next, multiply equation 86 by $\frac{24m(\gamma_r^l)^2}{s^l}$ and add the resulting inequality to equation 90

$$\begin{aligned}e_{r+1} + \tfrac{24m(\gamma_r^l)^2}{s^l}C_{r+1}^l \leq \left(1 - \tfrac{\mu_h\gamma_r^l}{2} + 200(\gamma_r^l)^2 L_{h1,y}^2\right)e_r + \left(\tfrac{24m(\gamma_r^l)^2}{s^l} - 12(\gamma_r^l)^2\right)C_r^l \\ - 2\gamma_r^l\mathbb{E}\left[h(x,\widehat{y}^r) - h(x,y(x))\right] - \left(2\gamma_r^l\rho_r^l - 400(\gamma_r^l)^2(\rho_r^l)^2\right)p_r \\ + A_r^l D_r^l + 8(\gamma_r^l)^2(G^*)^2 + \left(48 + \tfrac{8}{s^l}\right)\tfrac{(\gamma_r^l)^2\sigma^2}{K^l},\end{aligned} \tag{91}$$

Finally, substitute the client-drift estimate equation 88 into equation 91. Using the definitions of B_r^l and $\mathcal{R}_r$, we obtain

$$\begin{aligned}e_{r+1} + \tfrac{24m(\gamma_r^l)^2}{s^l}C_{r+1}^l \leq \left(1 - \tfrac{\mu_h\gamma_r^l}{2} + B_r^l\right)e_r + \left(\tfrac{24m(\gamma_r^l)^2}{s^l} - 12(\gamma_r^l)^2 + \tfrac{12(\gamma_r^l)^2 A_r^l}{s^l}\right)C_r^l \\ - 2\gamma_r^l\mathbb{E}\left[h(x,\widehat{y}^r) - h(x,y(x))\right] \\ - \left(2\gamma_r^l\rho_r^l - (\gamma_r^l)^2(\rho_r^l)^2\left(400 + \tfrac{48A_r^l}{s^l}\right)\right)p_r + \mathcal{R}_r.\end{aligned} \tag{92}$$

Then absorb the control and penalty terms. First,

$$\begin{aligned}\tfrac{24m(\gamma_r^l)^2}{s^l} - 12(\gamma_r^l)^2 + \tfrac{12(\gamma_r^l)^2 A_r^l}{s^l} &= \tfrac{24m(\gamma_r^l)^2}{s^l}\left(1 - \tfrac{s^l - A_r^l}{2m}\right) \\ &\leq \tfrac{24m(\gamma_r^l)^2}{s^l}\left(1 - \tfrac{\mu_h\gamma_r^l}{2}\right),\end{aligned}$$

where the last inequality follows from $A_r^l + \mu_h m\gamma_r^l \leq s^l$.

$$2\gamma_r^l\rho_r^l - (\gamma_r^l)^2(\rho_r^l)^2\left(400 + \tfrac{48A_r^l}{s^l}\right) = \gamma_r^l\rho_r^l\left(2 - \gamma_r^l\rho_r^l\left(400 + \tfrac{48A_r^l}{s^l}\right)\right) \geq \gamma_r^l\rho_r^l,$$

by equation 81.

Because $\gamma_{r+1}^l \leq \gamma_r^l$,

$$\Phi_{r+1} = e_{r+1} + \tfrac{24m(\gamma_{r+1}^l)^2}{s^l}C_{r+1}^l \leq e_{r+1} + \tfrac{24m(\gamma_r^l)^2}{s^l}C_{r+1}^l.$$

Moreover, since $B_r^l \geq 0$,

$$1 - \tfrac{\mu_h\gamma_r^l}{2} \leq 1 - \tfrac{\mu_h\gamma_r^l}{2} + B_r^l.$$

Therefore equation 92 implies

$$\begin{aligned}\Phi_{r+1} &\le \left(1-\tfrac{\mu_h\gamma_r^l}{2}+B_r^l\right)\left(e_r+\tfrac{24m(\gamma_r^l)^2}{s^l}C_r^l\right)\\ &\quad -2\gamma_r^l\mathbb{E}\left[h(x,\widehat{y}^r)-h(x,y(x))\right]-\gamma_r^l\rho_r^l p_r+\mathcal{R}_r\\ &=\left(1-\tfrac{\mu_h\gamma_r^l}{2}+B_r^l\right)\Phi_r-2\gamma_r^l\mathbb{E}\left[h(x,\widehat{y}^r)-h(x,y(x))\right]\\ &\quad -\gamma_r^l\rho_r^l p_r+\mathcal{R}_r,\end{aligned}$$

which is equation 80. □

Lemma A.29. Suppose that $\mu_h>0$, $L_{1,y}^h>0$, $1\le s^l\le m$, and $K^l\ge 1$. Choose

$$0<c_l\le\min\left\{\frac{\sqrt{s^l}}{32L_{1,y}^h},\frac{s^l}{2\mu_h m}\right\},\qquad 0<c_\rho\le\frac{1}{448c_l},\tag{93}$$

and set

$$\gamma_g^l=\sqrt{s^l},\qquad \gamma_{l,r}^l=\frac{c_l}{K^l\sqrt{s^l}\left(1+\frac{\mu_h c_l}{4}(r-1)\right)},\qquad \rho_r^l=c_\rho\left(1+\frac{\mu_h c_l}{4}(r-1)\right).$$

Recall that

$$\gamma_r^l=K^l\gamma_g^l\gamma_{l,r}^l,\qquad L_r^l=L_{1,y}^h+\rho_r^l,\qquad A_r^l=2\gamma_r^l L_r^l+104(\gamma_r^l)^2(L_r^l)^2.$$

Then $\{\gamma_r^l\}_{r\ge1}$ is positive and nonincreasing, and, for every $r\ge1$,

$$\begin{aligned}K^l\gamma_{l,r}^l L_r^l&\le 2, & 0<\mu_h\gamma_r^l&<2,\\ A_r^l+\mu_h m\gamma_r^l&\le s^l, & \gamma_r^l\rho_r^l\left(400+\frac{48A_r^l}{s^l}\right)&\le 1.\end{aligned}\tag{94}$$

Proof. The schedules give

$$\gamma_r^l=\frac{c_l}{1+\frac{\mu_h c_l}{4}(r-1)},\qquad \gamma_r^l\rho_r^l=c_l c_\rho\le\frac{1}{448}.$$

Since $c_l>0$ and $\mu_h>0$, the sequence $\{\gamma_r^l\}_{r\ge1}$ is positive and nonincreasing.

