Communication-Efficient Federated Hypergradient Computation via Aggregated Iterative Differentiation

Peiyao Xiao [1] **Kaiyi Ji** [1]

Abstract

Federated bilevel optimization has attracted increasing attention due to emerging machine learning and communication applications. The biggest challenge lies in computing the gradient of the upper-level objective function (i.e., hypergradient) in the federated setting due to the nonlinear and distributed construction of a series of global Hessian matrices. In this paper, we propose a novel communication-efficient federated hypergradient estimator via aggregated iterative differentiation (AggITD). AggITD is simple to implement and significantly reduces the communication cost by conducting the federated hypergradient estimation and the lower-level optimization simultaneously. We show that the proposed AggITD-based algorithm achieves the same sample complexity as existing approximate implicit differentiation (AID)-based approaches with much fewer communication rounds in the presence of data heterogeneity. Our results also shed light on the great advantage of ITD over AID in the federated/distributed hypergradient estimation. This differs from the comparison in the non-distributed bilevel optimization, where ITD is less efficient than AID. Our extensive experiments demonstrate the great effectiveness and communication efficiency of the proposed method.

1. Introduction

Bilevel optimization has drawn significant attention from the machine learning (ML) community due to its wide applications in ML including meta-learning (Finn et al., 2017; Rajeswaran et al., 2019), automated hyperparameter optimization (Franceschi et al., 2018; Feurer & Hutter, 2019), reinforcement learning (Konda & Tsitsiklis, 1999; Hong et al., 2020), adversarial learning (Zhang et al., 2022; Liu et al., 2021a), signal processing (Kunapuli et al., 2008) and AI-aware communication networks (Ji & Ying, 2023). Existing studies on bilevel optimization have mainly focused on the single-machine scenario. However, due to computational challenges such as the second-order hypergradient computation and the increasing scale of problem models (e.g., deep neural networks), learning on a single machine turns out to be inefficient and unscalable. In addition, data privacy has also arisen as a critical concern in the single-machine setting recently (McMahan et al., 2017). These challenges have greatly motivated the recent development of federated bilevel optimization, with emerging applications such as federated meta-learning (Tarzanagh et al., 2022), hyperparameter tuning for federated learning (Huang et al., 2022), resource allocation over edges (Ji & Ying, 2022) and graph-aided federated learning (Xing et al., 2022) etc.

Mathematically, federated bilevel optimization takes the following formulation with m clients.

$$\min_{x\in\mathbb{R}^{d_1}} \quad f(x)=\frac{1}{m}\sum_{i=1}^{m} f_i(x,y^*_{(x)})$$

$$\text{subject to}\ \ y^*_{(x)}\in\arg\min_{y\in\mathbb{R}^{d_2}}\frac{1}{m}\sum_{i=1}^{m} g_i(x,y), \tag{1}$$

where the upper- and lower-level functions $f_i(x,y)=\mathbb{E}_{\xi_i}F_i(x,y;\xi_i)$ and $g_i(x,y)=\mathbb{E}_{\zeta_i}G_i(x,y;\zeta_i)$ for each client i are jointly continuously differentiable. To efficiently solve the distributed nested problem in Equation (1), the biggest challenge lies in computing the gradient of the upper-level objective, i.e., the hypergradient $\nabla f(x)$, due to the approximation of a global Hessian inverse matrix and the client drift induced by the data heterogeneity (Karimireddy et al., 2020; Hsu et al., 2019). To overcome these issues, existing approaches all focus on the AID-based federated hypergradient estimation (Huang et al., 2022; Tarzanagh et al., 2022). However, the AID-based approaches naturally contain two consecutive loops at each outer iteration, each of which contains a large number of communication rounds, for minimizing the lower-level objective and constructing the federated hypergradient estimate, separately, as shown in the left illustration in Figure 1. This heavily complicates

[1] Department of Computer Science and Engineering, University at Buffalo, Baffalo, USA. Correspondence to: Kaiyi Ji <kaiyiji@buffalo.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

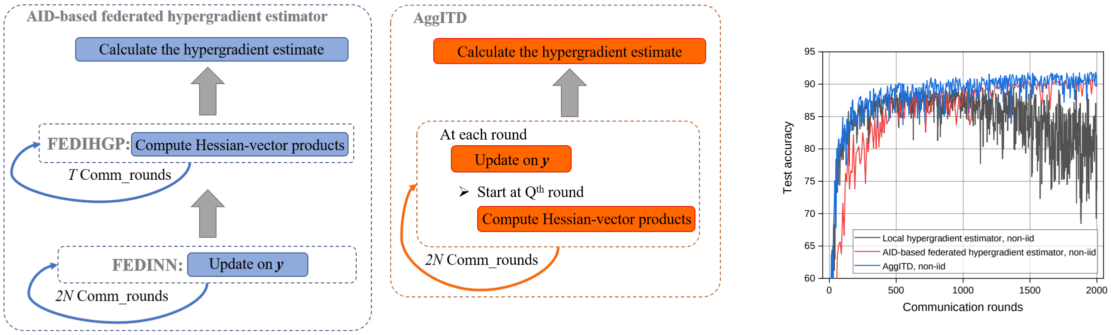

Figure 1. Comparison between AID-based FHE (left) in FedNest (Tarzanagh et al., 2022) and our proposed AggITD estimator (middle). The right plot compares the performance among fully local hypergradient estimator (i.e., using only local information), AID-based FHE and AggITD in federeated hyper-representation learning in the presence of data heterogeneity.

Hypergradient estimators	Comm_rounds/Outer_itr	Comm_loops/Outer_itr	Sample complexity
AID-based FHE (Tarzanagh et al., 2022)	$2N + T + 3$	2	$\widetilde{\mathcal{O}}(\epsilon^{-2})$
AggITD (this paper)	$2N + 3$	1	$\widetilde{\mathcal{O}}(\epsilon^{-2})$

Table 1. Comparison of AID-based FHE and the proposed AggITD in the presence of data heterogeneity. Communication round: the procedure that "**for** $i \in S$, in parallel **do**", where the participating clients send their local information (gradients or Hessian-vector products) to the server for aggregation, and the aggregated information is then broadcast back to clients. N and T: the number of iterations for optimizing the lower-level objective and approximating the global Hessian-inverse-vector product, respectively. Sample complexity: the total number of samples to achieve an ϵ-accurate stationary point. $\widetilde{\mathcal{O}}$: hide $\log$ factors.

the implementation and increases the communication cost.

1.1. Main contributions

In this paper, we propose a new federated hypergradient estimator (FHE) via aggregated iterative differentiation, which we refer to as AggITD. As shown in Figure 1, our AggITD estimator leverages intermediate iterates of the lower-level updates on y for the federated hypergradient estimation rather than the last iterate as in AID-based methods, and hence admits a simpler implementation and much fewer communication rounds by conducting the lower-level updates on y and the Hessian-vector-based hypergradient estimation simultaneously within the same communication loop. Our detailed contributions are summarized as below.

A new ITD scheme. We first show that existing ITD-based approaches in the non-distributed setting (Franceschi et al., 2018; Grazzi et al., 2020; Ji et al., 2021) rely on the accomplishment of the lower-level updates on y for the matrix-vector-based hypergradient estimation, and hence still requires two long communication loops for the federated hypergradient estimation (see Section 2.2 for more details). In contrast, we propose a new iterative differentiation process suitable for the efficient distributed implementation, which starts the matrix-vector based hypergradient estimation at a randomly sampled intermediate lower-level iterate, as illustrated in Figure 1. We anticipate that our estimator can be of independent interest to other distributed settings such as decentralized or asynchronous bilevel optimization.

Communication-efficient bilevel optimization. Building on the proposed AggITD, we further develop a federated bilevel optimization algorithm named FBO-AggITD, which incorporates the technique of federated variance reduction into the lower- and upper-level updates on y and x to mitigate the impact of the client drift on the hypergradient estimation accuracy. FBO-AggITD contains only a single communication loop, where only efficient matrix-vector products rather than Hessian or Hessian-inverse matrices are computed and communicated for the global Hessian-inverse-vector approximation.

New theoretical analysis. We provide a novel error and convergence analysis for the proposed AggITD estimator and FBO-AggITD algorithm, respectively. The analysis addresses two major challenges. First, differently from the AID-based estimator, the proposed AggITD depends on less accurate intermediate iterates $y^t, t = Q + 1, ..., N$ at a random index Q, which may introduce uncontrollable estimation errors due to the client drift. Second, the randomness from stochastic Hessian matrices and gradients further complicates the analysis. In fact, there has been no analysis even for non-distributed stochastic ITD-based estimators. To this end, a tighter recursion type of analysis is developed by decoupling the errors induced by the lower-level updates and the global Hessian-inverse-vector approximation. As shown in Table 1, AggITD achieves the same sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-2})$ as the AID-based FHE (Tarzanagh et al., 2022), with much fewer communication rounds.

Strong empirical performance. As shown in the right plot of Figure 1, AggITD admits a much faster convergence rate w.r.t. communication rounds and better test accuracy than AID-based FHE. In addition, compared to the fully local hypergradient estimator (which is computed using only local client data), AggITD achieves a much higher test accuracy with a comparable rate and is much more stable with lower variance. This demonstrates the importance of aggregation under the lower-level heterogeneity. Such comparisons are also observed in the experiments in Section 5.

1.2. Related work

Bilevel optimization. A large body of bilevel optimization methods have been proposed since the work in Bracken & McGill 1973. For example, Hansen et al. 1992; Gould et al. 2016; Shi et al. 2005; Sinha et al. 2017 reduced the bilevel problem to the single-level constraint-based problem. Gradient-based methods have drawn more attention in machine learning recently, which can be generally categorized into AID (Domke, 2012; Pedregosa, 2016; Liao et al., 2018; Arbel & Mairal, 2022) and ITD (Maclaurin et al., 2015; Franceschi et al., 2017; Finn et al., 2017; Shaban et al., 2019; Grazzi et al., 2020) based methods. Various stochastic bilevel optimizers have also been developed via momentum (Yang et al., 2021; Huang & Huang, 2021; Guo & Yang, 2021), variance reduction (Yang et al., 2021; Dagréou et al., 2022), Neumann series (Chen et al., 2021b; Ji et al., 2021). Theoretically, the convergence of bilevel optimization has been analyzed by Franceschi et al. 2018; Shaban et al. 2019; Liu et al. 2021b; Ghadimi & Wang 2018; Ji et al. 2021; Hong et al. 2020. More results and details can be found in the survey by Liu et al. 2021a. In this paper, we propose a new stochastic ITD-based hypergradient estimator, which is further extended to the federated setting.

Federated learning. Federated Learning was firstly introduced to allow different clients to train a model collaboratively without sharing data (Konečný et al., 2015; Shokri & Shmatikov, 2015; Mohri et al., 2019). As one of the earliest methods, FedAvg has been shown to effectively reduce the communication cost (McMahan et al., 2017). An increasing number of variants of FedAvg have been further proposed to address the issues such as the slow convergence and client drift via regularization (Li et al., 2020; Acar et al., 2021), variance reduction (Mitra et al., 2021; Karimireddy et al., 2020), proximal splitting (Pathak & Wainwright, 2020) and adaptive optimization (Reddi et al., 2020). In the homogeneous setting, FedAvg is relevant to local SGD, and has been analyzed in Stich 2019; Wang & Joshi 2018; Stich & Karimireddy 2019; Basu et al. 2019. In the heterogeneous setting, Li et al. 2020; Wang et al. 2020; Mitra et al. 2021; Li et al. 2019; Khaled et al. 2019 provided the convergence analysis of their methods.

Federated bilevel optimization. Recent works (Gao, 2022; Li et al., 2022) focused on the homogeneous setting, and proposed momentum-based methods with fully local hypergradient estimators. The most relevant work (Tarzanagh et al., 2022) proposed FedNest using an AID-based FHE, and further provided its convergence rate guarantee despite the data heterogeneity. This paper proposes a simple and communication-efficient method via an ITD-based FHE.

Bilevel optimization has also been studied in other distributed setups such as decentralized bilevel optimization (Chen et al., 2022; Yang et al., 2022; Lu et al., 2022) and asynchronous bilevel optimization over directed network (Yousefian, 2021). We anticipate that our proposed ITD-based estimator can be also applied to these scenarios.

Notations. We use $\partial f(x, y^*_{(x)})/\partial x$ to denote the gradient of f as a function of x, and $\nabla_x f$ and $\nabla_y f$ are partial derivatives of f with respect to x and y. For any vector v and matrix M, we denote $\|v\|$ and $\|M\|$ as Euclidean and spectral norms, respectively. We let $f(x,y) = \frac{1}{m}\sum_{i=1}^{m} f_i(x,y)$ and $g(x,y) = \frac{1}{m}\sum_{i=1}^{m} g_i(x,y)$ denote the averaged upper- and lower-level objective functions across all clients i. Finally, let $S = \{1, ..., m\}$ denote the set of all clients.

2. Federated Hypergradient Computation

2.1. Federated Hypergradient and Existing Approach

Federated hypergradient. The biggest challenge of federated bilevel optimization lies in computing the aggregated hypergradient $\nabla f(x) = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial f_i(x, y^*_{(x)})}{\partial x}$ due to the implicit dependence of the global lower-level solution $y^*_{(x)}$ on x. Using the implicit function theorem (Griewank & Walther, 2008) and if $g(\cdot)$ is twice differentiable and $\nabla_y^2 g(x, y^*_{(x)})$ is invertible, an explicit form of $\nabla f(x)$ is

$$\begin{aligned}\nabla f(x) =& \frac{1}{m}\sum_{i=1}^{m}\Big(\nabla_x f_i(x, y^*_{(x)}) - \nabla_x\nabla_y g(x, y^*_{(x)}) \\ &\times \left[\nabla_y^2 g(x, y^*_{(x)})\right]^{-1}\nabla_y f_i(x, y^*_{(x)})\Big),\end{aligned} \tag{2}$$

where the first and second terms on the right side are direct and indirect parts of the federated hypergradient. As shown by Equation (2), two challenges arise in the federated hypergradient computation. First, the second-order derivatives $\nabla_x\nabla_y g(x, y^*_{(x)})$ and $\nabla_y^2 g(x, y^*_{(x)})$ are all global information that is not accessible to each client i. This greatly complicates the design of an unbiased estimate of $\nabla f(x)$. For example, it can be seen that a straightforward estimator by replacing such two global quantities with their local counterparts, i.e., $\nabla_x\nabla_y g_i(x, y^*_{(x)})$ and $\nabla_y^2 g_i(x, y^*_{(x)})$ is a biased approximation of $\nabla f(x)$ due to the client drift. Second, it is highly infeasible to compute and communicate second-order information (such as Hessian inverse or even Hessian/Jacobian matrices) due to the restrictive computing and communication resource.

AID-based FHE. To address these challenges, Tarzanagh et al. 2022 recently proposed a matrix-vector-based FHE building on a non-federated AID-based estimate used in Ghadimi & Wang 2018, which takes the form of

$$\hat{h}^I(x) = \frac{1}{m}\sum_{i=1}^{m}\left[\nabla_x F_i(x, y^N; \xi_i) - \nabla_x \nabla_y G_i(x, y^N) p_{T'}\right]$$

where y^N is first obtained to estimate the global $y^*_{(x)}$ via a FedSVRG (Mitra et al., 2021; Konečný et al., 2016) type of method with $2N$ communication rounds and an aggregated Hessian-inverse-vector (HessIV) estimate

$$p_{T'} = \prod_{t=1}^{T'}\big(I - \lambda\frac{1}{|S_t|}\sum_{i\in S_t}\nabla_y^2 G_i(x, y^N; \zeta_{i,t})\big)p_0$$

$$\text{with} \quad p_0 = \frac{\lambda T}{|S|}\sum_{i\in S}\nabla_y F_i(x, y^N; \xi_{i,0})$$

is then constructed based on the inner output y^N using extra T' communication rounds as, for $t = 1, ..., T'$

$$\text{Local client } i: \quad p_{i,t} = (I - \lambda\nabla_y^2 G_i(x, y^N; \zeta_{i,t}))p_{t-1}$$

$$\text{Server aggregates:} \quad p_t = \frac{1}{|S_t|}\sum_{i\in S_t} p_{i,t},$$

where T' is chosen from $\{0, ..., N-1\}$ uniformly at random. However, several challenges still remain, as elaborated in the next Section 2.2.

2.2. Our Method: Aggregated Iterative Differentiation

Challenges in AID-based FHC. Note that at each outer iteration k, AID-based FedIHGP includes two major communication loops, i.e., $2N$ rounds for inner y updates and T' rounds for outer FHC, which introduce two challenges in practice. First, the construction of an AID-based hypergradient estimate is built on the output y^N is inherently separated from the inner y updating loop, and the resulting two communication and optimization loops complicate the implementation in practice. Second, the separate T' (which can be large at an order of $\kappa \log\frac{1}{\epsilon}$ in the worst case (Tarzanagh et al., 2022)) communication rounds for the HessIV estimation can add a non-trivial communication burden on the FL systems due to the limited communication bandwidth and resource (e.g., in wireless setting). Then, an important question here is: *Can we develop a new FHE that can address these implementation and communication challenges simultaneously, while achieving better communication and computational performance in theory and in practice?* In this section, we provide an affirmative answer to this question by developing a novel aggregated iterative differentiation (AggITD) for communication-efficient FHC.

