LOCAL STOCHASTIC BILEVEL OPTIMIZATION WITH MOMENTUM-BASED VARIANCE REDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bilevel Optimization has witnessed notable progress recently with new emerging efficient algorithms and has been applied to many machine learning tasks such as data cleaning, few-shot learning, and neural architecture search. However, little attention has been paid to solving the bilevel problems under distributed setting. Federated learning (FL) is an emerging paradigm that solves machine learning tasks over distributed-located data. FL problems are challenging to solve due to the heterogeneity and communication bottleneck. However, it is unclear how these challenges will affect the convergence of Bilevel Optimization algorithms. In this paper, we study Federated Bilevel Optimization problems. Specifically, we first propose the FedBiO, a deterministic gradient-based algorithm, and we show that it requires $O(\epsilon^{-1.5})$ number of steps/communication steps to reach an ϵ-stationary point. Then we propose FedBiOAcc to accelerate FedBiO with the momentum-based variance-reduction technique under the stochastic scenario. We show that FedBiOAcc needs $O(\epsilon^{-1.5})$ number of steps and $O(\epsilon^{-1})$ communication steps, this matches the best known rate for single-level stochastic federated algorithms. Finally, we validate our proposed algorithms via the important Fair Federated Learning task. More specifically, we define a bilevel-based group fair FL objective. Our algorithms show superior performance compared to other baselines in numerical experiments.

1 INTRODUCTION

Bilevel optimization problems Willoughby (1979); Solodov (2007) involve two levels of problems: an outer problem and an inner problem. The two problems are entangled: the outer problem is a function of the minimizer of the inner problem. Recently, great progress has been made to solve this type of problems, especially, efficient single loop algorithms have been developed based on various gradient approximation techniques Ji et al. (2020); Huang & Huang (2021). Bilevel optimization problems also frequently emerge in machine learning tasks, such as hyper-parameter optimization, meta-learning, neural architecture search *etc.*. However, most existing Bilevel Optimization work focuses on the standard non-distributed setting, and how to solve the Bilevel Optimization problems under distributed settings is underexplored. Federated learning is a recently promising distributed learning paradigm. In Federated Learning McMahan et al. (2017), a set of clients jointly solve a machine learning task under the coordination of a central server. To protect user privacy and reduce communication burden, clients perform multiple steps of local update before communication, but this slows down the convergence. Various algorithms Wang et al. (2019b); Yu et al. (2019); Haddadpour & Mahdavi (2019); Karimireddy et al. (2019); Bayoumi et al. (2020); Xing et al. were proposed to accelerate its training. However, most of these algorithms focus on standard single level optimization problems. In Xing et al., the author considered one type of bilevel formulation, but their algorithm needs to communicate the Hessian matrix for every iteration, which is impractical in practice due to high communication cost. So efficient algorithms designed for Federated Bilevel Optimization are still missing. In this work, we propose two novel algorithms for Federated Bilevel Optimization and aim to make one step forward to mitigate this gap.

More specifically, we propose **FedBiO** and **FedBiOAcc**. The FedBiO algorithm adapts single loop bilevel algorithms to the federated setting. More precisely, clients optimize their local bilevel problems with a single loop algorithm for several steps and then communicate with the server to average their local states. To further accelerate the FedBiO algorithm, we utilize the momentum-

based variance reduction technique to control the stochastic noise in the local updates of FedBiO. We denote this accelerated algorithm as FedBiOAcc. The convergence analysis of the two algorithms is very challenging and our first main contribution is to provide a rigorous analysis of the two algorithms. We need to carefully balance two types of errors to show the convergence. The first type of error is the 'consensus error': In Federated Learning, clients make updates locally, and local states drift away from each other. As a result, the gradient directions queried in these states do not represent the true descent directions. The consensus error also exists in Federated Bilevel Optimization problems where both inner and outer variables drift. The second type of error is '(hyper)-gradient bias': In bilevel optimization, exact (hyper)-gradient requires solving the inner problem which is computationally expensive. So a practical compromise is to solve the inner problem approximately and use a biased (hyper)-gradient in the training. In Federated Bilevel Optimization, the (hyper)-gradient bias entangles with the consensus error, which makes the analysis even more difficult. However, by carefully choosing a potential function and exploiting the recursive relations of the above errors, we successfully show the convergence of our algorithms.

Finally, to illustrate the application of our algorithms, we study the group fairness problem in federated learning through the lens of Bilevel Optimization. Fairness over sensitive groups is one of the most important desiderata in the development of machine learning models. However, Federated Learning by design does not learn group-fair models. Meanwhile, due to the fact that sensitive groups often spread across different clients and clients are not allowed to share data with each other. Fair algorithms developed in non-distributed setting can not be applied directly. Recently, several research works focus on group fairness in Federated Learning: Papadaki et al. (2021) exploited the notion of minimax fairness to learn group fair models, but requires access of the global statistics of sensitive groups; Cui et al. (2021) enforced the local group fairness with linear constraints, but a local fair model may not be global group fair as clients often have heterogeneous distributions. On top of these limitations, we propose a bilevel formulation to develop group fair models. More precisely, we use a small set of samples that are balanced group-wise to tune the groups weights; in other words, we find the optimal group weights such that the learned weighted model can perform well over the validation set. We solve this problem with our two proposed algorithms and validate them over real-world datasets. Finally, we highlight the main **contributions** of our paper as follows:

1. We propose two novel federated bilevel learning algorithms: FedBiO and FedBiOAcc. We theoretically show the convergence of both algorithms: FedBiO has iteration and communication complexity of $O(\epsilon^{-1.5})$ and FedBiOAcc has iteration complexity of $O(\epsilon^{-1.5})$, communication complexity of $O(\epsilon^{-1})$ and linear speed up *w.r.t* the number of clients. In particular, FedBiOAcc matches the optimal rate of single-level stochastic federated algorithms;
2. We propose a Bilevel Optimization Formulation to improve the group fairness in Federated Learning. We compare our algorithms with various baselines. Experimental results show superior performance of our new algorithms.

Notations We use ∇ to denote the full gradient, use ∇_x to denote the partial derivative for variable x, higher order derivatives follow similar rules. $||\cdot||$ represents l_2 norm for vectors and spectral norm for matrices. $[K]$ represents the sequence of integers from 1 to K.

2 Related Works

Bilevel Optimization Bilevel optimization dates back to at least 1960s when Willoughby (1979) proposed a regularization method, and then followed by many research works Ferris & Mangasarian (1991); Solodov (2007); Yamada et al. (2011); Sabach & Shtern (2017), while in machine learning community, similar ideas in the name of implicit differentiation were also used in Hyper-parameter Optimization Larsen et al. (1996); Chen & Hagan (1999); Bengio (2000); Do et al. (2007). Early algorithms for Bilevel Optimization solved the accurate inner problem solution for each outer variable. Recently, researchers developed algorithms which solve the inner problem with a fixed number of steps, and use 'back-propagation through time' technique to compute the hyper-gradient Domke (2012); Maclaurin et al. (2015); Franceschi et al. (2017); Pedregosa (2016); Shaban et al. (2018). Very Recently, it witnessed a surge of interest in using implicit differentiation to derive single loop algorithms Ghadimi & Wang (2018); Hong et al. (2020); Ji et al. (2020); Ji & Liang (2021); Khanduri et al. (2021); Chen et al. (2021); Yang et al. (2021); Huang & Huang (2021); Li et al. (2021a). The bilevel optimization has been widely applied to various machine learning applications, such as Hyper-parameter optimization Lorraine & Duvenaud (2018); Okuno et al. (2018); Franceschi et al.

(2018), meta learning Zintgraf et al. (2019); Song et al. (2019); Soh et al. (2020), neural architecture search Liu et al. (2018); Wong et al. (2018); Xu et al. (2019), adversarial learning Tian et al. (2020); Yin et al. (2020); Gao et al. (2020), deep reinforcement learning Yang et al. (2018); Tschiatschek et al. (2019), *etc.*

Federated Learning Federated learning McMahan et al. (2017) is a promising privacy-preserving learning paradigm over distributed data. A basic algorithm for FL is the FedAvg McMahan et al. (2017) algorithm, where clients receive the current model from the server at each synchronization step and then update the model locally for several steps and finally upload the new model back to the server. Compared to the traditional data-center distributed learning, Federated Learning poses new challenges including data heterogeneity, privacy concerns, high communication cost and unfairness. To deal with these challenges, some variants of FedAvg Karimireddy et al. (2019); Li et al. (2019b); Sahu et al. (2018); Zhao et al. (2018); Mohri et al. (2019); Li et al. (2021b) are proposed. For example, Li et al. (2018) added regularization terms over the client objective to reduce the client drift. Hsu et al. (2019); Karimireddy et al. (2019); Wang et al. (2019a) used variance reduction techniques to control variates. Fairness in Federated Learning has also drawn more attention recently. Some researchers Mohri et al. (2019); Deng et al. (2020); Li et al. (2019a; 2021b) focus on making models exhibit similar performance across different clients. More recently, group fairness is also studied in federated learning. One possible approach is to learn optimal group weights by formulating it as a minimax optimization problem Du et al. (2021); Papadaki et al. (2021). Another approach is to re-weight the sensitive groups based on local or global statistics Abay et al. (2020); Ezzeldin et al. (2021), this approach often involves the transfer of sensitive information. Then a recent work Cui et al. (2021) proposes FCFL which improves both client fairness and group fairness with multi-objective optimization approach. In our work, we formulate the group fairness as a bilevel optimization problem and solve it as an application of our algorithms.

3 PRELIMINARIES

Bilevel Optimization Bilevel Optimization problems are composed of two levels of entangled problems as defined in Eq. 1:

$$\min_{x\subseteq\mathbb{R}^p} h(x) \coloneqq f(x, y_x) \quad \text{s.t. } y_x = \arg\min_{y\in\mathbb{R}^d} g(x, y), \tag{1}$$

As shown in Eq. 1, the outer problem ($f(x, y_x)$) depends on the solution of the inner problem ($g(x, y)$. In machine learning, we usually consider the following stochastic formulation as shown in Eq. 2:

$$\min_{x\subseteq\mathbb{R}^p} h(x) \coloneqq \mathbb{E}[f(x, y_x; \mathcal{B}_f)] \text{ s.t. } y_x = \arg\min_{y\in\mathbb{R}^d} \mathbb{E}[g(x, y; \mathcal{B}_g)], \tag{2}$$

where both the outer and inner problems are defined as expectations of some random variables $\mathcal{B}_f$ (outer) and $\mathcal{B}_g$ (inner). Next, we state some assumptions about the problems Eq. 1 and Eq. 2:

Assumption 1. Function $f(x, y) \coloneqq \mathbb{E}[f(x, y; \mathcal{B}_f)]$ is possibly non-convex and $g(x, y) \coloneqq \mathbb{E}[g(x, y; \mathcal{B}_g)]$ is μ-strongly convex *w.r.t* y for any given x.

Assumption 2. Function $f(x, y)$ is L-Lipschitz and has B-bounded gradient;

Assumption 3. Function $g(x, y)$ is L-Lipschitz. For higher-order derivatives, we have:

a) $\|\nabla^2_{xy} g(x, y)\| \leq C_{g,xy}$ for some constant $C_{g,xy}$

b) $\nabla^2_{xy} g(x, y)$ and $\nabla^2_{y^2} g(x, y)$ are Lipschitz continuous with constant $L_{g,xy}$ and L_{g,y^2} respectively

Assumption 4. We have unbiased stochastic first order and second order derivative oracle with bounded variance, *e.g.* we assume $\mathbb{E}[\nabla_x f(x, y; \xi)] = \nabla_x f(x, y)$ and $var(\nabla_x f(x, y; \xi)) \leq \sigma^2$

Remark 1. As stated in Assumption 1, we study the non-convex-strongly-convex bilevel optimization problems, this special case is widely studied in the bilevel literature Ji & Liang (2021); Ghadimi & Wang (2018).

Remark 2. Assumptions 2 and 3, we assume the Lipschitz condition also holds for the stochastic query, *i.e.* $f(x, y; \mathcal{B}_f)$ and $g(x, y; \mathcal{B}_g)$. Furthermore, we require stronger conditions in Assumptions 2 and 3 than single level optimization problems: bounded gradients (for f) and second order smoothness (for g). But these conditions are necessary to derive the smoothness of $h(x)$ and some other basic properties and are widely used in bilevel literature Ghadimi & Wang (2018); Ji et al. (2020).

Remark 3. A more complete version of Assumption 4 is included in the appendix, where we state all properties we assume the unbiased and bounded-variance assumption holds.

Next, we define the properties the hyper-gradient $\nabla h(x)$, firstly, we denote $\Phi(x, y)$ as:

$$\Phi(x, y) = \nabla_x f(x, y) - \nabla^2_{xy} g(x, y) \times [\nabla^2_{y^2} g(x, y)]^{-1} \nabla_y f(x, y), \tag{3}$$

Then based on Assumption 1, we can derive $\Phi(x, y_x) = \nabla h(x)$ (we omit the proof of this fact; please refer to related bilevel literature such as Ghadimi & Wang (2018)). Eq. 3 involves Hessian inverse, we usually evaluate it approximately. An approach of approximation is through the Neumann series Lorraine et al. (2020). More precisely, suppose we have independent minibatches of samples $\mathcal{B}_x = \{\mathcal{B}_j (j = 1, \ldots, Q), \mathcal{B}_f, \mathcal{B}_g\}$, then we estimate $\Phi(x, y)$ as:

$$\Phi(x, y; \mathcal{B}_x) = \nabla_x f(x, y; \mathcal{B}_f) - \eta \nabla_{xy} g(x, y; \mathcal{B}_g) \sum_{q=-1}^{Q-1} \prod_{j=Q-q}^{Q} (I - \eta \nabla^2_{y^2} g(x, y; \mathcal{B}_j)) \nabla_y f(x, y; \mathcal{B}_f) \tag{4}$$

We have the following Proposition about the approximation property of $\Phi(x, y; \mathcal{B}_x)$:

Proposition 1. *(Combine Lemma 4 and Lemma 7 in Yang et al. (2021)) Suppose Assumptions 2, 3 and 4 hold and $\eta < \frac{1}{L}$, the hypergradient estimator $\Phi(x, y; \mathcal{B}_x)$ w.r.t. x based on a minibatch $\mathcal{B}_x$ has bounded variance and bias:*

a) $\mathbb{E}[\|\mathbb{E}[\Phi(x, y; \mathcal{B}_x)] - \Phi(x, y)\|^2] \leq G_1^2$, *where* $G_1 = (1 - \eta\mu)^{Q+1} ML/\mu$

b) $\mathbb{E}\|\Phi(x, y; \mathcal{B}_x) - \mathbb{E}[\Phi(x, y; \mathcal{B}_x)]\|^2 \leq G_2^2$, *where* $G_2 = 2M^2 + 12M^2L^2\eta^2(Q + 1)^2 + 4M^2L^2(Q + 2)(Q + 1)^2\eta^4\sigma^2$

Finally, we show some properties of smoothness in Proposition 2. A more formal version of the proposition can be found in Proposition 6 in the appendix.

Proposition 2. *Suppose Assumptions 2 and 3 hold, the following statements hold: For any given $x_1, x_2 \in X$, we have $\|y_{x_1} - y_{x_2}\| \leq L_h \|x_1 - x_2\|$, $\|\nabla h(x_2) - \nabla h(x_1)\| \leq L_h \|x_2 - x_1\|$, $\|\Phi(x_1; y_1) - \Phi(x_2; y_2)\|^2 \leq L_h^2 (\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2)$, furthermore, we have $\|\Phi(x; y) - \nabla h(x)\| \leq L_h \|y_x - y\|$.*

Federated Learning A general FL problem studies the following problem:

$$\min_{x \subset \mathbb{R}^p} h(x) \coloneqq \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_\xi [f^{(m)}(x; \xi)], \tag{5}$$

There is one server and M clients. A basic algorithm to solve this problem is FedAvg McMahan et al. (2017), where clients perform multiple gradient descent steps before communication with the server.

4 Federated Bilevel Optimization

In this section, we discuss Federated Bilevel Optimization problems. Following the standard FL setting, we assume that there is one server and multiple clients. Specifically, the optimization problem solved by each client is a Bilevel Optimization Problem. More formally, we consider the following Federated Bilevel Optimization problem $h(x)$:

$$\min_{x \subset \mathbb{R}^p} h(x) \coloneqq \frac{1}{M} \sum_{m=1}^{M} f^{(m)}(x, y_x^{(m)}) \quad \text{s.t. } y_x^{(m)} = \arg\min_{y \in \mathbb{R}^d} g^{(m)}(x, y), \tag{6}$$

where M is the number of clients and $f^{(m)}(x, y)$ and $g^{(m)}(x, y)$ are the outer and inner problems of the client m, respectively. $h(x)$ denotes the overall objective. For ease of discussion, we denote $h^{(m)}(x) = f^{(m)}(x, y_x^{(m)})$, while $\nabla h^{(m)}(x)$ denotes the gradient *w.r.t* x. Note that it is possible that both $f^{(m)}(x, y) \neq f^{(k)}(x, y)$ and $g^{(m)}(x, y) \neq g^{(k)}(x, y)$ for $m \neq k, m, k \in [M]$. In other words, we consider the heterogeneous case.

Algorithm 1 Federated Bilevel Optimization (**FedBiO**)

1: **Input:** Initial states x_1 and y_1; learning rates $\{\gamma_t, \eta_t\}_{t=1}^T$
2: **Initialization:** Set $x_1^{(m)} = x_1$ and $y_1^{(m)} = y_1$
3: **for** $t = 1$ **to** T **do**
4: Compute $\omega_t^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})$ and compute $\nu_t^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)})$ with Eq. (3);
5: Update $y_{t+1}^{(m)} = y_t^{(m)} - \gamma_t \omega_t^{(m)}$, $\hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta_t \nu_t^{(m)}$;
6: **if** $t + 1 \bmod \mathrm{I} = 0$ **then**
7: $x_{t+1}^{(m)} = \bar{x}_{t+1} = 1/M \sum_{m=1}^M \hat{x}_{t+1}^{(m)}$
8: **else**
9: $x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)}$
10: **end if**
11: **end for**

Remark 4. It is possible to consider a slightly different formulation where the inner problem is also distributed across clients; we leave this as a future direction to investigate. Note that bilevel optimization with multiple lower tasks are considered in Guo et al. (2021), however, they do not solve the problem under FL constraints and instead focus on sampling effect of lower problems.

In machine learning, we consider the stochastic case of Eq. 6 as follows:

$$\min_{x \subset \mathbb{R}^p} h(x) \coloneqq \frac{1}{M} \sum_{m=1}^M \mathbb{E}[f^{(m)}(x, y_x^{(m)}; \mathcal{B}_f)] \quad \text{s.t. } y_x^{(m)} = \arg\min_{y \in \mathbb{R}^d} \mathbb{E}[g^{(m)}(x, y; \mathcal{B}_g)]. \tag{7}$$

Federated Bilevel Optimization problems are more complicated than single level Federated Learning problems. In FedAvg, clients perform multiple steps of local gradient descent before communication with the server. For the deterministic case, Eq. 5, clients evaluate the exact gradient $\nabla f^{(m)}$ locally. However, the local hypergradient $\nabla h^{(m)}(x)$ of Eq. 6 has the following form:

$$\nabla h^{(m)}(x) = \nabla_x f^{(m)}(x, y_x) - \nabla^2_{xy} g^{(m)}(x, y_x)[\nabla^2_{y^2} g^{(m)}(x, y_x)]^{-1} \nabla_y f^{(m)}(x, y_x),$$

where y_x is the minimizer of the inner objective as defined in Eq 6. To get y_x, we need to solve the inner problem for each new state of x, this is computationally expensive to evaluate a gradient. So it is infeasible to evaluate the exact hypergradient for Federated Bilevel Optimization, thus, the FedAvg algorithm is not suitable for solving Federated Bilevel Optimization problems.

The recent progress in Bilevel Optimization Ji et al. (2020) shows that exact y_x is not necessary to solve the problem, instead, an alternative update of inner variable and outer variable is sufficient. So, we propose our first algorithm named FedBiO whose procedure is shown in Algorithm 1. In the algorithm, we start from two random states $x_1^{(m)}$ and $y_1^{(m)}$. For each local iteration, we update $x_t^{(m)}$ and $y_t^{(m)}$ with a gradient-like step with the gradients defined in line 5 of Algorithm 1. For every I iterations, we average the x states over clients. Note that we do not average over the y state, as in Eq. 6, $y_x^{(m)}$ only depends on the state x and $g^{(m)}(x, y)$, the average of outer state $x^{(m)}$ is sufficient.

In Algorithm 1, clients perform alternative updates of inner and outer variables locally and communicate with the server every I iterations. Compared with single level Federated Learning problems, its convergence analysis is much more challenging. More specifically, suppose we consider the virtual average state $\bar{x}_t = \frac{1}{M} \sum_{m=1}^M \hat{x}_t^{(m)}$, and measure its convergence with $\|\nabla h(\bar{x}_t)\|^2$. There are two sources of errors. The first one is the outer variable consensus error defined as $\frac{1}{M} \sum_{m=1}^M \|\hat{x}_t^{(m)} - \bar{x}_t\|^2$, and the other one is the inner variable estimation error $\frac{1}{M} \sum_{m=1}^M \|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2$. Note that the outer variable consensus error is often seen in the analysis in FedAvg-type algorithms, which is caused by local updates. As for the inner variable estimation error, it measures the imperfection of inner variable. In FedBiO, these two types of errors are entangled with each other. To see that, for $\bar{t}_s$ which satisfies $\bar{t}_s + 1 = s \times I$, we have:

$$\begin{aligned}\|y_{\bar{t}_s}^{(m)} - y_{x_{\bar{t}_s}^{(m)}}^{(m)}\|^2 = \|y_{\bar{t}_s}^{(m)} - y_{\bar{x}_{\bar{t}_s}}^{(m)}\|^2 &\leq 2\|y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)}\|^2 + 2\|y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)} - y_{\bar{x}_{\bar{t}_s}}^{(m)}\|^2 \\ &\leq 2\|y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)}\|^2 + 2\rho^2 \|\hat{x}_{\bar{t}_s}^{(m)} - \bar{x}_{\bar{t}_s}\|^2,\end{aligned}$$

Algorithm 2 Accelerated Federated Bilevel Optimization (**FedBiOAcc**)

1: **Input:** constants c_ω, c_ν, γ, η, δ, u, σ, initial state (x_1, y_1);
2: **Initialization:** Set $y_1^{(m)} = y_1$, $x_1^{(m)} = x_1$ for $m \in [M]$
3: **for** $t = 1$ **to** T **do**
4: Randomly sample minibatches $\mathcal{B}_y$ and $\mathcal{B}_x$
5: **if** $t = 1$ **then**
6: $\omega_t^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y)$, $\hat{\nu}_t^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_x)$
7: **else**
8: $\omega_t^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) + (1 - c_\omega \alpha_{t-1}^2)(\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y))$
9: $\mu_{t-1}^{(m)} = \Phi^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_x)$, $\mu_t^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_x)$
10: $\hat{\nu}_t^{(m)} = \mu_t^{(m)} + (1 - c_\nu \alpha_{t-1}^2)(\nu_{t-1}^{(m)} - \mu_{t-1}^{(m)})$
11: **end if**
12: Evaluate $\alpha_t = \frac{\delta}{(u + \sigma^2 \times t)^{1/3}}$
13: $y_{t+1}^{(m)} = y_t^{(m)} - \gamma \alpha_t \omega_t^{(m)}$, $\hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta \alpha_t \hat{\nu}_t^{(m)}$
14: **if** $t + 1 \bmod \mathrm{I} = 0$ **then**
15: $x_t^{(m)} = \bar{x}_t = \frac{1}{M} \sum_{j=1}^M x_t^{(j)}$, $\nu_t^{(m)} = \bar{\nu}_t = \frac{1}{M} \sum_{j=1}^M \hat{\nu}_t^{(j)}$, $x_{t+1}^{(m)} = \bar{x}_{t+1} = \frac{1}{M} \sum_{j=1}^M \hat{x}_{t+1}^{(j)}$
16: **else**
17: $\nu_t^{(m)} = \hat{\nu}_t^{(m)}$, $x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)}$
18: **end if**
19: **end for**

The equality is because we average the state $x^{(m)}$ at step $\bar{t}_s$, the first inequality follows the triangle inequality, and the second inequality follows Proposition 2. The inequality shows that the inner variable estimation error can be decomposed to two parts: estimation error to $y^{(m)}_{\hat{x}_t^{(m)}}$ (denoted by local variable $\hat{x}_t^{(m)}$) and $\|y^{(m)}_{\hat{x}_{\bar{t}_s}^{(m)}} - y^{(m)}_{\bar{x}_{\bar{t}_s}}\|^2$ which is related to outer variable consensus error. The first error can be bounded following the standard argument of the gradient descent step (Line 5 in Algorithm 1), while for the outer variable consensus error, it relies on accumulated past consensus error, inner variable estimation error and a term related to the client heterogeneity (please refer to Lemma 3 in the appendix for detailed expressions). In other words, the imperfect inner variable estimation in turn increases the consensus error. Although the two types of errors increase the analysis complexity with entanglement, we could bound them by exploiting their recursive relations. In the convergence analysis sections, we show that our FedBiO converges with rate $O(\epsilon^{-1.5})$.

Next, we consider the Federated Stochastic Bilevel Optimization as defined in Eq. 7. To control stochastic noise, we apply the idea of momentum-based variance reduction Cutkosky & Orabona (2019). The algorithm procedure is summarized in Algorithm 2. The main step of the algorithm is as follows:

$$\omega_t^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) + (1 - c_\omega \alpha_{t-1}^2)(\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y))$$
$$\hat{\nu}_t^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_x) + (1 - c_\nu \alpha_{t-1}^2)(\nu_{t-1}^{(m)} - \Phi^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_x)).$$

where $\Phi^{(m)}$ follows the definition in Eq. 4 by replacing f and g with $f^{(m)}$ and $g^{(m)}$, respectively. If $t \bmod I = 0$, we average $x_t^{(m)}$, $x_{t-1}^{(m)}$ and the momentum state $\hat{\nu}_t^{(m)}$ as in line 17 of Algorithm 2. The analysis of FedBiOAcc is more complicated than that of FedBiO. There are several types of errors we need bound to get the convergence: including the entangled inner variable estimation error and the outer variable consensus error as in FedBiO, but also the biases from the momentum terms, *i.e.* the outer momentum bias $\|\nu_t^{(m)} - \nabla h(x_t^{(m)})\|^2$ and the inner momentum bias $\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\|^2$. However, we still see the favorable $O(\epsilon^{-1.5})$ convergence rate of FedBiOAcc by balancing different sources of errors. In fact, for both types of momentum bias, we can derive similar recursive equations as its non-distributed counterpart Yang et al. (2021) but with additional terms related to the outer variable consensus error. As for the consensus error, we can bound it by carefully choosing the related hyper-parameters in Algorithm 2.