By equation 93 and $s^l\ge1$,

$$\begin{aligned}\frac{\gamma_r^l L_r^l}{\sqrt{s^l}}&=\frac{c_l L_{1,y}^h}{\sqrt{s^l}\left(1+\frac{\mu_h c_l}{4}(r-1)\right)}+\frac{c_l c_\rho}{\sqrt{s^l}}\\ &\le\frac{1}{32}+\frac{1}{448}=\frac{15}{448}<\frac{1}{24}.\end{aligned}$$

Consequently,

$$K^l\gamma_{l,r}^l L_r^l=\frac{\gamma_r^l L_r^l}{\sqrt{s^l}}\le\frac{1}{24}<2.$$

Moreover,

$$\begin{aligned}\frac{A_r^l}{s^l}&=\frac{2}{\sqrt{s^l}}\frac{\gamma_r^l L_r^l}{\sqrt{s^l}}+104\left(\frac{\gamma_r^l L_r^l}{\sqrt{s^l}}\right)^2\\ &\le\frac{2}{24\sqrt{s^l}}+\frac{104}{24^2}\le\frac{19}{72}<\frac{1}{2}.\end{aligned}$$

The bound $c_l\le s^l/(2\mu_h m)$ implies

$$0<\mu_h\gamma_r^l\le\mu_h c_l\le\frac{s^l}{2m}\le\frac{1}{2}<2.$$

Combining this bound with $A_r^l \le s^l/2$ yields

$$A_r^l + \mu_h m \gamma_r^l \le \frac{s^l}{2} + \mu_h m c_l \le s^l.$$

Finally,

$$\gamma_r^l \rho_r^l \left(400 + \frac{48 A_r^l}{s^l}\right) \le \frac{1}{448}(400 + 24) = \frac{53}{56} < 1.$$

Thus all inequalities in equation 94 hold for every $r \ge 1$. □

Theorem A.30. Consider Algorithm 2 for a fixed outer query x. Suppose Assumption 2.2 holds. Choose

$$\gamma_g^l = \sqrt{s^l}, \qquad \gamma_{l,r}^l = \frac{c_l}{K^l \sqrt{s^l}\left(1 + \frac{\mu_h c_l}{4}(r-1)\right)}, \qquad \rho_r^l = c_\rho \left(1 + \tfrac{\mu_h c_l}{4}(r-1)\right).$$

Let c_l and c_ρ satisfy the conditions in Lemma A.29. Define $G^* := \sup_{x \in \mathbb{R}^n} \left(\frac{1}{m}\sum_{i=1}^m \|\nabla_y h_i(x, y(x))\|^2\right)^{1/2}$. Then, for every $R^l \ge 1$,

$$\mathbb{E}\left[\|y_\varepsilon(x) - y(x)\|^2\right] = O\left(\frac{\exp\left(\frac{(L_{1,y}^h)^2}{\mu_h}\right)}{\mu_h^2}\left(\frac{m(G^*)^2}{(R^l)^2 s^l} + \frac{1}{R^l}\left((G^*)^2 + \left(1 + \tfrac{1}{s^l}\right)\frac{\sigma_h^2}{K^l}\right)\right)\right).$$

Proof. Throughout the proof, fix x and define $q_r := 1 + \frac{\mu_h c_l}{4}(r-1)$. Since $\gamma_g^l = \sqrt{s^l}$, the effective lower-level stepsize is

$$\gamma_r^l := K^l \gamma_g^l \gamma_{l,r}^l = \tfrac{c_l}{q_r},$$

and

$$\rho_r^l = c_\rho q_r.$$

Hence, γ_r^l is positive and nonincreasing, ρ_r^l is positive and nondecreasing, and

$$\gamma_r^l \rho_r^l = c_l c_\rho.$$

We first verify the conditions required by Lemma A.28. By the prescribed schedules,

$$\gamma_r^l L_r^l = c_l \left(\tfrac{L_{1,y}^h}{q_r} + c_\rho\right) \le c_l \left(L_{1,y}^h + c_\rho\right),$$

and therefore

$$A_r^l \le A_1^l.$$

By the choice of c_l and c_ρ,

$$K^l \gamma_{l,r}^l L_r^l \le \tfrac{1}{\sqrt{104}} < 2, \qquad 0 < \mu_h \gamma_r^l < 2,$$

and

$$A_r^l + \mu_h m \gamma_r^l \le s^l, \qquad \gamma_r^l \rho_r^l \left(400 + \tfrac{48 A_r^l}{s^l}\right) \le 1.$$

Thus Lemma A.28 applies and yields

$$\Phi_{r+1} \le \left(1 - \tfrac{\mu_h \gamma_r^l}{2} + B_r^l\right)\Phi_r - 2\gamma_r^l \mathbb{E}\left[h(x, \widehat{y}^r) - h(x, y(x))\right] - \gamma_r^l \rho_r^l p_r + \mathcal{R}_r. \tag{95}$$

By Lemma A.27, with $y = \widehat{y}^r$ and $\rho = \rho_r^l$,

$$2\left(h(x, \widehat{y}^r) - h(x, y(x))\right) + \rho_r^l g_x(\widehat{y}^r) \ge \mu_h \|\widehat{y}^r - y(x)\|^2 - \tfrac{2(G^*)^2}{\rho_r^l}.$$

Taking expectations and recalling

$$e_r = \mathbb{E}\left[\|\widehat{y}^r - y(x)\|^2\right], \qquad p_r = \mathbb{E}\left[g_x(\widehat{y}^r)\right],$$

we obtain

$$2\mathbb{E}\left[h(x,\widehat{y}^r)-h(x,y(x))\right]+\rho_r^l p_r \geq \mu_h e_r - \tfrac{2(G^*)^2}{\rho_r^l}. \tag{96}$$

Define

$$w_1 := 1, \qquad w_{r+1} := \tfrac{w_r}{1-\mu_h\gamma_r^l/2+B_r^l}, \tag{97}$$

and let

$$T_{R^l} := \textstyle\sum_{r=1}^{R^l} w_{r+1}\gamma_r^l.$$

Multiplying equation 95 by w_{r+1} and using equation 97 gives

$$w_{r+1}\gamma_r^l\left(2\mathbb{E}\left[h(x,\widehat{y}^r)-h(x,y(x))\right]+\rho_r^l p_r\right) \leq w_r\Phi_r - w_{r+1}\Phi_{r+1} + w_{r+1}\mathcal{R}_r. \tag{98}$$

Combining equation 96 and equation 98, we obtain

$$\mu_h w_{r+1}\gamma_r^l e_r \leq w_r\Phi_r - w_{r+1}\Phi_{r+1} + w_{r+1}\mathcal{R}_r + \tfrac{2(G^*)^2 w_{r+1}\gamma_r^l}{\rho_r^l}.$$