Our idea. Instead of constructing the federated hypergradient after obtaining the inner output y^N, our idea is to utilize the intermediate iterates $y^1, ..., y^N$ and communication rounds of the inner y loop also for the federated hypergradient approximation, and hence remove the expensive T' communication rounds. To do this, one possible solution is to use the idea of an ITD-based method from the non-federated bilevel optimization (Ji et al., 2021; Grazzi et al., 2020), which approximates the hypergradient $\frac{\partial f(x, y^*(x))}{\partial x}$ by computing $\frac{\partial f(x, y^N)}{\partial x}$ via the automatic differentiation, where y^N is the N-step output of gradient descent[1], i.e., $y_{t+1} = y_t - \alpha\nabla_y g(x, y_t)$ for $t = 0, ..., N-1$. The explicit form of the indirect part of $\frac{\partial f(x, y^N)}{\partial x}$ is then taken as

$$-\alpha\sum_{t=0}^{N-1}\nabla_x\nabla_y g(x, y^t) \times \prod_{j=t+1}^{N-1}(I - \alpha\nabla_y^2 g(x, y^j))\nabla_y f(x, y^N), \qquad (3)$$

which, however, still needs an extra communication loop for the construction because its matrix-vector computations require the information of $\nabla_y f(x, y^N)$ at the output y^N, and in addition, the N summations complicate the federated implementation. To this end, we next provide a novel aggregated ITD-based estimator for FHC, which uses the same communication loop for both the y updates and the federated hypergradient construction.

Algorithm 1 $\widetilde{h},\ y^N =$ **AggITD**(x, y, β)

```
Set $y^0 = y$ and choose $Q$ from $\{0, ..., N\}$ UAR
for $t = 0, 1, 2, ..., N$ do
    for $i \in S$ in parallel do
        Compute $q_i^t = \nabla_y G_i(x, y^t; \zeta_{i,t})$ for $y$ updates
        if $t = Q$, compute $r_i^t = \nabla_y F_i(x, y^t; \xi_{i,t})$
        if $t \geq Q + 1$ and $t \leq N$, compute $z_i^t = z^{t-1} - \frac{\partial\langle\nabla_y G_i(x, y^t; u_{i,t}), z^{t-1}\rangle}{\partial y^t}$ via autograd
    end for
    if $t \leq N - 1$ then
        Server aggregates and broadcasts $q^t = \frac{1}{|S|}\sum_{i\in S} q_i^t$
        $y^{t+1} =$ One-Round-Lower$(x, y^t, q^t, \beta)$
    end if
    if $t = Q$, aggregate $z^t := r^t = \frac{1}{|S|}\sum_{i\in S} r_i^t$
    if $t \geq Q+1$ and $t \leq N$, aggregate $z^t = \frac{1}{|S|}\sum_{i\in S} z_i^t$
end for
$p = \lambda(N+1)z^{N-1}$ if $Q < N$ or $\lambda(N+1)z^N$ otherwise.
for $i \in S$ in parallel do
    $\widetilde{h}_i = \nabla_x F_i(x, y^N; \xi_i) - \frac{\partial\langle\nabla_y G_i(x, y^N; \chi_i), p\rangle}{\partial x}$
end for
Server aggregates $\widetilde{h} = \frac{1}{|S|}\sum_{i\in S}\widetilde{h}_i$
```

[1] We take GD as an illustration example, and other solvers can also be used

Algorithm 2 $y_+ =$ **One-Round-Lower**(x, y, q, β)

```
for $i \in S$ in parallel do
    $y_0^i = y$ and choose $\beta^i \in (0, \beta]$
    for $\upsilon = 0, 1, 2, ..., \tau_i - 1$ do
        $q_\upsilon^i = \nabla_y G_i(x, y_\upsilon^i; \zeta_\upsilon^i) - \nabla_y G_i(x, y; \zeta_\upsilon^i) + q$
        $y_{\upsilon+1}^i = y_\upsilon^i - \beta^i q_\upsilon^i$
    end for
end for
$y_+ = \frac{1}{|S|} \sum_{i \in S} y_{\tau_i}^i$
```

Proposed AggITD. As shown in Algorithm 1 and the illustration in Figure 1, AggITD first samples an index Q from the set $\{0, ..., N\}$ uniformly at random, and then at each inner iteration t, each client i computes the local gradient $\nabla_y G_i(x, y^t; \zeta_{i,t})$, which are aggregated for optimizing the lower-level objective via the federated SVRG-type One-Round-Lower sub-procedure in Algorithm 2. The steps in lines 5-6 and 12-13 provide an efficient iterative way to construct a novel estimate of federated Hessian-inverse-vector product $(\nabla_y^2 g(x, y_{(x)}^*))^{-1} \nabla_y f(x, y_{(x)}^*)$, which is given by

$$\begin{aligned}\widehat{\text{HessIV}} =& \lambda(N+1) \prod_{t=N}^{Q+1} \big(I - \frac{\lambda}{|S|} \sum_{i \in S} \nabla_y^2 G_i(x, y^t; u_{i,t})\big) \\ & \times \Big[\frac{1}{|S|} \sum_{i \in S} \nabla_y F_i(x, y^Q; \xi_{i,Q})\Big],\end{aligned}$$

where we use $\prod_{j=N}^{N+1}(\cdot) = I$ for simplicity. Note that these steps for the FHC process compute and communicate only efficient Hessian-vector products $\frac{\partial \langle \nabla_y G_i(x, y^t; u_{i,t}), z^{t-1} \rangle}{\partial y^t} = \nabla_y^2 G_i(x, y^t; u_{i,t}) z^{t-1}$ using automatic differentiation (e.g., **torch.autograd**), rather than Hessian or Hessian inverse matrices. After broadcasting the global $\widehat{\text{HessIV}}$, each client i builds a local FHE $\widetilde{h}_i(x) = \widetilde{h}_i^D(x) - \widetilde{h}_i^I(x)$, where the direct and indirect parts are given by

$$\begin{aligned}\widetilde{h}_i^D(x) =& \nabla_x F_i(x, y^N; \xi_i) \\ \widetilde{h}_i^I(x) =& \nabla_x \nabla_y G_i(x, y^N; \chi_i) \widehat{\text{HessIV}}.\end{aligned}$$

Then, the aggregated hypergradient estimate is given by $\widetilde{h}(x) = \widetilde{h}^D(x) - \widetilde{h}^I(x) = \frac{1}{|S|} \sum_{i \in S} \widetilde{h}_i(x)$. Meanwhile, we would like to point out the differences between our method and distributed bilevel problems, such as (Yang et al., 2022). First, in our algorithm, the server is to aggregate the local weights from clients and broadcast the aggregated weights back to the clients. In contrast, for such decentralized methods, the server needs to compute the gradients or hypergradients. Then, our method runs multiple local updates to improve communication efficiency, whereas the decentralized methods do not have such operations. Third, all such decentralized methods use the AID-based hypergradient estimator, whereas our method uses the ITD-based scheme. However, to analyze this AggITD-based estimator, several technical challenges arise as below.

Technical challenges. First, differently from the AID-based FHE that is evaluated at the last iterate y^N, our proposed estimator depends on less accurate intermediate iterates $y^t, t = Q+1, ..., N$, which may introduce larger or even uncontrollable estimation errors given the client drift effect. Thus, a more careful and tighter analysis is required. Second, the randomness from stochastic Hessian matrices and gradients further complicates the analysis. In fact, there has been no analysis for even non-federated (i.e., $|S| = 1$) **stochastic** ITD-based estimators. Third, the aggregation $\frac{1}{|S|} \sum_{i \in S}$ complicates the bias and variance analysis.

3. Proposed Algorithm

We now develop a new federated bilevel optimizer named FBO-AggITD based on the proposed AggITD estimator. As shown in Algorithm 3, FBO-AggITD first obtains the federated hypergradient estimate $\widetilde{h}$ and the approximate $y_{k+1} = y_k^N$ of the lower-level solution y_k^* via the **AggITD** sub-procedure in Algorithm 1. Then, building on $\widetilde{h}$ and y_{k+1}, similarly to (Tarzanagh et al., 2022), we use a local SVRG-type **One-Round-Upper** sub-procedure for solving the upper-level problem w.r.t. x, where each client i runs τ_i steps based on the radient $h_{i,\upsilon}$ given by

$$\begin{aligned}h_{i,\upsilon} =& h - \nabla_x F_i(x, y_+; \xi_\upsilon^i) + \nabla_x F_i(x_\upsilon^i, y_+; \xi_\upsilon^i) \\ =& \widetilde{h}^D(x) - \widetilde{h}^I(x) - \nabla_x F_i(x, y_+; \xi_\upsilon^i) \\ & + \nabla_x F_i(x_\upsilon^i, y_+; \xi_\upsilon^i),\end{aligned}$$

where the direct part $\widetilde{h}^D(x) = \frac{1}{|S|} \sum_{i \in S} \nabla_x F_i(x, y_+; \xi_i)$ of the global hypergradient estimate $\widetilde{h}$ uses different samples ξ_i from ξ_υ^i of the local gradient $\nabla_x F_i(x, y_+; \xi_\upsilon^i), i \in S$ to provide an SVRG-type variance reduction effect on the direct part of the hypergradient. This is in contrast to the upper update in FedNest (Tarzanagh et al., 2022) where the data samples ξ_i and ξ_υ^i are chosen to be the same. Note that we do not apply the SVRG-type updates to the entire hypergradient but only the direct part because the indirect part requires the global Hessian information at iterates x_υ^i, which is infeasible at each client i.

Algorithm 3 FBO-AggITD

```
Input: $K, N \in \mathbb{N}, \alpha_k, \beta_k > 0$, initializations $x_0, y_0$.
for $k = 0, 1, 2, ..., K$ do
    $\widetilde{h},\ y_{k+1} = $ AggITD$(x_k, y_k, \beta_k)$
    $x_{k+1} = $ One-Round-Upper$(x_k, y_{k+1}, \widetilde{h}, \alpha_k)$
end for
```

Algorithm 4 $x_+ =$ **One-Round-Upper**(x, y_+, h, α)

```
for $i \in S$ in parallel do
    $x_0^i = x$ and choose $\alpha^i \in (0, \alpha]$
    for $\upsilon = 0, 1, 2, ..., \tau_i - 1$ do
        $h_{i,\upsilon} = h - \nabla_x F_i(x, y_+; \xi_\upsilon^i) + \nabla_x F_i(x_\upsilon^i, y_+; \xi_\upsilon^i)$
        $x_{\upsilon+1}^i = x_\upsilon^i - \alpha^i h_{i,\upsilon}$
    end for
end for
$x_+ = \frac{1}{|S|} \sum_{i \in S} x_{\tau_i}^i$
```

4. Main Results

4.1. Definitions and Assumptions

Let $z = (x, y) \in \mathbb{R}^{d_1 + d_2}$. Throughout this paper, we make the following definitions and standard assumptions on the lower- and upper-level objectives, as also adopted in stochastic bilevel optimization (Ji et al., 2021; Hong et al., 2020; Khanduri et al., 2021; Chen et al., 2021a) as well as in the federated bilevel optimization (Tarzanagh et al., 2022).

Definition 1. *A mapping f is L-Lipschitz continuous if for $\forall\, z, z'$, $\|f(z) - f(z')\| \le L\|z - z'\|$.*

Since the objective $f(x)$ is nonconvex, algorithms are expected to find an ϵ-accurate stationary point defined below.

Definition 2. *We say $\bar{x}$ is an ϵ-accurate stationary point of the objective function $f(x)$ if $\mathbb{E}\|\nabla f(\bar{x})\|^2 \le \epsilon$, where $\bar{x}$ is the output of an algorithm.*

Assumption 1. *The lower-level function $G_i(x, y; \zeta_i)$ is μ-strongly-convex w.r.t. y for any ζ_i.*

The following assumption imposes the Lipschitz conditions on the lower- and upper-level functions for each client i.

Assumption 2. *The objective functions satisfy*

- *The function $F_i(z; \xi_i)$ is M-Lipschitz continuous.*
- *The gradients $\nabla F_i(z; \xi_i)$ and $\nabla G_i(z; \zeta_i)$ are unbiased estimators of $\nabla f_i(z)$ and $\nabla g_i(z)$.*
- *The gradients $\nabla F_i(z; \xi_i)$ and $\nabla G_i(z; \zeta_i)$ are L_f- and L_g-Lipschitz continuous, respectively.*

Assumption 3. *The second-order derivatives satisfy*

- *The derivatives $\nabla_x \nabla_y G_i(z; \zeta_i)$ and $\nabla_y^2 G_i(z; \zeta_i)$ are unbiased estimators of $\nabla_x \nabla_y g_i(z)$ and $\nabla_y^2 g_i(z)$.*
- *The derivatives $\nabla_x \nabla_y G_i(z; \zeta_i)$ and $\nabla_y^2 G_i(z; \zeta_i)$ are ρ-Lipschitz continuous.*

Assumption 4. *The variances of gradients $\nabla F_i(z; \xi_i)$ and $\nabla G_i(z; \zeta_i)$ are bounded by σ_f^2 and σ_1^2. Moreover, the lower-level client dissimilarity $\mathbb{E}\|\nabla g_i(z) - \nabla g(z)\|^2 \le \sigma_2^2$.*

In this paper, let $\sigma_g^2 = \max\{\sigma_1^2, \sigma_2^2\}$ for notational simplicity. Assumption 4 is commonly adopted in the heterogeneous FL, and it is reduced to the homogeneous setting when $\sigma_2 = 0$. It is worth noting that our assumptions are exactly the same as existing AID-based federated/distributed bilevel studies such as (Tarzanagh et al., 2022).

4.2. Estimation Properties for AggITD

We analyze the estimation properties of AggITD. Let

$$\begin{aligned} B^I =& \mathbb{E}\left[\|\mathbb{E}[\widetilde{h}^I(x)] - \nabla_x \nabla_y g(x, y^N) \right. \\ & \left. \times (\nabla_y^2 g(x, y^N)^{-1}) \nabla_y f(x, y^N)\|^2 \,|\, x, y^N\right] \end{aligned}$$

denote the estimation error of the indirect part of $\widetilde{h}^I(x)$.

Proposition 1. *Suppose Assumptions 1-4 are satisfied and let $y_{(x)}^* = \arg\min_y g(x, y)$. Further, set $\lambda \le \min\{10, \frac{1}{L_g}\}$ and $\beta^i = \frac{\beta}{\tau_i}$ and any stepsize $\alpha > 0$, where $\beta \le \min\{1, \lambda, \frac{1}{6L_g}\}$. Then, we have*

$$\begin{aligned} B^I \le& [4\lambda^2 L_g^2 M^2 \alpha_1(N) + 4\lambda^2 L_f^2 L_g^2 \alpha_3(N)]\, \mathbb{E}\, \|y - y_{(x)}^*\|^2 \\ &+ \frac{4L_g^2 M^2 (1 - \lambda\mu)^{2N+2}}{\mu^2} + 400\lambda^2 \beta^2 L_g^2 M^2 \sigma_g^2 \rho^2 \alpha_2(N) \\ &+ \frac{200\lambda\beta^2 \sigma_g^2 L_f^2 L_g^2 N(N+1)}{\mu}, \end{aligned} \tag{4}$$

where $\alpha_1(N) = 4(N+1)(1 - \frac{\beta\mu}{2})^N[\frac{\rho^2}{\lambda\mu^3} + \frac{4\rho^2}{\beta\mu^3}]$, $\alpha_2(N) = \frac{N(N+1)(1+(1-\lambda\mu)^2)}{\lambda\mu^3}$ and $\alpha_3(N) = \frac{3(N+1)(1-\frac{\beta\mu}{2})^N}{\lambda\mu}$.