5 CONVERGENCE ANALYSIS

In this section, we provide a formal analysis of the convergence of our two algorithms, *i.e.* FedBiO and FedBiOAcc. Before diving into the convergence results, we introduce two additional assumptions:

5.1 ADDITIONAL ASSUMPTIONS

We first state the assumptions needed in our analysis. We assume $f^{(m)}(x,y)$ and $g^{(m)}(x,y)$ for $m \in [M]$ satisfy Assumption 1,Assumption 2, Assumption 3 and Assumption 4 as Defined in Section 3. Next, we also need to bound the differences among clients to get convergence results. More precisely, we assume Assumption 5 holds. Similar assumptions have been used in previous Federated Learning literature Khanduri et al. (2021); Woodworth (2021).

Assumption 5. For any $m, j \in [M]$ and x, we have:$\|\nabla_x f^{(m)}(x,y) - \nabla_x f^{(j)} f(x,y)\| \leq \zeta_f$, $\|\nabla_y f^{(m)}(x,y) - \nabla_y f^{(j)} f(x,y)\| \leq \zeta_f$, $\|\nabla_{xy} g^{(m)}(x,y) - \nabla_{xy} g^{(j)}(x,y)\| \leq \zeta_{g,xy}$, $\|\nabla_{y^2} g^{(m)}(x,y) - \nabla_{y^2} g^{(j)}(x,y)\| \leq \zeta_{g,yy}$, $\|y_x^{(m)} - y_x^{(j)}\| \leq \zeta_{g^*}$, where ζ_f, $\zeta_{g,xy}$, $\zeta_{g,yy}$, ζ_{g^*} are constants.

Based on the above assumption, we can bound the overall heterogeneity of the function $h^{(m)}(x)$, $m \in [M]$, *i.e.* $\|\nabla h^{(m)}(x) - \nabla h^{(j)}(x)\| \leq \zeta$ for some constant ζ. The proof of this result is summarized in Proposition 7 in the Appendix. Next, in addition to the bounded noise Assumption 4. We make the following assumption:

Assumption 6. The bias and variance of the stochastic hypergradient are bounded *i.e.* $\mathbb{E}[\|\mu_t^{(m)} - \mathbb{E}[\mu_t^{(m)}]\|^2] \leq \sigma^2$ and $\mathbb{E}[\|\mathbb{E}[\mu_t^{(m)}] - \Phi(x_t^{(m)}, y_t^{(m)})\|^2] \leq G^2$ for $m \in [M]$ and $t \in [T]$, where $\mu_t^{(m)}$ is the stochastic hyper-gradient denoted in Line 10 of Algorithm 2.

The assumption is reasonable due to Proposition 1, and we can choose $\sigma = G_1$ and $G = G_2$.

5.2 CONVERGENCE ANALYSIS FOR FEDBIO AND FEDBIOACC

In this subsection, we provide the convergence result for our FedBiO algorithm 1 and the FedBiOAcc algorithm 2. Firstly, for FedBiO, we have the following Theorem:

Theorem 5.1. *Suppose Assumption 1- 3, 5 hold,* $\delta < min\left(\frac{\sqrt{(1-q)(1-q_1 q^I)}}{2L_h^2 I\sqrt{\bar{q}_1 q_1 q^I}}, \frac{1}{12L_h I}, \frac{\mu\gamma}{2}, 1\right)$, $\gamma < \frac{1}{L}$ *and* $\eta = \frac{\delta}{T^{1/3}}$*, we have:*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla h(\bar{x}_t)\|^2 \leq \frac{2(h(\bar{x}_t) - h^*)}{\delta T^{2/3}} + \frac{L_h^2 B_{\bar{t}_0}}{(1-q)T} + \frac{2L_h^2 B_{\bar{t}_0}}{(1-q)(1-q_1 q^I)T} + \frac{M^{'}\delta^2}{T^{2/3}}$$

where $B_{\bar{t}_0} = \frac{1}{M}\sum_{m=1}^{M}\|y_1^{(m)} - y_{x_1^{(m)}}^{(m)}\|^2$, $q = (1 - \frac{\mu\gamma}{2})$, $q_1 = 1 + \frac{\mu\gamma}{4}$ *and* $\bar{q}_1 = 1 + \frac{4}{\mu\gamma}$, h^* *the optimal value,* $M^{'}$ *is some constant.*

Remark 5. We omit the exact form of some constants in Theorem 5.1 and the full version can be found in Theorem 9.1 in the Appendix. As shown by the Theorem, our FedBiO converges with the $O(\epsilon^{-1.5})$ number of steps (T). This is worse than the optimal rate $O(\epsilon^{-1})$ for non-distributed deterministic bilevel optimization. The consensus error is the source of this gap; we have to decrease the learning rate to bound this consensus error. Meanwhile, the communication complexity is $O(\epsilon^{-1})$.

Next we provide the convergence result for the FedBiOAcc algorithm. To prove the convergence of FedBiOAcc, we denote the potential function $\mathcal{G}_t$ as follows:

$$\mathcal{G}_t = h(\bar{x}_t) + \frac{\eta}{160L_h^2\alpha_t}\left\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\right\|^2 + \frac{1}{M}\sum_{m=1}^{M}\left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2$$
$$+ \frac{\gamma}{32\mu L^2\alpha_t}\sum_{m=1}^{M}\left\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\right\|^2.$$

Then we have the following result for FedBiOAcc:

Table 1: Performance comparison between FedBiO, FedBiOAcc and baselines

	Distribution	I.I.D			Non-I.I.D		
	Metrics	**Test Acc.**	**Train EqOpp.**	**Test EqOpp.**	**Test Acc.**	**Train EqOpp.**	**Test EqOpp.**
Adult	FedAvg	.8239±.0167	.0391± .0061	.0420±.0034	.8283±.0080	.0261±.0022	.0507±.0011
	FedReg	.8240±.0159	.0361±.0047	.0425±.0029	.8271±.0077	**.0244±.0013**	.0498±.0010
	FedMinMax	.8228±.0163	.0220 ±.0057	.0366±.0049	.8272±.0077	.0274±.0029	.0363±.0013
	FCFL	.8238±.0159	.0356±.0029	.0452±.0012	.8273±.0074	.0249±.0032	.0501±.0014
	FedBiO	.8228±.0163	.0238±.0058	.0337±.0012	**.8331±.0019**	.0263±.0010	**.0338±.0008**
	FedBiOAcc	**.8391±.0163**	**.0222±.0064**	**.0335±.0006**	.8204±.0013	.0289±.0005	.0356±.0055
Credit	FedAvg	.6873±.0314	.0788±.0136	.0599±.0122	**.7386±.0011**	.0832±.0248	.1354±.0128
	FedReg	.6870±.0374	.0836±.0015	.0575±.0114	.7303±.0097	.0735±.0216	.1341±.0088
	FedMinMax	.6759±.0757	.0857±.0042	.0722±.0013	.6966±.0104	**.0477±.0155**	.1222±.0024
	FCFL	.6864±.0237	.0727±.0073	**.0375±.0028**	.7266±.0026	.0777±.0162	.1463±.0014
	FedBiO	.7015±.0169	**.0548±.0072**	.0513±.0059	.7339±.0033	.0782±.0116	.1260±.0013
	FedBiOAcc	**.7067±.0121**	.0665±.0034	.0501±.0051	.7312±.0023	.0799±.0152	**.1021±.0011**

Theorem 5.2. *Suppose Assumption 1- 4, 5, 6 hold and the hyper-parameter c_ν, c_ω, η, γ, δ and u are chosen according to Theorem 10.1 in the Appendix. and the learning rate α_t is chosen as in Algorithm 2, then we have:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla h(\bar{x}_t)\|^2\right] \leq M'\left(\frac{16L_hI}{T} + \frac{L_h}{(MT)^{2/3}}\right)$$

where M' is some constant up to a logarithmic factor and the expectation is w.r.t. the stochasticity of the algorithm.

Remark 6. The full version of Theorem 5.2 is shown in Theorem 10.1 in the Appendix. Recall that T is the total number of running steps, I is the number of local steps before communication, and M is the number of clients. Suppose we choose $I = T^{1/3}M^{-2/3}$, the above bound is $O((MT)^{-2/3})$, thus we require $O(M^{-1}\epsilon^{-1.5})$ (up to a logarithm factor) number of running steps to reach an ϵ-stationary point. Meanwhile, we have a dependence of $O(M^{-1})$, so we achieve the linear speedup *w.r.t* the number of clients. Finally, we also get the number of communication steps, *i.e.* T/I is O(ϵ^{-1}). In summary, our FedBiOAcc matches the best known convergence rate for federated stochastic single level optimization Khanduri et al. (2021).

6 Fair Federated Bilevel Learning

In this section, we apply FedBiO and FedBiOAcc to solve the Fair Federated Learning tasks. The code of all experiments is written in Pytorch, and the Federated Learning environment is simulated via Pytorch.Distributed Package. We use servers with AMD EPYC 7763 64-Core CPU.

6.1 Group Fair Federated Learning

In this task, we investigate the group fairness in Federated Learning from the Bilevel Optimization's perspective. The basic idea is to exploit a small validation set that are group-balanced to learn a fair federated model. More specifically, we first assign a weight for each sensitive group and learn a group-weighted federated model. Then we test the performance of the learned model with our group-balanced validation set, based on the validation performance, we adjust the group weights. We repeat this process until we find optimal group weights such that the learned model performs equally well for all different groups in the validation set. This task can be formulated as a Federated bilevel problem of the form Eq. 7, an exact formulation of the Fair Federated Bilevel Learning is included in Section 11 of the appendix. Note our fair federated learning formulation is general and is compatible with various group fairness metrics such as Equal Opportunity (EqOpp) Hardt et al. (2016) and Equalized Odds (EqOdds) Hardt et al. (2016). Furthermore, it also does not require access to the global statistics of the groups, which is difficult to acquire in the Federated Learning setting.

Since the fair federated learning model has a bilevel formulation, we solve it with our FedBiO and FedBiOAcc algorithms. We compare with the following baselines: FedAvg McMahan et al. (2017),

FedReg, FedMinMax Papadaki et al. (2021) and FCFL Cui et al. (2021). The methods proposed in (Zhang et al., 2021) are similar to FedMinMax; we do not include it in the results. Furthermore, since our focus is group fairness, we do not include client fairness (robustness) focused models such as AFL Mohri et al. (2019) and q-FedAvg Li et al. (2019a). The FedReg baseline is to add a regularization term over the FedAvg objective, and the regularization term could be any fairness metrics such as EqOpp. Note that FedReg evaluates the metric only with local statistics.

We tested on real-world benchmark datasets: Credit Asuncion & Newman (2007) and Adult Kohavi et al. (1996). We pre-process the datasets with code provided by (Diana et al., 2021). For each dataset, we first split it into train and test splits with ratio 7:3, and we keep the group distribution the same for the train and test splits. Then for the train set, we consider both I.I.D and Non-I.I.D cases. For the I.I.D case, we uniformly randomly split the train-set into three subsets and distribute each subset to a client. For the Non-I.I.D case, we split the train-set by sensitive attributes and for each attribute, we split its data into three shares with ratio $2:2:6$ and then randomly distribute each share to one client. Finally, for our FedBiO and FedBiOAcc, we select a small subset of the local train set to create the group-fair validation set. We fit a logistic regression model over the benchmark datasets. For our methods, we perform a two-stage training procedure: we first estimate optimal group weights with the bilevel formulation, then we use the learned weight to fit a weighted logistic regression model with FedAvg. For FedReg and FCFL, we choose its regularization term as the EqOpp metric. The definition of EqOpp metric is included in the Appendix 11. Finally, we perform grid search for the hyper-parameters of all methods and hyper-parameter choices are introduced in the Appendix 11.

We summarize the results in Table 1, where we use Test accuracy and EqOpp as metrics, we run 10 runs for each case and report the mean and standard deviation in the table. The best result for each metric is highlighted. As shown in the table, either FedBiO or FedBiOAcc gets the best result for most cases. FedReg/FCFL are based on local group statistics to achieve fairness, and tend to perform worse in the Non-I.I.D case, *e.g.* for the Adult dataset, FCFL gets a much lower Train EqOpp in the Non-I.I.D case compared to the I.I.D one, but its Test EqOpp is worse. FedMinMax is a strong baseline and can get good performance in both settings. However, our algorithms have two advantages compared to FedMinMax. Firstly, we do not query global statistics; furthermore, our algorithms communicate every I iterations, while FedMinMax collects the model states from clients at every iteration.

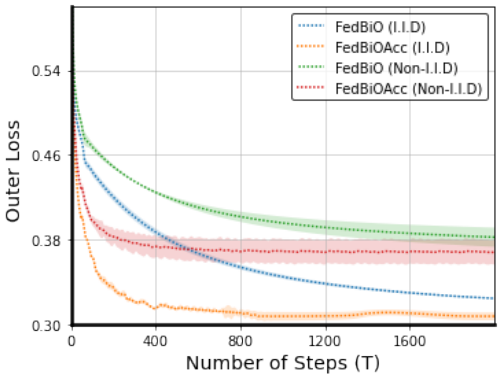

Figure 1: Outer objective Loss *w.r.t* Number of Communication Rounds for comparison between FedBiO and FedBiOAcc for I.I.D and Non-I.I.D cases. The results are for the Adult dataset.

Finally, we compare the convergence rate of the FedBiO and FedBiOAcc algorithms. The results for the Adult dataset are shown in Figure 1 (The results for Credit dataset is deferred to Appendix 11). As shown by the plots, FedBiOAcc converges much faster than FedBiO for both I.I.D and Non-I.I.D cases; furthermore, we observe that data heterogeneity slows down the convergence.

7 Conclusion

In this paper, we studied a class of novel Federated Bilevel Optimization problems and proposed two efficient algorithms, *i.e.*, FedBiO and FedBiOAcc, to solve these problems. In addition, we provided a rigorous convergence analysis framework for our proposed methods. Specifically, we proved that our FedBiO has iteration/communication complexity $O(\epsilon^{-1.5})$ and FedBiOAcc has iteration complexity $O(\epsilon^{-1.5})$ and communication complexity $O(\epsilon^{-1})$, meanwhile FedBiOAcc achieves linear speedup *w.r.t* the number of clients. Finally, we apply our new algorithms to solve the important Fair Federated Learning problem with using a new bilevel optimization formulation. The experimental results validate the efficacy of our algorithms.

REFERENCES

Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html`.

Ahmed Khaled Ragab Bayoumi, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529, 2020.

Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8): 1889–1900, 2000.

Dingding Chen and Martin T Hagan. Optimal use of regularization and cross-validation in neural network modeling. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, volume 2, pp. 1275–1280. IEEE, 1999.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. 2021.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pp. 15236–15245, 2019.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33, 2020.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 66–76, 2021.

Chuong B Do, Chuan-Sheng Foo, and Andrew Y Ng. Efficient multiple hyperparameter learning for log-linear models. In *NIPS*, volume 2007, pp. 377–384. Citeseer, 2007.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pp. 318–326. PMLR, 2012.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 181–189. SIAM, 2021.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *arXiv preprint arXiv:2110.00857*, 2021.

Michael C Ferris and Olvi L Mangasarian. Finite perturbation of convex programs. *Applied Mathematics and Optimization*, 23(1):263–273, 1991.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1165–1173. JMLR. org, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimilano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. *arXiv preprint arXiv:1806.04910*, 2018.

Chen Gao, Yunpeng Chen, Si Liu, Zhenxiong Tan, and Shuicheng Yan. Adversarialnas: Adversarial neural architecture search for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5680–5689, 2020.

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

Zhishuai Guo, Quanqi Hu, Lijun Zhang, and Tianbao Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

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

Feihu Huang and Heng Huang. Enhanced bilevel optimization via bregman distance. *arXiv preprint arXiv:2107.12301*, 2021.

Kaiyi Ji and Yingbin Liang. Lower bounds and accelerated algorithms for bilevel optimization. *arXiv preprint arXiv:2102.03926*, 2021.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Provably faster algorithms for bilevel optimization and applications to meta-learning. *arXiv preprint arXiv:2010.07962*, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

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

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *arXiv preprint arXiv:2102.07367*, 2021.

Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Jan Larsen, Lars Kai Hansen, Claus Svarer, and M Ohlsson. Design and regularization of neural networks: the optimal use of a validation set. In *Neural Networks for Signal Processing VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop*, pp. 62–71. IEEE, 1996.

Junyi Li, Bin Gu, and Heng Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. *arXiv preprint arXiv:2112.04660*, 2021a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019a.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019b.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Jonathan Lorraine and David Duvenaud. Stochastic hyperparameter optimization through hypernetworks. *arXiv preprint arXiv:1802.09419*, 2018.

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

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

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

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. *arXiv preprint arXiv:1902.00146*, 2019.

Takayuki Okuno, Akiko Takeda, and Akihiro Kawana. Hyperparameter learning via bilevel nonsmooth optimization. *arXiv preprint arXiv:1806.01520*, 2018.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Federating for learning group fair models. *arXiv preprint arXiv:2110.01999*, 2021.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. *arXiv preprint arXiv:1602.02355*, 2016.

Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3, 2018.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. *arXiv preprint arXiv:1810.10667*, 2018.

Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3516–3525, 2020.

Mikhail Solodov. An explicit descent method for bilevel convex optimization. *Journal of Convex Analysis*, 14(2):227, 2007.

Xingyou Song, Wenbo Gao, Yuxiang Yang, Krzysztof Choromanski, Aldo Pacchiano, and Yunhao Tang. Es-maml: Simple hessian-free meta learning. *arXiv preprint arXiv:1910.01215*, 2019.

Yuesong Tian, Li Shen, Guinan Su, Zhifeng Li, and Wei Liu. Alphagan: Fully differentiable architecture search for generative adversarial networks. *arXiv preprint arXiv:2006.09134*, 2020.

Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. *arXiv preprint arXiv:1906.00429*, 2019.

Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019a.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019b.

Ralph A Willoughby. Solutions of ill-posed problems (an tikhonov and vy arsenin). *SIAM Review*, 21 (2):266, 1979.

Catherine Wong, Neil Houlsby, Yifeng Lu, and Andrea Gesmundo. Transfer learning with neural automl. *arXiv preprint arXiv:1803.02780*, 2018.

Blake Woodworth. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021.

Pengwei Xing, Songtao Lu, Lingfei Wu, and Han Yu. Big-fed: Bilevel optimization enhanced graph-aided federated learning.

Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong. Pc-darts: Partial channel connections for memory-efficient architecture search. *arXiv preprint arXiv:1907.05737*, 2019.

Isao Yamada, Masahiro Yukawa, and Masao Yamagishi. Minimizing the moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 345–390. Springer, 2011.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *arXiv preprint arXiv:2106.04692*, 2021.

Zhuoran Yang, Zuyue Fu, Kaiqing Zhang, and Zhaoran Wang. Convergent reinforcement learning with function approximation: A bilevel optimization perspective. 2018.

Haiyan Yin, Dingcheng Li, Xu Li, and Ping Li. Meta-cotgan: A meta cooperative training paradigm for improving adversarial text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9466–9473, 2020.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5693–5700, 2019.

Fengda Zhang, Kun Kuang, Yuxuan Liu, Chao Wu, Fei Wu, Jiaxun Lu, Yunfeng Shao, and Jun Xiao. Unified group fairness on federated learning. *arXiv preprint arXiv:2111.04986*, 2021.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pp. 7693–7702. PMLR, 2019.

NOTATIONS AND ASSUMPTIONS

Before we start the proof, we first define some notations. We define $\bar{t}_s \coloneqq sI + 1$ with $s \in [S]$. Note at $\bar{t}_s$ iteration, we have $x_t^{(m)} = \bar{x}_t$ for $m \in [M]$.

Next, we restate more detailed assumptions needed in our proof:

Assumption 0.1. Function $f(x, y) \coloneqq \mathbb{E}[f(x, y; \mathcal{B}_f)]$ is possibly non-convex and $g(x, y) \coloneqq \mathbb{E}[g(x, y; \mathcal{B}_g)]$ is μ-strongly convex *w.r.t* y for any given x, *i.e.* for any $y_1, y_2 \in \mathbb{R}^d$, we have:

$$g(x, y_1) \geq g(x, y_2) + \langle \nabla g_y(x, y_2), y_2 - y_1 \rangle + \frac{\mu}{2}||y_2 - y_1||^2.$$

Assumption 0.2. Function $f(x, y)$ is L-Lipschitz, *i.e.* for for any $x_1, x_2 \in \mathcal{X}$ and for any y_1, $y_2 \in \mathbb{R}^d$, and we denote $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, then we have:

$$f(z_1) \leq f(z_2) + \langle \nabla f(z_2), z_1 - z_2 \rangle + \frac{L}{2}||z_1 - z_2||^2.$$

or equivalently: $||\nabla f(z_1) - \nabla f(z_2)|| \leq L||z_1 - z_2||$. We also assume and $f(x, y)$ has B-bounded gradient, *i.e.* for for any $x \in \mathcal{X}$ and any $y \in \mathbb{R}^d$, and we denote $z = (x, y)$, then we have $||\nabla f(z)|| \leq M$.

Assumption 0.3. Function $g(x, y)$ is L-Lipschitz. *i.e.* for for any x_1, $x_2 \in \mathcal{X}$ and for any y_1, $y_2 \in \mathbb{R}^d$, and we denote $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, then we have:

$$g(z_1) \leq g(z_2) + \langle \nabla g(z_2), z_1 - z_2 \rangle + \frac{L}{2}||z_1 - z_2||^2.$$

equivalently: $||\nabla g(z_1) - \nabla g(z_2)|| \leq L||z_1 - z_2||$. For higher-order derivatives, we have:

a) $\|\nabla^2_{xy} g(x, y)\| \leq C_{g,xy}$ for some constant $C_{g,xy}$;

b) $\nabla^2_{xy} g(x, y)$ and $\nabla^2_{y^2} g(x, y)$ are Lipschitz continuous with constant $L_{g,xy}$ and L_{g,y^2} respectively, *i.e.* for for any x_1, $x_2 \in \mathcal{X}$ and for any y_1, $y_2 \in \mathbb{R}^d$, and we denote $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, then we have: $||\nabla^2_{xy} g(z_1) - \nabla^2_{xy} g(z_2)|| \leq L_{g,xy}||z_1 - z_2||$ and $||\nabla^2_{y^2} g(z_1) - \nabla^2_{y^2} g(z_2)|| \leq L_{g,y^2}||z_1 - z_2||$.

Assumption 0.4. We have unbiased stochastic first order and second order derivative oracle with bounded variance, more specifically, we have:

a) we have $\nabla_x f(x, y; \xi)$, such that: $E[\nabla_x f(x, y; \xi)] = \nabla_x f(x, y)$ and $var(\nabla_x f(x, y; \xi)) \leq \sigma^2$.

a) we have $\nabla_y f(x, y; \xi)$, such that: $E[\nabla_y f(x, y; \xi)] = \nabla_y f(x, y)$ and $var(\nabla_y f(x, y; \xi)) \leq \sigma^2$.

c) we have $\nabla^2_{y^2} g(x, y, \xi)$, such that: $E[\nabla^2_{y^2} g(x, y; \xi)] = \nabla^2_{y^2} g(x, y)$, $E[\nabla^2_{y^2} g(x, y; \xi)v] = \nabla^2_{y^2} g(x, y)v$ and $var(\nabla^2_{y^2} g(x, y; \xi)) \leq \sigma^2$ and $var(\nabla^2_{y^2} g(x, y; \xi)v) \leq \sigma^2$ for any vector v.

d) we have $\nabla^2_{xy} g(x, y; \xi)$, such that: $E[\nabla^2_{xy} g(x, y; \xi)] = \nabla^2_{xy} g(x, y)$, $E[\nabla^2_{xy} g(x, y; \xi)v] = \nabla^2_{xy} g(x, y)v$ and $var(\nabla^2_{xy} g(x, y; \xi)) \leq \sigma^2$ and $var(\nabla^2_{xy} g(x, y; \xi)v) \leq \sigma^2$ for any vector v.

Assumption 0.5. For any $m, j \in [M]$ and x, we have:$\|\nabla_x f^{(m)}(x, y) - \nabla_x f^{(j)} f(x, y)\| \leq \zeta_f$, $\|\nabla_y f^{(m)}(x, y) - \nabla_y f^{(j)} f(x, y)\| \leq \zeta_f$, $\|\nabla_{xy} g^{(m)}(x, y) - \nabla_{xy} g^{(j)}(x, y)\| \leq \zeta_{g,xy}$, $\|\nabla_{y^2} g^{(m)}(x, y) - \nabla_{y^2} g^{(j)}(x, y)\| \leq \zeta_{g,yy}$, $\|y_x^{(m)} - y_x^{(j)}\| \leq \zeta_{g^*}$, where ζ_f, $\zeta_{g,xy}$, $\zeta_{g,yy}$, ζ_{g^*} are constants.

Assumption 0.6. The bias and variance of the stochastic hyper-gradient is bounded, *i.e.* $\mathbb{E}[\|\mu_t^{(m)} - \mathbb{E}[\mu_t^{(m)}]\|^2] \leq \sigma^2$ and $\mathbb{E}[\|\mathbb{E}[\mu_t^{(m)}] - \Phi(x_t^{(m)}, y_t^{(m)})\|^2] \leq G^2$ for $m \in [M]$ and $t \in [T]$, where $\mu_t^{(m)}$ is the stochastic hyper-gradient denoted in Line 10 of Algorithm 2.

8 PRELIMINARIES

In this section, we state some propositions useful in the proof:

Proposition 3 (Lemma 3 of Karimireddy et al. (2020))**.** *(generalized triangle inequality) Let $\{x_k\}, k \in K$ be K vectors. Then the following are true:*

1. $||x_i + x_j||^2 \leq (1 + a)||x_i||^2 + (1 + \frac{1}{a})||x_j||^2$ *for any $a > 0$, and*
2. $||\sum_{k=1}^K x_k||^2 \leq K \sum_{k=1}^K ||x_k||^2$

Proposition 4 (Lemma C.1 of Khanduri et al. (2021))**.** *For a finite sequence $x^{(k)} \in \mathbb{R}^d$ for $k \in [K]$ define $\bar{x} := \frac{1}{K} \sum_{k=1}^K x^{(k)}$, we then have $\sum_{k=1}^K \|x^{(k)} - \bar{x}\|^2 \leq \sum_{k=1}^K \|x^{(k)}\|^2$.*

Proposition 5 (Lemma C.2 of Khanduri et al. (2021))**.** *Let $a_0 > 0$ and $a_1, a_2, \ldots, a_T \geq 0$. We have*

$$\sum_{t=1}^T \frac{a_t}{a_0 + \sum_{i=t}^t a_i} \leq \ln\left(1 + \frac{\sum_{i=1}^t a_i}{a_0}\right).$$

Proposition 6. *(Proposition 2) Suppose Assumptions 2 and 3 hold, the following statements hold:*

a) $\|\Phi(x;y) - \nabla h(x)\| \leq C\|y_x - y\|$, *where* $C = L + LC_{g,xy}/\mu + B(L_{g,xy}/\mu + L_{g,y^2}C_{g,xy}/\mu^2)$.

b) y_x *is Lipschitz continuous in* x *with constant* $\rho = C_{g,xy}/\mu$.

c) $h(x)$ *is Lipschitz continuous in* x *with constant* $\bar{L}$ *i.e., for any given* $x_1, x_2 \in X$, *we have* $\|\nabla h(x_2) - \nabla h(x_1)\| \leq \bar{L}\|x_2 - x_1\|$ *where* $\bar{L} = (L + C)C_{g,xy}/\mu + L + B(L_{g,xy}B/\mu + L_{g,y^2}C_{g,xy}/\mu^2)$.

d) $\|\Phi(x_1;y_1) - \Phi(x_2;y_2)\|^2 \leq \Gamma^2(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2)$, *where* $\Gamma = L + BL_{g,xy}/\mu + C_{g,xy}(L/\mu + BL_{g,yy}/\mu^2)$.