Summing over $r=1,\dots,R^l$ and using $\Phi_{R^l+1}\geq 0$ yields

$$\mu_h \textstyle\sum_{r=1}^{R^l} w_{r+1}\gamma_r^l e_r \leq \Phi_1 + \sum_{r=1}^{R^l} w_{r+1}\mathcal{R}_r + 2(G^*)^2\sum_{r=1}^{R^l}\tfrac{w_{r+1}\gamma_r^l}{\rho_r^l}.$$

Recall that $y_\varepsilon(x) = \dfrac{\sum_{r=1}^{R^l} w_{r+1}\gamma_r^l\widehat{y}^r}{\sum_{r=1}^{R^l} w_{r+1}\gamma_r^l}$. By convexity of the squared norm,

$$\mathbb{E}\left[\|y_\varepsilon(x)-y(x)\|^2\right] \leq \tfrac{\sum_{r=1}^{R^l} w_{r+1}\gamma_r^l e_r}{T_{R^l}}.$$

Therefore,

$$\mathbb{E}\left[\|y_\varepsilon(x)-y(x)\|^2\right] \leq \tfrac{1}{\mu_h T_{R^l}}\left(\Phi_1 + \textstyle\sum_{r=1}^{R^l} w_{r+1}\mathcal{R}_r + 2(G^*)^2\sum_{r=1}^{R^l}\tfrac{w_{r+1}\gamma_r^l}{\rho_r^l}\right). \tag{99}$$

We next estimate the weights. Since $\gamma_r^l = \frac{c_l}{q_r}$, we have

$$1-\tfrac{\mu_h\gamma_r^l}{2} = \tfrac{1+\frac{\mu_h c_l}{4}(r-3)}{1+\frac{\mu_h c_l}{4}(r-1)}.$$

Factoring the recursion in equation 97 gives

$$w_{r+1} = \tfrac{\left(1+\frac{\mu_h c_l}{4}(r-2)\right)\left(1+\frac{\mu_h c_l}{4}(r-1)\right)}{\left(1-\frac{\mu_h c_l}{4}\right)\left(1-\frac{\mu_h c_l}{2}\right)} \times \textstyle\prod_{k=1}^{r}\left(1+\tfrac{B_k^l}{1-\mu_h\gamma_k^l/2}\right)^{-1}. \tag{100}$$

Using $A_r^l \leq A_1^l$ and the definition of B_r^l,

$$B_r^l \leq \left(200+\tfrac{24A_1^l}{s^l}\right)(L_{1,y}^h)^2\tfrac{c_l^2}{q_r^2}.$$

Moreover,

$$1-\tfrac{\mu_h\gamma_r^l}{2} \geq 1-\tfrac{\mu_h c_l}{2} > 0.$$

Hence

$$\textstyle\sum_{r=1}^{\infty}\frac{B_r^l}{1-\mu_h\gamma_r^l/2} \leq \frac{c_l^2(L_{1,y}^h)^2\left(200+24A_1^l/s^l\right)}{1-\mu_h c_l/2}\sum_{r=1}^{\infty}\frac{1}{q_r^2} = \mathcal{O}\left(\tfrac{(L_{1,y}^h)^2}{\mu_h}\right).$$

Using $\log(1+t)\leq t$ for $t\geq 0$, we obtain

$$\textstyle\prod_{k=1}^{r}\left(1+\frac{B_k^l}{1-\mu_h\gamma_k^l/2}\right)^{-1} \geq \exp\left(-\mathcal{O}\left(\tfrac{(L_{1,y}^h)^2}{\mu_h}\right)\right).$$

It follows from equation 100 that

$$T_{R^l} = \Omega\left(\mu_h c_l^2 \exp\left(-\mathcal{O}\left(\frac{(L_{1,y}^h)^2}{\mu_h}\right)\right)(R^l)^2\right). \tag{101}$$

Next, since $A_r^l \le A_1^l \le s^l$, the definition of $\mathcal{R}_r$ gives

$$\mathcal{R}_r = \mathcal{O}\left((\gamma_r^l)^2\left((G^*)^2 + \left(1+\tfrac{1}{s^l}\right)\tfrac{\sigma_h^2}{K^l}\right)\right).$$

Moreover, equation 100 implies

$$w_{r+1}(\gamma_r^l)^2 = \mathcal{O}(c_l^2).$$

Therefore,

$$\textstyle\sum_{r=1}^{R^l} w_{r+1}\mathcal{R}_r = \mathcal{O}\left(c_l^2 R^l\left((G^*)^2 + \left(1+\tfrac{1}{s^l}\right)\tfrac{\sigma_h^2}{K^l}\right)\right). \tag{102}$$

Similarly,

$$\tfrac{\gamma_r^l}{\rho_r^l} = \tfrac{c_l}{c_\rho q_r^2}.$$

Using equation 100,

$$\textstyle\sum_{r=1}^{R^l} \tfrac{w_{r+1}\gamma_r^l}{\rho_r^l} = \mathcal{O}\left(\tfrac{c_l R^l}{c_\rho}\right). \tag{103}$$

Since the control variates are initialized at zero,

$$C_1^l = \tfrac{1}{m}\textstyle\sum_{i=1}^m \|\nabla_y h_i(x, y(x))\|^2 = (G^*)^2.$$

Thus

$$\Phi_1 = \mathbb{E}\left[\|\widehat{y}^1 - y(x)\|^2\right] + \tfrac{24mc_l^2}{s^l}(G^*)^2. \tag{104}$$

Substituting equation 101, equation 102, equation 103, and equation 104 into equation 99 yields

$$\mathbb{E}\left[\|y_\varepsilon(x) - y(x)\|^2\right] = O\Bigg(\frac{\exp\left(\frac{(L_{1,y}^h)^2}{\mu_h}\right)}{\mu_h^2}\Bigg(\frac{\mathbb{E}\left[\|\widehat{y}^1 - y(x)\|^2\right]}{(R^l)^2} + \frac{m(G^*)^2}{(R^l)^2 s^l} + \frac{1}{R^l}\left((G^*)^2 + \left(1+\tfrac{1}{s^l}\right)\tfrac{\sigma_h^2}{K^l}\right)\Bigg)\Bigg).$$

This completes the proof. □

Corollary A.31. Let $\varepsilon > 0$. Then $\mathbb{E}\left[\|y_\varepsilon(x) - y(x)\|^2\right] \le \varepsilon$ is guaranteed whenever

$$\varepsilon = O\left(\max\left\{\frac{m(G^*)^2\exp\left(\frac{2(L_{1,y}^h)^2}{\mu_h}\right)}{\mu_h^2(R^l)^2 s^l}, \frac{\exp\left(\frac{(L_{1,y}^h)^2}{\mu_h}\right)}{\mu_h^2 R^l}\left((G^*)^2 + \left(1+\tfrac{1}{s^l}\right)\tfrac{\sigma_h^2}{K^l}\right)\right\}\right).$$