Proposition 1 provides an upper bound on the second moment of the estimation bias of the AggITD estimator. As shown in Equation (4), the first two terms $\mathcal{O}((1 - \frac{\beta\mu}{2})^N\, \mathbb{E}[\|y - y_{(x)}^*\|^2])$ and $\mathcal{O}((1 - \lambda\mu)^{2N+2})$ correspond to the estimation errors without the client drift, which can be made small by choosing N properly. In addition, the initialization gap $\mathbb{E}[\|y - y_{(x)}^*\|^2]$ further relaxes the requirement of N due to the warm start $y_k = y_{k-1}^N$ (see Algorithm 3), as shown in the final convergence analysis. It is worth mentioning that these two terms match the error bound of the stochastic AID-based hypergradient estimator in non-federated setting (Ji et al., 2021; Ghadimi & Wang, 2018; Chen et al., 2021a), and hence our analysis can be of independent interest to non-federated bilevel optimization. Also note that the last two error terms $\mathcal{O}(\lambda^2\beta^2)$ and $\mathcal{O}(\lambda\beta^2)$ are induced by the client drift in the y updates, which exists especially in the FL, can be addressed by choosing a sufficiently small stepsize β. Technically, we first show via a recursive analysis that the key approximation error between the expected indirect part of the AggITD estimator

$$\begin{aligned} \mathbb{E}[\widetilde{h}^I(x) | x, y^N] =& \lambda \nabla_x \nabla_y g(x, y^N) \\ & \times \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) \nabla_y f(x, y^Q) \end{aligned} \tag{5}$$

and the underlying truth is bounded by

$$\mathcal{O}\Big(\sum_{Q=0}^{N}(1-\lambda\mu)^{2N-2Q}\|y^Q-y^*_{(x)}\|^2+\|y^N-y^*_{(x)}\|^2\Big).$$

Note from Equation (5) that although the optimality gap $\|y^Q-y^*_{(x)}\|$ can be large for small Q (which is induced by our ITD-based construction), the coupling factor $(1-\lambda\mu)^{2N-2Q}$ still makes the overall bound to be small, and this validates the design principle of our AggITD estimator. Then, unconditioning on x, y^N, incorporating the convergence bounds on the iterates y^Q with intrinsic client drift, we derive the final estimation bounds on AggITD. The following proposition characterizes the estimation variance of the global indirect hypergradient estimate $\widetilde{h}^I_i(x)$ and the local hypergradient estimate at iteration v of client i.

Proposition 2. *Suppose Assumptions 1-3 are satisfied. Set $\lambda\le\min\{10,\frac{1}{L_g}\}$. Then, conditioning on x, y_+, we have*

$$\mathbb{E}\,\|\widetilde{h}^I_i(x)-\bar{h}^I_i(x)\|^2\le\sigma_h^2,$$

$$\mathbb{E}\,\|\widetilde{h}^D_i(x^i_v,y_+)-\widetilde{h}^D_i(x^i_0,y_+)+\widetilde{h}^D(x)-\widetilde{h}^I(x)\|^2\le D_h^2$$

where the constants are given by $\sigma_h^2=\frac{\lambda(N+1)L_g^2M^2}{\mu}$ and $D_h^2=12M^2+\frac{4\lambda(N+1)L_g^2M^2}{\mu}$.

Proposition 2 demonstrates that the varaince of our AggITD estimation is bounded. Based on the important bias and variance characterizations in Propositions 1 and 2, we next provide the total convergence and complexity analysis for the proposed FBO-AggITD algorithm.

4.3. Convergence and Complexity Analysis

We first provide a descent lemma on the total objective $f(x)$.

Lemma 1 (Objective descent)**.** *Suppose Assumptions 1-4 hold. Let $y^*=\arg\min_y g(x,y)$. Further, we set $\lambda\le\min\{10,\frac{1}{L_g}\}$, $\alpha^i=\frac{\alpha}{\tau_i}$ with $\tau_i\ge 1$ for some positive α and $\beta^i=\frac{\beta}{\tau_i}$, where $\beta\le\min\{1,\lambda,\frac{1}{6L_g}\}$ $\forall i\in S$. We have*

$$\begin{aligned}
&\mathbb{E}[f(x_+)]-\mathbb{E}[f(x)]\\
&\le-\frac{\alpha}{2}\,\mathbb{E}[\|\nabla f(x)\|^2]+4\alpha^2(\sigma_h^2+\sigma_f^2)L'_f+2\alpha^2M^2L'_f\\
&-\frac{\alpha}{2}(1-4\alpha L'_f)\,\mathbb{E}\,\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{v=0}^{\tau_i-1}\bar{h}^D_i(x^i_v,y_+)-\bar{h}^I(x)\Big\|^2\\
&+\frac{3\alpha}{2}\Big[B^I(x,y)+\frac{M_f^2}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{v=0}^{\tau_i-1}\mathbb{E}[\|x^i_v-x\|^2]\\
&\qquad+M_f^2\,\mathbb{E}[\|y_+-y^*\|^2]\Big] \qquad (6)
\end{aligned}$$

where the estimation bias $B^I(x,y)$ is defined in Proposition 1, and the expected quantities $\bar{h}^I(x)=\mathbb{E}[\widetilde{h}^I(x)|x,y_+]$, $\bar{h}^D_i(x^i_v,y_+)=\mathbb{E}[\widetilde{h}^D_i(x^i_v,y_+)|x^i_v]$

Note from Lemma 1 that the bound on the total objective descent contains three error terms including the FHC bias $B^I(x,y)$, which is handled by Proposition 1, the lower-level estimation error $\mathbb{E}\|y_+-y^*\|^2$, which is handled by the descent lemma on the lower-level objective function $g(x,\cdot)$, and the upper-level client drift $\sum_{i=1}^m\frac{1}{\tau_i}\sum_{v=0}^{\tau_i-1}\mathbb{E}[\|x^i_v-x\|^2]$. Also note that the bias error $B^I(x,y)$ contains the lower-level initialization gap $\mathbb{E}\|y-y^*\|^2$, which is characterized by the following lemma.

Lemma 2 (Lower-level initialization gap under warm start)**.** *Suppose Assumptions 1-4 hold. Let $y^*=\arg\min_y g(x,y)$ and $y^*_{(x_+)}=\arg\min_y g(x_+,y)$. Further, set $\alpha^i=\frac{\alpha}{\tau_i}$ with $\tau_i\ge 1$ with some $\alpha>0$, $\forall i\in S$. Then, we have*

$$\begin{aligned}
&\mathbb{E}[\|y_+-y^*_{(x_+)}\|^2]\\
&\le b_1(\alpha)\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{v=0}^{\tau_i-1}\big(\bar{h}^D_i(x^i_v,y_+)-\bar{h}^I(x)\big)\Big\|^2\Big]\\
&\qquad+b_2(\alpha)\,\mathbb{E}[\|y_+-y^*\|^2]+b_3(\alpha)(2\sigma_h^2+2\sigma_f^2+M^2)
\end{aligned}$$

where the constants are given by $b_1(\alpha)=4L_y^2\alpha^2+\frac{L_y^2\alpha^2}{4\gamma}+\frac{2L_{yx}\alpha^2}{\eta}$, $b_2(\alpha)=1+4\gamma+\frac{\eta L_{yx}D_h^2\alpha^2}{2}$, $b_3(\alpha)=4\alpha^2L_y^2+\frac{2L_{yx}\alpha^2}{\eta}$ with a flexible parameter $\gamma>0$.

As shown in the above Lemma 2, the lower-level initialization gap contains a hypergradient estimate norm $\mathcal{O}(\alpha^2)\,\mathbb{E}\,\big\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{v=0}^{\tau_i-1}\big(\bar{h}^D_i(x^i_v,y_+)-\bar{h}^I(x)\big)\big\|^2$, which is dominated by the same hypergradient norm with the factor $\Theta(-\alpha)$ in Lemma 1 for the stepsize α small enough. Then, the remaining step is to upper bound the upper-level client drift $\mathbb{E}\|x^i_v-x\|^2$.

Lemma 3 (Upper client drift)**.** *Suppose Assumptions 1-4 are satisfied. Set $\lambda\le\min\{10,\frac{1}{L_g}\}$, $\alpha^i=\frac{\alpha}{\tau_i}$ and $\beta^i=\frac{\beta}{\tau_i},\tau_i\ge 1$ where $\alpha\le\frac{1}{324M_f^2+6M_f}\le\frac{1}{6M_f}$, $\beta\le\min\{1,\lambda,\frac{1}{6L_g}\}$ $\forall i\in S$. Recall the definitions of $y^*=\arg\min_y g(x,y)$, $\bar{h}(x)=\mathbb{E}[\widetilde{h}(x)|x,y_+]$. Then, we have*

$$\begin{aligned}
\mathbb{E}[\|x^i_v-x\|^2]\le 18\tau_i^2(\alpha^i)^2\Big[&3M_f^2\,\mathbb{E}[\|y_+-y^*\|^2]\\
&+3\,\mathbb{E}[\|\nabla f(x)\|^2]+B^I(x,y)+3\sigma_h^2+6\sigma_f^2\Big]
\end{aligned}$$

where the bias $B^I(x,y)$ is defined in Proposition 1.

It can be seen from Lemma 3 that the upper-level client drift is bounded by the lower-level estimation error $\mathbb{E}\|y_+-y^*\|^2$, the total gradient norm $\mathbb{E}\|\nabla f(x)\|^2$ and the hypergradient estimation bias $B^I(x,y)$, which can be addressed by the descent lemmas on y and x (i.e., Lemma 1) and Proposition 1 for the stepsize α^i small enough. By combining the above lemmas, we next provide the general convergence analysis.

Theorem 1. *Suppose Assumptions 1-4 are satisfied. Set $\lambda\le\min\{10,\frac{1}{L_g}\}$, $\alpha^i_k=\frac{\alpha_k}{\tau_i}$ and $\beta^i_k=\frac{\beta_k}{\tau_i}$ for $i\in S$. Choose*

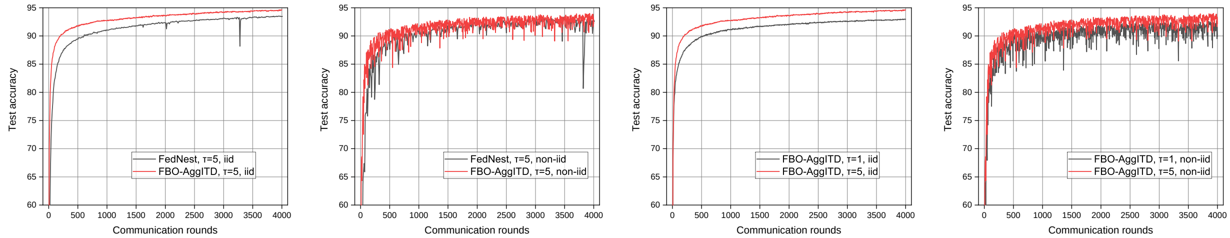

Figure 2. Hyper-representation on MNIST dataset with a 2-layer MLP with SVRG-type optimizer. Left two plots: comparison of FBO-AggITD and Fednest (Tarzanagh et al., 2022) in the i.i.d. and non-i.i.d. cases. Right two plots: the impact of the number τ of local update steps on FBO-AggITD.

Algorithm	Comm_rounds/Outer_itr	Data	Outer_ep	Comm_rounds (90%)	Final Accuracy
FedNest	2N+T+3	IID	τ=1	1630	91.68%
			τ=5	610	93.48%
		NON-IID	τ=1	1380	91.46%
			τ=5	760	92.87%
FBO-AggITD	2N+3	IID	τ=1	530	92.94%
			τ=5	195	94.61%
		NON-IID	τ=1	520	92.67%
			τ=5	305	93.88%

Table 2. Quantitative comparison between FBO-AggITD and FedNest.

parameters such that $\alpha_k = \min\{\bar\alpha_1, \bar\alpha_2, \bar\alpha_3, \frac{\bar\alpha}{\sqrt{K}}\}, \beta_k \in \big[\max\big\{\frac{\bar\beta\alpha_k}{N}, \frac{\lambda}{10}\big\}, \min\big\{1, \lambda, \frac{1}{6L_g}\big\}\big]$*, where* $\bar\alpha_1, \bar\alpha_2, \bar\alpha_3, \bar\alpha$ *and* $\bar\beta$ *are constants independent of K, whose specific forms are given in Appendix D.1. Then, the outputs of the proposed FBO-AggITD algorithms satisfy*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}\bigg(\frac{1}{\min\{\bar\alpha_1, \bar\alpha_2, \bar\alpha_3\}K} + \frac{1}{\bar\alpha\sqrt{K}} + \frac{\bar\alpha\max\{c_0, c_1\sigma_h^2, c_2, c_3\}}{\sqrt{K}} + (1-\lambda\mu)^{2N}\bigg),$$

where $c_0, c_1, c_2,$ *and* c_3 *are positive constants independent of K, whose complete forms are given in Appendix D.1.*

By specifying the parameters N and $\bar\alpha$ properly, we obtain the following complexity results.

Corollary 1. *Under the same setting as in Theorem 1, if we choose* $N = \mathcal{O}(\kappa_g)$, $\bar\alpha = \mathcal{O}(\kappa_g^{-4})$, *then we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}(\frac{\kappa_g^4}{K} + \frac{\kappa_g^4}{\sqrt{K}})$$

To achieve an ϵ*-accurate stationary point, the total number of samples required by FBO-AggITD is* $\mathcal{O}(\kappa_g^9\epsilon^{-2})$.

As shown in Corollary 1, the overall sample complexity (i.e., the total number of data samples required to achieve an ϵ-accurate stationary point) of our FBO-AggITD is $\mathcal{O}(\kappa^9\epsilon^{-2})$, which matches the sample complexities of stocBiO (Ji et al., 2021), BSA (Ghadimi & Wang, 2018) and ALSET (Chen et al., 2021a) in the non-federated bilevel optimization and FedNest (Tarzanagh et al., 2022) in the federated setting despite the data heterogeneity. Note that our method uses only $(2N+3)/(2N+T+3)$ communication rounds of FedNest (shown in Table 1) at each outer iteration. As a result, in theory, our method achieves a constant-level improvement over FedNest. To improve the dependence on ϵ, we suspect that the server-level variance reduction or periodic averaging can help, but this goes beyond the focus of this paper. We are happy to leave it for future study.

5. Experiments

In this section, we compare the performance of the proposed FBO-AggITD method with FedNest and LFedNest in Tarzanagh et al. 2022 on a hyper-representation problem. Following the problem setup in Franceschi et al. 2018, we use a 2-layer multilayer perceptron (MLP) as the backbone, where the hidden layer is optimized at the upper-level problem and the head is optimized at the lower-level problem. We study the impact of data heterogeneity on the comparison algorithms by considering both the i.i.d. and non-i.i.d. ways of data partitioning of MNIST, following the setup in McMahan et al. 2017.

The first two plots in Figure 2 compare our FBO-AggITD method with FedNest in both i.i.d. and non-i.i.d. setups with $\tau = 5$, respectively. It can be seen that FBO-AggITD converges much faster than FedNest, and achieves a higher test accuracy with much fewer communication rounds. In the non-i.i.d. case also shows that FBO-AggITD is more stable with lower variance than FedNest. The last two plots in

Figure 2 show that local updates are useful to improve communication efficiency and stabilize the training. In Table 2, it can be seen that to achieve an accuracy of 90%, our FBO-AggITD uses more than 2-3 times fewer communication rounds than FedNest, in both the i.i.d. and non-i.i.d. cases and in addition, for all four setups, FBO-AggITD achieves a higher final test accuracy than FedNest.

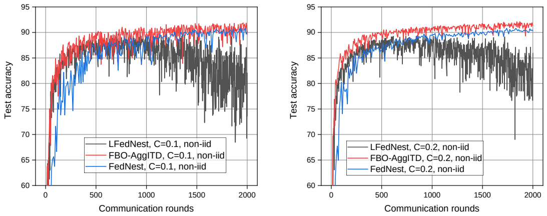

Figure 3. Comparison under different client participation ratios.

In Figure 1 and Figure 3, we compare the performance of our FBO-AggITD, FedNest, and LFedNest (which uses a fully local AID-based hypergradient estimator) given different client participation ratios (denoted as C) in the non-i.i.d. setting. It can be seen that FBO-AggITD outperforms the other two algorithms with higher communication efficiency and higher accuracy. Note that LFedNest has the largest variance and the lowest accuracy, and this validates the importance of federated hypergradient computation. All above experiments use SVRG-type optimizer which outperforms the SGD-type optimizer, shown in Figure 4.

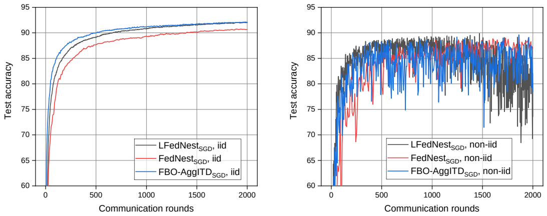

Figure 4. Performance using SGD-type optimizer.

Figure 4 compares the performance of FBO-AggITD, FedNest and LFedNest when the **One-Round-Lower** uses the SGD-type FedAvg methed. In the both i.i.d. and non-i.i.d. settings, our method (which is defined as FBO-AggITD$_{SGD}$) still performs the best with the fastest convergence rate w.r.t. the number of communication rounds. Another observation is that using the the SGD-type lower-level solver introduces a larger variance and fluctuation than the SVRG-type optimizer, by comparing Figure 2 and Figure 4. This validates the importance of variance reduction in mitigating the impact of the client drift on the convergence performance.

Finally, Figure 5 shows the performance of FBO-AggITD on CIFAR-10 with MLP/CNN network in the i.i.d. setting. We found that FedNest could not converge in this task after an extensive grid search on hyperparameters. However, our method can converge with both MLP and CNN backbones. However, the test accuracy is not satisfactory here. We suspect that it is because the objective function in hyper-representation is not good for federated setting, and a more careful network architecture should be designed for more challenging datasets. We would like to leave this for the future work.