We denote $L_h = max(\bar{L}, \Gamma, C, \rho, 1)$ *for convenience.*

Next if Case d) holds, it is straightforward to also get the stochastic version, *i.e.* $\|\Phi(x_1;y_1;\mathcal{B}) - \Phi(x_2;y_2,\mathcal{B})\|^2 \leq \Gamma^2(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2)$.

Proof. We only prove the Case d) here. Proof of other cases can be found in Lemma 2.2 of (Ghadimi & Wang, 2018).

$$
\begin{aligned}
&\|\Phi(x_1;y_1) - \Phi(x_2;y_2)\| \\
&= \Big\|\nabla_x f(x_1,y_1) - \nabla_{xy} g(x_1,y_1)\Big(\nabla_{yy} g(x_1,y_1)\Big)^{-1}\nabla_y f(x_1,y_1) \\
&\qquad - \nabla_x f(x_2,y_2) - \nabla_{xy} g(x_2,y_2)\Big(\nabla_{yy} g(x_2,y_2)\Big)^{-1}\nabla_y f(x_2,y_2)\Big\| \\
&\leq \Big\|\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\Big\| + \Big\|\nabla_{xy} g(x_1,y_1) \\
&\quad - \nabla_{xy} g(x_2,y_2)\Big\|\Big\|\Big(\nabla_{yy} g(x_2,y_2)\Big)^{-1}\nabla_y f(x_1,y_1)\Big\| \\
&\quad + \Big\|\nabla_{xy} g(x_2,y_2)\Big\|\Big\|\Big(\nabla_{yy} g(x_1,y_1)\Big)^{-1}\nabla_y f(x_1,y_1) - \Big(\nabla_{yy} g(x_2,y_2)\Big)^{-1}\nabla_y f(x_2,y_2)\Big\| \\
&\leq \left(L + \frac{BL_{g,xy}}{\mu} + C_{g.xy}\left(\frac{L}{\mu} + \frac{BL_{g,yy}}{\mu^2}\right)\right)\left(\Big\|x_1 - x_2\Big\|^2 + \Big\|y_1 - y_2\Big\|^2\right)^{1/2}
\end{aligned}
$$

which finishes the proof. □

Proposition 7. *With Assumption 1, 2, 3 and Assumption 5 hold, we have:*

$$
\begin{aligned}
\|\nabla h^{(m)}(x) - \nabla h^{(j)}(x)\| \leq &\left(1 + \frac{C_{g,xy}}{\mu}\right)\zeta_{f,x} + \frac{B}{\mu}\zeta_{g,xy} + \frac{BC_{g,xy}}{\mu^2}\zeta_{g,yy} \\
&+ \left(L + \frac{BL_{g,xy}}{\mu} + \frac{C_{g,xy}L}{\mu} + \frac{BC_{g,xy}L_{g,y^2}}{\mu^2}\right)\zeta_{g^*}
\end{aligned}
$$

Proof. Follow the formulation shown in Eq. 3 ($\Phi(x, y_x)$), we have:

$$
\begin{aligned}
\|\nabla h^{(m)}(x) - \nabla h^{(j)}(x)\| = &\|\nabla_x f^{(m)}(x, y_x^{(m)}) - \nabla^2_{xy} g^{(m)}(x, y_x^{(m)})[\nabla^2_{y^2} g^{(m)}(x, y_x^{(m)})]^{-1}\nabla_y f^{(m)}(x, y_x^{(m)}) \\
&- \Big(\nabla_x f^{(j)}(x, y_x^{(j)}) - \nabla^2_{xy} g^{(j)}(x, y_x^{(j)})[\nabla^2_{y^2} g^{(j)}(x, y_x^{(j)})]^{-1}\nabla_y f^{(j)}(x, y_x^{(j)})\Big)\| \\
\leq &\Big\|\nabla_x f^{(m)}(x, y_x^{(m)}) - \nabla_x f^{(j)}(x, y_x^{(j)})\Big\| + \Big\|\nabla_{xy} g^{(m)}(x, y_x^{(m)}) \\
&- \nabla_{xy} g^{(j)}(x, y_x^{(j)})\Big\|\Big\|\Big(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\Big)^{-1}\nabla_y f^{(m)}(x, y_x^{(m)})\Big\| \\
&+ \Big\|\nabla_{xy} g^{(j)}(x, y_x^{(j)})\Big\|\Big\|\Big(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\Big)^{-1}\nabla_y f^{(m)}(x, y_x^{(m)}) \\
&- \Big(\nabla_{yy} g^{(j)}(x, y_x^{(j)})\Big)^{-1}\nabla_y f^{(j)}(x, y_x^{(j)})\Big\|
\end{aligned}
$$

where the inequality is due to the triangle inequality. Next we bound the three terms separately. For the first term:

$$\begin{aligned}\left\|\nabla_x f^{(m)}(x, y_x^{(m)}) - \nabla_x f^{(j)}(x, y_x^{(j)})\right\| &\leq \left\|\nabla_x f^{(m)}(x, y_x^{(m)}) - \nabla_x f^{(j)}(x, y_x^{(m)})\right\| \\ &\quad + \left\|\nabla_x f^{(j)}(x, y_x^{(m)}) - \nabla_x f^{(j)}(x, y_x^{(j)})\right\| \\ &\leq \zeta_f + L\left\|y_x^{(m)} - y_x^{(j)}\right\| \leq \zeta_f + L\zeta_{g^*} \end{aligned} \tag{8}$$

where the second inequality is due to the Assumption 5 and smoothness assumption the Assumption 2. The last inequality also follows the Assumption 5. Next, for the second term, we have:

$$\begin{aligned}&\left\|\nabla_{xy} g^{(m)}(x, y_x^{(m)}) - \nabla_{xy} g^{(j)}(x, y_x^{(j)})\right\| \left\|\left(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\right)^{-1} \nabla_y f^{(m)}(x, y_x^{(m)})\right\| \\ &\leq \frac{B}{\mu}\left\|\nabla_{xy} g^{(m)}(x, y_x^{(m)}) - \nabla_{xy} g^{(j)}(x, y_x^{(j)})\right\| \\ &\leq \frac{B}{\mu}\left\|\nabla_{xy} g^{(m)}(x, y_x^{(m)}) - \nabla_{xy} g^{(j)}(x, y_x^{(m)})\right\| + \frac{B}{\mu}\left\|\nabla_{xy} g^{(j)}(x, y_x^{(m)}) - \nabla_{xy} g^{(j)}(x, y_x^{(j)})\right\| \\ &\leq \frac{B\zeta_{g,xy}}{\mu} + \frac{BL_{g,xy}}{\mu}\left\|y_x^{(m)} - y_x^{(j)})\right\| \leq \frac{B\zeta_{g,xy}}{\mu} + \frac{BL_{g,xy}\zeta_{g^*}}{\mu}\end{aligned}$$

where the first inequality follows from the Assumption 1, 2; the second inequality follows from triangle inequality; the third inequality follows from Assumption 5, 3, the last inequality follows from Assumption 5. Next, for the third term, we have:

$$\begin{aligned}&\left\|\nabla_{xy} g^{(j)}(x, y_x^{(j)})\right\| \left\|\left(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\right)^{-1} \nabla_y f^{(m)}(x, y_x^{(m)}) - \left(\nabla_{yy} g^{(j)}(x, y_x^{(j)})\right)^{-1} \nabla_y f^{(j)}(x, y_x^{(j)})\right\| \\ &\leq C_{g,xy}\left\|\left(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\right)^{-1} \nabla_y f^{(m)}(x, y_x^{(m)}) - \left(\nabla_{yy} g^{(j)}(x, y_x^{(j)})\right)^{-1} \nabla_y f^{(j)}(x, y_x^{(j)})\right\| \\ &\leq C_{g,xy}\left\|\left(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\right)^{-1}\right\| \left\|\nabla_y f^{(m)}(x, y_x^{(m)}) - \nabla_y f^{(j)}(x, y_x^{(j)})\right\| \\ &\quad + C_{g,xy}\left\|\left(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\right)^{-1} - \left(\nabla_{yy} g^{(j)}(x, y_x^{(j)})\right)^{-1}\right\| \left\|\nabla_y f^{(j)}(x, y_x^{(j)})\right\| \\ &\leq \frac{C_{g,xy}}{\mu}\left\|\nabla_y f^{(m)}(x, y_x^{(m)}) - \nabla_y f^{(j)}(x, y_x^{(j)})\right\| \\ &\quad + MC_{g,xy}\left\|\left(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\right)^{-1} - \left(\nabla_{yy} g^{(j)}(x, y_x^{(j)})\right)^{-1}\right\| \\ &\leq \frac{C_{g,xy}(\zeta_f + L\zeta_{g^*})}{\mu} + BC_{g,xy}\left\|\left(\nabla_{yy} g^{(m)}(x, y_x^{(m)})\right)^{-1}\right\| \times \\ &\qquad \left\|\nabla_{yy} g^{(m)}(x, y_x^{(m)}) - \nabla_{yy} g^{(j)}(x, y_x^{(j)})\right\| \left\|\left(\nabla_{yy} g^{(j)}(x, y_x^{(j)})\right)^{-1}\right\| \\ &\leq \frac{C_{g,xy}(\zeta_f + L\zeta_{g^*})}{\mu} + \frac{BC_{g,xy}(\zeta_{g,yy} + L_{g,y^2}\zeta_{g^*})}{\mu^2}\end{aligned}$$

where the first inequality is by Assumption 3; the second inequality is by triangle inequality; the third inequality is by Assumption 3, 2; the fourth inequality is by Cauchy Schwartz inequality; the last inequality is by Assumption 1, 3 and the result in Eq. 8. Combine everything together, we have:

$$\begin{aligned}\left\|\nabla_x f^{(m)}(x, y_x^{(m)}) - \nabla_x f^{(j)}(x, y_x^{(j)})\right\| &\leq \zeta_{f,x} + L\zeta_{g^*} + \frac{B\zeta_{g,xy}}{\mu} + \frac{BL_{g,xy}\zeta_{g^*}}{\mu} + \frac{C_{g,xy}(\zeta_{f,x} + L\zeta_{g^*})}{\mu} \\ &\quad + \frac{BC_{g,xy}(\zeta_{g,yy} + L_{g,y^2}\zeta_{g^*})}{\mu^2}\end{aligned}$$

which completes the proof. □

9 PROOF FOR THE FEDBIO ALGORITHM

In this section, we present the proofs for the FedBiO algorithm, we will focus on the deterministic case.

9.1 HYPER-GRADIENT BIAS

Lemma 1. *For all $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$, the iterates generated satisfy:*

$$\left\|\nabla h(\bar{x}_t) - \bar{\nu}_t\right\|^2 \leq \frac{L_h^2}{M}\sum_{m=1}^{M}\left(\left(1+2L_h^2\right)\left\|x_t^{(m)} - \bar{x}_t\right\|^2 + \left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2\right)$$

where $\bar{\nu}_t = \frac{1}{M}\sum_{m=1}^{M}\Phi^{(m)}(x_t^{(m)}, y_t^{(m)})$, and $\nabla h(\bar{x}_t) = \frac{1}{M}\sum_{m=1}^{M}\Phi^{(m)}(\bar{x}_t, y_{\bar{x}_t}^{(m)})$.

Proof. By definition of $\bar{\nu}_t$ and $\nabla h(\bar{x}_t)$, we have:

$$\begin{aligned}
\left\|\nabla h(\bar{x}_t) - \bar{\nu}_t\right\|^2 &= \left\|\frac{1}{M}\sum_{m=1}^{M}\left(\nu_t^{(m)} - \nabla h^{(m)}(\bar{x}_t)\right)\right\|^2 \overset{(a)}{\leq} \frac{1}{M}\sum_{m=1}^{M}\left\|\nu_t^{(m)} - \nabla h^{(m)}(\bar{x}_t)\right\|^2 \\
&\overset{(b)}{\leq} \frac{L_h^2}{M}\sum_{m=1}^{M}\left(\left\|x_t^{(m)} - \bar{x}_t\right\|^2 + \left\|y_t^{(m)} - y_{\bar{x}_t}^{(m)}\right\|^2\right) \\
&\leq \frac{L_h^2}{M}\sum_{m=1}^{M}\left(\left\|x_t^{(m)} - \bar{x}_t\right\|^2 + \left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)} + y_{x_t^{(m)}}^{(m)} - y_{\bar{x}_t}^{(m)}\right\|^2\right) \\
&\leq \frac{L_h^2}{M}\sum_{m=1}^{M}\left(\left(1+2L_h^2\right)\left\|x_t^{(m)} - \bar{x}_t\right\|^2 + \left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2\right)
\end{aligned}$$

□

where inequality (a) follows the generalized triangle inequality; inequality (b) follows the Proposition 2.

9.2 INNER VARIABLE DRIFT LEMMA

Lemma 2. *When $\gamma < \frac{1}{L}$, we have:*

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{M}\sum_{m=1}^{M}\left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2 \leq \frac{B_{\bar{t}_0}}{1-q} + \frac{B_{\bar{t}_0}}{(1-q)(1-q_1q^I)} + \frac{L_h^2M_h^2q_1\bar{q}(S-1)}{(1-q)^2(1-q_1q^I)}\eta^2 \\
+ \frac{\bar{q}_1q_1q^IL_h^2}{(1-q)(1-q_1q^I)}\sum_{j=1}^{S-1}\hat{A}_{\bar{t}_j} + \frac{L_h^2M_h^2\bar{q}T}{1-q}\eta^2
\end{aligned}$$

where $B_t = \frac{1}{M}\sum_{m=1}^{M}\left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2$, $\hat{A}_t = \frac{1}{M}\sum_{m=1}^{M}\left\|\hat{x}_t^{(m)} - \bar{x}_t\right\|^2$, $q = (1-\frac{\mu\gamma}{2})$, $\bar{q} = (1+\frac{2}{\mu\gamma})$, $q_1 = 1+\frac{\mu\gamma}{4}$ and $\bar{q}_1 = 1+\frac{4}{\mu\gamma}$, $M_h = \frac{B(\mu+C_{g,xy})}{\mu}$.

Proof. Note from Algorithm and the definition of $\bar{t}_s$ that at $t = \bar{t}_{s-1}$ with $s \in [S]$, $x_t^{(m)} = \bar{x}_t$, for all k. For $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$, with $s \in [S]$, we have:

$$\begin{aligned}
\left\| y_t^{(m)} - y_{x_t^{(m)}}^{(m)} \right\|^2 &\leq (1 + \frac{\mu\gamma}{2}) \left\| y_t^{(m)} - y_{x_{t-1}^{(m)}}^{(m)} \right\|^2 + (1 + \frac{2}{\mu\gamma}) \left\| y_{x_t^{(m)}}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)} \right\|^2 \\
&\leq (1 + \frac{\mu\gamma}{2})(1 - \mu\gamma) \left\| y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)} \right\|^2 + (1 + \frac{2}{\mu\gamma}) \left\| y_{x_t^{(m)}}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)} \right\|^2 \\
&\leq (1 - \frac{\mu\gamma}{2}) \left\| y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)} \right\|^2 + L_h^2 (1 + \frac{2}{\mu\gamma}) \left\| x_t^{(m)} - x_{t-1}^{(m)} \right\|^2 \\
&\leq (1 - \frac{\mu\gamma}{2}) \left\| y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)} \right\|^2 + L_h^2 \eta^2 (1 + \frac{2}{\mu\gamma}) \left\| \nu_{t-1}^{(m)} \right\|^2 \\
&\leq (1 - \frac{\mu\gamma}{2}) \left\| y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)} \right\|^2 + L_h^2 M_h^2 \eta^2 (1 + \frac{2}{\mu\gamma})
\end{aligned}$$

where the second inequality is due to the property of gradient descent for strongly convex function when $\gamma < 1/L$. For the last inequality, we use the fact that: $\|\nu_{t-1}^{(m)}\| = \|\Phi^{(m)}(x_t^{(m)}, y_t^{(m)})\| \leq B + BC_{g,xy}/\mu$ and we denote $M_h = B + BC_{g,xy}/\mu$. We also denote $q = (1 - \frac{\mu\gamma}{2})$ and $\bar{q} = (1 + \frac{2}{\mu\gamma})$ for ease of notation. By telescoping two sides, we have:

$$\begin{aligned}
\left\| y_t^{(m)} - y_{x_t^{(m)}}^{(m)} \right\|^2 &\leq q^{t - \bar{t}_{s-1}} \left\| y_{\bar{t}_{s-1}}^{(m)} - y_{x_{\bar{t}_{s-1}}^{(m)}}^{(m)} \right\|^2 + L_h^2 M_h^2 \eta^2 \bar{q} \sum_{l=\bar{t}_{s-1}}^{t-1} q^{t-l-1} \\
&\overset{(b)}{\leq} q^{t - \bar{t}_{s-1}} \left\| y_{\bar{t}_{s-1}}^{(m)} - y_{x_{\bar{t}_{s-1}}^{(m)}}^{(m)} \right\|^2 + \frac{L_h^2 M_h^2 \bar{q}}{1 - q} \eta^2
\end{aligned}$$

where in inequality (b), we use the fact $q^{t - \bar{t}_{s-1}} < 1$ for any t. Then we average over all M clients and have:

$$\frac{1}{M} \sum_{j=1}^{M} \left\| y_t^{(m)} - y_{x_t^{(m)}}^{(m)} \right\|^2 \leq \frac{1}{M} \sum_{m=1}^{M} q^{t - \bar{t}_{s-1}} \left\| y_{\bar{t}_{s-1}}^{(m)} - y_{x_{\bar{t}_{s-1}}^{(m)}}^{(m)} \right\|^2 + \frac{L_h^2 M_h^2 \bar{q}}{1 - q} \eta^2 \tag{9}$$

As for $t = \bar{t}_s$, we average variable x over the m clients and $x_{\bar{t}_s}^{(m)} = \bar{x}_{\bar{t}_s}$, while the inner variable error is related to the x variable before averaging, *i.e.* $\hat{x}_{\bar{t}_s}^{(m)}$. By the generalized triangle inequality, we have:

$$\begin{aligned}
\left\| y_{\bar{t}_s}^{(m)} - y_{\bar{x}_{\bar{t}_s}}^{(m)} \right\|^2 &\leq (1 + \frac{\mu\gamma}{4}) \left\| y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)} \right\|^2 + (1 + \frac{4}{\mu\gamma}) \left\| y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)} - y_{\bar{x}_{\bar{t}_s}}^{(m)} \right\|^2 \\
&\leq (1 + \frac{\mu\gamma}{4}) \left\| y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)} \right\|^2 + L_h^2 (1 + \frac{4}{\mu\gamma}) \left\| \hat{x}_{\bar{t}_s}^{(m)} - \bar{x}_{\bar{t}_s} \right\|^2
\end{aligned}$$

We denote $q_1 = 1 + \frac{\mu\gamma}{4}$ and $\bar{q}_1 = 1 + \frac{4}{\mu\gamma}$. By averaging over M clients, we have:

$$\frac{1}{M} \sum_{m=1}^{M} \left\| y_{\bar{t}_s}^{(m)} - y_{x_{\bar{t}_s}^{(m)}}^{(m)} \right\|^2 \leq \frac{q_1}{M} \sum_{m=1}^{M} \left\| y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)} \right\|^2 + \frac{L_h^2 \bar{q}_1}{M} \sum_{m=1}^{M} \left\| \hat{x}_{\bar{t}_s}^{(m)} - \bar{x}_{\bar{t}_s} \right\|^2$$

We can bound the first term with Eq. (9) by setting $t = \bar{t}_s$, and we have:

$$\frac{1}{M} \sum_{m=1}^{M} \left\| y_{\bar{t}_s}^{(m)} - y_{x_{\bar{t}_s}^{(m)}}^{(m)} \right\|^2 \leq \frac{q_1 q^I}{M} \sum_{m=1}^{M} \left\| y_{\bar{t}_{s-1}}^{(m)} - y_{x_{\bar{t}_{s-1}}^{(m)}}^{(m)} \right\|^2 + \frac{L_h^2 \bar{q}_1}{M} \sum_{m=1}^{M} \left\| \hat{x}_{\bar{t}_s}^{(m)} - \bar{x}_{\bar{t}_s} \right\|^2 + \frac{L_h^2 M_h^2 q_1 \bar{q}}{1 - q} \eta^2$$

For ease of notation, we denote $B_t = \frac{1}{M} \sum_{m=1}^{M} \|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2$ and $\hat{A}_t = \frac{1}{M} \sum_{m=1}^{M} \|\hat{x}_t^{(m)} - \bar{x}_t\|^2$. Then the above equations can be written as:

$$B_t \leq q^{t - \bar{t}_{s-1}} B_{\bar{t}_{s-1}} + \frac{L_h^2 M_h^2 \bar{q}}{1 - q} \eta^2, t \in [\bar{t}_{s-1} + 1, \bar{t}_s - 1] \tag{10}$$

and:

$$B_{\bar{t}_s} \leq q_1 q^I B_{\bar{t}_{s-1}} + \bar{q}_1 L_h^2 \hat{A}_{\bar{t}_s} + \frac{L_h^2 M_h^2 q_1 \bar{q}}{1-q}\eta^2, t = \bar{t}_s$$

By telescoping, for $s \geq 1$ we have:

$$\begin{aligned} B_{\bar{t}_s} &\leq q_1^s q^{sI} B_{\bar{t}_0} + \frac{L_h^2 M_h^2 q_1 \bar{q}}{1-q}\eta^2 \sum_{j=0}^{s-1} q_1^j q^{jI} + \bar{q}_1 L_h^2 \sum_{j=1}^{s} q_1^{s-j} q^{(s-j)I} \hat{A}_{\bar{t}_j} \\ &\leq q_1^s q^{sI} B_{\bar{t}_0} + \frac{L_h^2 M_h^2 q_1 \bar{q}}{(1-q)(1-q_1 q^I)}\eta^2 + \bar{q}_1 L_h^2 \sum_{j=1}^{s} q_1^{(s-j)} q^{(s-j)I} \hat{A}_{\bar{t}_j} \end{aligned} \tag{11}$$

Then by summing Eq. (10) from $\bar{t}_{s-1}$ to t, we have:

$$\sum_{l=\bar{t}_{s-1}}^{t} B_l \leq \sum_{l=\bar{t}_{s-1}}^{t} q^{l-\bar{t}_{s-1}} B_{\bar{t}_{s-1}} + \frac{L_h^2 M_h^2 \bar{q}(t-\bar{t}_{s-1}-1)}{1-q}\eta^2 \leq \frac{B_{\bar{t}_{s-1}}}{1-q} + \frac{L_h^2 M_h^2 \bar{q}(t-\bar{t}_{s-1}-1)}{1-q}\eta^2$$

Combine the above inequality with Eq. (11) and for $S \geq 2$, we have:

$$\begin{aligned} \sum_{t'=\bar{t}_{s-1}}^{t} B_{t'} \leq{}& \frac{q_1^{s-1} q^{(s-1)I} B_{\bar{t}_0}}{1-q} + \frac{L_h^2 M_h^2 q_1 \bar{q}}{(1-q)^2(1-q_1 q^I)}\eta^2 \\ &+ \frac{\bar{q}_1 L_h^2}{1-q} \sum_{j=1}^{s-1} q_1^{s-1-j} q^{(s-1-j)I} \hat{A}_{\bar{t}_j} + \frac{L_h^2 M_h^2 \bar{q}(t-\bar{t}_{s-1}-1)}{1-q}\eta^2. \end{aligned} \tag{12}$$

and for $s = 1$, we have:

$$\sum_{t'=\bar{t}_0}^{\bar{t}_1-1} B_{t'} \leq \frac{B_{\bar{t}_0}}{1-q} + \frac{L_h^2 M_h^2 \bar{q}(t-\bar{t}_{s-1}-1)}{1-q}\eta^2 \tag{13}$$

Finally, we sum t from $1 \to T$ and have:

$$\begin{aligned} \sum_{t=1}^{T} B_t \leq{}& \frac{B_{\bar{t}_0}}{1-q} + \sum_{s=2}^{S} \frac{q_1^{s-1} q^{(s-1)I} B_{\bar{t}_0}}{1-q} + \frac{L_h^2 M_h^2 q_1 \bar{q}(S-1)}{(1-q)^2(1-q_1 q^I)}\eta^2 \\ &+ \frac{\bar{q}_1 L_h^2}{1-q} \sum_{s=2}^{S} \sum_{j=1}^{s-1} q_1^{s-1-j} q^{(s-1-j)I} \hat{A}_{\bar{t}_j} + \frac{L_h^2 M_h^2 \bar{q} S(I-1)}{1-q}\eta^2 \\ \overset{(a)}{\leq}{}& \frac{B_{\bar{t}_0}}{1-q} + \sum_{s=2}^{S} \frac{q_1^{s-1} q^{(s-1)I} B_{\bar{t}_0}}{1-q} + \frac{L_h^2 M_h^2 q_1 \bar{q}(S-1)}{(1-q)^2(1-q_1 q^I)}\eta^2 \\ &+ \frac{\bar{q}_1 L_h^2}{1-q} \sum_{j=1}^{S-1} \sum_{s=1}^{S-j} q_1^s q^{sI} \hat{A}_{\bar{t}_j} + \frac{L_h^2 M_h^2 \bar{q} S(I-1)}{1-q}\eta^2 \\ \leq{}& \frac{B_{\bar{t}_0}}{1-q} + \frac{B_{\bar{t}_0}}{(1-q)(1-q_1 q^I)} + \frac{L_h^2 M_h^2 q_1 \bar{q}(S-1)}{(1-q)^2(1-q_1 q^I)}\eta^2 \\ &+ \frac{\bar{q}_1 q_1 q^I L_h^2}{(1-q)(1-q_1 q^I)} \sum_{j=1}^{S-1} \hat{A}_{\bar{t}_j} + \frac{L_h^2 M_h^2 \bar{q} T}{1-q}\eta^2. \end{aligned}$$

where inequality (a) rearranges the terms in the fourth sum term. This completes the proof. □