Proof. By Theorem A.30,

$$\mathbb{E}\left[\|y_\varepsilon(x) - y(x)\|^2\right] = O\left(\frac{\exp\left(\frac{(L_{1,y}^h)^2}{\mu_h}\right)}{\mu_h^2}\left(\frac{m(G^*)^2}{(R^l)^2 s^l} + \frac{1}{R^l}\left((G^*)^2 + \left(1+\tfrac{1}{s^l}\right)\tfrac{\sigma_h^2}{K^l}\right)\right)\right).$$

Define

$$A_{R^l} := \frac{m(G^*)^2\exp\left(\frac{(L_{1,y}^h)^2}{\mu_h}\right)}{\mu_h^2(R^l)^2 s^l},$$

and

$$B_{R^l} := \frac{\exp\left(\frac{(L_{1,y}^h)^2}{\mu_h}\right)}{\mu_h^2 R^l}\left[(G^*)^2 + \left(1+\tfrac{1}{s^l}\right)\tfrac{\sigma_h^2}{K^l}\right].$$

Then

$$\mathbb{E}\left[\|y_\varepsilon(x) - y(x)\|^2\right] = O(A_{R^l} + B_{R^l}).$$

Since

$$A_{R^l} + B_{R^l} \le 2\max\{A_{R^l}, B_{R^l}\},$$

we obtain

$$\mathbb{E}\big[\|y_\varepsilon(x) - y(x)\|^2\big] = O(\max\{A_{R^l}, B_{R^l}\}).$$

Therefore,

$$\mathbb{E}\big[\|y_\varepsilon(x) - y(x)\|^2\big] \le \varepsilon$$

whenever

$$\max\{A_{R^l}, B_{R^l}\} = O(\varepsilon).$$

This proves the result. □

Lemma A.32. Suppose Assumption 2.2 holds and $\sup_x \frac{1}{m}\sum_{i=1}^m \|\nabla_y h_i(x, y(x))\|^2 \le C_G$, for some constant $C_G > 0$ independent of m, r, R^l, K^l, and s^l. Let $\epsilon_r := \mathbb{E}\big[\|y_r(x) - y(x)\|^2\big]$ denote the lower-level solution error at outer round r. Choose $R_r^l = r+1,\ K_r^l = \sqrt{K},\ s_r^l = \sqrt{s}$. Then

$$\epsilon_r = O\Big(\frac{\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\mu_h^2}\big(\frac{m}{(r+1)^2\sqrt{s}} + \frac{1}{r+1}\big(1 + \frac{\sigma_h^2}{\sqrt{K}}\big)\big)\Big). \tag{105}$$

Moreover, for $R \ge 2$,

$$\frac{1}{R+1}\sum_{r=1}^{R+1}\epsilon_{r-1} = O\Big(\frac{\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\mu_h^2 R}\big(\frac{m}{\sqrt{s}} + \big(1 + \frac{\sigma_h^2}{\sqrt{K}}\big)\log R\big)\Big). \tag{106}$$

Proof. From the lower-level convergence result, for arbitrary R_r^l, K_r^l, and s_r^l, we have

$$\epsilon_r = O\Big(\max\Big\{\frac{m(G^*)^2\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\mu_h^2 (R_r^l)^2 s_r^l},\ \frac{\exp(\frac{(L_{1,y}^h)^2}{\mu_h})}{\mu_h^2 R_r^l}\big((G^*)^2 + \big(1 + \frac{1}{s_r^l}\big)\frac{\sigma_h^2}{K_r^l}\big)\Big\}\Big).$$

By assumption, $(G^*)^2 \le C_G$. Moreover, since $s_r^l \ge 1$, we have $1 + \frac{1}{s_r^l} \le 2$, and $\exp(\frac{(L_{1,y}^h)^2}{\mu_h}) \le \exp(\frac{2(L_{1,y}^h)^2}{\mu_h})$. Therefore, using $\max\{a,b\} \le a + b$ and absorbing C_G and numerical constants into the big-O constant gives

$$\epsilon_r = O\Big(\frac{\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\mu_h^2}\big(\frac{m}{(R_r^l)^2 s_r^l} + \frac{1}{R_r^l}\big(1 + \frac{\sigma_h^2}{K_r^l}\big)\big)\Big). \tag{107}$$

Substituting

$$R_r^l = r+1, \qquad K_r^l = \sqrt{K}, \qquad s_r^l = \sqrt{s}$$

into equation 107 gives equation 105.

For the averaged error, applying equation 105 to ϵ_{r-1} gives

$$\epsilon_{r-1} = O\Big(\frac{\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\mu_h^2}\big(\frac{m}{r^2\sqrt{s}} + \frac{1}{r}\big(1 + \frac{\sigma_h^2}{\sqrt{K}}\big)\big)\Big).$$

Hence,

$$\frac{1}{R+1}\sum_{r=1}^{R+1}\epsilon_{r-1} = O\Big(\frac{\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\mu_h^2 (R+1)}\big(\frac{m}{\sqrt{s}}\sum_{r=1}^{R+1}\frac{1}{r^2} + \big(1 + \frac{\sigma_h^2}{\sqrt{K}}\big)\sum_{r=1}^{R+1}\frac{1}{r}\big)\Big).$$

Using $\sum_{r=1}^{R+1}\frac{1}{r} = O(\log(R+1)),\quad \sum_{r=1}^{R+1}\frac{1}{r^2} = O(1)$, together with $\frac{1}{R+1} = O(\frac{1}{R}),\ \log(R+1) = O(\log R)$, for $R \ge 2$, yields equation 106. □

Proof for Theorem 3.3.

Proof. (i) For $r \geq 0$, let $a_r := \mathbb{E}\left[f^\eta(\hat{x}^r)\right] + y\tilde{\gamma}^2(\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta})^3\frac{m}{s}\mathcal{C}_r + u\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}\mathcal{D}_r$. By Lemma A.24, for every $r \geq 1$, we have

$$a_r \leq a_{r-1} + (c+g+o)\epsilon_{r-1} + (w-d+p)\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] + \sqrt{2\pi}\tilde{\gamma}^2\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2\left(1824 - \tfrac{1200}{K} + y\tfrac{m}{s}\rho^2(7296 - \tfrac{4800}{K}) + \tfrac{y-216}{1296}(10944 - \tfrac{6912}{K})\right).$$