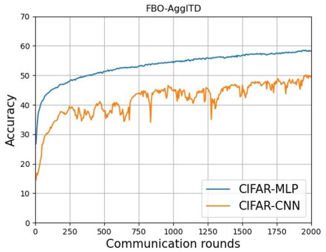

Figure 5. Performance on CIFAR-10 with MLP and CNN.

6. Conclusions

In this paper, we propose a simple and communication-efficient federated hypergradient estimator based on a novel aggregated iterative differentiation (AggITD). We show that the proposed AggITD-based algorithm achieves the same sample complexity as existing approaches with much fewer communication rounds on non-i.i.d. datasets. We anticipate our new estimator can be further applied to other distributed scenarios such as decentralized bilevel optimization.

References

Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., and Saligrama, V. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Arbel, M. and Mairal, J. Amortized implicit differentiation for stochastic bilevel optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Bracken, J. and McGill, J. T. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Chen, T., Sun, Y., and Yin, W. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021b.

Chen, X., Huang, M., and Ma, S. Decentralized bilevel optimization. *arXiv preprint arXiv:2206.05670*, 2022.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *arXiv preprint arXiv:2201.13409*, 2022.

Domke, J. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 318–326, 2012.

Feurer, M. and Hutter, F. Hyperparameter optimization. In *Automated Machine Learning*, pp. 3–33. Springer, Cham, 2019.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1126–1135, 2017.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning (ICML)*, pp. 1165–1173, 2017.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577. PMLR, 2018.

Gao, H. On the convergence of momentum-based algorithms for federated stochastic bilevel optimization problems. *arXiv preprint arXiv:2204.13299*, 2022.

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

Gould, S., Fernando, B., Cherian, A., Anderson, P., Cruz, R. S., and Guo, E. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

Griewank, A. and Walther, A. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

Guo, Z. and Yang, T. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

Hansen, P., Jaumard, B., and Savard, G. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing*, 13(5):1194–1217, 1992.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Huang, F. and Huang, H. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

Huang, Y., Lin, Q., Street, N., and Baek, S. Federated learning on adaptively weighted nodes by bilevel optimization. *arXiv preprint arXiv:2207.10751*, 2022.

Ji, K. and Ying, L. Network utility maximization with general and unknown utility functions: A distributed, data-driven bilevel optimization approach. *Submitted*, 2022.

Ji, K. and Ying, L. Network utility maximization with unknown utility functions: A distributed, data-driven bilevel optimization approach. *arXiv preprint arXiv:2301.01801*, 2023.

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

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Khaled, A., Mishchenko, K., and Richtárik, P. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 30271–30283, 2021.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Konečnỳ, J., McMahan, B., and Ramage, D. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Kunapuli, G., Bennett, K. P., Hu, J., and Pang, J.-S. Classification model selection via bilevel programming. *Optimization Methods & Software*, 23(4):475–489, 2008.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, J., Huang, F., and Huang, H. Local stochastic bilevel optimization with momentum-based variance reduction. *arXiv preprint arXiv:2205.01608*, 2022.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Liao, R., Xiong, Y., Fetaya, E., Zhang, L., Yoon, K., Pitkow, X., Urtasun, R., and Zemel, R. Reviving and improving recurrent back-propagation. In *Proc. International Conference on Machine Learning (ICML)*, 2018.

Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021a.

Liu, R., Liu, X., Yuan, X., Zeng, S., and Zhang, J. A value-function-based interior-point method for non-convex bi-level optimization. In *International Conference on Machine Learning (ICML)*, 2021b.

Lu, S., Cui, X., Squillante, M. S., Kingsbury, B., and Horesh, L. Decentralized bilevel optimization for personalized client learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5543–5547. IEEE, 2022.

Maclaurin, D., Duvenaud, D., and Adams, R. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning (ICML)*, pp. 2113–2122, 2015.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning (ICML)*, pp. 4615–4625. PMLR, 2019.

Pathak, R. and Wainwright, M. J. Fedsplit: An algorithmic framework for fast federated optimization. *Advances in Neural Information Processing Systems*, 33:7057–7066, 2020.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning (ICML)*, pp. 737–746, 2016.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 113–124, 2019.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2020.

Shaban, A., Cheng, C.-A., Hatch, N., and Boots, B. Truncated back-propagation for bilevel optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1723–1732, 2019.

Shi, C., Lu, J., and Zhang, G. An extended kuhn–tucker approach for linear bilevel programming. *Applied Mathematics and Computation*, 162(1):51–63, 2005.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Sinha, A., Malo, P., and Deb, K. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

Stich, S. U. Local sgd converges fast and communicates little. In *ICLR 2019-International Conference on Learning Representations (ICLR)*, 2019.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. Fednest: Federated bilevel, minimax, and compositional optimization. *arXiv preprint arXiv:2205.02215*, 2022.

Wang, J. and Joshi, G. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

Xing, P., Lu, S., Wu, L., and Yu, H. Big-fed: Bilevel optimization enhanced graph-aided federated learning. *IEEE Transactions on Big Data*, 2022.

Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:13670–13682, 2021.

Yang, S., Zhang, X., and Wang, M. Decentralized gossip-based stochastic bilevel optimization over communication networks. *arXiv preprint arXiv:2206.10870*, 2022.

Yousefian, F. Bilevel distributed optimization in directed networks. In *2021 American Control Conference (ACC)*, pp. 2230–2235. IEEE, 2021.

Zhang, Y., Zhang, G., Khanduri, P., Hong, M., Chang, S., and Liu, S. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning (ICML)*, pp. 26693–26712. PMLR, 2022.

Supplementary Materials

A. Further Specifications on Experiments

A.1. Additional experiments

Experiments on MNIST with CNN networks. Figure 6 compares the performance of FBO-AggITD, FedNest and LFedNest on MNIST when the backbone is chosen as CNN and One-Round-Lower uses the SVRG-type method. In the non-i.i.d. setting, it turns out that both FedNest and LFedNest failed to converge depsite of a grid search for stepsizes. The grid search on inner step sizes and outer step sizes of 4 settings are [(0.003, 0.01), (0.001, 0.005), (0.0005, 0.003), (0.0003, 0.001)]. However, our method (which is defined as FBO-AggITD) can have the ability to converge in both non-i.i.d. and i.i.d. cases with high training accuracies. The inner step szie and outer step size are chosen as [0.003, 0.01] after grid search. The training accuracies after 2000 communication rounds in i.i.d. and non-i.i.d. cases are 97.6% and 96.7%, respectively.

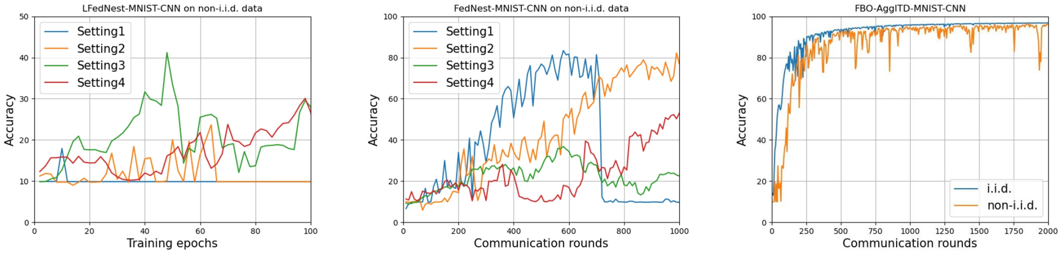

Figure 6. Performance of LFedNest, FedNest and FBO-AggITD on MNIST when the backbone is chosen as CNN and One-Round-Lower uses the SVRG-type method.

Running time comparison. The following Figure 7 shows the running time comparison between FedNest, FBO-AggITD and gossip-based method, Algorithm 2 in (Yang et al., 2022). Our FBO-AggITD archives a running time comparable to that of FedNest because both methods consume a similar number of gradient and Hessian-vector computations.However, our FBO-AggITD converges much faster than this gossip-based method, which is slower due to the computation of the Hessian and Jacobian matrices. Since no codes are provided in (Yang et al., 2022), we wrote a code for comparison.

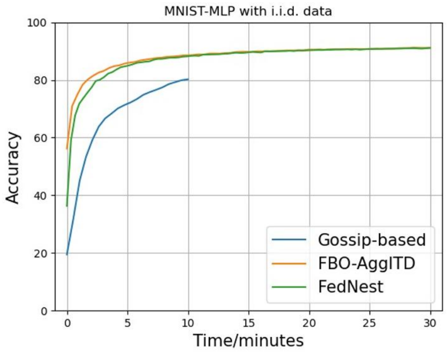

Figure 7. Performance of FedNest, FBO-AggITD and gossip-based method on MNIST when the backbone is chosen as MLP with i.i.d. data.

A.2. Model Architectures

We first follow the same experiment in (Tarzanagh et al., 2022), thus the model is a 2-layer multilayer perceptron (MLP) with 200 hidden units. The outer problem optimizes the hidden layer with 157,000 parameters, and the inner problem optimizes the output layer with 2,010 parameters. Additionally, for the CIFAR-10-CNN experiment, we use the 7-layer CNN (LeCun et al., 1998) model to train CIFAR-10. We optimize the last two fully connected layers' parameters for solving the lower-level problem and optimize the rest layers' parameters for solving the upper-level problem.

A.3. Hyperparameter settings

For all comparison methods, we optimize their hyperparameters via grid search guided by the default values in their source codes, to ensure the best performance given the algorithms are convergent.

Parameter selection for the experiments in Figure 2 and Figure 7. For FedNest and FBO-AggITD, we used the same hyperparameter configuration for both the i.i.d. and non-i.i.d. settings. In particular, the inner-stepsize is 0.003, the outer-loop stepsize is 0.01, the constant $\lambda = 0.01$ and the number of inner-loop steps is 5. The choice of the number τ of outer local epochs and the data setup are indicated in the figures. Then the default value for the client participation ratio is $C = 0.1$. Here, it is worth mentioning that for all comparison methods, we optimize their hyperparameters via grid search guided by the default values in their source codes, to ensure the best performance given the algorithms are convergent.

Parameters selection for the experiments in Figure 3 and Figure 4.

In Figure 3 and Figure 4, the choice of stepsizes and constant λ of FedNest and FBO-AggITD is the same as in Figure 2. For LFedNest, we choose the same hyperparameters as FedNest and FBO-AggITD, except that in the non-i.i.d. case, the inner- and outer-stepsizes are set smaller to be 0.001 and 0.005 to avoid the overfitting. The number τ of outer local epochs is set to be 1 for all cases. In Figure 4, the client participation ratio is $C = 0.1$, and the update optimizer in the inner loop is the SGD-type FedAvg method rather than FedSVRG. The choice of hyperparameters for Figure 6 is indicated above and for Figure 5 the choice of inner step size and the outer step size are 0.002 and 0.01, respectively while the other options keep the same.

B. Notations

For simplicity, we remove subscript k as long as the involved definitions are clear in the context. In some proof steps, we will use x and x_+ (similarly for y and y_+) to denote x_k and x_{k+1} (similarly y_k and y_{k+1}), where the definitions of x_k and y_k are given in Algorithm 3. Based on Algorithm 3, we also have the definition of $y_+ = y^N$. We recall and define useful notations for the ease of presentation.

$$\begin{aligned} \text{Direct parts.}\quad & \widetilde{h}_i^D(x_\upsilon^i, y_+) = \nabla_x F_i(x_\upsilon^i, y_+; \xi_\upsilon^i),\ \widetilde{h}_i^D(x) = \nabla_x F_i(x, y_+; \xi_i),\ \bar{\nabla} f_i^D(x,y) = \nabla_x f_i(x,y) \\ \text{Indirect parts.}\quad & \widetilde{h}_i^I(x) = \lambda(N+1)\nabla_x\nabla_y G_i(x, y^N; \chi_i) \prod_{t=N}^{Q+1}(I - \lambda\nabla_y^2 G(x, y^t; u_t))\nabla_y F(x, y^Q; \xi_Q) \\ & \bar{\nabla} f_i^I(x,y) = \nabla_x\nabla_y g_i(x,y)(\nabla_y^2 g(x,y))^{-1}\nabla_y f(x,y), \end{aligned} \tag{7}$$

where ξ_υ^i and ξ_i are different data samples and two crucial components are defined by

$$\nabla_y F(x, y^Q; \xi_Q) = \frac{1}{|S|}\sum_{i\in S}\nabla_y F_i(x, y^Q; \xi_{i,Q}),\ \ \nabla_y^2 G(x, y^t; u_t) = \frac{1}{|S|}\sum_{i\in S}\nabla_x\nabla_y G_i(x, y^t; u_{i,t}).$$

Based on the notations in Equation (7), we also recall the important forms of our stochastic hypergradient estimate $\widetilde{h}(x)$ constructed by the proposed AggITD method as well as its expectation form $\bar{h}(x) = \mathbb{E}[\widetilde{h}(x)|x, y_+]$, and an auxiliary hypergradient notation $\bar{\nabla} f(x, y_+)$, respectively.

$$\begin{aligned} \widetilde{h}(x) =& \frac{1}{|S|}\sum_{i\in S}\widetilde{h}_i(x) = \frac{1}{|S|}\sum_{i\in S}[\widetilde{h}_i^D(x) - \widetilde{h}_i^I(x)] = \widetilde{h}^D(x) - \widetilde{h}^I(x) \\ =& \nabla_x F(x, y_+; \xi) - \lambda(N+1)\nabla_x\nabla_y G(x, y^N; \chi)\prod_{t=N}^{Q+1}(I - \lambda\nabla_y^2 G(x, y^t; u_t))\nabla_y F(x, y^Q; \xi_Q) \\ \bar{h}(x) =& \frac{1}{|S|}\sum_{i\in S}\bar{h}_i(x) = \frac{1}{|S|}\sum_{i\in S}[\bar{h}_i^D(x) - \bar{h}_i^I(x)] = \bar{h}^D(x) - \bar{h}^I(x) \\ =& \nabla_x f(x, y_+) - \lambda\nabla_x\nabla_y g(x, y^N)\sum_{Q=0}^{N}\prod_{t=N}^{Q+1}(I - \lambda\nabla_y^2 g(x, y^t))\nabla_y f(x, y^Q) \end{aligned}$$

$$\begin{aligned}\bar{\nabla} f(x,y) =& \frac{1}{|S|}\sum_{i\in S}\bar{\nabla} f_i(x,y) = \frac{1}{|S|}\sum_{i\in S}[\bar{\nabla} f_i^D(x,y) - \bar{\nabla} f_i^I(x,y)] = \bar{\nabla} f^D(x,y) - \bar{\nabla} f^I(x,y)\\ =& \nabla_x f(x,y) - \nabla_x \nabla_y g(x,y)(\nabla_y^2 g(x,y))^{-1}\nabla_y f(x,y),\end{aligned} \tag{8}$$

Based on Equation (8), it is noted that the hypergradient $\nabla f(x) = \bar{\nabla} f(x, y^*_{(x)})$. By the analysis in Ghadimi & Wang 2018 and Chen et al. 2021a, the following lemma characterizes the continuity and smoothness properties of the inner and outer functions (f_i, g_i) for all $i \in S$.

Lemma 4. *Suppose Assumption 1-Assumption 3 hold, for all x_1 and x_2:*

$$\begin{aligned}\|\nabla f(x_1) - \nabla f(x_2)\| \leq& L'_f\|x_1 - x_2\|,\\ \|y^*_{(x_1)} - y^*_{(x_2)}\| \leq& L_y\|x_1 - x_2\|,\\ \|\nabla y^*_{(x_1)} - \nabla y^*_{(x_2)}\| \leq& L_{yx}\|x_1 - x_2\|.\end{aligned} \tag{9}$$

Besides, for all $i \in S, x_1, x_2$ and y, we have

$$\begin{aligned}\|\bar{\nabla} f_i(x_1, y) - \bar{\nabla} f_i(x_1, y^*_{(x_1)})\| \leq& M_f\|y^*_{(x_1)} - y\|\\ \|\bar{\nabla} f_i(x_2, y) - \bar{\nabla} f_i(x_1, y)\| \leq& M_f\|x_2 - x_1\|,\end{aligned}$$

where all constants are given by

$$\begin{aligned}L_y :=& \frac{L_g}{\mu} = \mathcal{O}(\kappa_g)\\ L_{yx} :=& \frac{\rho + \rho L_y}{\mu} + \frac{L_g(\rho + \rho L_y)}{\mu^2} = \mathcal{O}(\kappa_g^3)\\ M_f :=& L_f + \frac{L_g L_f}{\mu} + \frac{M}{\mu}(\rho + \frac{L_g \rho}{\mu}) = \mathcal{O}(\kappa_g^2)\\ L'_f :=& L_f + \frac{L_g(L_f + M_f)}{\mu} + \frac{M}{\mu}(\rho + \frac{L_g\rho}{\mu}) = \mathcal{O}(\kappa_g^3)\end{aligned} \tag{10}$$

where all other Lipschitzness constants are provided in Assumptions 1-4.