9.3 Bound for Client Drift

Lemma 3. *For* $\eta < min\left(\frac{\sqrt{(1-q)(1-q_1 q^I)}}{2L_h^2 I\sqrt{\bar{q}_1 q_1 q^I}}, \frac{1}{12L_h I}, \frac{\mu\gamma}{2}, 1\right)$ *and* $\gamma < \frac{1}{L}$*, then we have:*

$$\sum_{t=1}^{T} \hat{A}_t \leq \frac{6SL_h^2 I^2 B_{\bar{t}_0}}{1-q}\eta^2 + \frac{18(S-1)L_h^4 M_h^2 I^2}{(1-q)^2(1-q_1 q^I)}\eta^2 + \frac{12L_h^4 M_h^2 TI(I-1)}{1-q}\eta^2 + 18TI^2\zeta^2\eta^2$$

where $\hat{A}_t$*,* q*,* q_1*,* $\bar{q}_1$*,* M_h *are defined as in Lemma 2.*

Proof. Note from Algorithm and the definition of $\bar{t}_s$ that at $t = \bar{t}_{s-1}$ with $s \in [S]$, $x_t^{(m)} = \bar{x}_t$, for all k. For $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$, with $s \in [S]$, we have: $\hat{x}_t^{(m)} = \hat{x}_{t-1}^{(m)} - \eta\nu_{t-1}^{(m)}$, this implies that: $\hat{x}_t^{(m)} = x_{\bar{t}_{s-1}}^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\nu_\ell^{(m)}$ and $\bar{x}_t = \bar{x}_{\bar{t}_{s-1}} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\bar{\nu}_\ell$. So for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$, with $s \in [S]$ we have:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \|\hat{x}_t^{(m)} - \bar{x}_t\|^2 &= \frac{1}{M}\sum_{m=1}^{M} \Big\| x_{\bar{t}_{s-1}}^{(m)} - \bar{x}_{\bar{t}_{s-1}} - \Big(\sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\nu_\ell^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\bar{\nu}_\ell \Big) \Big\|^2 \\
&\overset{(a)}{=} \frac{1}{M}\sum_{m=1}^{M} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big) \Big\|^2 \\
&\overset{(b)}{\leq} \frac{2}{M}\sum_{m=1}^{M} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\Big(\nu_\ell^{(m)} - \nabla h^{(m)}(x_\ell^{(m)})\Big) - \frac{1}{M}\sum_{j=1}^{M}\Big(\nu_\ell^{(j)} - \nabla h^{(j)}(x_\ell^{(j)})\Big) \Big\|^2 \\
&\qquad + \frac{2}{M}\sum_{m=1}^{M} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\Big(\nabla h^{(m)}(x_\ell^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_\ell^{(j)})\Big) \Big\|^2 \\
&\overset{(c)}{\leq} \frac{2}{M}\sum_{m=1}^{M} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\Big(\nu_\ell^{(m)} - \nabla h^{(m)}(x_\ell^{(m)})\Big) \Big\|^2 \\
&\qquad + \frac{2}{M}\sum_{m=1}^{M} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\Big(\nabla h^{(m)}(x_\ell^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_\ell^{(j)})\Big) \Big\|^2.
\end{aligned} \tag{14}$$

where the equality (a) follows from the fact that $x_{\bar{t}_{s-1}}^{(m)} = \bar{x}_{\bar{t}_{s-1}}$ for $t = \bar{t}_{s-1}$; (b) uses triangle inequality and (c) follows from the application of Proposition 4. Then for the first term of 14, we have:

$$\begin{aligned}
\Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\Big(\nu_\ell^{(m)} - \nabla h^{(m)}(x_\ell^{(m)})\Big) \Big\|^2 &\leq I\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big\| \Big(\nu_\ell^{(m)} - \nabla h^{(m)}(x_\ell^{(m)})\Big) \Big\|^2 \\
&\overset{(a)}{\leq} L_h^2 I\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \Big\| y_\ell^{(m)} - y_{x_\ell^{(m)}}^{(m)} \Big\|^2.
\end{aligned} \tag{15}$$

where (a) Follows Proposition 2. Next, for the second term of 14 we have:

$$\begin{aligned}
&\sum_{m=1}^{M} \Big\| \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\Big(\nabla h^{(m)}(x_\ell^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_\ell^{(j)})\Big) \Big\|^2 \\
&\overset{(a)}{\leq} I \sum_{\ell=\bar{t}_{s-1}}^{t-1}\sum_{m=1}^{M} \Big\| \eta\Big(\nabla h^{(m)}(x_\ell^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_\ell^{(j)})\Big) \Big\|^2 \\
&\overset{(b)}{\leq} I \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta^2 \Big[3\sum_{m=1}^{M} \big\|\nabla h^{(m)}(x_\ell^{(m)}) - \nabla h^{(m)}(\bar{x}_\ell)\big\|^2 + 3M \Big\| \nabla h(\bar{x}_\ell) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_\ell^{(j)}) \Big\|^2 \\
&\qquad + 3\sum_{m=1}^{M} \big\|\nabla h^{(m)}(\bar{x}_\ell) - \nabla h(\bar{x}_\ell)\big\|^2 \Big] \\
&\overset{(c)}{\leq} I \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta^2 \Big[6L_h^2 \sum_{m=1}^{M} \big\|x_\ell^{(m)} - \bar{x}_\ell\big\|^2 + 3\sum_{m=1}^{M} \Big\| \nabla h^{(m)}(\bar{x}_\ell) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(\bar{x}_\ell) \Big\|^2 \Big] \\
&\overset{(d)}{\leq} 6L_h^2 I\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1}\sum_{m=1}^{M} \big\|x_\ell^{(m)} - \bar{x}_\ell\big\|^2 + 3MI^2\eta^2\zeta^2.
\end{aligned} \tag{16}$$

where (a) utilizes the fact that $t - \bar{t}_{s-1} \leq I$ for $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$ and the generalized triangle inequality; (b) follows the generalized triangle inequality; (c) follows from the L_h lipschitzness of h; and (d) utilizes the heterogeneity Assumption 5 and also the fact that $t - \bar{t}_{s-1} \leq I$ for $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$.

Substituting 15 and 16 in 14 we get:

$$\frac{1}{M}\sum_{m=1}^{M}\|\hat{x}_t^{(m)} - \bar{x}_t\|^2 \leq \frac{2L_h^2 I\eta^2}{M}\sum_{m=1}^{M}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\left\|y_\ell^{(m)} - y_{x_\ell^{(m)}}^{(m)}\right\|^2 + 12L_h^2 I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\frac{1}{M}\sum_{m=1}^{M}\|x_\ell^{(m)} - \bar{x}_\ell\|^2 + 6I^2\zeta^2\eta^2.$$

Next we use $\hat{A}_t = \frac{1}{M}\sum_{m=1}^{M}\|\hat{x}_t^{(m)} - \bar{x}_t\|^2$ and $B_t = \frac{1}{M}\sum_{m=1}^{M}\left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2$ as in Lemma 2, then the above inequality can be simplified to:

$$\hat{A}_t \leq 2L_h^2 I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1} B_l + 12L_h^2 I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\hat{A}_l + 6I^2\zeta^2\eta^2.$$

For $s \geq 2$, we substitute Eq. (12) to get:

$$\begin{aligned}
\hat{A}_t &\leq 2L_h^2 I\eta^2\sum_{l=\bar{t}_{s-1}}^{t-1} B_l + 12L^2 I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\hat{A}_l + 6I^2\zeta^2\eta^2 \\
&\leq \frac{2L_h^2 I q_1^{s-1}q^{(s-1)I}B_{\bar{t}_0}}{1-q}\eta^2 + \frac{2L_h^4 M_h^2 I q_1\bar{q}}{(1-q)^2(1-q_1q^I)}\eta^4 + 2L_h^4 I\bar{q}_1\eta^2\sum_{j=1}^{s-1}\frac{q_1^{s-1-j}q^{(s-1-j)I}\hat{A}_{\bar{t}_j}}{1-q} \\
&\quad + \frac{2L_h^4 M_h^2 I(I-1)\bar{q}}{1-q}\eta^4 + 6I^2\zeta^2\eta^2 + 12L_h^2 I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\hat{A}_l
\end{aligned}$$

Summing both sides from $t = \bar{t}_{s-1}+1$ to $\bar{t}_s$, we get:

$$\begin{aligned}
\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s}\hat{A}_t &\leq \frac{2L_h^2 I^2 q_1^{s-1}q^{(s-1)I}B_{\bar{t}_0}\eta^2}{1-q} + \frac{2L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1q^I)}\eta^4 + \frac{2L_h^4 M_h^2 I^2\bar{q}(I-1)}{1-q}\eta^4 + 6I^3\zeta^2\eta^2 \\
&\quad + 2L_h^4 I^2\bar{q}_1\eta^2\sum_{j=1}^{s-1}\frac{q_1^{s-1-j}q^{(s-1-j)I}\hat{A}_{\bar{t}_j}}{1-q} + 12L_h^2 I\eta^2\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\hat{A}_l \\
&\leq \frac{2L_h^2 I^2 q_1^{s-1}q^{(s-1)I}B_{\bar{t}_0}\eta^2}{1-q} + \frac{2L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1q^I)}\eta^4 + \frac{2L_h^4 M_h^2 I^2\bar{q}(I-1)}{1-q}\eta^4 + 6I^3\zeta^2\eta^2 \\
&\quad + 2L_h^4 I^2\bar{q}_1\eta^2\sum_{j=1}^{s-1}\frac{q_1^{s-1-j}q^{(s-1-j)I}\hat{A}_{\bar{t}_j}}{1-q} + 12L_h^2 I^2\eta^2\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s}\hat{A}_l
\end{aligned}$$

For ease of notation, we denote $C_s = \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s}\hat{A}_t$, then we have:

$$\begin{aligned}
C_s &\leq \frac{2L_h^2 I^2 q_1^{s-1}q^{(s-1)I}B_{\bar{t}_0}\eta^2}{1-q} + \frac{2L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1q^I)}\eta^4 + \frac{2L_h^4 M_h^2 I^2\bar{q}(I-1)}{1-q}\eta^4 + 6I^3\zeta^2\eta^2 \\
&\quad + 2L_h^4 I^2\bar{q}_1\eta^2\sum_{j=1}^{s-1}\frac{q_1^{s-1-j}q^{(s-1-j)I}\hat{A}_{\bar{t}_j}}{1-q} + 12L_h^2 I^2\eta^2 C_s \\
&\leq \frac{2L_h^2 I^2 B_{\bar{t}_0}\eta^2}{1-q} + \frac{2L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1q^I)}\eta^4 + \frac{2L_h^4 M_h^2 I^2\bar{q}(I-1)}{1-q}\eta^4 + 6I^3\zeta^2\eta^2 \\
&\quad + 2L_h^4 I^2\bar{q}_1\eta^2\sum_{j=1}^{s-1}\frac{q_1^{s-1-j}q^{(s-1-j)I}\hat{A}_{\bar{t}_j}}{1-q} + 12L_h^2 I^2\eta^2 C_s. \qquad (17)
\end{aligned}$$

The second inequality follows that $q_1 q^I = (1+\mu\gamma/4)(1-\mu\gamma/2)^I \leq (1+\mu\gamma/4)(1-\mu\gamma/2) \leq 1-\mu\gamma/4 < 1$.

Next for $s=1$, substitute Eq. 13, we have:

$$\begin{aligned}\hat{A}_t &\leq 2L_h^2 I\eta^2 \sum_{\ell=\bar{t}_0}^{t-1} B_l + 12L_h^2 I\eta^2 \sum_{\ell=\bar{t}_0}^{t-1} \hat{A}_l + 6I^2\zeta^2\eta^2 \\ &\leq \frac{2L_h^2 I B_{\bar{t}_0}}{1-q}\eta^2 + \frac{2L_h^4 M_h^2 I(I-1)\bar{q}}{1-q}\eta^4 + 12L_h^2 I\eta^2 \sum_{\ell=\bar{t}_0}^{t-1} \hat{A}_l + 6I^2\zeta^2\eta^2.\end{aligned}$$

Summing both sides from $t=\bar{t}_0+1$ to $\bar{t}_s$, we get:

$$C_1 = \sum_{t=\bar{t}_0+1}^{\bar{t}_s} \hat{A}_t \leq \frac{2L_h^2 I^2 B_{\bar{t}_0}}{1-q}\eta^2 + \frac{2L_h^4 M_h^2 I^2(I-1)\bar{q}}{1-q}\eta^4 + 6I^3\zeta^2\eta^2 + 12L_h^2 I^2\eta^2 C_1 \tag{18}$$

Then we combine 17 and 18 to have:

$$\begin{aligned}\sum_{s=1}^{S} C_s &\leq \frac{2SL_h^2 I^2 B_{\bar{t}_0}}{1-q}\eta^2 + \frac{2(S-1)L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1 q^I)}\eta^4 + \frac{2SL_h^4 M_h^2 I^2(I-1)\bar{q}}{1-q}\eta^4 + 6SI^3\zeta^2\eta^2 \\ &\quad + 2L_h^4 I^2\bar{q}_1\eta^2 \sum_{s=2}^{S}\sum_{j=1}^{s-1} \frac{(q_1 q^I)^{s-1-j} C_j}{1-q} + 12L_h^2 I^2\eta^2 \sum_{s=1}^{S} C_s \\ &\leq \frac{2SL_h^2 I^2 B_{\bar{t}_0}}{1-q}\eta^2 + \frac{2(S-1)L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1 q^I)}\eta^4 + \frac{2SL_h^4 M_h^2 I^2(I-1)\bar{q}}{1-q}\eta^4 + 6SI^3\zeta^2\eta^2 \\ &\quad + \frac{2L_h^4 I^2\bar{q}_1 q_1 q^I}{(1-q)(1-q_1 q^I)}\eta^2 \sum_{s=1}^{S} C_s + 12L_h^2 I^2\eta^2 \sum_{s=1}^{S} C_s.\end{aligned}$$

where in the first inequality, we use the fact that $\hat{A}_{\bar{t}_j} \leq C_j$, then by rearranging the terms, we have:

$$\begin{aligned}\left(1 - \frac{2L_h^4 I^2\bar{q}_1 q_1 q^I}{(1-q)(1-q_1 q^I)}\eta^2 - 12L_h^2 I^2\eta^2\right)\sum_{t=1}^{T}\hat{A}_t &\leq \frac{2SL_h^2 I^2 B_{\bar{t}_0}}{1-q}\eta^2 + \frac{2(S-1)L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1 q^I)}\eta^4 \\ &\quad + \frac{2SL_h^4 M_h^2 I^2(I-1)\bar{q}}{1-q}\eta^4 + 6SI^3\zeta^2\eta^2\end{aligned}$$

Suppose $\eta < min\left(\frac{\sqrt{(1-q)(1-q_1 q^I)}}{2L_h^2 I\sqrt{\bar{q}_1 q_1 q^I}}, \frac{1}{12L_h I}\right)$, then we have

$$1 - \frac{2\bar{q}_1 L_h^2 I^2\eta^2\rho^2 q_1 q^I}{(1-q)(1-q_1 q^I)} - 12L_h^2 I^2\eta^2 \geq 1 - \frac{1}{2} - \frac{1}{12} > \frac{1}{3}$$

So we have:

$$\sum_{t=1}^{T}\hat{A}_t \leq \frac{6SL_h^2 I^2 B_{\bar{t}_0}}{1-q}\eta^2 + \frac{6(S-1)L_h^4 M_h^2 I^2 q_1\bar{q}}{(1-q)^2(1-q_1 q^I)}\eta^4 + \frac{6SL_h^4 M_h^2 I^2(I-1)\bar{q}}{1-q}\eta^4 + 18SI^3\zeta^2\eta^2$$

Note that we have

$$q_1\bar{q} = (1+\frac{\mu\gamma}{4})(1+\frac{2}{\mu\gamma}) = \frac{3}{2} + \frac{\mu\gamma}{4} + \frac{2}{\mu\gamma} < 2 + \frac{2}{\mu\gamma}$$

and by the assumption that $\eta < min(1, \frac{\mu\gamma}{2})$, we simplify the above inequality as:

$$\sum_{t=1}^{T}\hat{A}_t \leq \frac{6SL_h^2 I^2 B_{\bar{t}_0}}{1-q}\eta^2 + \frac{18(S-1)L_h^4 M_h^2 I^2}{(1-q)^2(1-q_1 q^I)}\eta^2 + \frac{12L_h^4 M_h^2 TI(I-1)}{1-q}\eta^2 + 18TI^2\zeta^2\eta^2$$

Therefore, the lemma is proved. □

9.4 DESCENT LEMMA

Lemma 4. *For all $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ and $s \in [S]$, the iterates generated satisfy:*

$$h(\bar{x}_{t+1}) \leq h(\bar{x}_t) - \frac{\eta}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{L_h^2(1+2L_h^2)\eta}{2M}\sum_{m=1}^{M}\|x_t^{(m)} - \bar{x}_t\|^2 + \frac{L_h^2\eta}{2M}\sum_{m=1}^{M}\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. Using the smoothness of f we have:

$$\begin{aligned}
h(\bar{x}_{t+1}) &\leq h(\bar{x}_t) + \langle \nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{L_h}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&\overset{(a)}{=} h(\bar{x}_t) - \eta\langle \nabla h(\bar{x}_t), \bar{\nu}_t\rangle + \frac{\eta^2 L_h}{2}\|\bar{\nu}_t\|^2 \\
&\overset{(b)}{=} h(\bar{x}_t) - \frac{\eta}{2}\left\|\bar{\nu}_t\right\|^2 - \frac{\eta}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta}{2}\left\|\nabla h(\bar{x}_t) - \bar{\nu}_t\right\|^2 + \frac{\eta_t^2 L_h}{2}\left\|\bar{\nu}_t\right\|^2 \\
&= h(\bar{x}_t) - \left(\frac{\eta}{2} - \frac{\eta^2 L_h}{2}\right)\left\|\bar{\nu}_t\right\|^2 - \frac{\eta}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta}{2}\left\|\nabla h(\bar{x}_t) - \bar{\nu}_t\right\|^2 \\
&\overset{(c)}{\leq} h(\bar{x}_t) - \frac{\eta}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta}{2}\left\|\nabla h(\bar{x}_t) - \bar{\nu}_t\right\|^2 \\
&\overset{(d)}{\leq} h(\bar{x}_t) - \frac{\eta}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{L_h^2\eta}{2M}\sum_{m=1}^{M}\left(\left(1+2L_h^2\right)\left\|x_t^{(m)} - \bar{x}_t\right\|^2 + \left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2\right)
\end{aligned}$$

where equality (a) follows from the iterate update given in Step 6 of Algorithm 1; (b) uses $\langle a, b\rangle = \frac{1}{2}[\|a\|^2 + \|b\|^2 - \|a-b\|^2]$; (c) follows the assumption that $\eta < 1/L_h$; (d) follows lemma 1. Hence, the lemma is proved. □

9.5 PROOF OF CONVERGENCE THEOREM

Theorem 9.1. *For $\delta < min\left(\frac{\sqrt{(1-q)(1-q_1q^I)}}{2L_h^2 I\sqrt{\bar{q}_1 q_1 q^I}}, \frac{1}{12L_h I}, \frac{\mu\gamma}{2}, 1\right)$, $\gamma < \frac{1}{L}$ and $\eta = \frac{\delta}{T^{1/3}}$, we have:*

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla h(\bar{x}_t)\|^2 \leq{}& \frac{2(h(\bar{x}_1) - h^*)}{\delta T^{2/3}} + \frac{L_h^2 B_{\bar{t}_0}}{(1-q)T} + \frac{2L_h^2 B_{\bar{t}_0}}{(1-q)(1-q_1q^I)T} \\
&+ \frac{\delta^2 L_h^4 M_h^2 q_1 \bar{q}}{(1-q)^2(1-q^I)IT^{2/3}} + \frac{\delta^2 L_h^4 \bar{q} M_h^2}{(1-q)T^{2/3}} + \left(\frac{\bar{q}_1 L_h^4 q_1 q^I}{(1-q)(1-q_1q^I)} + L_h^2(1+2L_h^2)\right) \\
&\times\left(\frac{6L_h^2 I B_{\bar{t}_0}}{1-q} + \frac{18L_h^4 M_h^2 I}{(1-q)^2(1-q_1q^I)} + \frac{12L_h^4 M_h^2 I(I-1)}{1-q} + 18I^2\zeta^2\right)\frac{\delta^2}{T^{2/3}}
\end{aligned}$$

where $B_{\bar{t}_0} = \frac{1}{M}\sum_{m=1}^{M}\|y_1^{(m)} - y_{x_1^{(m)}}^{(m)}\|^2$, $q = (1 - \frac{\mu\gamma}{2})$, $\bar{q} = (1 + \frac{2}{\mu\gamma})$, $q_1 = 1 + \frac{\mu\gamma}{4}$ and $\bar{q}_1 = 1 + \frac{4}{\mu\gamma}$

Proof. From the result of Lemma 4, summarize for $t = [T]$ and multiply both sides by $2/\eta T$ we get

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla h(\bar{x}_t)\|^2 &\leq \sum_{t=1}^{T}\frac{2(h(\bar{x}_t)-h(\bar{x}_{t+1}))}{\eta T} + \frac{L_h^2(1+2L_h^2)}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\|x_t^{(m)}-\bar{x}_t\|^2 \\
&\quad + \frac{L_h^2}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 \\
&\leq \frac{2(h(\bar{x}_1)-h^*)}{\eta T} + \frac{L_h^2(1+2L_h^2)}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\|\hat{x}_t^{(m)}-\bar{x}_t\|^2 \\
&\quad + \frac{L_h^2}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 \\
&\leq \frac{2(h(\bar{x}_1)-h^*)}{\eta T} + \frac{L_h^2 B_{\bar{t}_0}}{(1-q)T} + \frac{L_h^2 B_{\bar{t}_0}}{(1-q)(1-q_1q^I)T} + \frac{L_h^4 M_h^2(S-1)q_1\bar{q}}{(1-q)^2(1-q_1q^I)T}\eta^2 \\
&\quad + \frac{L_h^4\bar{q}_1 q_1 q^I}{MT(1-q)(1-q_1q^I)}\sum_{t=1}^{T}\sum_{m=1}^{M}\|\hat{x}_t^{(m)}-\bar{x}_t\|^2 + \frac{L_h^4 M_h^2\bar{q}}{1-q}\eta^2 \\
&\quad + \frac{L_h^2(1+2L_h^2)}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\|\hat{x}_t^{(m)}-\bar{x}_t\|^2 \\
&\leq \frac{2(h(\bar{x}_1)-h^*)}{\eta T} + \frac{L_h^2 B_{\bar{t}_0}}{(1-q)T} + \frac{L_h^2 B_{\bar{t}_0}}{(1-q)(1-q_1q^I)T} + \frac{L_h^4 M_h^2(S-1)q_1\bar{q}}{(1-q)^2(1-q_1q^I)T}\eta^2 \\
&\quad + \frac{L_h^4 M_h^2\bar{q}}{1-q}\eta^2 + \left(\frac{L_h^4\bar{q}_1 q_1 q^I}{(1-q)(1-q_1q^I)} + L_h^2(1+2L_h^2)\right)\times \\
&\quad \left(\frac{6L_h^2 I B_{\bar{t}_0}}{1-q} + \frac{18L_h^4 M_h^2 I}{(1-q)^2(1-q_1q^I)} + \frac{12L_h^4 M_h^2 I(I-1)}{1-q} + 18I^2\zeta^2\right)\eta^2.
\end{aligned}$$

where the second inequality uses $f(\bar{x}_t) \geq f^*$ and the fact $\sum_{m=1}^{M}\|x_t^{(m)}-\bar{x}_t\|^2 \leq \sum_{m=1}^{M}\|\hat{x}_t^{(m)}-\bar{x}_t\|^2$ for all t. The third inequality uses Lemma 2 and the fourth inequality uses 3. Finally, choice of $\eta = \frac{\delta}{T^{1/3}}$, δ is a constant such that $\delta < min\left(\frac{\sqrt{(1-q)(1-q_1q^I)}}{2L_h^2 I\sqrt{\bar{q}_1 q_1 q^I}}, \frac{1}{12L_h I}, \frac{\mu\gamma}{2}, 1\right)$ we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\|\nabla h(\bar{x}_t)\|^2 \\
&\leq \frac{2(h(\bar{x}_1)-h^*)}{\delta T^{2/3}} + \frac{L_h^2 B_{\bar{t}_0}}{(1-q)T} + \frac{2L_h^2 B_{\bar{t}_0}}{(1-q)(1-q_1q^I)T} + \frac{\delta^2 L_h^4 M_h^2 q_1\bar{q}}{(1-q)^2(1-q^I)IT^{2/3}} + \frac{\delta^2 L_h^4 M_h^2\bar{q}}{(1-q)T^{2/3}} \\
&\quad + \left(\frac{\bar{q}_1 L_h^2\rho^2 q_1 q^I}{(1-q)(1-q_1q^I)} + L_h^2(1+2\rho^2)\right)\left(\frac{6L_h^2 I B_{\bar{t}_0}}{1-q} + \frac{18L_h^4 M_h^2 I}{(1-q)^2(1-q_1q^I)}\right. \\
&\quad \left. + \frac{12L_h^4 M_h^2 I(I-1)}{1-q} + 18I^2\zeta^2\right)\frac{\delta^2}{T^{2/3}}
\end{aligned}$$

Therefore, we have the theorem. □

10 PROOF FOR THE FEDBIOACC ALGORITHM

In this section, we prove the convergence of the FedBiOAcc Algorithm.