We first derive a lower bound for $d-w-p$. By the definitions of w, d, and p, $\frac{d-w-p}{\tilde{\gamma}} = \frac{K+1}{4K} - \frac{y}{72}\rho - 12y\frac{m}{s}\rho^3$. The proof of Lemma A.24 gives $216 < y < 648$. Moreover, the stepsize conditions imply $\rho < \frac{1}{4A_0}$, $\quad \frac{m}{s}\rho^3 < \frac{1}{62208}$. Therefore, $\frac{y}{72}\rho < \frac{9}{4A_0}$ and $12y\frac{m}{s}\rho^3 < 12\cdot 648\cdot\frac{1}{62208} = \frac{1}{8}$, and hence $\frac{d-w-p}{\tilde{\gamma}} > \frac{K+1}{4K} - \frac{9}{4A_0} - \frac{1}{8}$. Define $c_2 := \frac{K+1}{4K} - \frac{1}{8} - \frac{9}{4A_0}$. Since $K \geq 2$ and $A_0 = 21(n+1) \geq 42$, $c_2 \geq \frac{1}{4} - \frac{1}{8} - \frac{9}{4\cdot 42} = \frac{1}{14} > 0$. Hence, we obtain

$$d - w - p \geq c_2\tilde{\gamma}. \tag{108}$$

Next, we bound $c+g+o$. From their definitions,

$$c+g+o = \tilde{\gamma}\frac{n^2(L_{0,y}^f)^2}{\eta^2}\left(\tfrac{3}{2} + 21\rho + 84y\tfrac{m}{s}\rho^3 + \tfrac{7(y-216)}{72}\rho\right).$$

Since $y < 648$ and $\rho < 1/(4A_0)$, $21\rho + \frac{7(y-216)}{72}\rho < 63\rho < \frac{63}{4A_0} = \frac{3}{4(n+1)} \leq \frac{3}{8}$. Also, the stepsize condition $\tilde{\gamma} \leq \frac{\eta}{\sqrt{288}\,nL_0^{\mathrm{imp}}}\sqrt{\frac{s}{3m}}$ implies $\frac{m}{s}\rho^2 \leq \frac{1}{864n}$, and hence $84y\frac{m}{s}\rho^3 < 84\cdot 648\cdot\frac{1}{864n}\cdot\frac{1}{4A_0} = \frac{3}{4n(n+1)} \leq \frac{3}{8}$. Consequently, $c+g+o \leq \frac{9\tilde{\gamma}}{4}\frac{n^2(L_{0,y}^f)^2}{\eta^2}$.

We next bound the constant term. Since $\frac{m}{s}\rho^2 \leq \frac{1}{864n} \leq \frac{1}{864}$ and $y < 648$, $y\frac{m}{s}\rho^2 < \frac{3}{4}$, $\frac{y-216}{1296} < \frac{1}{3}$. Therefore,

$$\begin{aligned}&1824 - \tfrac{1200}{K} + y\tfrac{m}{s}\rho^2(7296 - \tfrac{4800}{K}) + \tfrac{y-216}{1296}(10944 - \tfrac{6912}{K})\\ &\leq 1824 - \tfrac{1200}{K} + \tfrac{3}{4}(7296 - \tfrac{4800}{K}) + \tfrac{1}{3}(10944 - \tfrac{6912}{K}) = 10944 - \tfrac{7104}{K}.\end{aligned}$$

Thus,

$$\begin{aligned}&\sqrt{2\pi}\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2\left(1824 - \tfrac{1200}{K} + y\tfrac{m}{s}\rho^2(7296 - \tfrac{4800}{K}) + \tfrac{y-216}{1296}(10944 - \tfrac{6912}{K})\right)\\ &\leq (10944 - \tfrac{7104}{K})\sqrt{2\pi}\tilde{\gamma}^2\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2.\end{aligned} \tag{109}$$

Using equation 108–equation 109, we obtain

$$\mathbb{E}\left[\|\nabla f^\eta(\hat{x}^{r-1})\|^2\right] \leq \tfrac{a_{r-1}-a_r}{c_2\tilde{\gamma}} + \tfrac{9}{4c_2}\tfrac{n^2(L_{0,y}^f)^2}{\eta^2}\epsilon_{r-1} + \tfrac{10944-\frac{7104}{K}}{c_2}\sqrt{2\pi}\tilde{\gamma}\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2.$$

Summing over $r = 1, \ldots, R+1$, dividing by $R+1$, and using the definition of $\bar{x}^R$, we have

$$\begin{aligned}\mathbb{E}[\|\nabla f^\eta(\bar{x}^R)\|^2] \leq \tfrac{1}{R+1}\textstyle\sum_{r=1}^{R+1}&\left(\tfrac{a_{r-1}-a_r}{c_2\tilde{\gamma}} + \tfrac{10944-\frac{7104}{K}}{c_2}\sqrt{2\pi}\tilde{\gamma}\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2\right)\\ &+ \tfrac{9}{4c_2}\tfrac{n^2(L_{0,y}^f)^2}{\eta^2}\tfrac{1}{R+1}\textstyle\sum_{r=1}^{R+1}\epsilon_{r-1}.\end{aligned}$$

By the initialization of Algorithm 1, $\mathcal{C}_0 = \mathcal{D}_0 = 0$. Hence, $a_0 - f^\eta(x^*) = f^\eta(\hat{x}^0) - f^\eta(x^*)$. Applying equation 77 in Lemma A.25 to the nonnegative sequence $\{(a_r - f^\eta(x^*))/c_2\}$ with

$$\lambda_1 = \tfrac{10944-\frac{7104}{K}}{c_2}\sqrt{2\pi}\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2, \qquad \lambda_2 = 0,$$

we obtain

$$\begin{aligned}\mathbb{E}\left[\|\nabla f^\eta(\bar{x}^R)\|^2\right] \leq\ &\tfrac{f^\eta(\hat{x}^0)-f^\eta(x^*)}{c_2\gamma_{\max}(R+1)}\\ &+ \tfrac{2}{c_2\sqrt{R+1}}\sqrt{(10944 - \tfrac{7104}{K})\sqrt{2\pi}\tfrac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2(f^\eta(\hat{x}^0) - f^\eta(x^*))}\\ &+ \tfrac{9}{4c_2}\tfrac{n^2(L_{0,y}^f)^2}{\eta^2}\tfrac{1}{R+1}\textstyle\sum_{r=1}^{R+1}\epsilon_{r-1}.\end{aligned} \tag{110}$$

We now characterize $\gamma_{\max}$. From the stepsize conditions in Lemma A.24,

$$\tfrac{1}{\gamma_{\max}} = \mathcal{O}\left(\tfrac{L_0^{\mathrm{imp}}}{\eta}\max\left\{n, n\sqrt{\tfrac{m}{s}}, \sqrt{n}A_0, \sqrt{nB}, \sqrt{n}(A_1\tfrac{m}{s})^{1/4}, \sqrt{n}(\tfrac{m}{s})^{1/3}\right\}\right).$$