C. Proof of Proposition 1 and Proposition 2

For the estimator, recall from Equation (8) that the indirect part is given by

$$\widetilde{h}^I(x) = \lambda(N+1)\nabla_x\nabla_y G(x, y^N; \chi)\prod_{t=N}^{Q+1}(I - \lambda\nabla_y^2 G(x, y^t; u_t))\nabla_y F(x, y^Q; \xi_Q),$$

where Q is drawn form $\{0, ..., N\}$ uniformly at random.

C.1. Proof of Proposition 1

Proof. First, based on the definition of $\widetilde{h}^I(x)$ in Equation (8) and conditioning on x, y^N, we have

$$\begin{aligned}\mathbb{E}[\widetilde{h}^I(x)] =& \mathbb{E}\Big[\lambda(N+1)\nabla_x\nabla_y G(x, y^N; \chi)\prod_{t=N}^{Q+1}(I - \lambda\nabla_y^2 G(x, y^t; u_t))\nabla_y F(x, y^Q; \xi_Q)\Big]\\ \overset{(i)}{=}& \lambda\nabla_x\nabla_y g(x, y^N)\sum_{Q=0}^{N}\prod_{t=N}^{Q+1}(I - \lambda\nabla_y^2 g(x, y^t))\nabla_y f(x, y^Q),\end{aligned} \tag{11}$$

where (i) follows from the fact that Q is drawn from $\{0, ..., N\}$ uniformly at random and from the independence among χ, u_t, ξ_Q for $t = 1, ..., N$. Then the estimation bias of $\widetilde{h}^I(x)$ is bounded by

$$\|\mathbb{E}[\widetilde{h}^I(x)] - \nabla_x\nabla_y g(x, y^N)(\nabla_y^2 g(x, y^N))^{-1}\nabla_y f(x, y^N)\|^2$$

$$
\begin{aligned}
\leq & \Big[\|\nabla_x \nabla_y g(x, y^N)\|^2 \Big\|\lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) \nabla_y f(x, y^Q) - (\nabla_y^2 g(x, y^N))^{-1} \nabla_y f(x, y^N)\Big\|^2\Big] \\
\overset{(i)}{\leq} & L_g^2 \Big[\Big\|\lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) \nabla_y f(x, y^Q) - (\nabla_y^2 g(x, y^N))^{-1} \nabla_y f(x, y^N)\Big\|^2\Big] \\
= & L_g^2 \Big[\Big\|\lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) \nabla_y f(x, y^Q) - \lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) \nabla_y f(x, y^N) \\
& + \lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) \nabla_y f(x, y^N) - (\nabla_y^2 g(x, y^N))^{-1} \nabla_y f(x, y^N)\Big\|^2\Big] \\
\overset{(ii)}{\leq} & 2\lambda^2 L_g^2 \Big[\Big\| \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) [\nabla_y f(x, y^Q) - \nabla_y f(x, y^N)]\Big\|^2\Big] \\
& + 2L_g^2 \Big[\Big\|\lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) - (\nabla_y g(x, y^N))^{-1}\Big\|^2 \|\nabla_y f(x, y^N)\|^2\Big] \\
\overset{(iii)}{\leq} & 2\lambda^2 L_f^2 L_g^2 \underbrace{(N+1) \sum_{Q=0}^{N} (1 - \lambda\mu)^{2N-2Q} [\|y^Q - y^N\|^2]}_{\textcircled{1}} \\
& + 2L_g^2 M^2 \Big\|\lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) - (\nabla_y g(x, y^N))^{-1}\Big\|^2,
\end{aligned}
\tag{12}
$$

where (i) uses Assumption 2, (ii) follows from Young's inequality, and (iii) follows from Lemma 4 and Assumption 2. Then, unconditioning on x_k, y_k^N yields

$$
\begin{aligned}
& \mathbb{E}\left[\|\mathbb{E}[\widetilde{h}^I(x)] - \nabla_x \nabla_y g(x, y^N)(\nabla_y^2 g(x, y^N))^{-1} \nabla_y f(x, y^N)\|^2 \,|\, x, y^N\right] \\
\leq & 2\lambda^2 L_f^2 L_g^2 \underbrace{(N+1) \sum_{Q=0}^{N} (1 - \lambda\mu)^{2N-2Q}\, \mathbb{E}[\|y^Q - y^N\|^2]}_{\textcircled{1}} \\
& + 2L_g^2 M^2\, \mathbb{E}\left[\Big\|\lambda \sum_{Q=0}^{N} \prod_{t=N}^{Q+1} (I - \lambda \nabla_y^2 g(x, y^t)) - (\nabla_y g(x, y^N))^{-1}\Big\|^2\right].
\end{aligned}
\tag{13}
$$

Based on Theorem 4 in Mitra et al. 2021, for all $t \in [0, ..., N-1]$, we obtain

$$
\mathbb{E}[\|y^{t+1} - y_{(x)}^*\|^2] \leq (1 - \frac{\beta\mu}{2})\, \mathbb{E}[\|y^t - y_{(x)}^*\|^2] + 25\beta^2\sigma_g^2 \tag{14}
$$

which, by telescoping over t from 0 to $Q-1$ for any $Q \in \{0, ..., N\}$, yields

$$
\mathbb{E}[\|y^Q - y_{(x)}^*\|^2] \leq (1 - \frac{\beta\mu}{2})^Q\, \mathbb{E}[\|y - y_{(x)}^*\|^2] + 25N\beta^2\sigma_g^2. \tag{15}
$$

Now we provide the upper bound of the first term on the RHS of Equation (12) as

$$
\begin{aligned}
\textcircled{1} \leq & (N+1) \sum_{Q=0}^{N} (1 - \lambda\mu)^{2N-2Q} \Big[2\,\mathbb{E}[\|y^Q - y_{(x)}^*\|^2] + 2\,\mathbb{E}[\|y^N - y_{(x)}^*\|^2]\Big] \\
\overset{(i)}{\leq} & 2(N+1) \sum_{Q=0}^{N} (1 - \lambda\mu)^{2N-2Q} \Big[(1 - \frac{\beta\mu}{2})^N\, \mathbb{E}[\|y - y_{(x)}^*\|^2] + (1 - \frac{\beta\mu}{2})^Q\, \mathbb{E}[\|y - y_{(x)}^*\|^2] + 50N\beta^2\sigma_g^2\Big]
\end{aligned}
$$

$$\begin{aligned}
&\overset{(ii)}{\le} 2(N+1)\Big(\frac{(1-\frac{\beta\mu}{2})^N}{\lambda\mu} + \frac{(1-\frac{\beta\mu}{2})^N}{1-\frac{(1-\lambda\mu)^2}{1-\frac{\beta\mu}{2}}}\Big)\mathbb{E}[\|y-y^*_{(x)}\|^2] + \frac{100N(N+1)\beta^2\sigma_g^2}{\lambda\mu} \\
&\le 2\underbrace{(N+1)\frac{3(1-\frac{\beta\mu}{2})^N}{\lambda\mu}}_{\alpha_3(N)}\mathbb{E}[\|y-y^*_{(x)}\|^2] + \frac{100N(N+1)\beta^2\sigma_g^2}{\lambda\mu},
\end{aligned} \tag{16}$$

where (i) follows from Equation (15), (ii) follows because $\frac{(1-\lambda\mu)^2}{1-\frac{\beta\mu}{2}} \le \frac{1-\lambda\mu}{1-\frac{\beta\mu}{2}} \le \frac{1-\lambda\mu}{1-\frac{\lambda\mu}{2}} \le 1$ as the selection that $\beta < \lambda \le \frac{1}{L_g}$. Then we provide the upper bound of the second term in Equation (12) as

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\lambda\sum_{Q=0}^{N}\prod_{t=N}^{Q+1}(I-\lambda\nabla_y^2 g(x,y^t)) - (\nabla_y^2 g(x,y^N))^{-1}\Big\|^2\Big] \\
&=\lambda^2\,\mathbb{E}\Big[\Big\|\sum_{Q=0}^{N}\prod_{t=N}^{Q+1}(I-\lambda\nabla_y^2 g(x,y^t)) - \sum_{Q=0}^{N}(I-\lambda\nabla_y^2 g(x,y^N))^{N-Q} - \sum_{Q=N+1}^{\infty}(I-\lambda\nabla_y^2 g(x,y^N))^{Q}\Big\|^2\Big] \\
&\overset{(i)}{\le} 2\lambda^2(N+1)\sum_{Q=0}^{N}\mathbb{E}\Big[\underbrace{\Big\|\prod_{t=N}^{Q+1}(I-\lambda\nabla_y^2 g(x,y^t)) - (I-\lambda\nabla_y^2 g(x,y^N))^{N-Q}\Big\|^2}_{M_{N-Q}^2}\Big] + \frac{2(1-\lambda\mu)^{2N+2}}{\mu^2}
\end{aligned} \tag{17}$$

where (i) follows from the Young's inequality. Now we provide the upper bound of the term M_{N-Q} as

$$\begin{aligned}
M_{N-Q} =& \Big\|\prod_{t=N}^{Q+1}(I-\lambda\nabla_y^2 g(x,y^t)) - (I-\lambda\nabla_y^2 g(x,y^N))^{N-Q}\Big\| \\
=& \Big\|(I-\lambda\nabla_y^2 g(x,y^N))\Big[\prod_{t=N}^{Q+2}(I-\lambda\nabla_y^2 g(x,y^t)) - (I-\lambda\nabla_y^2 g(x,y^N))^{N-Q-1}\Big] \\
&\quad + (\lambda\nabla_y^2 g(x,y^N) - \lambda\nabla_y^2 g(x,y^{Q+1}))\prod_{t=N}^{Q+2}(I-\lambda\nabla_y^2 g(x,y^t))\Big\| \\
\overset{(i)}{\le}& (1-\lambda\mu)\underbrace{\Big\|\prod_{t=N}^{Q+2}(I-\lambda\nabla_y^2 g(x,y^t)) - (I-\lambda\nabla_y^2 g(x,y^N))^{N-Q-1}\Big\|}_{M_{N-Q-1}} + \lambda\rho(1-\lambda\mu)^{N-Q-1}\|y^N - y^{Q+1}\| \\
\overset{(ii)}{\le}& (1-\lambda\mu)^{N-Q}M_0 + \lambda\rho(1-\lambda\mu)^{N-Q-1}\sum_{\tau=Q+1}^{N}\|y^\tau - y^N\| \\
\overset{(iii)}{\le}& \lambda\rho(1-\lambda\mu)^{N-Q-1}\sum_{\tau=Q+1}^{N}\|y^\tau - y^N\|,
\end{aligned} \tag{18}$$

where (i) follows from the Assumption 1 and Assumption 3, (ii) can be obtained after telescoping over t from 0 to $N-1$ and (iii) follows from that $M_0 = 0$. Then substitute Equation (18) into Equation (17), we obtain,

$$\begin{aligned}
(N+1)\sum_{Q=0}^{N}\mathbb{E}[M_{N-Q}^2] \le& \lambda^2\rho^2(N+1)\sum_{Q=0}^{N}\Big[(1-\lambda\mu)^{2N-2Q-2}\Big](N-Q)\sum_{\tau=Q+1}^{N}\Big[2\Big(1-\frac{\beta\mu}{2}\Big)^{\tau}\mathbb{E}[\|y-y^*_{(x)}\|^2] \\
&+ 2\Big(1-\frac{\beta\mu}{2}\Big)^{N}\mathbb{E}[\|y-y^*_{(x)}\|^2] + 50N\beta^2\sigma_g^2 + 50\tau\beta^2\sigma_g^2\Big] \\
\le& \lambda^2\rho^2(N+1)\sum_{Q=0}^{N}(1-\lambda\mu)^{2N-2Q-2}(N-Q)\Big[\frac{4(1-\frac{\beta\mu}{2})^{Q+1}}{\beta\mu}\mathbb{E}[\|y-y^*_{(x)}\|^2]
\end{aligned}$$

$$
\begin{aligned}
&+2(N-Q)\Big(1-\frac{\beta\mu}{2}\Big)^N\,\mathbb{E}[\|y-y^*_{(x)}\|^2+100N(N-Q)\beta^2\sigma_g^2\Big]\\
=&2(N+1)\lambda^2\rho^2\sum_{Q=0}^{N}(1-\lambda\mu)^{2N-2Q-2}(N-Q)^2(1-\frac{\beta\mu}{2})^N\,\mathbb{E}[\|y-y^*_{(x)}\|^2]\\
&+4(N+1)\lambda^2\rho^2\sum_{Q=0}^{N}(1-\lambda\mu)^{2N-2Q-2}(N-Q)\frac{(1-\frac{\beta\mu}{2})^{Q+1}}{\beta\mu}\,\mathbb{E}[\|y-y^*_{(x)}\|^2]\\
&+100\beta^2\sigma_g^2\lambda^2\rho^2N(N+1)\sum_{Q=0}^{N}(1-\lambda\mu)^{2N-2Q-2}(N-Q)^2\\
<&\underbrace{4(N+1)(1-\frac{\beta\mu}{2})^N\Big(\frac{\rho^2}{\lambda\mu^3}+\frac{4\rho^2}{\beta\mu^3}\Big)}_{\alpha_1(N)}\mathbb{E}[\|y-y^*_{(x)}\|^2]\\
&+100\beta^2\rho^2\sigma_g^2\underbrace{\Big[\frac{N(N+1)(1+(1-\lambda\mu)^2)}{\lambda\mu^3}\Big]}_{\alpha_2(N)},
\end{aligned}
\tag{19}
$$

where the last inequality follows because $\sum_{t=0}^{N}(1-\lambda\mu)^{2N-2t-2}(N-t)^2<\frac{1+(1-\lambda\mu)^2}{\lambda^3\mu^3}$ and $\sum_{t=0}^{N}(1-\lambda\mu)^{2N-2t-2}(N-t)(1-\frac{\beta\mu}{2})^{t+1}<\frac{1}{\left(1-\frac{(1-\lambda\mu)^2}{1-\frac{\beta\mu}{2}}\right)^2}\le\frac{(2-\lambda\mu)^2}{\lambda^2\mu^2}$. Substituting Equation (19) into Equation (17), and applying Equation (19) and Equation (16) to Equation (12), we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbb{E}[\widetilde{h}^I(x)]-\nabla_x\nabla_y g(x,y^N)(\nabla_y^2 g(x,y^N)^{-1})\nabla_y f(x,y^N)\|^2\,|\,x,y^N\right]\\
\le&4\lambda^2L_f^2L_g^2\alpha_3(N)\,\mathbb{E}[\|y-y^*_{(x)}\|^2]+\frac{200\lambda\beta^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu}\\
&+\frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}+4\lambda^2L_g^2M^2\alpha_1(N)\,\mathbb{E}[\|y-y^*_{(x)}\|^2]+400\lambda^2\beta^2L_g^2M^2\sigma_g^2\rho^2\alpha_2(N)\\
=&[4\lambda^2L_g^2M^2\alpha_1(N)+4\lambda^2L_f^2L_g^2\alpha_3(N)]\,\mathbb{E}[\|y-y^*_{(x)}\|^2]+\frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}\\
&+400\lambda^2\beta^2L_g^2M^2\sigma_g^2\rho^2\alpha_2(N)+\frac{200\lambda\beta^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu},
\end{aligned}
$$

which completes the proof. □

C.2. Proof of Proposition 2

Based on the definition of $\widetilde{h}^I(x)$ and $\bar{h}_i^I(x)$, using the fact that $\mathrm{Var}(X)\le\mathbb{E}[X^2]$, and conditioning on x,y^N, we have

$$
\begin{aligned}
&\mathbb{E}\,\|\widetilde{h}_i^I(x)-\bar{h}_i^I(x)\|^2\le\mathbb{E}\,\|\widetilde{h}_i^I(x)\|^2\\
\le&\,\mathbb{E}\left\|\lambda(N+1)\nabla_x\nabla_y G_i(x,y^N;\chi)\prod_{t=N}^{Q+1}(I-\lambda\nabla_y^2G(x,y^t;u_t))\nabla_y F(x,y^Q;\xi_Q)\right\|^2\\
\overset{(i)}{\le}&\lambda^2(N+1)^2L_g^2M^2\mathbb{E}\left\|\prod_{t=N}^{Q+1}(I-\lambda\nabla_y^2G(x,y^t;u_t))\right\|^2\\
\overset{(ii)}{\le}&\lambda^2(N+1)^2L_g^2M^2\mathbb{E}_Q(1-\lambda\mu)^{2(N-Q)}\\
=&\lambda^2(N+1)L_g^2M^2\sum_{Q=0}^{N}(1-\lambda\mu)^{2Q}=\lambda^2(N+1)L_g^2M^2\frac{1-(1-\lambda\mu)^{2N}}{1-(1-\lambda\mu)^2}\\
\overset{(iii)}{\le}&\frac{\lambda(N+1)L_g^2M^2}{\mu},
\end{aligned}
\tag{20}
$$

where (i) follows from Assumption 2, (ii) follows from Assumption 1 and (iii) follows from $\lambda \leq \frac{1}{L_g}$. Then, the first part is proved. For the second part, conditioning on x, y_+, we have

$$
\begin{aligned}
&\mathbb{E}\,\|\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\|^2 \\
&\leq 4\,\mathbb{E}\,\|\widetilde{h}_i^D(x_\upsilon^i, y_+)\|^2 + 4\,\mathbb{E}\,\|\widetilde{h}_i^D(x_0^i, y_+)\|^2 + 4\,\mathbb{E}\,\Big\|\frac{1}{|S|}\sum_{i\in S}\widetilde{h}_i^D(x)\Big\|^2 + 4\,\mathbb{E}\,\Big\|\frac{1}{|S|}\sum_{i\in S}\widetilde{h}_i^I(x)\Big\|^2 \\
&\overset{(i)}{\leq} 8M^2 + 4\,\mathbb{E}\,\|\widetilde{h}_i^D(x)\|^2 + 4\,\mathbb{E}\,\|\widetilde{h}_i^I(x)\|^2 \\
&\overset{(ii)}{\leq} 12M^2 + \frac{4\lambda(N+1)L_g^2M^2}{\mu},
\end{aligned}
$$

where (i) follows from Assumption 2 and (ii) follows from Equation (20). Then, the proof is complete.