10.1 BOUND FOR HYPER-GRADIENT BIAS

Lemma 5. *With all assumptions hold and $c_\nu \alpha_t^2 < 1$, then for all $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$, we have:*

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\Big\|^2\Big] &\le (1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\Big[\Big\|\bar{\nu}_{t-1} - \frac{1}{M}\sum_{m=1}^{M}\nabla h^{(m)}(\bar{x}_{t-1})\Big\|^2\Big] \\
&+ \frac{4(c_\nu\alpha_{t-1}^2)^2\sigma^2}{M} + 8(c_\nu\alpha_{t-1}^2)^2G^2 + \frac{8L_h^2(c_\nu\alpha_{t-1}^2)^2}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\Big\|^2\Big] \\
&+ \frac{40L_h^2\eta^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\Big\|^2\Big] + \frac{40L_h^2\eta^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|\bar{\nu}_{t-1}\Big\|^2\Big] \\
&+ \frac{12L_h^2\gamma^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|\omega_{t-1}^{(m)}\Big\|^2\Big]
\end{aligned}$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. We have:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\big(\hat{\nu}_t^{(m)} - \nabla h^{(m)}(x_t^{(m)})\big)\Big\|^2\Big] \\
&\le \mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\Big(\mu_t^{(m)} + (1-c_\nu\alpha_{t-1}^2)(\nu_{t-1}^{(m)} - \mu_{t-1}^{(m)}) - \nabla h^{(m)}(x_t^{(m)})\Big)\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-c_\nu\alpha_{t-1}^2)\Big(\bar{\nu}_{t-1} - \frac{1}{M}\sum_{m=1}^{M}\nabla h^{(m)}(x_{t-1}^{(m)})\Big) \\
&\quad + \frac{1}{M}\sum_{m=1}^{M}\Big(\mu_t^{(m)} - \nabla h^{(m)}(x_t^{(m)}) + (1-c_\nu\alpha_{t-1}^2)(\nabla h^{(m)}(x_{t-1}^{(m)}) - \mu_{t-1}^{(m)})\Big)\Big\|^2\Big] \\
&\overset{(a)}{\le} (1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\Big[\Big\|\bar{\nu}_{t-1} - \frac{1}{M}\sum_{m=1}^{M}\nabla h^{(m)}(\bar{x}_{t-1})\Big\|^2\Big] \\
&\quad + \mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\Big(\mu_t^{(m)} - \nabla h^{(m)}(x_t^{(m)}) + (1-c_\nu\alpha_{t-1}^2)(\nabla h^{(m)}(x_{t-1}^{(m)}) - \mu_{t-1}^{(m)})\Big)\Big\|^2\Big]
\end{aligned} \tag{19}$$

where inequality (a) uses the fact that the cross product is zero in expectation; Next for the second term of the above equation. Now suppose we denote $\tilde{\mu}_t^{(m)} = \mathbb{E}[\mu_t^{(m)}]$, then by the triangle inequality,

we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\left(\mu_t^{(m)}-\nabla h^{(m)}(x_t^{(m)})+(1-c_\nu\alpha_{t-1}^2)(\nabla h^{(m)}(x_{t-1}^{(m)})-\mu_{t-1}^{(m)})\right)\right\|^2\right]\\
&\le 2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\left(\mu_t^{(m)}-\tilde{\mu}_t^{(m)}+(1-c_\nu\alpha_{t-1}^2)(\tilde{\mu}_{t-1}^{(m)}-\mu_{t-1}^{(m)})\right)\right\|^2\right]\\
&\qquad+2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\left(\tilde{\mu}_t^{(m)}-\nabla h^{(m)}(x_t^{(m)})+(1-c_\nu\alpha_{t-1}^2)(\nabla h^{(m)}(x_{t-1}^{(m)})-\tilde{\mu}_{t-1}^{(m)})\right)\right\|^2\right]\\
&\le \frac{2}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\left(\mu_t^{(m)}-\tilde{\mu}_t^{(m)}+(1-c_\nu\alpha_{t-1}^2)(\tilde{\mu}_{t-1}^{(m)}-\mu_{t-1}^{(m)})\right)\right\|^2\right]\\
&\qquad+\frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\left(\tilde{\mu}_t^{(m)}-\nabla h^{(m)}(x_t^{(m)})+(1-c_\nu\alpha_{t-1}^2)(\nabla h^{(m)}(x_{t-1}^{(m)})-\tilde{\mu}_{t-1}^{(m)})\right)\right\|^2\right]
\end{aligned}
\tag{20}
$$

The last inequality is by the generalized triangle inequality for the second term, the first term uses the fact that the cross product is zero in expectation. We bound the two terms in the above inequality separately. For the first term, we have:

$$
\begin{aligned}
&2\mathbb{E}\left[\left\|\mu_t^{(m)}-\tilde{\mu}_t^{(m)}+(1-c_\nu\alpha_{t-1}^2)(\tilde{\mu}_{t-1}^{(m)}-\mu_{t-1}^{(m)})\right\|^2\right]\\
&\overset{(a)}{\le} 4(c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\mu_t^{(m)}-\tilde{\mu}_t^{(m)}\right\|^2\right]+4(1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\mu_t^{(m)}-\mu_{t-1}^{(m)}-\tilde{\mu}_t^{(m)}+\tilde{\mu}_{t-1}^{(m)}\right\|^2\right]\\
&\overset{(b)}{\le} 4(c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\mu_t^{(m)}-\tilde{\mu}_t^{(m)}\right\|^2\right]+4(1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\mu_t^{(m)}-\mu_{t-1}^{(m)}\right\|^2\right]\\
&\overset{(c)}{\le} 4(c_\nu\alpha_{t-1}^2)^2\sigma^2+4L_h^2(1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|x_t^{(m)}-x_{t-1}^{(m)}\right\|^2+\left\|y_t^{(m)}-y_{t-1}^{(m)}\right\|^2\right]\\
&\le 4(c_\nu\alpha_{t-1}^2)^2\sigma^2+4L_h^2(1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\eta\alpha_{t-1}\nu_{t-1}^{(m)}\right\|^2+\left\|\gamma\alpha_{t-1}\omega_{t-1}^{(m)}\right\|^2\right]\\
&\overset{(d)}{\le} 4(c_\nu\alpha_{t-1}^2)^2\sigma^2+4L_h^2\mathbb{E}\left[\left\|\eta\alpha_{t-1}\nu_{t-1}^{(m)}\right\|^2+\left\|\gamma\alpha_{t-1}\omega_{t-1}^{(m)}\right\|^2\right]
\end{aligned}
$$

where inequality (a) follows the triangle inequality Proposition 3; (b) follows Propostion 4 due to the definition of $\tilde{\mu}_t^{(m)}$; (c) follows the smoothness property of L_h and the bounded variance assumption 6; (d) follows the fact that $c_\nu\alpha_t^2<1$. Next for the second term, we have:

$$
\begin{aligned}
&2\mathbb{E}\left[\left\|\tilde{\mu}_t^{(m)}-\nabla h^{(m)}(x_t^{(m)})+(1-c_\nu\alpha_{t-1}^2)(\nabla h^{(m)}(x_{t-1}^{(m)})-\tilde{\mu}_{t-1}^{(m)})\right\|^2\right]\\
&\overset{(a)}{\le} 4(c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\tilde{\mu}_{t-1}^{(m)}-\nabla h^{(m)}(x_{t-1}^{(m)})\right\|\right]+8\mathbb{E}\left[\left\|\tilde{\mu}_t^{(m)}-\tilde{\mu}_{t-1}^{(m)}\right\|^2\right]+8\mathbb{E}\left[\left\|\nabla h^{(m)}(x_t^{(m)})-\nabla h^{(m)}(x_{t-1}^{(m)})\right\|^2\right]\\
&\overset{(b)}{\le} 8(c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\tilde{\mu}_{t-1}^{(m)}-\Phi^{(m)}(x_{t-1}^{(m)},y_{t-1}^{(m)})\right\|\right]+8(c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\Phi^{(m)}(x_{t-1}^{(m)},y_{t-1}^{(m)})-\nabla h^{(m)}(x_{t-1}^{(m)})\right\|\right]\\
&\qquad+8\mathbb{E}\left[\left\|\tilde{\mu}_t^{(m)}-\tilde{\mu}_{t-1}^{(m)}\right\|^2\right]+8\mathbb{E}\left[\left\|\nabla h^{(m)}(x_t^{(m)})-\nabla h^{(m)}(x_{t-1}^{(m)})\right\|^2\right]\\
&\overset{(c)}{\le} 8(c_\nu\alpha_{t-1}^2)^2G^2+8L_h^2(c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left(\left\|y_{t-1}^{(m)}-y_{x_{t-1}^{(m)}}^{(m)}\right\|^2\right)\right]\\
&\qquad+8L_h^2\mathbb{E}\left[\left\|\eta\alpha_{t-1}\nu_{t-1}^{(m)}\right\|^2+\left\|\gamma\alpha_{t-1}\omega_{t-1}^{(m)}\right\|^2\right]+8L_h^2\mathbb{E}\left[\left\|\eta\alpha_{t-1}\nu_{t-1}^{(m)}\right\|^2\right]
\end{aligned}
$$

where inequality (a) and (b) follows the generalized triangle inequality; (c) follows the smoothness of $h(x)$ and the bounded bias assumption 6. Combine everything together, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\right\|^2\right] \\
&\leq (1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\bar{\nu}_{t-1} - \frac{1}{M}\sum_{m=1}^{M}\nabla h^{(m)}(\bar{x}_{t-1})\right\|^2\right] + \frac{4(c_\nu\alpha_{t-1}^2)^2\sigma^2}{M} + 8(c_\nu\alpha_{t-1}^2)^2G^2 \\
&\quad + \frac{8L_h^2(c_\nu\alpha_{t-1}^2)^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\right\|^2\right] + \frac{20L_h^2\eta^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\nu_{t-1}^{(m)}\right\|^2\right] \\
&\quad + \frac{12L_h^2\gamma^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\omega_{t-1}^{(m)}\right\|^2\right] \\
&\overset{(a)}{\leq} (1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\left[\left\|\bar{\nu}_{t-1} - \frac{1}{M}\sum_{m=1}^{M}\nabla h^{(m)}(\bar{x}_{t-1})\right\|^2\right] + \frac{4(c_\nu\alpha_{t-1}^2)^2\sigma^2}{M} + 8(c_\nu\alpha_{t-1}^2)^2G^2 \\
&\quad + \frac{8L_h^2(c_\nu\alpha_{t-1}^2)^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\right\|^2\right] \\
&\quad + \frac{40L_h^2\eta^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\right\|^2\right] + \frac{40L_h^2\eta^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\bar{\nu}_{t-1}\right\|^2\right] \\
&\quad + \frac{12L_h^2\gamma^2\alpha_{t-1}^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\omega_{t-1}^{(m)}\right\|^2\right]
\end{aligned}
$$

In inequality (a) we use the generalized triangle inequality 3. This completes the proof. □

10.2 Bound for Inner Variable Drift

Lemma 6. *Suppose $c_\omega\alpha_{t-1}^2 < 1$, then for $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$, with $s \in [S]$, we have:*

$$
\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\right\|^2\right] \\
&\quad \leq (1-c_\omega\alpha_{t-1}^2)^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\right\|^2\right] + 2(c_\omega\alpha_{t-1}^2)^2\sigma^2 \\
&\quad + 2L^2\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[2\eta^2\alpha_{t-1}^2\left(\left\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\right\|^2 + \left\|\bar{\nu}_{t-1}\right\|^2\right) + \gamma^2\alpha_{t-1}^2\left\|\omega_{t-1}^{(m)}\right\|^2\right]
\end{aligned}
$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. For $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$, with $s \in [S]$, we follow similar derivation as in Eq. (19) and get:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) + (1 - c_\omega \alpha_{t-1}^2)(\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y)) - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\right\|^2\right] \\
&= \mathbb{E}\Bigg[\Bigg\|(1 - c_\omega \alpha_{t-1}^2)(\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})) + \nabla g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \nabla g^{(m)}(x_t^{(m)}, y_t^{(m)}) \\
&\qquad + (1 - c_\omega \alpha_{t-1}^2)(\nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y))\Bigg\|^2\Bigg] \\
&\overset{(a)}{\leq} (1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\left[\left\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\right\|^2\right] \\
&\quad + \mathbb{E}\Bigg[\Bigg\|\nabla g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \nabla g^{(m)}(x_t^{(m)}, y_t^{(m)}) \\
&\quad + (1 - c_\omega \alpha_{t-1}^2)(\nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y))\Bigg\|^2\Bigg] \\
&\overset{(b)}{\leq} (1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\left[\left\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\right\|^2\right] \\
&\quad + 2(c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\left[\left\|\nabla g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \nabla g^{(m)}(x_t^{(m)}, y_t^{(m)})\right\|^2\right] \\
&\quad + 2(1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\Bigg[\Bigg\| - \nabla g^{(m)}(x_t^{(m)}, y_t^{(m)}) \\
&\quad + \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) + \nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y)\Bigg\|^2\Bigg] \\
&\overset{(c)}{\leq} (1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\left[\left\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\right\|^2\right] \\
&\quad + 2(c_\omega \alpha_{t-1}^2)^2 \sigma^2 + 2(1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\left[\left\|\nabla g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y)\right\|^2\right] \\
&\overset{(d)}{\leq} (1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\left[\left\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\right\|^2\right] \\
&\quad + 2(c_\omega \alpha_{t-1}^2)^2 \sigma^2 + 2(1 - c_\omega \alpha_{t-1}^2)^2 L^2 \mathbb{E}\left[\left\|\eta \alpha_{t-1} \nu_{t-1}^{(m)}\right\|^2 + \left\|\gamma \alpha_{t-1} \omega_{t-1}^{(m)}\right\|^2\right] \\
&\overset{(e)}{\leq} (1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\left[\left\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\right\|^2\right] + 2(c_\omega \alpha_{t-1}^2)^2 \sigma^2 \\
&\quad + 2L^2 \mathbb{E}\left[2\eta^2 \alpha_{t-1}^2 \left(\left\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\right\|^2 + \left\|\bar{\nu}_{t-1}\right\|^2\right) + \gamma^2 \alpha_{t-1}^2 \left\|\omega_{t-1}^{(m)}\right\|^2\right] \qquad (21)
\end{aligned}
$$

where inequality (a) uses the fact that the cross product term is zero in expectation; inequality (b) uses the generalized triangle inequality; inequality (c) follows the bounded variance assumption 4 and Proposition 4; inequality (d) uses the smoothness assumption 3; inequality (e) uses the generalized triangle inequality and the fact $(1 - c_\omega \alpha_{t-1}^2)^2 < 1$.

When $t = \bar{t}_s$, the only difference is that we use $\bar{x}_{\bar{t}_s-1}$ in Line 8 of the algorithm 2 to evaluate $\omega^{(m)}_{\bar{t}_s}$ instead of $x^{(m)}_{\bar{t}_s-1}$ when $t < \bar{t}_s$. We follow similar derivation as in Eq 21 and get:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\omega^{(m)}_{\bar{t}_s} - \nabla_y g^{(m)}(x^{(m)}_{\bar{t}_s}, y^{(m)}_{\bar{t}_s})\right\|^2\right] \\
&\leq (1 - c_\omega \alpha^2_{\bar{t}_s-1})^2 \mathbb{E}\left[\left\|\omega^{(m)}_{\bar{t}_s-1} - \nabla_y g^{(m)}(x^{(m)}_{\bar{t}_s-1}, y^{(m)}_{\bar{t}_s-1})\right\|^2\right] \\
&\quad + 2(c_\omega \alpha^2_{\bar{t}_s-1})^2\sigma^2 + 2(1 - c_\omega \alpha^2_{\bar{t}_s-1})^2 L^2 \mathbb{E}\left[\left\|x^{(m)}_{\bar{t}_s} - \bar{x}_{\bar{t}_s-1}\right\|^2 + \left\|\gamma\alpha_{\bar{t}_s-1}\omega^{(m)}_{\bar{t}_s-1}\right\|^2\right] \\
&\leq (1 - c_\omega \alpha^2_{\bar{t}_s-1})^2 \mathbb{E}\left[\left\|\omega^{(m)}_{\bar{t}_s-1} - \nabla_y g^{(m)}(x^{(m)}_{\bar{t}_s-1}, y^{(m)}_{\bar{t}_s-1})\right\|^2\right] \\
&\quad + 2(c_\omega \alpha^2_{\bar{t}_s-1})^2\sigma^2 + 2(1 - c_\omega \alpha^2_{\bar{t}_s-1})^2 L^2 \mathbb{E}\left[\left\|\bar{x}_{\bar{t}_s} - \bar{x}_{\bar{t}_s-1}\right\|^2 + \left\|\gamma\alpha_{\bar{t}_s-1}\omega^{(m)}_{\bar{t}_s-1}\right\|^2\right] \\
&\leq (1 - c_\omega \alpha^2_{\bar{t}_s-1})^2 \mathbb{E}\left[\left\|\omega^{(m)}_{\bar{t}_s-1} - \nabla_y g^{(m)}(x^{(m)}_{\bar{t}_s-1}, y^{(m)}_{\bar{t}_s-1})\right\|^2\right] \\
&\quad + 2(c_\omega \alpha^2_{\bar{t}_s-1})^2\sigma^2 + 2L^2 \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M}\left\|\eta\alpha_{\bar{t}_s-1}\nu^{(j)}_{\bar{t}_s-1}\right\|^2 + \left\|\gamma\alpha_{\bar{t}_s-1}\omega^{(m)}_{\bar{t}_s-1}\right\|^2\right] \qquad (22)
\end{aligned}
$$

The second inequality follows the fact that $x^{(m)}_{\bar{t}_s} = \bar{x}_{\bar{t}_s}$; the last inequality follows the generalized triangle inequality and the fact $(1 - c_\omega \alpha^2_{\bar{t}_s-1})^2 < 1$. Finally, combine Eq. 21 and 22 and average over all M clients finish the proof. □

Lemma 7. *For $\gamma \leq \frac{1}{6L_h}$ and $0 < \alpha_t \leq \frac{1}{16L_h I}$, we have for $t \in [\bar{t}_{s-1} + 1, \bar{t}_s - 1]$:*

$$
\begin{aligned}
\mathbb{E}\left[\left\|y^{(m)}_t - y^{(m)}_{x^{(m)}_t}\right\|^2\right] \leq (1 - \frac{\mu\gamma\alpha_{t-1}}{8})\mathbb{E}\left[\left\|y^{(m)}_{t-1} - y^{(m)}_{x^{(m)}_{t-1}}\right\|^2\right] - \frac{3\gamma^2\alpha_{t-1}}{4}\mathbb{E}\left[\left\|\omega^{(m)}_{t-1}\right\|^2\right] \\
+ \frac{5\gamma\alpha_{t-1}}{\mu}\mathbb{E}\left[\left\|\omega^{(m)}_{t-1} - \nabla_y g^{(m)}(x^{(m)}_{t-1}, y^{(m)}_{t-1})\right\|^2\right] \\
+ \frac{5L^2_h\eta^2\alpha_{t-1}}{\mu\gamma}\mathbb{E}\left[\left\|\nu^{(m)}_{t-1}\right\|^2\right].
\end{aligned}
$$

and when $t = \bar{t}_s$, we have:

$$
\begin{aligned}
\mathbb{E}\left[\left\|y^{(m)}_{\bar{t}_s} - y^{(m)}_{\bar{x}_{\bar{t}_s}}\right\|^2\right] \leq (1 - \frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\mathbb{E}\left[\left\|y^{(m)}_{\bar{t}_s-1} - y^{(m)}_{x^{(m)}_{\bar{t}_s-1}}\right\|^2\right] - \frac{3\gamma^2\alpha_{t-1}}{4}\mathbb{E}\left[\left\|\omega^{(m)}_{\bar{t}_s-1}\right\|^2\right] \\
+ \frac{5\gamma\alpha_{\bar{t}_s-1}}{\mu}\mathbb{E}\left[\left\|\omega^{(m)}_{\bar{t}_s-1} - \nabla_y g^{(m)}(x^{(m)}_{\bar{t}_s-1}, y^{(m)}_{\bar{t}_s-1})\right\|^2\right] \\
+ \frac{5L^2_h\eta^2\alpha_{\bar{t}_s-1}}{\mu\gamma}\mathbb{E}\left[\left\|\nu^{(m)}_{\bar{t}_s-1}\right\|^2\right] + (1 + \frac{8}{\mu\gamma\alpha_{\bar{t}_s-1}})L^2_h\mathbb{E}\left[\left\|\hat{x}^{(m)}_{\bar{t}_s} - \bar{x}_{\bar{t}_s}\right\|^2\right].
\end{aligned}
$$

Proof. Suppose we denote $\tilde{y}^{(m)}_{t+1} = y^{(m)}_t - \gamma\omega^{(m)}_t$, then we have $y^{(m)}_{t+1} = y^{(m)}_t + \alpha(\tilde{y}^{(m)}_{t+1} - y^{(m)}_t)$. We start the proof, firstly, by the strong convexity of of function $g^{(m)}(x, y)$, we have:

$$
\begin{aligned}
g^{(m)}(x^{(m)}_t, y)) &\geq g(x^{(m)}_t, y^{(m)}_t) + \langle\nabla_y g(x^{(m)}_t, y^{(m)}_t), y - y^{(m)}_t\rangle + \frac{\mu}{2}\|y - y^{(m)}_t\|^2 \\
&= g(x^{(m)}_t, y^{(m)}_t) + \langle w^{(m)}_t, y - \tilde{y}^{(m)}_{t+1}\rangle + \langle\nabla_y g(x^{(m)}_t, y^{(m)}_t) - w^{(m)}_t, y - \tilde{y}^{(m)}_{t+1}\rangle \\
&\quad \langle\nabla_y g(x^{(m)}_t, y^{(m)}_t), \tilde{y}^{(m)}_{t+1} - y^{(m)}_t\rangle + \frac{\mu}{2}\|y - y^{(m)}_t\|^2. \qquad (23)
\end{aligned}
$$

According to the smoothness assumption of $g^{(m)}(x,y)$, i.e., the function $g^{(m)}(x,y)$ is L-smooth, we have

$$\frac{L}{2}\|\tilde{y}_{t+1}^{(m)}-y_t^{(m)}\|^2 \geq g(x_t^{(m)},\tilde{y}_{t+1}^{(m)}) - g(x_t^{(m)},y_t^{(m)}) - \langle \nabla_y g(x_t^{(m)},y_t^{(m)}), \tilde{y}_{t+1}^{(m)}-y_t^{(m)}\rangle. \quad (24)$$

Combining the inequalities 23 with 24, we have

$$\begin{aligned} g^{(m)}(x_t^{(m)},y)) \geq g(x_t^{(m)},\tilde{y}_{t+1}^{(m)}) + \langle w_t^{(m)}, y-\tilde{y}_{t+1}^{(m)}\rangle + \langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y-\tilde{y}_{t+1}^{(m)}\rangle \\ + \frac{\mu}{2}\|y-y_t^{(m)}\|^2 - \frac{L}{2}\|\tilde{y}_{t+1}^{(m)}-y_t^{(m)}\|^2. \end{aligned} \quad (25)$$

Then for second term of the above inequality, we have:

$$\begin{aligned} \langle w_t^{(m)}, y-\tilde{y}_{t+1}^{(m)}\rangle &= \langle w_t^{(m)}, y_t^{(m)}-\tilde{y}_{t+1}^{(m)}\rangle + \langle w_t^{(m)}, y-y_t^{(m)}\rangle \\ &= \gamma\|w_t^{(m)}\|^2 + \langle w_t^{(m)}, y-y_t^{(m)}\rangle. \end{aligned} \quad (26)$$

Combining the inequalities 25 with 26, we have

$$\begin{aligned} g^{(m)}(x_t^{(m)},y)) \geq g^{(m)}(x_t^{(m)},\tilde{y}_{t+1}^{(m)}) + \langle w_t^{(m)}, y-y_t^{(m)}\rangle \\ + \langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y-\tilde{y}_{t+1}^{(m)}\rangle \\ + \Big(\gamma - \frac{L\gamma^2}{2}\Big)\|w_t^{(m)}\|^2 + \frac{\mu}{2}\|y-y_t^{(m)}\|^2. \end{aligned} \quad (27)$$

Let $y = y_{x_t^{(m)}}^{(m)}$ and we obtain

$$\begin{aligned} g^{(m)}(x_t^{(m)},y_{x_t^{(m)}}^{(m)})) \geq g^{(m)}(x_t^{(m)},\tilde{y}_{t+1}^{(m)}) + \langle w_t^{(m)}, y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\rangle \\ + \langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y_{x_t^{(m)}}^{(m)}-\tilde{y}_{t+1}^{(m)}\rangle \\ + \Big(\gamma - \frac{L\gamma^2}{2}\Big)\|w_t^{(m)}\|^2 + \frac{\mu}{2}\|y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\|^2. \end{aligned} \quad (28)$$

By definition of $y_{x_t^{(m)}}^{(m)}$, we have $g^{(m)}(x_t^{(m)},y_{x_t^{(m)}}^{(m)})) \leq g^{(m)}(x_t^{(m)},\tilde{y}_{t+1}^{(m)})$. Thus, we obtain

$$\begin{aligned} 0 \geq \langle w_t^{(m)}, y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\rangle + \langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y_{x_t^{(m)}}^{(m)}-y_{t+1}^{(m)}\rangle \\ + \Big(\gamma - \frac{L\gamma^2}{2}\Big)\|w_t^{(m)}\|^2 + \frac{\mu}{2}\|y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\|^2. \end{aligned} \quad (29)$$

By $y_{t+1}^{(m)} = y_t^{(m)} - \gamma\alpha_t\omega_t^{(m)}$, we have

$$\begin{aligned} \|y_{t+1}^{(m)}-y_{x_t^{(m)}}^{(m)}\|^2 &= \|y_t^{(m)}-\gamma\alpha_t\omega_t^{(m)}-y_{x_t^{(m)}}^{(m)}\|^2 \\ &= \|y_t^{(m)}-y_{x_t^{(m)}}^{(m)}\|^2 - 2\gamma\alpha_t\langle\omega_t^{(m)}, y_t^{(m)}-y_{x_t^{(m)}}^{(m)}\rangle + \gamma^2\alpha_t^2\|\omega_t^{(m)}\|^2. \end{aligned} \quad (30)$$

By rearranging terms, we have:

$$\langle\omega_t^{(m)}, y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\rangle = -\frac{1}{2\gamma\alpha_t}\|y_t^{(m)}-y_{x_t^{(m)}}^{(m)}\|^2 - \frac{\gamma\alpha_t}{2}\|\omega_t^{(m)}\|^2 + \frac{1}{2\gamma\alpha_t}\|y_{t+1}^{(m)}-y_{x_t^{(m)}}^{(m)}\|^2. \quad (31)$$

Considering the upper bound of the term $\langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y_{x_t^{(m)}}^{(m)}-y_{t+1}^{(m)}\rangle$, we have

$$\begin{aligned} &- \langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y_{x_t^{(m)}}^{(m)}-y_{t+1}^{(m)}\rangle \\ &= -\langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\rangle - \langle \nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}, y_t^{(m)}-y_{t+1}^{(m)}\rangle \\ &\leq \frac{1}{\mu}\|\nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}\|^2 + \frac{\mu}{4}\|y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\|^2 \\ &\qquad + \frac{1}{\mu}\|\nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}\|^2 + \frac{\mu}{4}\|y_t^{(m)}-y_{t+1}^{(m)}\|^2 \\ &= \frac{2}{\mu}\|\nabla_y g(x_t^{(m)},y_t^{(m)}) - w_t^{(m)}\|^2 + \frac{\mu}{4}\|y_{x_t^{(m)}}^{(m)}-y_t^{(m)}\|^2 + \frac{\mu\gamma^2\alpha_t^2}{4}\|\omega_t^{(m)}\|^2. \end{aligned}$$