Since $A_0 = 21(n+1) = \mathcal{O}(n)$, $A_1 = 126n+108 = \mathcal{O}(n)$, $B = 6A_1+1296\Theta = \mathcal{O}(n+\tfrac{m}{s})$, $n \geq 1$, and $m/s \geq 1$, we have $\sqrt{nB} = \mathcal{O}(n + \sqrt{n\tfrac{m}{s}})$, $\sqrt{n}(A_1\tfrac{m}{s})^{1/4} = \mathcal{O}(n^{3/4}(\tfrac{m}{s})^{1/4}) \leq \mathcal{O}(n\sqrt{\tfrac{m}{s}})$, and $\sqrt{n}(\tfrac{m}{s})^{1/3} \leq n\sqrt{\tfrac{m}{s}}$. Therefore, using the preceding bounds and $\max\{a,b\} \leq a+b$, we obtain

$$\tfrac{1}{\gamma_{\max}} = \mathcal{O}\left(\tfrac{L_0^{\mathrm{imp}}}{\eta}\left(n^{3/2} + n\sqrt{\tfrac{m}{s}}\right)\right). \tag{111}$$

Moreover, $c_2 \geq \frac{1}{14}$, $0 < 10944 - \frac{7104}{K} \leq 10944$, and $\sqrt{\frac{L_0^{\mathrm{imp}}\sqrt{n}}{\eta}n(L_0^{\mathrm{imp}})^2} = \frac{(L_0^{\mathrm{imp}})^{3/2}n^{3/4}}{\sqrt{\eta}}$. Therefore, using equation 111 in equation 110 and $R \leq R+1 \leq 2R$ for $R \geq 1$, we obtain

$$\mathbb{E}[\|\nabla f^\eta(\bar{x}^R)\|^2] \leq \mathcal{O}\left(\tfrac{L_0^{\mathrm{imp}}}{\eta R}\left(n^{3/2} + n\sqrt{\tfrac{m}{s}}\right)\right) + \mathcal{O}\left(\tfrac{(L_0^{\mathrm{imp}})^{3/2}n^{3/4}}{\sqrt{\eta R}}\right) + \mathcal{O}\left(\tfrac{n^2(L_{0,y}^f)^2}{\eta^2}\tfrac{1}{R+1}\textstyle\sum_{r=1}^{R+1}\epsilon_{r-1}\right).$$

By (Ebrahimi et al., 2025, Proposition 2(ii)), for any $x \in \mathbb{R}^n$, $\mathrm{dist}(0, \partial_\eta f(x)) \leq \|\nabla f^\eta(x)\|$. Therefore, evaluating the above inequality at $\bar{x}^R$, squaring both sides, and taking expectations, we obtain

$$\begin{aligned}\mathbb{E}\big[\mathrm{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] &\leq \mathbb{E}\big[\|\nabla f^\eta(\bar{x}^R)\|^2\big] \\ &\leq \mathcal{O}\Big(\tfrac{L_0^{\mathrm{imp}}}{\eta R}\left(n^{3/2} + n\sqrt{\tfrac{m}{s}}\right)\Big) + \mathcal{O}\Big(\tfrac{(L_0^{\mathrm{imp}})^{3/2}n^{3/4}}{\sqrt{\eta R}}\Big) \\ &\quad + \mathcal{O}\left(\tfrac{n^2(L_{0,y}^f)^2}{\eta^2}\tfrac{1}{R+1}\sum_{r=1}^{R+1}\epsilon_{r-1}\right),\end{aligned}$$

which proves part (i).

(ii) Suppose we employ Algorithm 3 for each of the two constrained lower-level calls. Choosing $R_r^l = r+1$, $K_r^l = \sqrt{K}$, and $s_r^l = \sqrt{s}$, Lemma A.32 gives equation 106.

(ii.a) Ignoring logarithmic factors and substituting the preceding bound into part (i), we obtain

$$\begin{aligned}\mathbb{E}\big[\mathrm{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] &\leq \mathcal{O}\left(\tfrac{L_0^{\mathrm{imp}}}{\eta R}\left(n^{3/2} + n\sqrt{\tfrac{m}{s}}\right)\right) + \mathcal{O}\left(\tfrac{(L_0^{\mathrm{imp}})^{3/2}n^{3/4}}{\sqrt{\eta R}}\right) \\ &\quad + \mathcal{O}\left(\tfrac{n^2(L_{0,y}^f)^2\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\eta^2\mu_h^2 R}\left(\tfrac{m}{\sqrt{s}} + \tfrac{\sigma_h^2}{\sqrt{K}}\right)\right).\end{aligned}$$

(ii.b) To guarantee $\mathbb{E}\big[\mathrm{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] \leq \varepsilon$, it is sufficient to choose

$$R_\varepsilon = \max\left\{\mathcal{O}\left(\tfrac{L_0^{\mathrm{imp}}}{\eta\varepsilon}\left(n^{3/2} + n\sqrt{\tfrac{m}{s}}\right)\right), \mathcal{O}\left(\tfrac{(L_0^{\mathrm{imp}})^3 n^{3/2}}{\eta\varepsilon^2}\right), \mathcal{O}\left(\tfrac{n^2(L_{0,y}^f)^2\exp(\frac{2(L_{1,y}^h)^2}{\mu_h})}{\eta^2\mu_h^2\varepsilon}\left(\tfrac{m}{\sqrt{s}} + \tfrac{\sigma_h^2}{\sqrt{K}}\right)\right)\right\}.$$

(ii.c) Since the lower-level algorithm is called twice at each outer round and $R_r^l = r+1$, summing the communication cost of the two lower-level calls over $\mathcal{O}(R_\varepsilon)$ outer rounds gives

$$\bar{R}_\varepsilon = \mathcal{O}\left(\textstyle\sum_{r=0}^{R_\varepsilon}(r+1)\right) = \mathcal{O}(R_\varepsilon^2).$$

Using $(a+b)^2 \leq 2(a^2+b^2)$, we therefore obtain

$$\bar{R}_\varepsilon = \mathcal{O}\left(\max\left\{\tfrac{(L_0^{\mathrm{imp}})^2}{\eta^2\varepsilon^2}\left(n^3 + n^2\tfrac{m}{s}\right), \tfrac{(L_0^{\mathrm{imp}})^6 n^3}{\eta^2\varepsilon^4}, \tfrac{n^4(L_{0,y}^f)^4\exp(\frac{4(L_{1,y}^h)^2}{\mu_h})}{\eta^4\mu_h^4\varepsilon^2}\left(\tfrac{m}{\sqrt{s}} + \tfrac{\sigma_h^2}{\sqrt{K}}\right)^2\right\}\right).$$