D. Proof of Theorem 1 and Corollary 1

We now provide some auxiliary lemmas to characterize the Theorem 1 and Corollary 1.

Lemma 5 (Restatement of Lemma 1)**.** *Suppose Assumptions 1-4 are satisfied. Let $y^* = \arg\min_y g(x,y)$. Further, we set $\lambda \leq \min\{10, \frac{1}{L_g}\}$, $\alpha^i = \frac{\alpha}{\tau_i}$ with $\tau_i \geq 1$ for some positive α and $\beta^i = \frac{\beta}{\tau_i}$, where $\beta \leq \min\{1, \lambda, \frac{1}{6L_g}\}$ $\forall i \in S$. Then, we have the following inequality*

$$
\begin{aligned}
\mathbb{E}[f(x_+)] - \mathbb{E}[f(x)] \leq &-\frac{\alpha}{2}\,\mathbb{E}[\|\nabla f(x)\|^2] + 4\alpha^2\sigma_h^2L_f' + 4\alpha^2\sigma_f^2L_f' + 2\alpha^2M^2L_f' \\
&- \frac{\alpha}{2}(1-4\alpha L_f')\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big\|^2\Big] \\
&+ \frac{3\alpha}{2}\Big[\big(4\lambda^2L_g^2M^2\alpha_1(N) + 4\lambda^2L_f^2L_g^2\alpha_3(N)\big)\,\mathbb{E}[\|y-y^*\|^2] + \frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2} \\
&+ 400\lambda^2\beta^2L_g^2M^2\sigma_g^2\rho^2\alpha_2(N) + \frac{200\lambda\beta^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu} \\
&+ \frac{M_f^2}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\mathbb{E}[\|x_\upsilon^i - x\|^2] + M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2]\Big]
\end{aligned}
\tag{21}
$$

where $\bar{h}^I(x) = \mathbb{E}[\widetilde{h}^I(x)|x, y_+]$, $\bar{h}_i^D(x_\upsilon^i, y_+) = \mathbb{E}[\widetilde{h}_i^D(x_\upsilon^i, y_+)|x_\upsilon^i]$ and $\alpha_1(N), \alpha_2(N), \alpha_3(N)$ are defined in Proposition 1.

Proof. From Algorithm 4, we have, $\forall i \in S$

$$
x_+ = x - \frac{1}{m}\sum_{i=1}^m\alpha^i\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\big),
$$

where $x_0^i = x$ and the data samples for $\widetilde{h}_i^D(x_0^i, y_+)$ and $\widetilde{h}^D(x)$ are different. Using the descent lemma yields

$$
\begin{aligned}
\mathbb{E}[f(x_+)] - \mathbb{E}[f(x)] \leq &\,\mathbb{E}[\langle x_+ - x, \nabla f(x)\rangle] + \frac{L_f'}{2}\,\mathbb{E}[\|x_+ - x\|^2] \\
= &-\mathbb{E}\Big[\Big\langle\frac{1}{m}\sum_{i=1}^m\alpha^i\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\big), \nabla f(x)\Big\rangle\Big] \\
&+ \frac{L_f'}{2}\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^m\alpha^i\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\big)\Big\|^2\Big].
\end{aligned}
\tag{22}
$$

We next bound each term of the right hand side (RHS) of Equation (22). In specific, for the first term, we have

$$
-\mathbb{E}\Big[\Big\langle\frac{1}{m}\sum_{i=1}^m\alpha^i\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\big), \nabla f(x)\Big\rangle\Big]
$$

$$
\begin{aligned}
&= - \mathbb{E}\Big[\mathbb{E}\Big[\Big\langle \frac{1}{m}\sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\big), \nabla f(x)\Big\rangle\Big| x, y_+\Big]\Big] \\
&\overset{(i)}{=} - \mathbb{E}\Big[\mathbb{E}\Big[\Big\langle \frac{1}{m}\sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big), \nabla f(x)\Big\rangle\Big| x_\upsilon^i\Big]\Big] \\
&\overset{(ii)}{=} - \alpha\,\mathbb{E}\Big[\Big\langle \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big), \nabla f(x)\Big\rangle\Big] \\
&= - \frac{\alpha}{2}\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big\|^2\Big] - \frac{\alpha}{2}\,\mathbb{E}[\|\nabla f(x)\|^2] \\
&\quad + \frac{\alpha}{2}\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big) - \nabla f(x)\Big\|^2\Big], \qquad (23)
\end{aligned}
$$

where (i) follows because $\bar{h}^I(x) = \mathbb{E}[\widetilde{h}^I(x)|x, y_+]$ and $\mathbb{E}\Big[\frac{1}{m}\sum_{i=1}^{m}\alpha^i\sum_{\upsilon=0}^{\tau_i-1}\big(-\widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x)\big)\Big| x, y_+\Big] = 0$, (ii) follows because $\bar{h}_i^D(x_\upsilon^i, y_+) = \mathbb{E}[\widetilde{h}_i^D(x_\upsilon^i, y_+)|x_\upsilon^i]$. The next step is to upper bound the last term of RHS of Equation (23). Based on the notations in Equation (7) and Equation (8), we have

$$
\begin{aligned}
&\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big) - \nabla f(x)\Big\|^2 \\
&\quad = \Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big) - \bar{\nabla} f(x, y_+) + \bar{\nabla} f(x, y_+) - \nabla f(x)\Big\|^2 \\
&\quad = \Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{\nabla} f_i^D(x, y_+)\big) - \bar{h}^I(x) + \bar{\nabla} f^I(x, y_+) + \bar{\nabla} f(x, y_+) - \nabla f(x)\Big\|^2 \\
&\quad \overset{(i)}{\leq} 3\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{\nabla} f_i^D(x, y_+)\big)\|^2 + 3\Big\|\bar{h}^I(x) - \bar{\nabla} f^I(x, y_+)\Big\|^2 + 3\|\bar{\nabla} f(x, y_+) - \nabla f(x)\|^2 \\
&\quad \overset{(ii)}{\leq} \frac{3M_f^2}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\|x_\upsilon^i - x\|^2 + 3M_f^2\|y_+ - y^*\|^2 + 3\|\bar{h}^I(x) - \bar{\nabla} f^I(x, y_+)\|^2 \qquad (24)
\end{aligned}
$$

where (i) follows from the Young's inequality and (ii) follows from Lemma 4 and Assumption 2. Then applying Proposition 1 to Equation (24), we can obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big) - \nabla f(x)\Big\|^2\Big] \\
&\qquad \leq \frac{12L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2} + [12\lambda^2L_g^2M^2\alpha_1(N) + 12\lambda^2L_f^2L_g^3\alpha_3(N)]\,\mathbb{E}[\|y - y^*\|^2] \\
&\qquad\quad + 1200\lambda^2\beta^2L_g^2M^2\sigma_g^2\rho^2\alpha_2(N) + \frac{600\lambda\beta^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu} \\
&\qquad\quad + \frac{3M_f^2}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\mathbb{E}[\|x_\upsilon^i - x\|^2] + 3M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2]. \qquad (25)
\end{aligned}
$$

Then for the second term of Equation (22), we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\big)\Big\|^2\Big] \\
&\quad \overset{(i)}{\leq} 2\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}^I(x)\big)\Big\|^2\Big] + 2\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i} \sum_{\upsilon=0}^{\tau_i-1}\big(-\widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x)\big)\Big\|^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
=&2\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i,y_+)-\bar{h}_i^D(x_\upsilon^i,y_+)+\bar{h}^I(x)-\widetilde{h}^I(x)+\bar{h}_i^D(x_\upsilon^i,y_+)-\bar{h}^I(x)\big)\Big\|^2\Big] \\
&+2\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_0^i,y_+)\big)\Big\|^2\Big]+2\alpha^2\,\mathbb{E}[\|\widetilde{h}^D(x)\|^2] \\
\overset{(ii)}{\leq}&4\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\bar{h}_i^D(x_\upsilon^i,y_+)-\bar{h}^I(x)\Big\|^2\Big] \\
&+4\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i,y_+)-\bar{h}_i^D(x_\upsilon^i,y_+)+\bar{h}^I(x)-\widetilde{h}^I(x)\big)\Big\|^2\Big]+4\alpha^2M^2 \\
\overset{(iii)}{\leq}&4\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\bar{h}_i^D(x_\upsilon^i,y_+)-\bar{h}^I(x)\Big\|^2\Big]+8\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i,y_+)-\bar{h}_i^D(x_\upsilon^i,y_+)\big)\Big\|^2 \\
&+8\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}^I(x)-\widetilde{h}^I(x)\big)\Big\|^2\Big]+4\alpha^2M^2 \\
\overset{(iv)}{\leq}&4\alpha^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\bar{h}_i^D(x_\upsilon^i,y_+)-\bar{h}^I(x)\Big\|^2\Big]+8\alpha^2\sigma_f^2+8\alpha^2\sigma_h^2+4\alpha^2M^2 \qquad (26)
\end{aligned}
$$

where (i) and (iii) follow from the Young's inequality, (ii) follows from Young's inequality and Assumption 2 and (iv) follows from Assumption 4 and lemma 4. Plugging Equation (25) and Equation (26) into Equation (22) completes the proof. □

Lemma 6 (Restatement of Lemma 2)**.** *Suppose Assumptions 1-4 are satisfied. Let* $y^* = \arg\min_y g(x,y)$ *and* $y^*_{(x_+)} = \arg\min_y g(x_+,y)$. *Further, set* $\alpha^i = \frac{\alpha}{\tau_i}$ *with* $\tau_i \geq 1$ *with some positive* α, $\forall i \in S$. *Then, we have*

$$
\begin{aligned}
\mathbb{E}[\|y_+-y^*_{(x_+)}\|^2] \leq& b_1(\alpha)\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i,y_+)-\bar{h}^I(x)\big)\Big\|^2\Big]+b_2(\alpha)\,\mathbb{E}[\|y_+-y^*\|^2] \\
&+b_3(\alpha)(2\sigma_h^2+2\sigma_f^2+M^2)
\end{aligned}
$$

where the constants are given by

$$
b_1(\alpha) :=4L_y^2\alpha^2+\frac{L_y^2\alpha^2}{4\gamma}+\frac{2L_{yx}\alpha^2}{\eta},\ b_2(\alpha):=1+4\gamma+\frac{\eta L_{yx}D_h^2\alpha^2}{2},\ b_3(\alpha):=4\alpha^2L_y^2+\frac{2L_{yx}\alpha^2}{\eta}
$$

with a flexible parameter $\gamma > 0$ *decided later.*

Proof. First note that

$$
\mathbb{E}[\|y_+-y^*_{(x_+)}\|^2] =\mathbb{E}[\|y_+-y^*\|^2]+\mathbb{E}[\|y^*_{(x_+)}-y^*\|^2]+2\,\mathbb{E}[\langle y_+-y^*,y^*-y^*_{(x_+)}\rangle]. \qquad (27)
$$

In Equation (27), we bound the second term using Lemma 4 and Equation (26) as

$$
\begin{aligned}
\mathbb{E}[\|y^*_{(x_+)}-y^*\|^2] \leq& L_y^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\alpha^i\sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i,y_+)-\widetilde{h}_i^D(x_0^i,y_+)+\widetilde{h}^D(x)-\widetilde{h}^I(x)\big)\Big\|^2\Big] \\
\leq& 4\alpha^2L_y^2\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i,y_+)-\bar{h}^I(x)\big)\Big\|^2\Big]+8\alpha^2L_y^2\sigma_h^2+8\alpha^2L_y^2\sigma_f^2+4\alpha^2L_y^2M^2,
\end{aligned}
$$

and for the third term, we have

$$
\begin{aligned}
\mathbb{E}[\langle y_+-y^*,y^*-y^*_{(x_+)}\rangle] =& -\mathbb{E}[\langle y_+-y^*,\nabla y^*(x_+-x)\rangle] \\
&-\mathbb{E}[\langle y_+-y^*,y^*_{(x_+)}-y^*-\nabla y^*(x_+-x)\rangle]. \qquad (28)
\end{aligned}
$$

For the first term on the RHS of the above Equation (28), we have

$$
\begin{aligned}
-\mathbb{E}[\langle y_+{-}y^*, \nabla y^*(x_+ - x)\rangle]
&= -\mathbb{E}\Big[\Big\langle y_+ - y^*, \mathbb{E}\Big[\frac{1}{m}\nabla y^* \sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\big)\Big| x, y_+\Big]\Big\rangle\Big] \\
&= -\mathbb{E}\Big[\Big\langle y_+ - y^*, \mathbb{E}\Big[\frac{1}{m}\nabla y^* \sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\widetilde{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big| x_\upsilon^i\Big]\Big\rangle\Big] \\
&= -\mathbb{E}\Big[\Big\langle y_+ - y^*, \frac{1}{m}\nabla y^* \sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big\rangle\Big] \\
&\overset{(i)}{\leq} \mathbb{E}\Big[\|y_+ - y^*\|\Big\|\frac{1}{m}\nabla y^* \sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big\|\Big] \\
&\overset{(ii)}{\leq} L_y\,\mathbb{E}\Big[\|y_+ - y^*\|\Big\|\frac{1}{m}\sum_{i=1}^{m}\alpha^i \sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big\|\Big] \\
&\overset{(iii)}{\leq} 2\gamma\,\mathbb{E}[\|y_+ - y^*\|^2] + \frac{L_y^2\alpha^2}{8\gamma}\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big\|^2\Big] \qquad (29)
\end{aligned}
$$

where (i) follows from the Cauchy–Schwarz inequality, (ii) follows from Lemma 4, and (iii) follows from Young's inequality that $ab \leq 2\gamma a^2 + \frac{b^2}{2\gamma}$. For the second term of RHS of Equation (28), we have

$$
\begin{aligned}
&-\mathbb{E}[\langle y_+ - y^*, y^*_{(x_+)} - y^* - \nabla y^*(x_+ - x)\rangle] \\
&\quad\leq \mathbb{E}[\|y_+ - y^*\|\|y^*_{(x_+)} - y^* - \nabla y^*(x_+ - x)\|] \\
&\quad\overset{(i)}{\leq} \frac{L_{yx}}{2}\,\mathbb{E}[\|y_+ - y^*\|\|x_+ - x\|^2] \\
&\quad\overset{(ii)}{\leq} \frac{\eta L_{yx}}{4}\,\mathbb{E}[\|y_+ - y^*\|^2\|x_+ - x\|^2] + \frac{L_{yx}}{4\eta}\,\mathbb{E}[\|x_+ - x\|^2] \\
&\quad\leq \frac{\eta L_{yx}}{4}\frac{1}{m}\sum_{i=1}^{m}\frac{\alpha^2}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\mathbb{E}\left[\|y_+ - y^*\|^2\,\mathbb{E}[\|\widetilde{h}_i^D(x_\upsilon^i, y_+) - \widetilde{h}_i^D(x_0^i, y_+) + \widetilde{h}^D(x) - \widetilde{h}^I(x)\|^2 | x, y_+]\right] \\
&\qquad + \frac{L_{yx}}{4\eta}\,\mathbb{E}[\|x_+ - x\|^2] \\
&\quad\overset{(iii)}{\leq} \frac{\eta L_{yx} D_h^2\alpha^2}{4}\,\mathbb{E}[\|y_+ - y^*\|^2] + \frac{L_{yx}\alpha^2}{\eta}\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x_\upsilon^i, y_+) - \bar{h}^I(x)\big)\Big\|^2\Big] \\
&\qquad + \frac{L_{yx}\alpha^2}{\eta}(2\sigma_h^2 + 2\sigma_f^2 + M^2) \qquad (30)
\end{aligned}
$$

where (i) follows from the decent lemma by the smoothness of $y^*(\cdot)$, (ii) follows from the Young's inequality, and (iii) follows from Proposition 1 and Equation (26). Substituting Equation (29) and Equation (30) into Equation (28), and using Equation (27), we complete the proof. □