So we have:

$$\begin{aligned}
&\langle \nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}, y_{x_t^{(m)}}^{(m)} - y_{t+1}^{(m)} \rangle \\
&\geq -\frac{2}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2 - \frac{\mu}{4}\|y_{x_t^{(m)}}^{(m)} - y_t^{(m)}\|^2 - \frac{\mu\gamma^2\alpha_t^2}{4}\|\omega_t^{(m)}\|^2. \qquad (32)
\end{aligned}$$

Next, combining the inequalities 29, 31 with 32, we have

$$\begin{aligned}
&-\frac{1}{2\gamma\alpha_t}\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 \\
&\geq \left(\frac{\mu}{4} - \frac{1}{2\gamma\alpha_t}\right)\|y_{x_t^{(m)}}^{(m)} - y_t^{(m)}\|^2 + \left(\gamma - \frac{\gamma\alpha_t}{2} - \frac{\mu\gamma^2\alpha_t^2}{4} - \frac{L\gamma^2}{2}\right)\|\omega_t^{(m)}\|^2 \\
&\quad - \frac{2}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2 \\
&\geq \left(\frac{\mu}{4} - \frac{1}{2\gamma\alpha_t}\right)\|y_{x_t^{(m)}}^{(m)} - y_t^{(m)}\|^2 + \left(\frac{\gamma}{2} - \frac{3L\gamma^2}{4}\right)\|\omega_t^{(m)}\|^2 \\
&\quad - \frac{2}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2 \\
&\geq \left(\frac{\mu}{4} - \frac{1}{2\gamma\alpha_t}\right)\|y_{x_t^{(m)}}^{(m)} - y_t^{(m)}\|^2 + \left(\frac{3\gamma}{8} + \frac{\gamma}{8} - \frac{3L\gamma^2}{4}\right)\|\omega_t^{(m)}\|^2 \\
&\quad - \frac{2}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2 \\
&\geq \left(\frac{\mu}{4} - \frac{1}{2\gamma\alpha_t}\right)\|y_{x_t^{(m)}}^{(m)} - y_t^{(m)}\|^2 + \frac{3\gamma}{8}\|\omega_t^{(m)}\|^2 \\
&\quad - \frac{2}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2. \qquad (33)
\end{aligned}$$

where the first inequality is due to $0 < \alpha_t \leq \frac{1}{16L_h I} < 1$, the second inequality holds by $L \geq \mu$, and the last inequality is due to and $\gamma \leq \frac{1}{6L_h} \leq \frac{1}{6L}$. It implies that

$$\begin{aligned}
\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 \leq (1 - \frac{\mu\gamma\alpha_t}{2})\|y_{x_t^{(m)}}^{(m)} - y_t^{(m)}\|^2 - \frac{3\gamma^2\alpha_t}{4}\|\omega_t^{(m)}\|^2 \\
+ \frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2. \qquad (34)
\end{aligned}$$

Next, we decompose the term $\|y_{t+1}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)}\|^2$ as follows:

$$\begin{aligned}
\|y_{t+1}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)}\|^2 &= \|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)} + y_{x_t^{(m)}}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)}\|^2 \\
&\leq (1 + \frac{\mu\gamma\alpha_t}{4})\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\|y_{x_t^{(m)}}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)})\|^2 \\
&\leq (1 + \frac{\mu\gamma\alpha_t}{4})\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\rho^2\|x_t^{(m)} - \hat{x}_{t+1}^{(m)}\|^2 \\
&= (1 + \frac{\mu\gamma\alpha_t}{4})\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})L_h^2\eta^2\alpha_t^2\|\hat{\nu}_t^{(m)}\|^2. \qquad (35)
\end{aligned}$$

where the first inequality holds by the Cauchy-Schwartz inequality and Young's inequality, and the second inequality is due to case b) of Proposition 3.9, and the last equality holds by Line 13 of Algorithm 2 and the definition that $\rho \leq L_h$. Combining the above inequalities 34 and 35, we have

$$\begin{aligned}
\|y_{t+1}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)})\|^2 \leq (1 + \frac{\mu\gamma\alpha_t}{4})(1 - \frac{\mu\gamma\alpha_t}{2})\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)})\|^2 - (1 + \frac{\mu\gamma\alpha_t}{4})\frac{3\gamma^2\alpha_t}{4}\|\omega_t^{(m)}\|^2 \\
+ (1 + \frac{\mu\gamma\alpha_t}{4})\frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})L_h\eta^2\alpha_t^2\|\hat{\nu}_t^{(m)}\|^2.
\end{aligned}$$

Since $0 < \alpha_t \leq \frac{1}{16L_hI} \leq \frac{1}{16\mu I}$, $0 < \gamma \leq \frac{1}{6L_h} \leq 1$ and $L_h \geq L \geq \mu$, and $I \geq 1$ we have:

$$
\begin{aligned}
(1+\frac{\mu\gamma\alpha_t}{4})(1-\frac{\mu\gamma\alpha_t}{2}) &= 1-\frac{\mu\gamma\alpha_t}{4}-\frac{\mu^2\gamma^2\alpha_t^2}{8} \leq 1-\frac{\mu\gamma\alpha_t}{4},\\
-(1+\frac{\mu\gamma\alpha_t}{4})\frac{3\gamma^2\alpha_t}{4} &\leq -\frac{3\gamma^2\alpha_t}{4},\\
(1+\frac{\mu\gamma\alpha_t}{4})\frac{4\gamma\alpha_t}{\mu} &\leq (1+\frac{1}{64I})\frac{4\gamma\alpha_t}{\mu} = \frac{65\gamma\alpha_t}{16\mu},\\
(1+\frac{4}{\mu\gamma\alpha_t})L_h^2\eta^2\alpha_t^2 &= \left(\alpha_t^2+\frac{4\alpha_t}{\mu\gamma}\right)L_h^2\eta^2 \leq \left(\frac{\alpha_t}{16\mu\gamma I}+\frac{4\alpha_t}{\mu\gamma}\right)L_h^2\eta^2 = \frac{65L_h^2\eta^2\alpha_t}{16\mu\gamma}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\|y_{t+1}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)})\|^2 \leq \left(1-\frac{\mu\gamma\alpha_t}{4}\right)\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)})\|^2 - \frac{3\gamma^2\alpha_t}{4}\|\omega_t^{(m)}\|^2\\
+ \frac{65\gamma\alpha_t}{16\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t\|^2 + \frac{65L_h^2\eta^2\alpha_t}{16\mu\gamma}\|\hat{\nu}_t^{(m)}\|^2.
\end{aligned}
$$

For $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$, with $s \in [S]$, we have:

$$
\begin{aligned}
\mathbb{E}\left[\left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2\right] \leq (1-\frac{\mu\gamma\alpha_{t-1}}{4})\mathbb{E}\left[\left\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\right\|^2\right] - \frac{3\gamma^2\alpha_{t-1}}{4}\mathbb{E}\left[\left\|\omega_{t-1}^{(m)}\right\|^2\right]\\
+ \frac{65\gamma\alpha_{t-1}}{16\mu}\mathbb{E}\left[\left\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\right\|^2\right]\\
+ \frac{65L_h^2\eta^2\alpha_{t-1}}{16\mu\gamma}\mathbb{E}\left[\left\|\nu_{t-1}^{(m)}\right\|^2\right]. \qquad (36)
\end{aligned}
$$

It is straightforward to verify the claim in the Lemma as we have $65/16 < 5$.

When $t = \bar{t}_s$, we average variable x over the m clients, *i.e.* $x_{\bar{t}_s}^{(m)} = \bar{x}_{\bar{t}_s}$. For $\left\|y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)}\right\|^2$, we can get similar recursive relation as:

$$
\begin{aligned}
\mathbb{E}\left[\left\|y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)}\right\|^2\right] \leq (1-\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{4})\mathbb{E}\left[\left\|y_{\bar{t}_s-1}^{(m)} - y_{x_{\bar{t}_s-1}^{(m)}}^{(m)}\right\|^2\right] - \frac{3\gamma^2\alpha_{t-1}}{4}\mathbb{E}\left[\left\|\omega_{\bar{t}_s-1}^{(m)}\right\|^2\right]\\
+ \frac{65\gamma\alpha_{\bar{t}_s-1}}{16\mu}\mathbb{E}\left[\left\|\omega_{\bar{t}_s-1}^{(m)} - \nabla_y g^{(m)}(x_{\bar{t}_s-1}^{(m)}, y_{\bar{t}_s-1}^{(m)})\right\|^2\right]\\
+ \frac{65L_h\eta^2\alpha_{\bar{t}_s-1}}{16\mu\gamma}\mathbb{E}\left[\left\|\nu_{\bar{t}_s-1}^{(m)}\right\|^2\right]. \qquad (37)
\end{aligned}
$$

while for $\left\|y_{\bar{t}_s}^{(m)} - y_{x_{\bar{t}_s}^{(m)}}^{(m)}\right\|^2 = \left\|y_{\bar{t}_s}^{(m)} - y_{\bar{x}_{\bar{t}_s}}^{(m)}\right\|^2$, by generalized triangle inequality, we have:

$$
\begin{aligned}
\mathbb{E}\left[\left\|y_{\bar{t}_s}^{(m)} - y_{\bar{x}_{\bar{t}_s}}^{(m)}\right\|^2\right] &\leq (1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\mathbb{E}\left[\left\|y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)}\right\|^2\right] + (1+\frac{8}{\mu\gamma\alpha_{\bar{t}_s-1}})\mathbb{E}\left[\left\|y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)} - y_{\bar{x}_{\bar{t}_s}}^{(m)}\right\|^2\right]\\
&\leq (1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\mathbb{E}\left[\left\|y_{\bar{t}_s}^{(m)} - y_{\hat{x}_{\bar{t}_s}^{(m)}}^{(m)}\right\|^2\right] + (1+\frac{8}{\mu\gamma\alpha_{\bar{t}_s-1}})L_h^2\mathbb{E}\left[\left\|\hat{x}_{\bar{t}_s}^{(m)} - \bar{x}_{\bar{t}_s}\right\|^2\right] \qquad (38)
\end{aligned}
$$

Combine Eq. 37 and Eq. 38 together, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|y_{\bar{t}_s}^{(m)}-y_{\bar{x}_{\bar{t}_s}}^{(m)}\right\|^2\right] \leq (1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})(1-\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{4})\mathbb{E}\left[\left\|y_{\bar{t}_s-1}^{(m)}-y_{x_{\bar{t}_s-1}^{(m)}}^{(m)}\right\|^2\right] \\ -(1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\frac{3\gamma^2\alpha_{\bar{t}_s-1}}{4}\mathbb{E}\left[\left\|\omega_{\bar{t}_s-1}^{(m)}\right\|^2\right] \\ +(1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\frac{65\gamma\alpha_{\bar{t}_s-1}}{16\mu}\mathbb{E}\left[\left\|\omega_{\bar{t}_s-1}^{(m)}-\nabla_y g^{(m)}(x_{\bar{t}_s-1}^{(m)},y_{\bar{t}_s-1}^{(m)})\right\|^2\right] \\ +(1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\frac{65L_h^2\eta^2\alpha_{\bar{t}_s-1}}{16\mu\gamma}\mathbb{E}\left[\left\|\nu_{\bar{t}_s-1}^{(m)}\right\|^2\right] \\ +(1+\frac{8}{\mu\gamma\alpha_{\bar{t}_s-1}})L_h^2\mathbb{E}\left[\left\|\hat{x}_{\bar{t}_s}^{(m)}-\bar{x}_{\bar{t}_s}\right\|^2\right]\end{aligned}$$

For the coefficients, since we set $\gamma \leq \frac{1}{6L_h} < 1$ and $0 < \alpha_t < \frac{1}{16L_hI} < \frac{1}{16\mu I}$, it is straightforward to verify the following inequalities hold:

$$\begin{aligned}(1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})(1-\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{4}) &= 1-\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8}-\frac{\mu^2\gamma^2\alpha_{\bar{t}_s-1}^2}{32} \leq 1-\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8} \\ -(1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\frac{3\gamma^2\alpha_{\bar{t}_s-1}}{4} &\leq -\frac{3\gamma^2\alpha_{\bar{t}_s-1}}{4} \\ (1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\frac{65\gamma\alpha_{\bar{t}_s-1}}{16\mu} &\leq \frac{33\gamma\alpha_{\bar{t}_s-1}}{8\mu} \leq \frac{5\gamma\alpha_{\bar{t}_s-1}}{\mu} \\ (1+\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\frac{65L_h^2\eta^2\alpha_{\bar{t}_s-1}}{16\mu\gamma} &\leq \frac{33L_h^2\eta^2\alpha_{\bar{t}_s-1}}{8\mu\gamma} \leq \frac{5L_h^2\eta^2\alpha_{\bar{t}_s-1}}{\mu\gamma}\end{aligned}$$

So we have for $t=\bar{t}_s$:

$$\begin{aligned}\mathbb{E}\left[\left\|y_{\bar{t}_s}^{(m)}-y_{\bar{x}_{\bar{t}_s}}^{(m)}\right\|^2\right] \leq (1-\frac{\mu\gamma\alpha_{\bar{t}_s-1}}{8})\mathbb{E}\left[\left\|y_{\bar{t}_s-1}^{(m)}-y_{x_{\bar{t}_s-1}^{(m)}}^{(m)}\right\|^2\right]-\frac{3\gamma^2\alpha_{\bar{t}_s-1}}{4}\mathbb{E}\left[\left\|\omega_{\bar{t}_s-1}^{(m)}\right\|^2\right] \\ +\frac{5\gamma\alpha_{\bar{t}_s-1}}{\mu}\mathbb{E}\left[\left\|\omega_{\bar{t}_s-1}^{(m)}-\nabla_y g^{(m)}(x_{\bar{t}_s-1}^{(m)},y_{\bar{t}_s-1}^{(m)})\right\|^2\right] \\ +\frac{5L_h^2\eta^2\alpha_{\bar{t}_s-1}}{\mu\gamma}\mathbb{E}\left[\left\|\nu_{\bar{t}_s-1}^{(m)}\right\|^2\right]+(1+\frac{8}{\mu\gamma\alpha_{\bar{t}_s-1}})L_h^2\mathbb{E}\left[\left\|\hat{x}_{\bar{t}_s}^{(m)}-\bar{x}_{\bar{t}_s}\right\|^2\right]\end{aligned}$$

Combine with cases when $t \in [\bar{t}_{s-1}+1, \bar{t}_s-1]$ in Eq. 36 completes the proof. □

10.3 Bound for Outer Variable Drift

Lemma 8. *For $\alpha < \frac{1}{16IL_h}$ and $0<\eta<1$, we have for $t \in [\bar{t}_{s-1}+1, \bar{t}_s-1]$:*

$$\begin{aligned}\sum_{m=1}^{M}\mathbb{E}\|\nu_t^{(m)}-\bar{\nu}_t\|^2 \leq \left(1+\frac{33}{32I}\right)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)}-\bar{\nu}_{t-1}\|^2+4IL_h^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\left[2\|\eta\bar{\nu}_{t-1}\|^2+\|\gamma\omega_{t-1}^{(m)}\|^2\right] \\ +\frac{IM(c_\nu\alpha_{t-1}^2)^2\sigma^2}{L_h}+\frac{8IM(c_\nu\alpha_{t-1}^2)^2G^2}{2L_h}+\frac{Mc_\nu^2\alpha_{t-1}^3\zeta^2}{L_h} \\ +\frac{\eta^2c_\nu^2\alpha_{t-1}^2(1+\rho^2)}{2}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\left\|\left(\nu_\ell^{(m)}-\bar{\nu}_\ell\right)\right\|^2\end{aligned}$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. For $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$, with $s \in [S]$, we have: $\hat{x}_t^{(m)} = \hat{x}_{t-1}^{(m)} - \eta\alpha_{t-1}\nu_{t-1}^{(m)}$, this implies that:

$$\hat{x}_t^{(m)} = x_{\bar{t}_{s-1}}^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\nu_\ell^{(m)} \quad \text{and} \quad \bar{x}_t = \bar{x}_{\bar{t}_{s-1}} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\bar{\nu}_\ell.$$

So for $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$, with $s \in [S]$ we have:

$$\frac{1}{M}\sum_{m=1}^{M}\|\hat{x}_t^{(m)} - \bar{x}_t\|^2 = \frac{1}{M}\sum_{m=1}^{M}\Big\|x_{\bar{t}_{s-1}}^{(m)} - \bar{x}_{\bar{t}_{s-1}} - \Big(\sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\alpha_\ell\nu_\ell^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\alpha_\ell\bar{\nu}_\ell\Big)\Big\|^2$$
$$\overset{(a)}{=} \frac{1}{M}\sum_{m=1}^{M}\Big\|\sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\alpha_\ell\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\Big\|^2 \overset{(b)}{\leq} \sum_{\ell=\bar{t}_{s-1}}^{t-1}\frac{I\eta^2\alpha_l^2}{M}\sum_{m=1}^{M}\Big\|\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\Big\|^2. \tag{39}$$

where the equality (a) follows from the fact that $x_{\bar{t}_{s-1}}^{(m)} = \bar{x}_{\bar{t}_{s-1}}$ for $t = \bar{t}_{s-1}$; inequality (b) is due to $t - \bar{t}_{s-1} \leq I$ and the generalized triangle inequality.

Next, we bound the term $\|\nu_t^{(m)} - \bar{\nu}_t\|^2$, for $t \in [\bar{t}_{s-1}+1, \bar{t}_s - 1]$, with $s \in [S]$:

$$\sum_{m=1}^{M}\mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2 = \sum_{m=1}^{M}\mathbb{E}\Big\|\mu_t^{(m)} + (1-c_\nu\alpha_{t-1}^2)\big(\nu_{t-1}^{(m)} - \mu_{t-1}^{(m)}\big)$$
$$- \Big(\frac{1}{M}\sum_{j=1}^{M}\mu_t^{(j)} + (1-c_\nu\alpha_{t-1}^2)\big(\bar{\nu}_{t-1} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\big)\Big)\Big\|^2$$
$$= \sum_{m=1}^{M}\mathbb{E}\Big\|(1-c_\nu\alpha_{t-1}^2)\big(\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\big) + \mu_t^{(m)}$$
$$- \frac{1}{M}\sum_{j=1}^{M}\mu_t^{(j)} - (1-c_\nu\alpha_{t-1}^2)\Big(\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big)\Big\|^2$$
$$\overset{(a)}{\leq} (1+\beta)(1-c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2$$
$$+ \Big(1+\frac{1}{\beta}\Big)\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_t^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_t^{(j)} - (1-c_\nu\alpha_{t-1}^2)\Big(\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big)\Big\|^2.$$

where (a) follows from the the generalized triangle inequality for some $\beta > 0$. Next we bound the second term:

$$\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_t^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_t^{(j)} - (1-c_\nu\alpha_{t-1}^2)\Big(\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big)\Big\|^2$$
$$= \sum_{m=1}^{M}\mathbb{E}\Big\|\mu_t^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_t^{(j)} - \Big(\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big) + c_\nu\alpha_{t-1}^2\Big(\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big)\Big\|^2$$
$$\overset{(a)}{\leq} 2\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_t^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_t^{(j)} - \Big(\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big)\Big\|^2$$
$$+ 2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big\|^2$$
$$\overset{(b)}{\leq} 2\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_t^{(m)} - \mu_{t-1}^{(m)}\Big\|^2 + 2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big\|^2$$
$$\overset{(c)}{\leq} 2L_h^2\sum_{m=1}^{M}\mathbb{E}\Big[\|x_t^{(m)} - x_{t-1}^{(m)}\|^2 + \|y_t^{(m)} - y_{t-1}^{(m)}\|^2\Big] + 2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big\|^2$$
$$\leq 2L_h^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\Big[\|\eta\nu_{t-1}^{(m)}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\Big] + 2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\Big\|\mu_{t-1}^{(m)} - \frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\Big\|^2.$$

where inequality (a) is from the triangle inequality, (b) follows Proposition 4; (c) follows from the Lipschitz-smoothness of the h. Next for the second term of the above equation:

$$
\begin{aligned}
&\sum_{m=1}^{M}\mathbb{E}\left\|\mu_{t-1}^{(m)}-\frac{1}{M}\sum_{j=1}^{M}\mu_{t-1}^{(j)}\right\|^2\\
&=\sum_{m=1}^{M}\mathbb{E}\left\|\mu_{t-1}^{(m)}-\tilde{\mu}_{t-1}^{(m)}-\frac{1}{M}\sum_{j=1}^{M}\left(\mu_{t-1}^{(j)}-\tilde{\mu}_{t-1}^{(j)}\right)+\tilde{\mu}_{t-1}^{(m)}-\frac{1}{M}\sum_{j=1}^{M}\tilde{\mu}_{t-1}^{(j)}\right\|^2\\
&\overset{(a)}{\leq} 2\sum_{m=1}^{M}\mathbb{E}\left\|\mu_{t-1}^{(m)}-\tilde{\mu}_{t-1}^{(m)}-\frac{1}{M}\sum_{j=1}^{M}\left(\mu_{t-1}^{(j)}-\tilde{\mu}_{t-1}^{(j)}\right)\right\|^2\\
&\qquad\qquad+2\sum_{k=1}^{K}\mathbb{E}\left\|\tilde{\mu}_{t-1}^{(m)}-\frac{1}{M}\sum_{j=1}^{M}\tilde{\mu}_{t-1}^{(j)}\right\|^2\\
&\overset{(b)}{\leq} 2\sum_{m=1}^{M}\mathbb{E}\left\|\mu_{t-1}^{(m)}-\tilde{\mu}_{t-1}^{(m)}\right\|^2+2\sum_{m=1}^{M}\mathbb{E}\left\|\tilde{\mu}_{t-1}^{(m)}-\frac{1}{M}\sum_{j=1}^{M}\tilde{\mu}_{t-1}^{(j)}\right\|^2\\
&\overset{(c)}{\leq} 2M\sigma^2+4\sum_{m=1}^{M}\mathbb{E}\left\|\nabla h^{(m)}(\bar{x}_{t-1})-\nabla h(\bar{x}_{t-1})\right\|^2\\
&\qquad+8\sum_{m=1}^{M}\mathbb{E}\left\|\tilde{\mu}_{t-1}^{(m)}-\nabla h^{(m)}(\bar{x}_{t-1})\right\|^2+8\sum_{m=1}^{M}\mathbb{E}\left\|\nabla h(\bar{x}_{t-1})-\frac{1}{M}\sum_{j=1}^{M}\tilde{\mu}_{t-1}^{(j)}\right\|^2\\
&\overset{(d)}{\leq} 2M\sigma^2+4\sum_{m=1}^{M}\mathbb{E}\left\|\nabla h^{(m)}(\bar{x}_{t-1})-\nabla h(\bar{x}_{t-1})\right\|^2+16\sum_{m=1}^{M}\mathbb{E}\left\|\tilde{\mu}_{t-1}^{(m)}-\nabla h^{(m)}(\bar{x}_{t-1})\right\|^2\\
&\overset{(e)}{\leq} 2M\sigma^2+4\sum_{m=1}^{M}\mathbb{E}\left\|\nabla h^{(m)}(\bar{x}_{t-1})-\nabla h(\bar{x}_{t-1})\right\|^2\\
&\qquad+32\sum_{m=1}^{M}\mathbb{E}\left\|\tilde{\mu}_{t-1}^{(m)}-\Phi^{(m)}(x_{t-1}^{(m)},y_{t-1}^{(m)})\right\|^2+32\sum_{m=1}^{M}\mathbb{E}\left\|\Phi^{(m)}(x_{t-1}^{(m)},y_{t-1}^{(m)})-\nabla h^{(m)}(\bar{x}_{t-1})\right\|^2\\
&\overset{(f)}{\leq} 2M\sigma^2+32MG^2+4\sum_{m=1}^{M}\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}\|\nabla h^{(m)}(\bar{x}_{t-1})-\nabla h^{(j)}(\bar{x}_{t-1})\|^2\\
&\qquad\qquad+32L_h^2\sum_{m=1}^{M}\mathbb{E}\left[\|x_{t-1}^{(m)}-\bar{x}_{t-1}\|^2+\|y_{t-1}^{(m)}-y_{\bar{x}_{t-1}}^{(m)}\|^2\right]\\
&\overset{(g)}{\leq} 2M\sigma^2+32MG^2+4M\zeta^2+32L_h^2(1+\rho^2)\sum_{m=1}^{M}\mathbb{E}\left[\|x_{t-1}^{(m)}-\bar{x}_{t-1}\|^2\right]
\end{aligned}
$$

inequality (a) uses triangle inequality; inequality (b) follows Proposition 4; inequality (c) follow Assumption 6 and generalized triangle inequality; inequality (d) and (e) follows the generalized inequality; inequality (f) follows the Assumption 6; inequality (g) utilizes intra-node heterogeneity assumption and Proposition 1.