□

Theorem A.33. Consider Algorithm 1 for addressing the problem in GFBO, where $Y(x) = \mathbb{R}^{\bar{n}}$. Suppose Assumption 2.1 holds. Let $\gamma_{\max} > 0$ satisfy $\gamma_{\max} \leq N$ and $\frac{\eta}{\gamma_{\max}} > L_0^{\text{imp}}\sqrt{n}\,M$, where N and M are defined in Definition A.21. Assume that the lower-level problem is solved inexactly, and let ϵ_r be the corresponding error level at round $r \geq 1$, satisfying $\mathbb{E}\left[\|y_{\epsilon_r}(\bullet) - y(\bullet)\|^2\right] \leq \epsilon_r$. Let $\epsilon_0 \geq \epsilon_1$ be a finite initial reference error, and suppose $\epsilon_r \leq \epsilon_{r-1}$ for all $r \geq 2$. Let $\bar{x}^R$ be selected uniformly at random from $\{\hat{x}^0, \ldots, \hat{x}^R\}$, such that $\mathbb{P}(\bar{x}^R = \hat{x}^{r-1}) = \frac{1}{R+1}$ for $r = 1, \ldots, R+1$. Then, there exists a constant effective stepsize $\tilde{\gamma} := K\gamma_g\gamma_l$, where $\gamma_g \geq 1$, satisfying $0 < \tilde{\gamma} \leq \gamma_{\max}$, such that the following statements hold true for $K \geq 2$ and $R \geq 1$.
At each round r of Algorithm 1, we employ SCAFFOLD for each of the two lower-level calls, where R_r^l, K_r^l, and s_r^l denote the number of communication rounds, the number of local iterations per client per round, and the number of selected clients, respectively. Then, $\epsilon_r = \mathcal{O}(1/(R_r^l K_r^l s_r^l))$. Let $R_r^l = r+1$, $K_r^l = \sqrt{K}$, and $s_r^l = \sqrt{s}$. Set $\epsilon_0 := \epsilon_1$.

(a) [Error bound] Ignoring logarithmic factors, the upper bound for the error metric $\mathbb{E}[\|\nabla f^\eta(\bar{x}^R)\|^2]$ is as follows.

$$\mathbb{E}[\|\nabla f^\eta(\bar{x}^R)\|^2] \leq \mathcal{O}\left(\frac{L_0^{\text{imp}}}{\eta R}\left(n^{3/2} + n\sqrt{\frac{m}{s}}\right)\right) + \mathcal{O}\left(\frac{(L_0^{\text{imp}})^{3/2} n^{3/4}}{\sqrt{\eta R}}\right) + \mathcal{O}\left(\frac{n^2 (L_{0,y}^f)^2}{\eta^2 \sqrt{K s} R}\right).$$

(b) [Upper-level communication complexity] The number of upper-level communication rounds to achieve $\mathbb{E}[\|\nabla f^\eta(\bar{x}^R)\|^2] \leq \varepsilon$ is given by

$$R_\varepsilon = \max\left\{\mathcal{O}\left(\frac{L_0^{\text{imp}}}{\eta\varepsilon}\left(n^{3/2} + n\sqrt{\frac{m}{s}}\right)\right), \mathcal{O}\left(\frac{(L_0^{\text{imp}})^3 n^{3/2}}{\eta\varepsilon^2}\right), \mathcal{O}\left(\frac{n^2 (L_{0,y}^f)^2}{\eta^2 \sqrt{K s}\varepsilon}\right)\right\}.$$

(c) [Total communication complexity] Since the lower-level algorithm is called twice at each outer round, the total number of communication rounds to achieve $\mathbb{E}[\|\nabla f^\eta(\bar{x}^R)\|^2] \leq \varepsilon$ is given by

$$\bar{R}_\varepsilon = \mathcal{O}\left(\max\left\{\frac{(L_0^{\text{imp}})^2}{\eta^2\varepsilon^2}\left(n^3 + n^2\frac{m}{s}\right), \frac{(L_0^{\text{imp}})^6 n^3}{\eta^2\varepsilon^4}, \frac{n^4 (L_{0,y}^f)^4}{\eta^4 K s \varepsilon^2}\right\}\right).$$

Proof. Suppose SCAFFOLD is employed for each of the two lower-level calls. Since $\epsilon_r = \mathcal{O}(1/(R_r^l K_r^l s_r^l))$, choosing $R_r^l = r+1$, $K_r^l = \sqrt{K}$, and $s_r^l = \sqrt{s}$ gives $\epsilon_r = \mathcal{O}\left(\frac{1}{(r+1)\sqrt{Ks}}\right)$. Set $\epsilon_0 := \epsilon_1$. For $r = 1$, $\epsilon_{r-1} = \epsilon_0 = \epsilon_1 = \mathcal{O}(\frac{1}{\sqrt{Ks}}) = \mathcal{O}(\frac{1}{r\sqrt{Ks}})$, while, for $r \geq 2$, $\epsilon_{r-1} = \mathcal{O}(\frac{1}{r\sqrt{Ks}})$. Therefore, we obtain

$$\frac{1}{R+1}\sum_{r=1}^{R+1}\epsilon_{r-1} = \mathcal{O}\left(\frac{1}{\sqrt{Ks}(R+1)}\sum_{r=1}^{R+1}\frac{1}{r}\right) = \mathcal{O}\left(\frac{1+\log(R+1)}{\sqrt{Ks}(R+1)}\right).$$

(a) Ignoring the logarithmic factor and substituting the preceding bound into part (i), we obtain

$$\mathbb{E}\big[\text{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] \leq \mathcal{O}\left(\frac{L_0^{\text{imp}}}{\eta R}\left(n^{3/2} + n\sqrt{\frac{m}{s}}\right)\right) + \mathcal{O}\left(\frac{(L_0^{\text{imp}})^{3/2} n^{3/4}}{\sqrt{\eta R}}\right) + \mathcal{O}\left(\frac{n^2 (L_{0,y}^f)^2}{\eta^2 \sqrt{K s} R}\right).$$

(b) To guarantee $\mathbb{E}\big[\text{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] \leq \varepsilon$, it is sufficient to take

$$R_\varepsilon = \max\left\{\mathcal{O}\left(\frac{L_0^{\text{imp}}}{\eta\varepsilon}\left(n^{3/2} + n\sqrt{\frac{m}{s}}\right)\right), \mathcal{O}\left(\frac{(L_0^{\text{imp}})^3 n^{3/2}}{\eta\varepsilon^2}\right), \mathcal{O}\left(\frac{n^2 (L_{0,y}^f)^2}{\eta^2 \sqrt{K s}\varepsilon}\right)\right\}.$$

(c) Since the lower-level algorithm is called twice at each outer round and $R_r^l = r+1$, summing the communication cost of the two lower-level calls over $\mathcal{O}(R_\varepsilon)$ outer rounds yields

$$\bar{R}_\varepsilon = \mathcal{O}\left(\sum_{r=0}^{R_\varepsilon}(r+1)\right) = \mathcal{O}(R_\varepsilon^2).$$

Using $(a+b)^2 \leq 2(a^2+b^2)$, we therefore obtain

$$\bar{R}_\varepsilon = \mathcal{O}\left(\max\left\{\frac{(L_0^{\text{imp}})^2}{\eta^2\varepsilon^2}\left(n^3 + n^2\frac{m}{s}\right), \frac{(L_0^{\text{imp}})^6 n^3}{\eta^2\varepsilon^4}, \frac{n^4 (L_{0,y}^f)^4}{\eta^4 K s \varepsilon^2}\right\}\right).$$

□

Remark A.34. We summarize several observations based on Theorem A.33.