Lemma 7 (Restatement of Lemma 3)**.** *Suppose Assumptions 1-4 are satisfied. Set* $\lambda \leq \min\{10, \frac{1}{L_g}\}$*,* $\alpha^i = \frac{\alpha}{\tau_i}$ *and* $\beta^i = \frac{\beta}{\tau_i}, \tau_i \geq 1$ *where* $\alpha \leq \frac{1}{324M_f^2+6M_f} \leq \frac{1}{6M_f}$*,* $\beta \leq \min\{1, \lambda, \frac{1}{6L_g}\}\ \forall i \in S$*. Recall the definitions of* $y^* = \arg\min_y g(x, y)$*,* $\bar{h}(x) = \mathbb{E}[\widetilde{h}(x)|x, y_+]$*. Then, we have*

$$
\begin{aligned}
\mathbb{E}[\|x_\upsilon^i - x\|^2] \leq{}& 18\tau_i^2(\alpha^i)^2\Big[3M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2] + 3\,\mathbb{E}[\|\nabla f(x)\|^2] + \frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2} \\
&+ [4\lambda^2L_g^2M^2\alpha_1(N) + 4\lambda^2L_f^2L_g^2\alpha_3(N)]\,\mathbb{E}[\|y - y^*\|^2]
\end{aligned}
$$

$$+ 400\lambda^2\beta^2 L_g^2 M^2\rho^2\sigma_g^2\alpha_2(N) + \frac{200\lambda\beta^2 L_f^2 L_g^2 N(N+1)\sigma_g^2}{\mu} + 3\sigma_h^2 + 6\sigma_f^2\Big] \tag{31}$$

where $\alpha_1(N), \alpha_2(N), \alpha_3(N)$ *are defined in Proposition 1.*

Proof. The result holds for $\tau_i = 1$ according to line 2 in Algorithm 4 where $x_0^i = x$, and hence we consider the case when $\tau_i > 1$. Based on the notations in Equation (8), we define

$$\begin{aligned}
v_\upsilon^i :=& \bar{h}(x) - \bar{\nabla} f(x, y_+),\\
\omega_\upsilon^i :=& \nabla_x F_i(x_\upsilon^i, y_+; \xi_{i,\upsilon}) - \nabla_x f_i(x_\upsilon^i, y_+) + \nabla_x f_i(x, y_+)\\
&- \nabla_x F_i(x, y_+; \xi_{i,\upsilon}) + \widetilde{h}(x) - \bar{h}(x),\\
z_\upsilon^i :=& \nabla_x f_i(x_\upsilon^i, y_+) - \nabla_x f_i(x, y_+) + \bar{\nabla} f(x, y_+) - \nabla f(x) + \nabla f(x).
\end{aligned} \tag{32}$$

Based on Algorithm 4, for each $i \in S$, and $\forall \upsilon \in 0, ..., \tau_i - 1$, we have,

$$x_{\upsilon+1}^i - x = x_\upsilon^i - x - \alpha^i(v_\upsilon^i + \omega_\upsilon^i + z_\upsilon^i). \tag{33}$$

Based on Lemma 4 and Proposition 1, we bound v_υ^i, ω_υ^i, and z_υ^i as

$$\begin{aligned}
\mathbb{E}[\|v_\upsilon^i\|^2] \leq& \frac{4L_g^2 M^2(1-\lambda\mu)^{2N+2}}{\mu^2} + [4\lambda^2 L_g^2 M^2\alpha_1(N) + 4\lambda^2 L_g^2 L_f^2\alpha_3(N)]\,\mathbb{E}[\|y - y^*\|^2]\\
&+ 400\lambda^2\beta^2 L_g^2 M^2\rho^2\sigma_g^2\alpha_2(N) + \frac{200\lambda\beta^2\sigma_g^2 L_f^2 L_g^2 N(N+1)}{\mu},\\
\mathbb{E}[\|\omega_\upsilon^i\|^2] \leq& 3\,\mathbb{E}[\|\nabla_x F_i(x_\upsilon^i, y_+; \xi_{i,\upsilon}) - \nabla_x f_i(x_\upsilon^i, y_+)\|^2\\
&+ \|\nabla_x f_i(x, y_+) - \nabla_x F_i(x, y_+; \xi_{i,\upsilon})\|^2 + \|\widetilde{h}(x) - \bar{h}(x)\|^2]\\
\leq& 6\sigma_f^2 + 3\sigma_h^2,\\
\mathbb{E}[\|z_\upsilon^i\|^2] \leq& 3\,\mathbb{E}[\|\nabla_x f_i(x_\upsilon^i, y_+) - \nabla_x f_i(x, y_+)\|^2\\
&+ \|\bar{\nabla} f(x, y_+) - \nabla f(x)\|^2 + \mathbb{E}\,\|\nabla f(x)\|^2]\\
\leq& 3(M_f^2\,\mathbb{E}[\|x_\upsilon^i - x\|^2] + M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2] + \mathbb{E}[\|\nabla f(x)\|^2]).
\end{aligned} \tag{34}$$

Now, we bound RHS of Equation (33) as

$$\begin{aligned}
\mathbb{E}[\|x_\upsilon^i &- x - \alpha^i(v_\upsilon^i + \omega_\upsilon^i + z_\upsilon^i)\|^2]\\
&\overset{(i)}{\leq} (1 + \frac{1}{2\tau_i - 1})\,\mathbb{E}[\|x_\upsilon^i - x\|^2] + 2\tau_i\,\mathbb{E}[\|\alpha^i(v_\upsilon^i + \omega_\upsilon^i + z_\upsilon^i)\|^2]\\
&\overset{(ii)}{\leq} (1 + \frac{1}{2\tau_i - 1})\,\mathbb{E}[\|x_\upsilon^i - x\|^2] + 6\tau_i(\alpha^i)^2\,\mathbb{E}[\|v_\upsilon^i\|^2 + \|\omega_\upsilon^i\|^2 + \|z_\upsilon^i\|^2]\\
&\overset{(iii)}{\leq} (1 + \frac{1}{2\tau_i - 1} + 18\tau_i(\alpha^i)^2 M_f^2)\,\mathbb{E}[\|x_\upsilon^i - x\|^2]\\
&\quad + 6\tau_i(\alpha^i)^2\Big[3M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2] + 3\,\mathbb{E}[\|\nabla f(x)\|^2] + \frac{4L_g^2 M^2(1-\lambda\mu)^{2N+2}}{\mu^2}\\
&\quad + [4\lambda^2 L_g^2 M^2\alpha_1(N) + 4\lambda^2 L_g^2 L_f^2\alpha_3(N)]\,\mathbb{E}[\|y - y^*\|^2] + 400\lambda^2\beta^2 L_g^2 M^2\rho^2\sigma_g^2\alpha_2(N)\\
&\quad + \frac{200\lambda\beta^2\sigma_g^2 L_f^2 L_g^2 N(N+1)}{\mu} + 6\sigma_f^2 + 3\sigma_h^2\Big]
\end{aligned} \tag{35}$$

where (i) follows from $\|x + y\|^2 \leq (1 + c)\|x\|^2 + (1 + \frac{1}{c})\|y\|^2$, (ii) follows from the Young's inequality, and (iii) uses Equation (34). Substituting Equation (35) into Equation (33) yields

$$\mathbb{E}[\|x_{\upsilon+1}^i - x\|^2] \leq (1 + \frac{1}{2\tau_i - 1} + 18\tau_i(\alpha^i)^2 M_f^2)\,\mathbb{E}[\|x_\upsilon^i - x\|^2]$$

$$
\begin{aligned}
&+ 6\tau_i(\alpha^i)^2\Big[3M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2] + 3\,\mathbb{E}[\|\nabla f(x)\|^2] + \frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}\\
&+ [4\lambda^2L_g^2M^2\alpha_1(N) + 4\lambda^2L_g^2L_f^2\alpha_3(N)]\,\mathbb{E}[\|y-y^*\|^2] + 400\lambda^2\beta^2L_g^2M^2\rho^2\sigma_g^2\alpha_2(N)\\
&+ \frac{200\lambda\beta^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu} + 3\sigma_h^2 + 6\sigma_f^2\Big]\\
\le&(1+\frac{1}{\tau_i-1})\,\mathbb{E}[\|x_v^i - x\|^2]\\
&+ 6\tau_i(\alpha^i)^2\Big[3M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2] + 3\,\mathbb{E}[\|\nabla f(x)\|^2] + \frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}\\
&+ [4\lambda^2L_g^2M^2\alpha_1(N) + 4\lambda^2L_g^2L_f^2\alpha_3(N)]\,\mathbb{E}[\|y-y^*\|^2] + 400\lambda^2\beta^2L_g^2M^2\rho^2\sigma_g^2\alpha_2(N)\\
&+ \frac{200\lambda\beta^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu} + 3\sigma_h^2 + 6\sigma_f^2\Big],
\end{aligned}
\tag{36}
$$

where the last inequality follows because $\alpha^i \le 1/(6M_f\tau_i)$. For all $\tau_i > 1$, we have

$$
\sum_{j=0}^{v-1}(1+\frac{1}{\tau_i-1})^j = \frac{(1+\frac{1}{\tau_i-1})^v - 1}{(1+\frac{1}{\tau_i-1}) - 1} \le \tau_i(1+\frac{1}{\tau_i})^v \le \tau_i(1+\frac{1}{\tau_i})^{\tau_i} \le \exp(1)\tau_i < 3\tau_i.
\tag{37}
$$

Finally, telescoping Equation (36) and using Equation (37), we have

$$
\begin{aligned}
\mathbb{E}[\|x_v^i - x\|^2] \le&18\tau_i^2(\alpha^i)^2\Big[3M_f^2\,\mathbb{E}[\|y_+ - y^*\|^2] + 3\,\mathbb{E}[\|\nabla f(x)\|^2] + \frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}\\
&+ [4\lambda^2L_g^2M^2\alpha_1(N) + 4\lambda^2L_g^2L_f^2\alpha_3(N)]\,\mathbb{E}[\|y-y^*\|^2] + 400\lambda^2\beta^2L_g^2M^2\rho^2\sigma_g^2\alpha_2(N)\\
&+ \frac{200\lambda\beta^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu} + 3\sigma_h^2 + 6\sigma_f^2\Big].
\end{aligned}
$$

Then, the proof is complete. □

D.1. Proof of Theorem 1

Theorem 2 (Restatement of Theorem 1)**.** *Suppose Assumption 1-4 hold. Further set* $\lambda \le \min\{10, \frac{1}{L_g}\}$, $\alpha_k^i = \frac{\alpha_k}{\tau_i}$ *an* $\beta_k^i = \frac{\beta_k}{\tau_i}$ *for all* $i \in S$*. Define* $\bar\beta = \Big(\frac{M_fL_y}{2}\bar\alpha_2 + 11M_fL_y + \eta L_{yx}D_h^2\bar\alpha_2 + \frac{(6+\frac{\bar\alpha_2}{3})(N+1)\lambda L_yL_g^2}{M_f}\Big(\frac{328\rho^2M^2}{\mu^3} + \frac{6L_f^2}{\mu}\Big)\Big)\frac{1}{\mu}$, $\bar\alpha_1 = \frac{1}{8L_f'+16M_fL_y+\frac{8M_fL_{yx}}{\eta L_y}}$, $\bar\alpha_2 = \frac{1}{324M_f^2+6M_f}$, $\bar\alpha_3 = \frac{N\min\{1,\lambda,\frac{1}{6L_g}\}}{2\bar\beta}$, *and* $\sigma_h^2 = \frac{\lambda(N+1)L_g^2M^2}{\mu}$ *, where* $L_f' = L_f + \frac{L_g(L_f+M_f)}{\mu} + \frac{M}{\mu}(\rho + \frac{L_g\rho}{\mu})$, $M_f = L_f + \frac{L_gL_f}{\mu} + \frac{M}{\mu}(\rho + \frac{L_g\rho}{\mu})$, $L_y = \frac{L_g}{\mu}$, *and* $L_{yx} = \frac{\rho+\rho L_y}{\mu} + \frac{L_g(\rho+\rho L_y)}{\mu^2}$*. Besides, define*

$$
\begin{aligned}
c_0 =&2L_f' + 4M_fL_y + \frac{2L_{yx}M_f}{\eta L_y},\\
c_1 =&\frac{1}{4} + 4L_f' + 8M_fL_y + \frac{4L_{yx}M_f}{\eta L_y},\\
c_2 =&\frac{1}{2} + 4L_f' + 8M_fL_y + \frac{4L_{yx}M_f}{\eta L_y},\\
c_3 =&\frac{25M_f}{L_y}\Big[1 + (12 + \frac{2\alpha_k}{3})(\frac{2\bar\alpha_2\lambda^2L_g^2M^2\rho^2L_y}{NM_f}\alpha_2(N) + \frac{\bar\alpha_2\lambda L_f^2L_g^2(N+1)L_y}{\mu M_f}) + \frac{M_fL_y\bar\alpha_2^2}{4}\\
&+ \frac{11\bar\alpha_2M_fL_y}{2} + \frac{\eta L_{yx}D_h^2\bar\alpha_2^2}{2}\Big]\frac{\bar\beta^2}{N},
\end{aligned}
$$

where $\eta = \frac{M_f}{L_y}$ *and* $D_h^2 = 8M^2 + \frac{4\lambda(N+1)L_g^2M^2}{\mu}$. *Choose parameters such that* $\alpha_k = \min\{\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3, \frac{\bar{\alpha}}{\sqrt{K}}\}$, $\beta_k \in [\max\{\frac{\bar{\beta}\alpha_k}{N}, \frac{\lambda}{10}\}, \min\{1, \lambda, \frac{1}{6L_g}\}]$, *where* $\bar{\alpha}$ *is a parameter that can be tuned. Then we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}\Big(\frac{1}{\min\{\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3\}K} + \frac{1}{\bar{\alpha}\sqrt{K}} + \frac{\bar{\alpha}\max\{c_0, c_1\sigma_h^2, c_2, c_3\}}{\sqrt{K}} + (1-\lambda\mu)^{2N}\Big). \tag{38}$$

Proof. Now, we define a Lyapunov function

$$\mathbb{W}_k := f(x_k, y^*_{(x_k)}) + \frac{M_f}{L_y}\|y_k - y^*_{(x_k)}\|^2.$$

Motivated by (Chen et al., 2021a), we bound the difference between two Lyapunov functions as

$$\mathbb{W}_{k+1} - \mathbb{W}_k = f(x_{k+1}, y^*_{(x_{k+1})}) - f(x_k, y^*_{(x_k)}) + \frac{M_f}{L_y}(\|y_{k+1} - y^*_{(x_{k+1})}\|^2 - \|y_k - y^*_{(x_k)}\|^2). \tag{39}$$

Recall that $\alpha_k^i = \frac{\alpha_k}{\tau_i}$, $\beta_k^i = \frac{\beta_k}{\tau_i}, \forall i \in S$. Using such stepsizes and substituting Lemma 5 into Equation (39), we have

$$\begin{aligned}
&\mathbb{E}[\mathbb{W}_{k+1}] - \mathbb{E}[\mathbb{W}_k] \\
&\quad \leq -\frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(x_k)] + 4\alpha_k^2\sigma_h^2L'_f + 4\alpha_k^2\sigma_f^2L'_f + 2\alpha_k^2M^2L'_f \\
&\qquad - \frac{\alpha_k}{2}(1 - 4\alpha_kL'_f)\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i^D(x^i_{k,\upsilon}, y_+) - \bar{h}^I(x)\big)\Big\|^2\Big] \\
&\qquad + \frac{3\alpha_k}{2}\Big[\big(4\lambda^2L_g^2M^2\alpha_1(N) + 4\lambda^2L_f^2L_g^2\alpha_3(N)\big)\,\mathbb{E}[\|y_k - y^*_{(x_k)}\|^2] + \frac{4L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2} \\
&\qquad + 400\lambda^2\beta_k^2L_g^2M^2\sigma_g^2\rho^2\alpha_2(N) + \frac{200\lambda\beta_k^2\sigma_g^2L_f^2L_g^2N(N+1)}{\mu} + \frac{M_f^2}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\mathbb{E}[\|x^i_{k,\upsilon} - x_k\|^2] \\
&\qquad + M_f^2\,\mathbb{E}[\|y_{k+1} - y^*_{(x_k)}\|^2]\Big] + \frac{M_f}{L_y}\mathbb{E}[\|y_{k+1} - y^*_{(x_{k+1})}\|^2 - \|y_k - y^*_{(x_k)}\|^2].
\end{aligned} \tag{40}$$