Finally, combine everything together, we have:

$$
\begin{aligned}
&\sum_{m=1}^{M}\mathbb{E}\|\nu_t^{(m)}-\bar{\nu}_t\|^2\\
&\leq (1+\beta)(1-c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)}-\bar{\nu}_{t-1}\|^2+2L_h^2\left(1+\frac{1}{\beta}\right)\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\left[\|\eta\nu_{t-1}^{(m)}\|^2+\|\gamma\omega_{t-1}^{(m)}\|^2\right]\\
&\quad+4M\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2\sigma^2+64M\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2G^2+8M\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2\zeta^2\\
&\quad+64\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2L_h^2(1+\rho^2)\sum_{m=1}^{M}\mathbb{E}\left[\|x_{t-1}^{(m)}-\bar{x}_{t-1}\|^2\right]\\
&\overset{(a)}{\leq}(1+\beta)(1-c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)}-\bar{\nu}_{t-1}\|^2+2L_h^2\left(1+\frac{1}{\beta}\right)\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\left[\|\eta\nu_{t-1}^{(m)}\|^2+\|\gamma\omega_{t-1}^{(m)}\|^2\right]\\
&\quad+4M\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2\sigma^2+64M\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2G^2+8M\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2\zeta^2\\
&\quad+64\left(1+\frac{1}{\beta}\right)(c_\nu\alpha_{t-1}^2)^2L_h^2(1+\rho^2)\sum_{\ell=\bar{t}_{s-1}}^{t-1}I\eta^2\alpha_l^2\sum_{m=1}^{M}\left\|\left(\nu_\ell^{(m)}-\bar{\nu}_\ell\right)\right\|^2\\
&\overset{(b)}{\leq}\left(1+\frac{1}{I}\right)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)}-\bar{\nu}_{t-1}\|^2+4IL_h^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\left[\|\eta\nu_{t-1}^{(m)}\|^2+\|\gamma\omega_{t-1}^{(m)}\|^2\right]\\
&\quad+8IM(c_\nu\alpha_{t-1}^2)^2\sigma^2+128IM(c_\nu\alpha_{t-1}^2)^2G^2+16IM(c_\nu\alpha_{t-1}^2)^2\zeta^2\\
&\quad+128I^2\eta^2(c_\nu\alpha_{t-1}^2)^2L_h^2(1+\rho^2)\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\left\|\left(\nu_\ell^{(m)}-\bar{\nu}_\ell\right)\right\|^2\\
&\overset{(c)}{\leq}\left(1+\frac{1}{I}\right)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)}-\bar{\nu}_{t-1}\|^2+4IL_h^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\left[2\|\eta\nu_{t-1}^{(m)}-\eta\bar{\nu}_{t-1}\|^2+2\|\eta\bar{\nu}_{t-1}\|^2+\|\gamma\omega_{t-1}^{(m)}\|^2\right]\\
&\quad+8IM(c_\nu\alpha_{t-1}^2)^2\sigma^2+128IM(c_\nu\alpha_{t-1}^2)^2G^2+16IM(c_\nu\alpha_{t-1}^2)^2\zeta^2\\
&\quad+128I^2\eta^2(c_\nu\alpha_{t-1}^2)^2L_h^2(1+\rho^2)\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\left\|\left(\nu_\ell^{(m)}-\bar{\nu}_\ell\right)\right\|^2\\
&\leq\left(1+\frac{1}{I}+8IL_h^2\eta^2\alpha_{t-1}^2\right)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)}-\bar{\nu}_{t-1}\|^2+4IL_h^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\left[2\|\eta\bar{\nu}_{t-1}\|^2+\|\gamma\omega_{t-1}^{(m)}\|^2\right]\\
&\quad+8IM(c_\nu\alpha_{t-1}^2)^2\sigma^2+128IM(c_\nu\alpha_{t-1}^2)^2G^2+16IM(c_\nu\alpha_{t-1}^2)^2\zeta^2\\
&\quad+128I^2\eta^2(c_\nu\alpha_{t-1}^2)^2L_h^2(1+\rho^2)\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\left\|\left(\nu_\ell^{(m)}-\bar{\nu}_\ell\right)\right\|^2\\
&\overset{(d)}{\leq}\left(1+\frac{33}{32I}\right)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)}-\bar{\nu}_{t-1}\|^2+4IL_h^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\left[2\|\eta\bar{\nu}_{t-1}\|^2+\|\gamma\omega_{t-1}^{(m)}\|^2\right]\\
&\quad+\frac{IM(c_\nu\alpha_{t-1}^2)^2\sigma^2}{L_h}+\frac{8IM(c_\nu\alpha_{t-1}^2)^2G^2}{2L_h}+\frac{Mc_\nu^2\alpha_{t-1}^3\zeta^2}{L_h}\\
&\quad+\frac{\eta^2c_\nu^2\alpha_{t-1}^2(1+\rho^2)}{2}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\left\|\left(\nu_\ell^{(m)}-\bar{\nu}_\ell\right)\right\|^2
\end{aligned}
$$

where inequality (a) follows Eq. 39; in inequality (b), we set $\beta = 1/I$ and use $I \geq 1$; Inequality (c) uses the generalized triangle inequality. The inequality (d) uses $\alpha_t < \frac{1}{16L_hI}$ and $\eta < 1$ Therefore, the lemma is proved. □

Lemma 9. *For $\alpha_t < \frac{1}{16L_hI}$, we have:*

$$\left(1 - \frac{3\eta^2 c_\nu^2(1+\rho^2)}{16^3 * 32 I^2 L_h^4}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t \leq \frac{3\eta^2}{32} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t E_t + \frac{3\gamma^2}{64} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t F_t + \left(\frac{3c_\nu^2\sigma^2}{32L_h} + \frac{3c_\nu^2 G^2}{2L_h} + \frac{3c_\nu^2\zeta^2}{16L_h}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3$$

where the terms D_t, E_t and F_t are denoted below.

Proof. To simplify the notation, we denote $A_t = \mathbb{E}\left[\left\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^M \nabla h^{(m)}(x_t^{(m)})\right\|^2\right]$, $B_t = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^M \left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2\right]$, $C_t = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^M \left\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\right\|^2\right]$, $D_t = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^M \left\|\nu_t^{(m)} - \bar{\nu}_t\right\|^2\right]$, $E_t = \mathbb{E}\left[\left\|\bar{\nu}_t\right\|^2\right]$, $F_t = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^M \left\|\omega_t^{(m)}\right\|^2\right]$. Then we rewrite Lemma 8 with our new notation as follows:

$$D_t \leq \left(1 + \frac{33}{32I}\right) D_{t-1} + 8IL_h^2\alpha_{t-1}^2\eta^2 E_{t-1} + 4IL_h^2\alpha_{t-1}^2\gamma^2 F_{t-1} + \frac{c_\nu^2\alpha_{t-1}^3\sigma^2}{2L_h} + \frac{8c_\nu^2\alpha_{t-1}^3 G^2}{L_h} + \frac{c_\nu^2\alpha_{t-1}^3\zeta^2}{L_h} + \frac{\eta^2 c_\nu^2\alpha_{t-1}^2(1+\rho^2)}{2} \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_l^2 D_l$$

Next apply the above equation recursively from $\bar{t}_{s-1} + 1$ to t. Note that $D_{\bar{t}_{s-1}} = 1/M \sum_{m=1}^M \mathbb{E}\|\nu_{\bar{t}_{s-1}}^{(m)} - \bar{\nu}_{\bar{t}_{s-1}}\|^2 = 0$, so we have:

$$\begin{aligned}
D_t &\leq 8IL_h^2\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \left(1 + \frac{33}{32I}\right)^{t-\ell} \alpha_\ell^2 E_\ell + 4IL_h^2\gamma^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \left(1 + \frac{33}{32I}\right)^{t-\ell} \alpha_\ell^2 F_\ell \\
&\quad + \left(\frac{c_\nu^2\sigma^2}{2L_h} + \frac{8c_\nu^2 G^2}{L_h} + \frac{c_\nu^2\zeta^2}{L_h}\right) \sum_{\ell=\bar{t}_{s-1}}^{t-1} \left(1 + \frac{33}{32I}\right)^{t-\ell} \alpha_\ell^3 \\
&\quad + \frac{\eta^2 c_\nu^2(1+\rho^2)}{2} \sum_{\ell=\bar{t}_{s-1}}^{t-1} \left(1 + \frac{33}{32I}\right)^{t-\ell} \alpha_\ell^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell} \alpha_{\bar{\ell}}^2 D_{\bar{\ell}} \\
&\leq 24IL_h^2\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 E_\ell + 12IL_h^2\gamma^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 F_\ell + \left(\frac{3c_\nu^2\sigma^2}{2L_h} + \frac{24c_\nu^2 G^2}{L_h} + \frac{3c_\nu^2\zeta^2}{L_h}\right) \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^3 \\
&\quad + \frac{3\eta^2 c_\nu^2(1+\rho^2)}{2} \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell} \alpha_{\bar{\ell}}^2 D_{\bar{\ell}}
\end{aligned}$$

The second inequality uses the fact that $t - l \leq I$ and the inequality $log(1 + a/x) \leq a/x$ for $x > -a$, so we have $(1 + a/x)^x \leq e^{a/x}$, Then we choose $a = 33/32$ and $x = I$. Finally, we use the fact that $e^{33/(32I)} \leq e^{33/32} \leq 3$.

Next we multiply α_t over both sides and take sum from $\bar{t}_{s-1}+1$ to $\bar{t}_s$, we have:

$$
\begin{aligned}
\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s} \alpha_t D_t \leq\ & 24IL_h^2\eta^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 E_\ell + 12IL_h^2\gamma^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 F_\ell \\
& + \left(\frac{3c_\nu^2\sigma^2}{2L_h} + \frac{24c_\nu^2G^2}{L_h} + \frac{3c_\nu^2\zeta^2}{L_h}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^3 \\
& + \frac{3\eta^2c_\nu^2(1+\rho^2)}{2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell} \alpha_{\bar{\ell}}^2 D_{\bar{\ell}} \\
\leq\ & 24IL_h^2\eta^2 \Big(\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t\Big) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^2 E_t + 12IL_h^2\gamma^2 \Big(\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t\Big) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^2 F_t \\
& + \left(\frac{3c_\nu^2\sigma^2}{2L_h} + \frac{24c_\nu^2G^2}{L_h} + \frac{3c_\nu^2\zeta^2}{L_h}\right)\Big(\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t\Big) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3 \\
& + \frac{3\eta^2c_\nu^2(1+\rho^2)}{2}\Big(\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t\Big) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{t} \alpha_{\bar{\ell}}^2 D_{\bar{\ell}} \\
\overset{(a)}{\leq}\ & \frac{3IL_h\eta^2}{2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^2 E_t + \frac{3IL_h\gamma^2}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^2 F_t \\
& + \left(\frac{3c_\nu^2\sigma^2}{32L_h} + \frac{3c_\nu^2G^2}{2L_h} + \frac{3c_\nu^2\zeta^2}{16L_h}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3 \\
& + \frac{3\eta^2c_\nu^2(1+\rho^2)}{32L_h} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{t} \alpha_{\bar{\ell}}^2 D_{\bar{\ell}} \\
\overset{(b)}{\leq}\ & \frac{3\eta^2}{32} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t E_t + \frac{3\gamma^2}{64} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t F_t + \left(\frac{3c_\nu^2\sigma^2}{32L_h} + \frac{3c_\nu^2G^2}{2L_h} + \frac{3c_\nu^2\zeta^2}{16L_h}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3 \\
& + \frac{3\eta^2c_\nu^2(1+\rho^2)}{16^3*32I^2L_h^4} \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_{\bar{\ell}} D_{\bar{\ell}}
\end{aligned}
$$

In inequalities (a) and (b), we use $\alpha_t < \frac{1}{16L_hI}$ multiple times. next notice that $\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s} \alpha_t D_t = \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t$ as $D_{\bar{t}_s} = D_{\bar{t}_{s-1}} = 0$, so we have:

$$
\begin{aligned}
\left(1 - \frac{3\eta^2c_\nu^2(1+\rho^2)}{16^3*32I^2L_h^4}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t \leq\ & \frac{3\eta^2}{32} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t E_t + \frac{3\gamma^2}{64} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t F_t \\
& + \left(\frac{3c_\nu^2\sigma^2}{32L_h} + \frac{3c_\nu^2G^2}{2L_h} + \frac{3c_\nu^2\zeta^2}{16L_h}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3
\end{aligned}
$$

This completes the proof. □

10.4 DESCENT LEMMA

Lemma 10 (Descent Lemma)**.** *For all $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$ and $s \in [S]$, the iterates generated satisfy:*

$$\mathbb{E}\Big[h(\bar{x}_{t+1})\Big] \leq \mathbb{E}\Big[h(\bar{x}_t)\Big] - \left(\frac{\eta\alpha_t}{2} - \frac{\eta^2\alpha_t^2 L}{2}\right)\mathbb{E}\Big[\Big\|\bar{\nu}_t\Big\|^2\Big] - \frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] + \frac{L_h^2 I\eta^3\alpha_t}{M}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\Big\|\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\Big\|^2 + \eta\alpha_t\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \bar{\nu}_t\Big\|^2\Big]$$

where the expectation is w.r.t the stochasticity of the algorithm.

Proof. Using the smoothness of $h(x)$ we have:

$$\begin{aligned}
\mathbb{E}[h(\bar{x}_{t+1})] &\leq \mathbb{E}\Big[h(\bar{x}_t) + \langle\nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{L_h}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[h(\bar{x}_t) - \eta\alpha_t\langle\nabla h(\bar{x}_t), \bar{\nu}_t\rangle + \frac{\eta^2\alpha_t^2 L_h}{2}\|\bar{\nu}_t\|^2\Big] \\
&\overset{(b)}{=} \mathbb{E}\Big[h(\bar{x}_t) - \frac{\eta\alpha_t}{2}\Big\|\bar{\nu}_t\Big\|^2 - \frac{\eta\alpha_t}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta\alpha_t}{2}\Big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\Big\|^2 + \frac{\eta\alpha_t^2 L_h}{2}\Big\|\bar{\nu}_t\Big\|^2\Big] \\
&= \mathbb{E}\Big[h(\bar{x}_t) - \left(\frac{\eta\alpha_t}{2} - \frac{\eta^2\alpha_t^2 L_h}{2}\right)\Big\|\bar{\nu}_t\Big\|^2 - \frac{\eta\alpha_t}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta\alpha_t}{2}\Big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\Big\|^2\Big]
\end{aligned}$$

where equality (a) follows from the iterate update given in Line 15 of Algorithm 2; (b) uses $\langle a, b\rangle = \frac{1}{2}[\|a\|^2 + \|b\|^2 - \|a-b\|^2]$; For the last term, we have:

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\Big\|^2\Big] &\overset{(a)}{\leq} 2\mathbb{E}\Big[\Big\|\nabla h(\bar{x}_t) - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\Big\|^2\Big] + 2\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \bar{\nu}_t\Big\|^2\Big] \\
&\overset{(b)}{\leq} \frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|\nabla h(\bar{x}_t) - \nabla h(x_t^{(m)})\Big\|^2\Big] + 2\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \bar{\nu}_t\Big\|^2\Big] \\
&\overset{(c)}{\leq} \frac{2L_h^2}{M}\sum_{m=1}^{M}\mathbb{E}\Big[\Big\|\bar{x}_t - x_t^{(m)}\Big\|^2\Big] + 2\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \bar{\nu}_t\Big\|^2\Big] \\
&\overset{(d)}{\leq} \frac{2L_h^2 I\eta^2}{M}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\Big\|\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\Big\|^2 + 2\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \bar{\nu}_t\Big\|^2\Big]
\end{aligned}$$

where inequality (a) uses triangle inequality, (b) uses the generalized triangle inequality, (c) uses the smoothness of $h(x)$, (d) uses Eq. 39. Combine the above two equations together, we get:

$$\mathbb{E}\Big[h(\bar{x}_{t+1})\Big] \leq \mathbb{E}\Big[h(\bar{x}_t)\Big] - \left(\frac{\eta\alpha_t}{2} - \frac{\eta^2\alpha_t^2 L_h}{2}\right)\mathbb{E}\Big[\Big\|\bar{\nu}_t\Big\|^2\Big] - \frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] + \frac{L_h^2 I\eta^3\alpha_t}{M}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\Big\|\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\Big\|^2 + \eta\alpha_t\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \bar{\nu}_t\Big\|^2\Big]$$

Hence, the lemma is proved. □

10.5 DESCENT IN POTENTIAL FUNCTION

We first denote the following potential function $\mathcal{G}(t)$:

$$\mathcal{G}_t = h(\bar{x}_t) + \frac{\eta}{160L_h^2\alpha_t}\Big\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\Big\|^2 + \frac{1}{M}\sum_{m=1}^{M}\Big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\Big\|^2 + \frac{\gamma}{32L^2\mu\alpha_t}\times\frac{1}{M}\sum_{m=1}^{M}\Big\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\Big\|^2$$

In the above definition, we correct the coefficients of the second and fourth term of potential function defined in Section 5.2. in the original main-text.

Lemma 11. *Suppose* $\frac{1}{\gamma} > max(\frac{1}{\mu}, 6L_h)$, $\frac{1}{\eta} > max(\frac{4\gamma}{\mu} + \frac{1}{4I}, \frac{12(1+\rho^2)}{I^2} + \frac{97}{256} + \frac{\gamma^2}{2\mu^2} + (1 + \frac{16}{\mu\gamma})I, \frac{240L_h^2}{\mu\gamma}, \frac{\mu}{\gamma})$, $c_\nu = 160L_h^2 + \frac{\sigma^2}{24\delta^3 L_h I}$, $c_\omega = 160L^2 + \frac{\sigma^2}{24\delta^3 L_h I}$, $u = max(2\sigma^2, \delta^3, c_\nu^{3/2}\delta^3, c_\omega^{3/2}\delta^3, 16^3 I^3 M^2 \sigma^2)$, $\delta = \frac{M^{2/3}\sigma^{2/3}}{L_h}$ *then we have:*

$$\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1} \frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s - 1}\alpha_t^3$$

where the expectation is w.r.t the stochasticity of the algorithm.

Take expectation for both sides of the potential function and we use the notation used in Lemma 9, the potential function has the following form:

$$\mathbb{E}[\mathcal{G}_t] = \mathbb{E}[h(\bar{x}_t)] + \frac{\eta A_t}{\hat{c}_\nu\alpha_t} + B_t + \frac{\gamma C_t}{\hat{c}_\omega\alpha_t}$$

We first bound the term $A_t/\alpha_{t-1} - A_{t-1}/\alpha_{t-2}$. For $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$. By the condition that $u \geq c_\nu^{3/2}\delta^3$, it is straightforward to verify that $c_\nu\alpha^2 < 1$. Then we rewrite Lemma 5 as follows using our new notation:

$$\begin{aligned}A_t \leq{}& (1 - c_\nu\alpha_{t-1}^2)^2 A_{t-1} + 4(c_\nu\alpha_{t-1}^2)^2\sigma^2/M + 8(c_\nu\alpha_{t-1}^2)^2 G^2 + 8L_h^2(c_\nu\alpha_{t-1}^2)^2 B_{t-1}\\ &+ 40L_h^2\eta^2\alpha_{t-1}^2 D_{t-1} + 40L_h^2\eta^2\alpha_{t-1}^2 E_{t-1} + 12L_h^2\gamma^2\alpha_{t-1}^2 F_{t-1}\end{aligned}$$

Naturally, we get:

$$\begin{aligned}\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \leq{}& \left(\frac{(1 - c_\nu\alpha_{t-1}^2)^2}{\alpha_{t-1}} - \frac{1}{\alpha_{t-2}}\right)A_{t-1} + 4c_\nu^2\alpha_{t-1}^3\sigma^2/M + 8c_\nu^2\alpha_{t-1}^3 G^2 + 8L_h^2 c_\nu^2\alpha_{t-1}^3 B_{t-1}\\ &+ 40L_h^2\eta^2\alpha_{t-1}D_{t-1} + 40L_h^2\alpha_{t-1}E_{t-1} + 12L_h^2\gamma^2\alpha_{t-1}F_{t-1}\\ \leq{}& \left(\alpha_{t-1}^{-1} - \alpha_{t-2}^{-1} - c_\nu\alpha_{t-1}\right)A_{t-1} + 4c_\nu^2\alpha_{t-1}^3\sigma^2/M + 8c_\nu^2\alpha_{t-1}^3 G^2 + 8L_h^2 c_\nu^2\alpha_{t-1}^3 B_{t-1}\\ &+ 40L_h^2\eta^2\alpha_{t-1}D_{t-1} + 40L_h^2\eta^2\alpha_{t-1}E_{t-1} + 12L_h^2(1 - c_\nu\alpha_{t-1}^2)^2\gamma^2\alpha_{t-1}F_{t-1}\end{aligned}$$

where the inequality is due to the fact that $(1 - c_\nu a_{t-1}^2)^2 \leq 1 - c_\nu a_{t-1}^2 \leq 1$ for all $t \in [T]$. Next for the term $\alpha_{t-1}^{-1} - \alpha_{t-2}^{-1}$ we have:

$$\begin{aligned}\alpha_t^{-1} - \alpha_{t-1}^{-1} = \frac{(u + \sigma^2 t)^{1/3}}{\delta} - \frac{(u + \sigma^2(t-1))^{1/3}}{\delta} &\overset{(a)}{\leq} \frac{\sigma^2}{3\delta(u + \sigma^2(t-1))^{2/3}}\\ &\overset{(b)}{\leq} \frac{2^{2/3}\sigma^2\delta^2}{3\delta^3(u + \sigma^2 t)^{2/3}} \overset{(c)}{=} \frac{2^{2/3}\sigma^2}{3\delta^3}\alpha_t^2 \overset{(d)}{\leq} \frac{\sigma^2}{24\delta^3 LI}\alpha_t\end{aligned}$$

where inequality (a) results from the concavity of $x^{1/3}$ as: $(x+y)^{1/3} - x^{1/3} \leq y/3x^{2/3}$, inequality (b) used the fact that $u_t \geq 2\sigma^2$, inequality (c) uses the definition of α_t, inequality (d) uses $u \geq 16^3 I^3 M^2$, so that $\alpha_t \leq \frac{1}{16L_h I}$ for all $t \in [T]$. Since we have $c_\nu = \hat{c}_\nu + \frac{\sigma^2}{24\delta^3 L_h I}$, where $\hat{c}_\nu = 160L_h^2$ is some constant. It is straightforward to verify that if we set $\delta = \frac{M^{2/3}\sigma^{2/3}}{L_h}$, we have $c_\nu \leq 2\hat{c}_\nu$. Next, we have:

$$\begin{aligned}\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \leq{}& -\hat{c}_\nu\alpha_{t-1}A_{t-1} + 4c_\nu^2\alpha_{t-1}^3\sigma^2/M + 8c_\nu^2\alpha_{t-1}^3 G^2 + 8L_h^2 c_\nu^2\alpha_{t-1}^3 B_{t-1} + 40L_h^2\eta^2\alpha_{t-1}D_{t-1}\\ &+ 40L_h^2\eta^2\alpha_{t-1}E_{t-1} + 12L_h^2\gamma^2\alpha_{t-1}F_{t-1}\end{aligned}$$

Then We multiply $\eta/\hat{c}_\nu$ on both sides and have:

$$\begin{aligned}\frac{\eta}{\hat{c}_\nu}\left(\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}}\right) \leq{}& -\eta\alpha_{t-1}A_{t-1} + 4c_\nu^2\alpha_{t-1}^3\sigma^2/(\hat{c}_\nu M) + 8c_\nu^2\eta\alpha_{t-1}^3 G^2/\hat{c}_\nu + 8L_h^2 c_\nu^2\eta\alpha_{t-1}^3 B_{t-1}/\hat{c}_\nu\\ &+ 40L_h^2\eta^3\alpha_{t-1}D_{t-1}/\hat{c}_\nu + 40L_h^2\eta^3\alpha_{t-1}E_{t-1}/\hat{c}_\nu + 12L_h^2\gamma^2\eta\alpha_{t-1}F_{t-1}/\hat{c}_\nu\end{aligned}$$

By telescoping from $\bar{t}_{s-1}+1$ to $\bar{t}_s$, we have:

$$\frac{\eta}{\hat{c}_\nu}\left(\frac{A_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}}-\frac{A_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}}\right) \leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\eta\alpha_t A_t + 4\eta c_\nu^2\sigma^2/(\hat{c}_\nu M)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + 8\eta c_\nu^2 G^2/\hat{c}_\nu\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + 8L_h^2\eta c_\nu^2/\hat{c}_\nu\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 B_t + 40L_h^2\eta^3/\hat{c}_\nu\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + 40L_h^2\eta^3/\hat{c}_\nu\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + 12L_h^2\eta\gamma^2/\hat{c}_\nu\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t. \tag{40}$$

Similarly, by the condition $u \geq c_\omega^{3/2}\delta^3$, the condition of Lemma 6 satisfies. For $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$, we have:

$$C_t \leq (1-c_\omega\alpha_{t-1}^2)^2 C_{t-1} + 2(c_\omega\alpha_{t-1}^2)^2\sigma^2 + 4L^2\eta^2\alpha_{t-1}^2(D_{t-1}+E_{t-1}) + 2L^2\gamma^2\alpha_{t-1}^2 F_{t-1}$$

We bound the term $C_t/\alpha_{t-1} - C_{t-1}/\alpha_{t-2}$ and follow similar derivation as $A_t/\alpha_{t-1} - A_{t-1}/\alpha_{t-2}$ and since we have $c_\omega = 5\hat{c}_\omega + \frac{\sigma^2}{24\delta^3 L_h I}$, where $\hat{c}_\omega = 32L^2$ is some constant. we get:

$$\frac{C_t}{\alpha_{t-1}} - \frac{C_{t-1}}{\alpha_{t-2}} \leq -5\hat{c}_\omega\alpha_{t-1}C_{t-1} + 2c_\omega^2\alpha_{t-1}^3\sigma^2 + 4L^2\eta^2\alpha_{t-1}(D_{t-1}+E_{t-1}) + 2L^2\gamma^2\alpha_{t-1}F_{t-1}$$

Divide $\gamma/\hat{c}_\omega$ for both sides and then telescope from $\bar{t}_{s-1}+1$ to $\bar{t}_s$, we have:

$$\frac{\gamma}{\mu\hat{c}_\omega}\left(\frac{C_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}}-\frac{C_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}}\right) \leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{5\gamma}{\mu}\alpha_t C_t + \frac{2\gamma c_\omega^2\sigma^2}{\mu\hat{c}_\omega}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + \frac{4L^2\gamma\eta^2}{\mu\hat{c}_\omega}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t(D_t+E_t) + \frac{2L^2\gamma^3}{\mu\hat{c}_\omega}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t. \tag{41}$$

Next from Lemma 7, we write it as follows, first for $t \in [\bar{t}_{s-1}+1, \bar{t}_s-1]$, we have:

$$B_t \leq \left(1-\frac{\mu\gamma\alpha_{t-1}}{8}\right)B_{t-1} - \frac{3\gamma^2\alpha_{t-1}F_{t-1}}{4} + \frac{5\gamma\alpha_{t-1}C_{t-1}}{\mu} + \frac{10L_h^2\eta^2\alpha_{t-1}D_{t-1}}{\mu\gamma} + \frac{10L_h^2\eta^2\alpha_{t-1}E_{t-1}}{\mu\gamma}$$

and when $t = \bar{t}_s$, we have:

$$B_t \leq \left(1-\frac{\mu\gamma\alpha_{t-1}}{8}\right)B_{t-1} - \frac{3\gamma^2\alpha_{t-1}F_{t-1}}{4} + \frac{5\gamma\alpha_{t-1}C_{t-1}}{\mu} + \frac{10L_h^2\eta^2\alpha_{t-1}D_{t-1}}{\mu\gamma} + \frac{10L_h^2\eta^2\alpha_{t-1}E_{t-1}}{\mu\gamma} + (1+\frac{8}{\mu\gamma\alpha_{t-1}})IL_h^2\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2 D_l$$

We telescope from $\bar{t}_{s-1}+1$ to $\bar{t}_s$ and have:

$$B_{\bar{t}_s} - B_{\bar{t}_{s-1}} \leq -\frac{\mu\gamma}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{3\gamma^2}{4}\sum_{t=\bar{t}_{s-1}^2}^{\bar{t}_s-1}\alpha_t F_t + \frac{5\gamma}{\mu}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t + \frac{10L_h^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + \frac{10L_h^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + (1+\frac{8}{\mu\gamma\alpha_{\bar{t}_s-1}})IL_h^2\eta^2\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^2 D_t$$

Next for $\alpha_t/\alpha_{\bar{t}_s-1}$, we have:

$$\frac{\alpha_t}{\alpha_{\bar{t}_s-1}} = \frac{(u_{\bar{t}_s-1}+\sigma^2(\bar{t}_s-1))^{1/3}}{(u_t+\sigma^2 t)^{1/3}} = \left(1+\frac{u_{\bar{t}_s-1}-u_t+\sigma^2(\bar{t}_s-1-t)}{u_t+\sigma^2 t}\right)^{1/3} \leq \left(1+\frac{(I-1)\sigma^2}{u_t+\sigma^2 t}\right)^{1/3} \leq 1+\frac{(I-1)}{3(t+I+1)} \leq 2$$

where we use the condition $u_t \geq (I+1)\sigma^2$. Then for the coefficient of the last term in the above inequality, we have $(1+\frac{8}{\mu\gamma\alpha_{\bar{t}_s-1}})\alpha_t^2 = \alpha_t^2 + \frac{8\alpha_t^2}{\mu\gamma\alpha_{\bar{t}_s-1}} < \alpha_t + \frac{16\alpha_t}{\mu\gamma} = (1+\frac{16}{\mu\gamma})\alpha_t$. Finally, we have:

$$
\begin{aligned}
B_{\bar{t}_s} - B_{\bar{t}_{s-1}} \leq & -\frac{\mu\gamma}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{3\gamma^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t + \frac{5\gamma}{\mu}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t + \frac{10L_h^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \\
& + \frac{10L_h^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + (1+\frac{16}{\mu\gamma})IL_h^2\eta^2\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \\
\leq & -\frac{\mu\gamma}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{3\gamma^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t + \frac{5\gamma}{\mu}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t \\
& + \frac{10L_h^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + \left(\left(1+\frac{16}{\mu\gamma}\right)I + \frac{10}{\mu\gamma}\right)L_h^2\eta^2\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t. \qquad (42)
\end{aligned}
$$

Next we rewrite Lemma 10 as follows:

$$
\begin{aligned}
\mathbb{E}\Big[h(\bar{x}_{t+1})\Big] \leq & \mathbb{E}\Big[h(\bar{x}_t)\Big] - \left(\frac{\eta\alpha_t}{2} - \frac{\eta^2\alpha_t^2 L_h}{2}\right)E_t - \frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] \\
& + L_h^2 I\eta^3\alpha_t\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2 D_l + \eta\alpha_t A_t
\end{aligned}
$$

We telescope from $\bar{t}_{s-1}$ to $\bar{t}_s - 1$ to have:

$$
\begin{aligned}
\mathbb{E}\Big[h(\bar{x}_{\bar{t}_s}) - h(\bar{x}_{\bar{t}_{s-1}})\Big] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\left(\frac{\eta\alpha_t}{2} - \frac{\eta^2\alpha_t^2 L}{2}\right)E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] \\
& + L_h^2 I\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2 D_l + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\eta\alpha_t A_t \\
\leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\left(\frac{\eta\alpha_t}{2} - \frac{\eta^2\alpha_t^2 L}{2}\right)E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] \\
& + L_h^2 I\eta^3\left(\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^2 D_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\eta\alpha_t A_t \\
\leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\left(\frac{\eta\alpha_t}{2} - \frac{\eta^2\alpha_t^2 L}{2}\right)E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] \\
& + \frac{\eta^3}{256}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\eta\alpha_t A_t. \qquad (43)
\end{aligned}
$$

In the last inequality, we use the fact that $\bar{t}_s - \bar{t}_{s-1} \leq I$ and $\alpha_t < \frac{1}{16L_h I}$.