(i) Relation to single-level SCAFFOLD. In the single-level nonconvex smooth setting, SCAFFOLD achieves communication complexity $\mathcal{O}\big(\frac{1}{sK\epsilon^2} + \frac{1}{\epsilon}(\frac{m}{s})^{2/3}\big)$ (Karimireddy et al., 2020, Theorem III). Suppressing the dependence on n, η, and L_0^{imp}, Theorem A.33 (b) yields $\mathcal{O}\big(\max\{\frac{1}{\sqrt{Ks\epsilon}}, \frac{1}{\epsilon}\sqrt{\frac{m}{s}}, \frac{1}{\epsilon^2}\}\big)$ for the nonsmooth nonconvex implicit upper-level objective. Since $m/s \geq 1$, $\sqrt{m/s} \leq (m/s)^{2/3}$, and hence our partial-participation term has an improved dependence on m/s. However, the two results are not directly comparable because our bound additionally contains the terms $1/(\sqrt{Ks\epsilon})$ and $1/\epsilon^2$, which arise from the inexact lower-level solutions and the nonsmooth bilevel zeroth-order setting, respectively.
(ii) Effect of local steps on communication complexity. The dependence on K in Theorem A.33 (b) appears through the term $1/(\sqrt{Ks\epsilon})$. Therefore, increasing K reduces the contribution of the lower-level inexactness to the upper-level communication complexity. However, the terms $\sqrt{m/s}/\epsilon$ and $1/\epsilon^2$ are independent of K. Hence, increasing the number of local steps reduces the overall upper-level communication bound when $1/(\sqrt{Ks\epsilon})$ is an active term, but it does not affect the terms that are independent of K.

Algorithm 4 Client i's local steps

1: **Input:** client i's initial point $y_{i,0}^r$ (warm-started from the most recently updated local state y_i), query point x, local update steps $H_{i,r}$, and local step size $\gamma_{i,t}$
2: **for** $t = 0, 1, \ldots, H_{i,r} - 1$ **do**
3: client i generates a random sample $\zeta_{i,t}^r \in \mathcal{D}_i$
4: $y_{i,t+1}^r(x) = \Pi_{Y_i(x)}[y_{i,t}^r - \gamma_{i,t}\tilde{h}_i(x, y_{i,t}^r, \zeta_{i,t}^r)]$
5: **end for**
6: **return** $y_{i,H_{i,r}}^r$ as $y_{\epsilon_{i,r}}$

In the next remark, we explain how the proof of Theorem 3.3 (i) extends to Algorithm 1 for addressing problem PFBO.

Remark A.35. For problem PFBO, the lower-level problems are solved locally at each client and do not involve federated communication. Accordingly, the delayed zeroth-order gradient approximation used in problem GFBO is not required, and only the inexactness of the lower-level solution remains. In this setting, each client computes $y_{\epsilon_{i,r}}(z_{i,k-1}^r \pm v^r)$ using local data, which implies $\hat{g}_{i,k-1}^{\eta,r,z} = g_{i,k-1}^{\eta,r,z}$ and $\hat{w}_{i,k-1}^{\eta,r,z} = 0$. Consequently, the analysis reduces to bounding the inexactness error only, and the result in Theorem 3.3 (i) extends to Theorem 3.5 (i) by setting the delay-related terms to zero.

Proof for Theorem 3.5.

Proof. (i) By Remark A.35, the proof follows from the proof of Theorem 3.3 (i) by setting the delay-related terms to zero.

(ii.a) We assume that each client employs the projected SGD method given in Algorithm 4 to address its lower-level problem. Therefore, $\epsilon_r = \mathcal{O}(\frac{1}{H_r})$, where H_r denotes the number of local iterations performed by each client at round r of Algorithm 1 for addressing problem PFBO. Let $H_r = r$ and set $\epsilon_0 := \epsilon_1$. For $r = 1$, $\epsilon_{r-1} = \epsilon_0 = \epsilon_1 = \mathcal{O}(1) = \mathcal{O}(\frac{1}{r})$. For $r \geq 2$, $\epsilon_{r-1} = \mathcal{O}(\frac{1}{r-1}) = \mathcal{O}(\frac{1}{r})$, where the last equality follows from $\frac{1}{r-1} \leq \frac{2}{r}$. Therefore, invoking Lemma A.26 (i), we have

$$\frac{1}{R+1}\sum_{r=1}^{R+1}\epsilon_{r-1} = \mathcal{O}\left(\frac{1}{R+1}\sum_{r=1}^{R+1}\frac{1}{r}\right) = \mathcal{O}\left(\frac{1+\log(R+1)}{R+1}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{R+1}\right).$$

Therefore, invoking part (i), we obtain

$$\mathbb{E}\big[\text{dist}^2\big(0, \partial_\eta f(\bar{x}^R)\big)\big] \leq \mathcal{O}\left(\frac{L_0^{\text{imp}}}{\eta R}\left(n^{3/2} + n\sqrt{\frac{m}{s}}\right)\right) + \mathcal{O}\left(\frac{(L_0^{\text{imp}})^{3/2}n^{3/4}}{\sqrt{\eta R}}\right) + \mathcal{O}\left(\frac{n^2(L_{0,y}^f)^2}{\eta^2 R}\right).$$

(ii.b) Regarding part (ii.a), in order to guarantee that $\mathbb{E}\left[\operatorname{dist}^2\left(0, \partial_\eta f(\bar{x}^R)\right)\right] \leq \varepsilon$, it suffices to choose the number of communication rounds as

$$R_\varepsilon = \max\left\{\mathcal{O}\left(\frac{L_0^{\mathrm{imp}}}{\eta\varepsilon}\left(n^{3/2} + n\sqrt{\frac{m}{s}}\right)\right), \mathcal{O}\left(\frac{(L_0^{\mathrm{imp}})^3 n^{3/2}}{\eta\varepsilon^2}\right), \mathcal{O}\left(\frac{n^2 (L_{0,y}^f)^2}{\eta^2\varepsilon}\right)\right\}.$$

(ii.c) The lower-level problems are solved locally and therefore do not require communication between the clients and the server. Consequently, the total number of communication rounds is determined by the upper-level communication complexity in part (ii.b). Hence, $\bar{R}_\varepsilon = R_\varepsilon$. □