Then, following Lemma 6, Equation (40) can be rewritten as

$$\begin{aligned}
&\mathbb{E}[\mathbb{W}_{k+1}] - \mathbb{E}[\mathbb{W}_k] \\
&\quad \leq 4\alpha_k^2\sigma_h^2L'_f + 4\alpha_k^2\sigma_f^2L'_f + 2\alpha_k^2M^2L'_f + \frac{M_f}{L_y}b_3(\alpha_k)(2\sigma_h^2 + 2\sigma_f^2 + M^2) \\
&\qquad + \frac{300\alpha_k\lambda\beta_k^2L_f^2L_g^2N(N+1)\sigma_g^2}{\mu} + \frac{6\alpha_kL_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2} + 600\alpha_k\lambda^2\beta_k^2L_g^2M^2\rho^2\sigma_g^2\alpha_2(N) \\
&\qquad - \frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{3\alpha_kM_f^2}{2m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\mathbb{E}[\|x^i_{k,\upsilon} - x_k\|^2] && \text{(41a)} \\
&\qquad - \big(\frac{\alpha_k}{2} - 2\alpha_k^2L'_f - \frac{M_f}{L_y}b_1(\alpha_k)\big)\,\mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\upsilon=0}^{\tau_i-1}\big(\bar{h}_i(x^i_{k,\upsilon}, y_+) - \bar{h}(x)\big)\Big\|^2\Big] && \text{(41b)} \\
&\qquad + \big(\frac{3\alpha_kM_f^2}{2} + \frac{M_f}{L_y}b_2(\alpha_k)\big)\,\mathbb{E}[\|y_{k+1} - y^*_{(x_k)}\|^2] \\
&\qquad + \big(6\alpha_k\lambda^2L_g^2M^2\alpha_1(N) + 6\alpha_k\lambda^2L_f^2L_g^2\alpha_3(N) - \frac{M_f}{L_y}\big)\,\mathbb{E}[\|y_k - y^*_{(x_k)}\|^2]. && \text{(41c)}
\end{aligned}$$

Set $\gamma = M_fL_y\alpha_k$. Then according to the selections in Theorem 2 that $\alpha_k \leq \frac{1}{324M_f^2 + 6M_f}$, $\alpha_k \leq \frac{1}{8L'_f + 16M_fL_y + \frac{8M_fL_{yx}}{\eta L_y}}$, and substituting Equation (15) in Equation (41c), the following results can be obtained.

$$\text{(41a)} \leq -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(x_k)\|^2] + \frac{\alpha_k^2}{4}\sigma_h^2 + \frac{\alpha_k^2}{2}\sigma_f^2 + \frac{\alpha_k^2L_g^2M^2(1-\lambda\mu)^{2N+2}}{3\mu^2}$$

$$+\frac{100\alpha_k^2\lambda^2\beta_k^2L_g^2M^2\rho^2\sigma_g^2}{3}\alpha_2(N)+\frac{50\alpha_k^2\lambda\beta_k^2L_f^2L_g^2N(N+1)\sigma_g^2}{3\mu}+\frac{25M_f}{L_y}(\frac{M_fL_y}{4}\alpha_k^2)N\beta_k^2\sigma_g^2$$
$$+\frac{M_f}{L_y}\Big[(\frac{M_fL_y}{4}\alpha_k^2)(1-\frac{\beta_k\mu}{2})^N+\frac{\alpha_k^2\lambda^2L_yL_g^2M^2}{3M_f}\alpha_1(N)+\frac{\alpha_k^2\lambda^2L_yL_f^2L_g^2}{3M_f}\alpha_3(N)\Big]\mathbb{E}[\|y_k-y^*_{(x_k)}\|^2], \tag{42}$$

In (41b), we have $\frac{\alpha_k}{2}-2\alpha_k^2L_f'-\frac{M_f}{L_y}b_1(\alpha_k)\geq 0,$ (43)

$$\text{(41c)}\leq\frac{25M_f}{L_y}\big(\frac{3\alpha_kM_fL_y}{2}+b_2(\alpha_k)\big)N\beta_k^2\sigma_g^2+\frac{M_f}{L_y}\Big[\big(\frac{3\alpha_kM_fL_y}{2}+b_2(\alpha_k)\big)(1-\frac{\beta_k\mu}{2})^N$$
$$+\frac{6\alpha_k\lambda^2L_g^2L_yM^2\alpha_1(N)}{M_f}+\frac{6\alpha_k\lambda^2L_f^2L_yL_g^2}{M_f}\alpha_3(N)-1\Big]\mathbb{E}[\|y_k-y^*_{(x_k)}\|^2]. \tag{44}$$

Then, adding Equation (42), Equation (43) and Equation (44) together, we have

$$\begin{aligned}
&\mathbb{E}[\mathbb{W}_{k+1}]-\mathbb{E}[\mathbb{W}_k]\\
&\leq-\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(x_k^*)\|^2]+\frac{\alpha_k^2\sigma_f^2}{2}+\frac{\alpha_k^2\sigma_h^2}{4}+\frac{50\alpha_k\lambda\beta_k^2L_f^2L_g^2N(N+1)\sigma_g^2}{\mu}(6+\frac{\alpha_k}{3})\\
&\quad+2\alpha_k^2L_f'(2\sigma_f^2+2\sigma_h^2+M^2)+100\alpha_k\lambda^2\beta_k^2L_g^2M^2\rho^2\sigma_g^2(6+\frac{\alpha_k}{3})\alpha_2(N)+\frac{M_f(2\sigma_f^2+2\sigma_h^2+M^2)}{L_y}b_3(\alpha_k)\\
&\quad+\frac{2\alpha_kL_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}(3+\frac{\alpha_k}{6})+\frac{25M_f}{L_y}(\frac{M_fL_y}{4}\alpha_k^2+\frac{3M_fL_y\alpha_k}{2}+b_2(\alpha_k))N\beta_k^2\sigma_g^2\\
&\quad+\frac{M_f}{L_y}\Big(\big(\frac{M_fL_y}{4}\alpha_k^2+\frac{3M_fL_y\alpha_k}{2}+b_2(\alpha_k)\big)(1-\frac{\beta_k\mu}{2})^N-1+\frac{2\alpha_k\lambda^2L_yL_g^2M^2}{M_f}\alpha_1(N)(3+\frac{\alpha_k}{6})\\
&\qquad+\frac{2\alpha_k\lambda^2L_yL_f^2L_g^2}{M_f}\alpha_3(N)(3+\frac{\alpha_k}{6})\Big)\mathbb{E}[\|y_k-y^*_{(x_k)}\|^2]\\
&\leq-\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(x_k^*)\|^2]+\frac{\alpha_k^2\sigma_f^2}{2}+\frac{\alpha_k^2\sigma_h^2}{4}+\frac{50\alpha_k\lambda\beta_k^2L_f^2L_g^2N(N+1)\sigma_g^2}{\mu}(6+\frac{\alpha_k}{3})\\
&\quad+2\alpha_k^2L_f'(2\sigma_f^2+2\sigma_h^2+M^2)+100\alpha_k\lambda^2\beta_k^2L_g^2M^2\rho^2\sigma_g^2(6+\frac{\alpha_k}{3})\alpha_2(N)+\frac{M_f(2\sigma_f^2+2\sigma_h^2+M^2)}{L_y}b_3(\alpha_k)\\
&\quad+\frac{2\alpha_kL_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}(3+\frac{\alpha_k}{6})+\frac{25M_f}{L_y}(\frac{M_fL_y}{4}\alpha_k^2+\frac{3M_fL_y\alpha_k}{2}+b_2(\alpha_k))N\beta_k^2\sigma_g^2\\
&\quad+\frac{M_f}{L_y}\Big[\big(\frac{M_fL_y}{4}\alpha_k^2+\frac{3M_fL_y\alpha_k}{2}+b_2(\alpha_k)\big)(1-\frac{\beta_k\mu}{2})^N-1\\
&\qquad+\frac{2\alpha_k\lambda^2L_yL_g^2M^2}{M_f}(6+\frac{\alpha_k}{3})(N+1)(1-\frac{\beta_k\mu}{2})^N\Big(\frac{2\rho^2}{\lambda\mu^3}+\frac{80\rho^2}{\lambda\mu^3}\Big)\\
&\qquad+\frac{3\alpha_k\lambda L_yL_f^2L_g^2}{\mu M_f}(6+\frac{\alpha_k}{3})(N+1)(1-\frac{\beta_k\mu}{2})^N\Big]\mathbb{E}[\|y_k-y^*_{(x_k)}\|^2],
\end{aligned} \tag{45}$$

where in the last inequality, recalling from Lemma 6 and Proposition 1 that $b_2(\alpha):=1+4\gamma+\frac{\eta L_{yx}D_h^2\alpha_k^2}{2}$, $\alpha_1(N)=4(N+1)(1-\frac{\beta_k\mu}{2})^N\Big(\frac{\rho^2}{\lambda\mu^3}+\frac{4\rho^2}{\beta_k\mu^3}\Big)$, $\alpha_3(N)=3(N+1)\frac{(1-\beta_k\mu)^N}{\lambda\mu}$, we choose $\beta_k\geq\frac{\lambda}{10}$. Based on the parameters selections in Theorem 2 that $\beta_k\geq\Big(\frac{M_fL_y}{2}\alpha_k+11M_fL_y+\eta L_{yx}D_h^2\alpha_k+\frac{(6+\frac{\alpha_k}{3})(N+1)\lambda L_yL_g^2}{M_f}\Big(\frac{328\rho^2M^2}{\mu^3}+\frac{6L_f^2}{\mu}\Big)\Big)\frac{\alpha_k}{\mu N}$ and $\gamma=M_fL_y\alpha_k$, we have

$$\Rightarrow\exp\Big(\frac{M_fL_y}{4}\alpha_k^2+\frac{11M_fL_y\alpha_k}{2}+\frac{\eta L_{yx}D_h^2\alpha_k^2}{2}+\frac{\alpha_k(6+\frac{\alpha_k}{3})(N+1)\lambda^2L_yL_g^2}{M_f}\times$$
$$\Big(\frac{164\rho^2M^2}{\lambda\mu^3}+\frac{6L_f^2}{\lambda\mu}\Big)\Big)\exp(-\frac{N\beta_k\mu}{2})\leq 1$$

$$
\begin{aligned}
\Rightarrow \big(&\frac{M_fL_y}{4}\alpha_k^2+\frac{3M_fL_y\alpha_k}{2}+b_2(\alpha_k)\big)(1-\frac{\beta_k\mu}{2})^N+\frac{2\alpha_k\lambda^2L_yL_g^2M^2}{M_f}\alpha_1(N)(3+\frac{\alpha_k}{6})\\
&+\frac{2\alpha_k\lambda^2L_yL_f^2L_g^2}{M_f}\alpha_3(N)(3+\frac{\alpha_k}{6})-1\le 0. \qquad (46)
\end{aligned}
$$

Then plugging Equation (46) into Equation (45), we can obtain that

$$
\begin{aligned}
&\mathbb{E}[\mathbb{W}_{k+1}]-\mathbb{E}[\mathbb{W}_k]\\
&\le-\frac{\alpha_k}{4}\,\mathbb{E}[\|\nabla f(x_k)\|^2]+\frac{2\alpha_kL_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}(3+\frac{\alpha_k}{6})+\big(2\alpha_k^2L_f'+\frac{M_f}{L_y}b_3(\alpha_k)\big)M^2\\
&\quad+\big(4\alpha_k^2L_f'+\frac{\alpha_k^2}{4}+\frac{2M_f}{L_y}b_3(\alpha_k)\big)\sigma_h^2+\big(4\alpha_k^2L_f'+\frac{\alpha_k^2}{2}+\frac{2M_f}{L_y}b_3(\alpha_k)\big)\sigma_f^2\\
&\quad+\frac{25M_f}{L_y}\big(\frac{\alpha_k\lambda^2L_g^2M^2\rho^2L_y}{NM_f}(24+\frac{4\alpha_k}{3})\alpha_2(N)+\frac{\alpha_k\lambda L_f^2L_g^2(N+1)L_y}{\mu M_f}(12+\frac{2\alpha_k}{3})\\
&\quad+\frac{M_fL_y\alpha_k^2}{4}+\frac{3\alpha_kM_fL_y}{2}+b_2(\alpha_k)\big)\frac{\bar\beta^2}{N}\alpha_k^2\sigma_g^2\\
&\le-\frac{\alpha_k}{4}\,\mathbb{E}[\|\nabla f(x_k)\|^2]+\frac{2\alpha_kL_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}(3+\frac{\alpha_k}{6})+c_0\alpha_k^2M^2+c_1\alpha_k^2\sigma_h^2+c_2\alpha_k^2\sigma_f^2+c_3\alpha_k^2\sigma_g^2 \qquad (47)
\end{aligned}
$$

where c_0,c_1,c_2,c_3 are defined in Theorem 2. Finally, telescoping Equation (47) yields

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k)\|^2]\le&\frac{4\mathbb{W}^0}{\sum_{k=0}^{K-1}\alpha_k}+\frac{4c_0\sum_{k=0}^{K-1}\alpha_k^2}{\sum_{k=0}^{K-1}\alpha_k}M^2+\frac{4c_1\sum_{k=0}^{K-1}\alpha_k^2}{\sum_{k=0}^{K-1}\alpha_k}\sigma_h^2+\frac{4c_2\sum_{k=0}^{K-1}\alpha_k^2}{\sum_{k=0}^{K-1}\alpha_k}\sigma_f^2\\
&+\frac{4c_3\sum_{k=0}^{K-1}\alpha_k^2}{\sum_{k=0}^{K-1}\alpha_k}\sigma_g^2+\frac{8L_g^2M^2\sum_{k=0}^{K-1}\alpha_k(1-\lambda\mu)^{2N+2}}{\mu^2\sum_{k=0}^{K-1}\alpha_k}(3+\frac{\alpha_k}{6})\\
\le&\frac{4\mathbb{W}^0}{\min\{\bar\alpha_1,\bar\alpha_2,\bar\alpha_3,\frac{\bar\alpha}{\sqrt{K}}\}K}+\frac{4c_0\bar\alpha}{\sqrt{K}}M^2+\frac{4c_1\bar\alpha}{\sqrt{K}}\sigma_h^2+\frac{4c_2\bar\alpha}{\sqrt{K}}\sigma_f^2+\frac{4c_3\bar\alpha}{\sqrt{K}}\sigma_g^2\\
&+\frac{8L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}(3+\frac{\alpha_k}{6})\\
\le&\frac{4\mathbb{W}^0}{\min\{\bar\alpha_1,\bar\alpha_2,\bar\alpha_3\}K}+\frac{4\mathbb{W}^0}{\bar\alpha\sqrt{K}}+\frac{4c_0\bar\alpha}{\sqrt{K}}M^2+\frac{4c_1\bar\alpha}{\sqrt{K}}\sigma_h^2+\frac{4c_2\bar\alpha}{\sqrt{K}}\sigma_f^2+\frac{4c_3\bar\alpha}{\sqrt{K}}\sigma_g^2\\
&+\frac{8L_g^2M^2(1-\lambda\mu)^{2N+2}}{\mu^2}(3+\frac{\alpha_k}{6})\\
=&\mathcal{O}\Big(\frac{1}{\min\{\bar\alpha_1,\bar\alpha_2,\bar\alpha_3\}K}+\frac{1}{\bar\alpha\sqrt{K}}+\frac{\bar\alpha\max\{c_0,c_1\sigma_h^2,c_2,c_3\}}{\sqrt{K}}+(1-\lambda\mu)^{2N}\Big). \qquad (48)
\end{aligned}
$$

The proof is complete. □

D.2. Proof of Corollary 1

Proof. Let $\eta=\frac{M_f}{L_y}=\mathcal{O}(\kappa_g)$. It follows from Lemma 4 and Theorem 2 that

$$
\begin{aligned}
L_y=&\mathcal{O}(\kappa_g),\ L_{yx}=\mathcal{O}(\kappa_g^3),\ M_f=\mathcal{O}(\kappa_g^2),\ L_f'=\mathcal{O}(\kappa_g^3),\ \sigma_h^2=\mathcal{O}(N\kappa_g),\\
\bar\alpha_1=&\mathcal{O}(\kappa_g^{-3}),\ \bar\alpha_2=\mathcal{O}(\kappa_g^{-4}),\ \bar\alpha_3=\mathcal{O}(N\kappa_g^{-4}+\kappa_g^{-3}),\ \bar\beta=\mathcal{O}(\kappa_g^4+N\kappa_g^3),\\
c_0=&\mathcal{O}(\kappa_g^3),\ c_1=\mathcal{O}(\kappa_g^3),\ c_2=\mathcal{O}(\kappa_g^3),\ c_3=\mathcal{O}\Big(\big(\frac{\kappa_g^8}{N}+N\kappa_g^6\big)\big(\kappa_g+N\kappa_g^{-1}\big)\Big).
\end{aligned} \qquad (49)
$$

Now, if we select $N=\mathcal{O}(\kappa_g)$, $\bar\alpha=\mathcal{O}(\kappa_g^{-4})$, we obtain from Equation (48) that

$$
\bar\beta=\mathcal{O}(\kappa_g^4),\ c_3=\mathcal{O}(\kappa_g^8),\ \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x_k)\|^2]=\mathcal{O}(\frac{\kappa_g^4}{K}+\frac{\kappa_g^4}{\sqrt{K}}).
$$

To achieve an ϵ-stationary point, it requires $K = \mathcal{O}(\kappa_g^8 \epsilon^{-2})$ and the number of samples in ξ and ζ are both $\mathcal{O}(\kappa_g^9 \epsilon^{-2})$. Then the proof is complete. □