Recall that Potential function is defined as:

$$
\mathbb{E}[\mathcal{G}_t] = \mathbb{E}[h(\bar{x}_t)] + \frac{\eta A_t}{\hat{c}_\nu\alpha_t} + B_t + \frac{\gamma C_t}{\mu\hat{c}_\omega\alpha_t}
$$

Combine Eq. (40),Eq. (41), Eq. (42) and Eq. (43) and we have:

$$\begin{aligned}\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{-1}s}] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] \\ & + \left(\frac{2c_\omega^2\gamma\sigma^2}{\mu\hat{c}_\omega} + \frac{4\eta c_\nu^2\sigma^2}{\hat{c}_\nu M} + \frac{8\eta c_\nu^2 G^2}{\hat{c}_\nu}\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 \\ & - \left(\frac{3}{4} - \frac{2L^2\gamma}{\mu\hat{c}_\omega} - \frac{12\eta L_h^2}{\hat{c}_\nu}\right)\gamma^2\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t \\ & - \left(\frac{\eta}{2} - \frac{\eta^2\alpha_t L_h}{2} - \frac{10L_h^2\eta^2}{\mu\gamma} - \frac{4L^2\gamma\eta^2}{\mu\hat{c}_\omega} - \frac{40L_h^2\eta^3}{\hat{c}_\nu}\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t \\ & - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\left(\frac{\mu\gamma}{8} - \frac{12L_h^2c_\nu^2\eta\alpha_t^2}{\hat{c}_\nu}\right)\alpha_t B_t \\ & + \left(\frac{\eta}{256} + (1+\frac{16}{\mu\gamma})I + \frac{10L_h^2}{\mu\gamma} + \frac{4L^2\gamma}{\mu\hat{c}_\omega} + \frac{40\eta L_h^2}{\hat{c}_\nu}\right)\eta^2\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t\end{aligned}$$

Since we take $\hat{c}_\omega = 32L^2$, $\hat{c}_\nu = 160L_h^2$, $\alpha_t < \frac{1}{16L_hI}$, $\frac{1}{\gamma} = max(\frac{1}{\mu}, 6L_h)$, $\frac{1}{\eta} > max(\frac{4\gamma}{\mu} + \frac{1}{4I}, \frac{12(1+\rho^2)}{I^2} + \frac{97}{256} + \frac{\gamma^2}{2\mu^2} + (1+\frac{16}{\mu\gamma})I, \frac{240L_h^2}{\mu\gamma}, \frac{\mu}{\gamma})$, then we have:

$$\begin{aligned}\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] + \left(\frac{c_\omega^2\sigma^2}{16L^2} + \frac{3c_\nu^2\sigma^2}{80L_h^2}\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 \\ & - \frac{5}{8}\gamma^2\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t - \frac{\eta}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t - \frac{\mu\gamma}{16}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t \\ & + \left(1 - \frac{6(1+\rho^2)}{I^2}\eta^2\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \qquad (44)\end{aligned}$$

For the term related to F_t, we have:

$$\frac{3}{4} - \frac{2L^2\gamma}{\mu\hat{c}_\omega} - \frac{12\eta L_h^2}{\hat{c}_\nu} = \frac{3}{4} - \frac{\gamma}{16\mu} - \frac{3\eta}{40} \geq \frac{3}{4} - \frac{\gamma}{4\mu} \geq \frac{1}{2}$$

where we use $\eta \leq \gamma/\mu$ and $\gamma < \mu$; Next for the term related to E_t, we have:

$$\frac{\eta}{2} - \frac{\eta^2\alpha_t L_h}{2} - \frac{10L_h^2\eta^2}{\mu\gamma} - \frac{4L^2\gamma\eta^2}{\mu\hat{c}_\omega} - \frac{40L_h^2\eta^3}{\hat{c}_\nu} \geq \frac{\eta}{2} - \frac{\eta^2}{32I} - \frac{\eta}{4} - \frac{\gamma\eta^2}{2\mu} \geq \frac{\eta}{4} - \frac{\eta^2}{8}\left(\frac{1}{4I} + \frac{4\gamma}{\mu}\right) \geq \frac{\eta}{8}$$

where the first inequality is because $\hat{c}_\omega = 32L^2$, $\hat{c}_\nu = 160L_h^2$, $\alpha_t < \frac{1}{16L_hI}$, $\eta < \frac{\mu\gamma}{40L_h^2}$, the last inequality is due to $\frac{1}{\eta} \geq \frac{4\gamma}{\mu} + \frac{1}{4I}$. Next for the term related to B_t, we have:

$$\frac{\mu\gamma}{8} - \frac{12L_h^2c_\nu^2\eta\alpha_t^2}{\hat{c}_\nu} \geq \frac{\mu\gamma}{8} - \frac{160 * 24L_h^4\eta}{16^2L_h^2I^2} = \frac{\mu\gamma}{8} - \frac{15L_h^2\eta}{I^2} \geq \frac{\mu\gamma}{8} - \frac{\mu\gamma}{16I^2} \geq \frac{\mu\gamma}{16}$$

The first inequality is by $c_\nu \leq 2\hat{c}_\nu$, $\hat{c}_\nu = 160L_h^2$ and $\alpha_t < \frac{1}{16L_hI}$; the third last inequality is by $\eta < \frac{\mu\gamma}{240L_h^2}$. Lastly, for the term related to D_t, we have:

$$\begin{aligned}&\left(\frac{\eta^2}{256} + (1+\frac{16}{\mu\gamma})I + \frac{10L_h^2\eta}{\mu\gamma} + \frac{4L^2\gamma\eta}{\mu\hat{c}_\omega} + \frac{40L_h^2\eta^2}{\hat{c}_\nu}\right)\eta \\ &\quad \leq \left(\frac{\eta^2}{256} + (1+\frac{16}{\mu\gamma})I + \frac{1}{4} + \frac{\gamma\eta}{2\mu}\right)\eta \leq \left(\frac{97}{256} + \frac{\gamma^2}{2\mu^2} + (1+\frac{16}{\mu\gamma})I\right)\eta \leq 1 - \frac{12(1+\rho^2)}{I^2}\eta\end{aligned}$$

The first inequality is by $\hat{c}_\omega = 32L^2$, $\hat{c}_\nu = 160L_h^2$ and $\eta < \frac{\mu\gamma}{40L_h^2}$; the second inequality is by $\eta < 1$ and $\eta < \frac{\gamma}{\mu}$; The last inequality is by $\frac{1}{\eta} \geq \frac{12(1+\rho^2)}{I^2} + \frac{97}{256} + \frac{\gamma^2}{2\mu^2} + (1 + \frac{16}{\mu\gamma})I$. Next, by Lemma 9, we have:

$$\left(1 - \frac{3\eta^2 c_\nu^2(1+\rho^2)}{16^3 * 32I^2L_h^4}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t \leq \frac{3\eta^2}{32} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t E_t + \frac{3\gamma^2}{64} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t F_t + \left(\frac{3c_\nu^2\sigma^2}{32L_h} + \frac{3c_\nu^2 G^2}{2L_h} + \frac{3c_\nu^2\zeta^2}{16L_h}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3 \tag{45}$$

since we have:

$$\frac{c_\nu^2}{16^3 * 32L_h^4} \leq \frac{4\hat{c}_\nu^2}{16^3 * 32L_h^4} \leq \frac{4 * 320^2 L_h^4}{16^3 * 32L_h^4} = 25/8 < 4$$

The first inequality is by $c_\nu \leq 2\hat{c}_\nu$. So we have:

$$\frac{3\eta^2 c_\nu^2(1+\rho^2)}{16^3 * 64I^2L_h^4} \leq \frac{12\eta^2(1+\rho^2)}{I^2} \leq \frac{12\eta(1+\rho^2)}{I^2}$$

Combine Eq. (44) and Eq. (45) and use $\gamma < 1$ and $\eta < 1$, we have:

$$\begin{aligned}
\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2} \mathbb{E}\left[\|\nabla h(\bar{x}_t)\|^2\right] + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{40L_h^2} + \frac{c_\nu^2 G^2}{20L_h^2}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3 \\
& + \left(\frac{3c_\nu^2\sigma^2}{32L_h} + \frac{3c_\nu^2 G^2}{2L_h} + \frac{3c_\nu^2\zeta^2}{16L_h}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3 \\
& - \frac{1}{4}\gamma^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t F_t - \frac{\eta}{32} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t E_t - \frac{\mu\gamma}{16} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t B_t \\
\leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2} \mathbb{E}\left[\|\nabla h(\bar{x}_t)\|^2\right] \\
& + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3
\end{aligned} \tag{46}$$

which completes the proof.

Theorem 10.1. *Suppose* $\frac{1}{\gamma} > max(\frac{1}{\mu}, 6L_h)$, $\frac{1}{\eta} > max(\frac{4\gamma}{\mu} + \frac{1}{4I}, \frac{12(1+\rho^2)}{I^2} + \frac{97}{256} + \frac{\gamma^2}{2\mu^2} + (1 + \frac{16}{\mu\gamma})I, \frac{240L_h^2}{\mu\gamma}, \frac{\mu}{\gamma})$, $c_\nu = 160L_h^2 + \frac{\sigma^2}{24\delta^3 L_h I}$, $c_\omega = 160L^2 + \frac{\sigma^2}{24\delta^3 L_h I}$, $u = max(2\sigma^2, \delta^3, c_\nu^{3/2}\delta^3, c_\omega^{3/2}\delta^3, 16^3 I^3 M^2\sigma^2)$, $\delta = \frac{M^{2/3}\sigma^{2/3}}{L_h}$, *then we have:*

$$\frac{1}{T}\sum_{t=1}^{T-1} \mathbb{E}\left[\|\nabla h(\bar{x}_t)\|^2\right] \leq \left(\frac{2(h(\bar{x}_1) - h^*)}{\eta} + \frac{2\sigma^2 u^{1/3}}{\hat{c}_\nu\delta} + \frac{2\sigma^{8/3}}{\hat{c}_\nu\delta} + \frac{2C_{y,1}^2}{\eta} + \frac{2\gamma u^{1/3}\sigma^2}{\eta\hat{c}_\omega\delta} + \frac{2\gamma\sigma^{8/3}}{\eta\hat{c}_\omega\delta} + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\frac{2\delta^3\ln(T)}{\eta\sigma^2}\right)\left(\frac{16L_h I}{T} + \frac{L_h}{(MT)^{2/3}}\right)$$

where the expectation is w.r.t the stochasticity of the algorithm.

First, based on Lemma 11, we have:

$$\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2} \mathbb{E}\left[\|\nabla h(\bar{x}_t)\|^2\right] + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t^3$$

Next we sum for all $s \in [S]$ and assume $T = SI + 1$, we have:

$$\mathbb{E}[\mathcal{G}_T] - \mathbb{E}[\mathcal{G}_1] \leq -\sum_{t=1}^{T-1} \frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\sum_{t=1}^{T-1}\alpha_t^3$$

So we have:

$$\begin{aligned}\sum_{t=1}^{T-1} \frac{\eta\alpha_t}{2}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] &\leq \mathbb{E}[\mathcal{G}_1] - \mathbb{E}[\mathcal{G}_T] + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\sum_{t=1}^{T-1}\alpha_t^3 \\ &\leq h(\bar{x}_1) - h^* + \frac{\eta A_1}{\hat{c}_\nu\alpha_1} + B_1 + \frac{\gamma C_1}{\hat{c}_\omega\alpha_1} \\ &\quad + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\sum_{t=1}^{T-1}\alpha_t^3\end{aligned}$$

where we use $\mathcal{G}_T \geq h^*$ and h^* denotes the optimal value of h. Then for the last term:

$$\sum_{t=1}^{T}\alpha_t^3 = \sum_{t=1}^{T}\frac{\delta^3}{u+\sigma^2 t} \leq \sum_{t=1}^{T}\frac{\delta^3}{\sigma^2+\sigma^2 t} = \frac{\delta^3}{\sigma^2}\sum_{t=1}^{T}\frac{1}{1+t} \leq \frac{\delta^3}{\sigma^2}\ln(T+1). \tag{47}$$

where the first inequality follows $u_t \geq (I+1)\sigma^2 > \sigma^2$, the last inequality follows Proposition 5. Next use the fact that α_t is non-increasing, we have:

$$\begin{aligned}\frac{\eta\alpha_T}{2}\sum_{t=1}^{T-1}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] &\leq h(\bar{x}_1) - h^* + \frac{\eta A_1}{\hat{c}_\nu\alpha_1} + B_1 + \frac{\gamma C_1}{\hat{c}_\omega\alpha_1} \\ &\quad + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\frac{\delta^3}{\sigma^2}\ln(T)\end{aligned}$$

Divide both sides by $2T/\eta\alpha_T$, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] &\leq \frac{2(h(\bar{x}_1) - h^*)}{\eta\alpha_T T} + \frac{2A_1}{\hat{c}_\nu\alpha_1\alpha_T T} + \frac{2B_1}{\eta\alpha_T T} + \frac{2\gamma C_1}{\eta\hat{c}_\omega\alpha_1\alpha_T T} \\ &\quad + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\frac{2\delta^3\ln(T)}{\eta\sigma^2\alpha_T T}\end{aligned}$$

Next we have

$$A_1 = \mathbb{E}\left[\left\|\bar{\nu}_1 - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_1^{(m)})\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\left(\mathcal{G}^{(m)}(x_1^{(m)}, y_1^{(m)}; \mathcal{B}_x) - \nabla h(x_1^{(m)})\right)\right\|^2\right] \leq \sigma^2$$

and $B_1 = \frac{1}{M}\sum_{m=1}^{M}\|y_1^{(m)} - y_{x_1^{(m)}}^{(m)}\|^2 \leq C_{y,1}^2$, $C_1 = \mathbb{E}[\frac{1}{M}\sum_{m=1}^{M}\|\omega_1^{(m)} - \nabla_y g^{(m)}(x_1^{(m)}, y_1^{(m)})\|^2] \leq \sigma^2$. Then we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] &\leq \frac{2(h(\bar{x}_1) - h^*)}{\eta\alpha_T T} + \frac{2\sigma^2}{\hat{c}_\nu\alpha_1\alpha_T T} + \frac{2C_{y,1}^2}{\eta\alpha_T T} + \frac{2\gamma\sigma^2}{\eta\hat{c}_\omega\alpha_1\alpha_T T} \\ &\quad + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\frac{2\delta^3\ln(T)}{\eta\sigma^2\alpha_T T}\end{aligned}$$

Note that we have:

$$\frac{1}{\alpha_t t} = \frac{(u+\sigma^2 t)^{1/3}}{\delta t} \leq \frac{u^{1/3}}{\delta t} + \frac{\sigma^{2/3}}{\delta t^{2/3}}$$

where the inequality uses the fact that $(x+y)^{1/3} \leq x^{1/3} + y^{1/3}$. Consider $t = 1$ and $t = T$ and we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] &\leq \Bigg(\frac{2(h(\bar{x}_1)-h^*)}{\eta} + \frac{2\sigma^2 u^{1/3}}{\hat{c}_\nu \delta} + \frac{2\sigma^{8/3}}{\hat{c}_\nu \delta} + \frac{2C_{y,1}^2}{\eta} + \frac{2\gamma u^{1/3}\sigma^2}{\eta \hat{c}_\omega \delta} + \frac{2\gamma\sigma^{8/3}}{\eta\hat{c}_\omega\delta} \\ &\quad + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\frac{2\delta^3\ln(T)}{\eta\sigma^2}\Bigg)\left(\frac{u^{1/3}}{\delta T} + \frac{\sigma^{2/3}}{\delta T^{2/3}}\right) \\ &\leq \Bigg(\frac{2(h(\bar{x}_1)-h^*)}{\eta} + \frac{2\sigma^2 u^{1/3}}{\hat{c}_\nu \delta} + \frac{2\sigma^{8/3}}{\hat{c}_\nu \delta} + \frac{2C_{y,1}^2}{\eta} + \frac{2\gamma u^{1/3}\sigma^2}{\eta \hat{c}_\omega \delta} + \frac{2\gamma\sigma^{8/3}}{\eta\hat{c}_\omega\delta} \\ &\quad + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\frac{2\delta^3\ln(T)}{\eta\sigma^2}\Bigg)\left(\frac{u^{1/3}}{\delta T} + \frac{\sigma^{2/3}}{\delta T^{2/3}}\right)\end{aligned}$$

Finally, we have $c_\nu = 160L_h^2 + \frac{\sigma^2}{24\delta^3 L_h I} = 160L_h^2 + \frac{L_h^2}{24M^2 I} \leq 200L_h^2$, $c_\omega = 160L^2 + \frac{\sigma^2}{24\delta^3 L_h I} \leq 160L^2 + \frac{L_h^2}{24M^2 I} \leq 200L_h^2 \leq 200L_h^2$, $\delta^3 = \frac{M^2\sigma^2}{L_h^3} \leq M^2\sigma^2$, then we can verify that $2\sigma^2 \leq 16^3 I^3 M^2\sigma^2$, $\delta^3 \leq 16^3 I^3 M^2\sigma^2$, $c_\nu^{3/2}\delta^3 \leq 200^{3/2} M^2\sigma^2 \leq 16^3 I^3 M^2\sigma^2$ and $c_\omega^{3/2}\delta^3 \leq 200^{3/2} M^2\sigma^2 \leq 16^3 I^3 M^2\sigma^2$, in other words, we have $u \leq 16^3 I^3 M^2\sigma^2$, then we can get:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\Big[\|\nabla h(\bar{x}_t)\|^2\Big] &\leq \Bigg(\frac{2(h(\bar{x}_1)-h^*)}{\eta} + \frac{2\sigma^2 u^{1/3}}{\hat{c}_\nu \delta} + \frac{2\sigma^{8/3}}{\hat{c}_\nu \delta} + \frac{2C_{y,1}^2}{\eta} + \frac{2\gamma u^{1/3}\sigma^2}{\eta \hat{c}_\omega \delta} + \frac{2\gamma\sigma^{8/3}}{\eta\hat{c}_\omega\delta} \\ &\quad + \left(\frac{c_\omega^2\sigma^2}{16\mu L^2} + \frac{c_\nu^2\sigma^2}{5L_h^2} + \frac{2c_\nu^2 G^2}{L_h^2} + \frac{3c_\nu^2\zeta^2}{16L_h^2}\right)\frac{2\delta^3\ln(T)}{\eta\sigma^2}\Bigg)\left(\frac{16L_h I}{T} + \frac{L_h}{(MT)^{2/3}}\right)\end{aligned}$$

which completes the proof.

11 More Experimental Details

In this task, we investigate the group fairness in Federated Learning from the Bilevel Optimization's perspective.

We first introduce some notations for ease of discussion. We denote a sample as (o,s,a), where $o \in \mathbb{R}^d$, $s \in [I]$, $a \in [K]$ are input attributes, predictive attributes (we use classification as a demonstration) and sensitive attributes, respectively. Furthermore, we denote $D_t^{(m)} = \{o_i^{(m,t)}, s_i^{(m,t)}, a_i^{(m,t)}\}$ and $D_v^{(m)} = \{o_i^{(m,v)}, s_i^{(m,v)}, a_i^{(m,v)}\}$ as training and the validation set at the m_{th} client, respectively. Furthermore, we denote $n_{s,a}^{(m,v)}$ as the number of samples which has label s and sensitive attribute label a over the validation set of the m_{th} client, and we denote $n_{s,*}^{(m,v)}$ as the number of samples with label s over the validation set of the m_{th} client and $n_{*,a}^{(m,v)}$ as the number of samples with sensitive attributes a over the validation set of the m_{th} client. Similarly, we define $n_{s,a}^{(m,t)}$, $n_{s,*}^{(m,t)}$ and $n_{*,a}^{(m,t)}$ for the training set over the m_{th} client.

In our task, we assume the validation set of each client is group-balanced, *i.e.* we assume the validation sets have the follow properties: $n_{*,a_1}^{(m,v)} = n_{*,a_2}^{(m,v)}$, for $a_1, a_2 \in [K]$. Then we optimize the following objective to learn a group fair model:

$$\begin{aligned}&\min_{\omega\in\Omega}\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n^{(m,v)}}\sum_{i=1}^{n^{(m,v)}} f(\theta_\omega^{(m)}; o_i^{(m,v)}, s_i^{(m,v)}) \\ &\textit{s.t. } \theta_\omega^{(m)} = \arg\min_{\theta\in\mathbb{R}^d}\frac{1}{n^{(m,t)}}\sum_{i=1}^{n^{(m,t)}}\omega_{a_i^{(m,t)}} f(\theta; o_i^{(m,t)}, s_i^{(m,t)})\end{aligned} \tag{48}$$

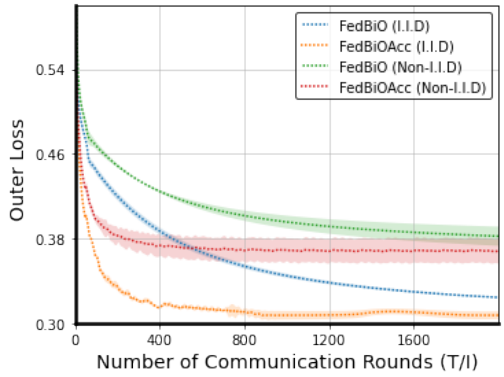

Figure 2: Outer objective Loss *w.r.t* Number of Communication Rounds for comparison between FedBiO and FedBiOAcc for I.I.D and Non-I.I.D cases. Results are for the Credit dataset.

where f denotes the model to fit, we use a two-layer fully connected neural network with L_2 regularization in the experiments. $\omega = \{\omega_a\}, a \in [K]$ are weights for sensitive groups, which correspond to the outer variable x for the problem 6 and θ denotes the model parameter and corresponds to the inner variable of the problem 6.

Intuitively, the objective Eq. 48 achieves fairness through tuning the group weights such that the learned model $\theta_\omega^{(m)}$ performs well over the validation set $D_t^{(m)}$. Since $D_t^{(m)}$ has balanced samples from all sensitive groups, the model $\theta_\omega^{(m)}$ has to perform equally well for all different groups to get small loss over the validation set.

In Table 1, we use the well known Equal Opportunity as our main fairness metrics. More precisely, it is defined as the $max_{z_1,z_2\in[K]}\|\mathbb{P}(\hat{s}=1|s=1,a=z_1)-\mathbb{P}(\hat{s}=1|s=1,a=z_2)\|$, where $\hat{s}$ is is predication made by the model, $\mathbb{P}$ is the probability notation.

For the hyper-parameters, we use $I = 5$ for FedAvg, FedReg and our algorithms, and $I = 1$ for FedMinMax and FCFL. For learning rates, we search over the range of $\{0.001, 0.01, 0.1, 1\}$, most algorithms get best performance at 0.1 or 1. Then for special hyper-parameters of each algorithms: For FedReg the regularization coefficient is set as 0.1; For FedMinMax, the stepsize for the group weights are 0.1; For FCFL, we use hyper-parameters provided from the original paper Cui et al. (2021). For our FedBiO, we set the outer learning rate η as 0.1; For our FedBiOAcc, we set δ as 0.1, u as 1 and c_ν as 1, c_ω as 1. The results reported in Table 1 are run for 2000 steps with batchsize 128 for the Adult Dataset and 32 for the Credit Dataset. The two datasets we used in experiments are widely used benchmarks for fair machine learning. More specifically, the Adult dataset aims to predict income level based on around 200 features, and the race/gender are sensitive groups. We use the race as the sensitive attribute in experiments, which include 5 different racial groups. Next the (German) Credit dataset aims to predict good and bad credit risks. the gender and marital status are sensitive attributes. Lastly, we show the convergence results of FedBiO and FedBiOAcc for the Credit dataset in Figure 2.