Communication-Efficient Federated Bilevel Optimization with Global and Local Lower Level Problems

Junyi Li
Computer Science
University of Maryland
College Park, MD 20742
junyili.ai@gmail.com

Feihu Huang
ECE
University of Pittsburgh
Pittsburgh, PA 15261
huangfeihu2018@gmail.com

Heng Huang *
Computer Science
University of Maryland
College Park, MD 20742
henghuanghh@gmail.com

Abstract

Bilevel Optimization has witnessed notable progress recently with new emerging efficient algorithms. However, its application in the Federated Learning setting remains relatively underexplored, and the impact of Federated Learning’s inherent challenges on the convergence of bilevel algorithms remain obscure. In this work, we investigate Federated Bilevel Optimization problems and propose a communication-efficient algorithm, named FedBiOAcc. The algorithm leverages an efficient estimation of the hyper-gradient in the distributed setting and utilizes the momentum-based variance-reduction acceleration. Remarkably, FedBiOAcc achieves a communication complexity \(O(\epsilon^{-1})\), a sample complexity \(O(\epsilon^{-1.5})\) and the linear speed up with respect to the number of clients. We also analyze a special case of the Federated Bilevel Optimization problems, where lower level problems are locally managed by clients. We prove that FedBiOAcc-Local, a modified version of FedBiOAcc, converges at the same rate for this type of problems. Finally, we validate the proposed algorithms through two real-world tasks: Federated Data-cleaning and Federated Hyper-representation Learning. Empirical results show superior performance of our algorithms.

1 Introduction

Bilevel optimization [54, 50] has increasingly drawn attention due to its wide-ranging applications in numerous machine learning tasks, including hyper-parameter optimization [44], meta-learning [64] and neural architecture search [38]. A bilevel optimization problem involves an upper problem and a lower problem, wherein the upper problem is a function of the minimizer of the lower problem. Recently, great progress has been made to solve this type of problems, particularly through the development of efficient single-loop algorithms that rely on diverse gradient approximation techniques [24]. However, the majority of existing bilevel optimization research concentrates on standard, non-distributed settings, and how to solve the bilevel optimization problems under distributed settings have received much less attention. Federated learning (FL) [42] is a recently promising distributed learning paradigm. In FL, a set of clients jointly solve a machine learning task under the coordination of a central server. To protect user privacy and mitigate communication overhead, clients perform multiple steps of local update before communicating with the server. A variety of algorithms [53, 62, 17, 28, 11] have been proposed to accelerate this training process. However, most of these algorithms primarily address standard single-level optimization problems. In this work, we study the bilevel optimization problems in the Federated Learning setting and investigate the

---

*This work was partially supported by NSF IIS 1838627, 1837956, 1956002, 2211492, CNS 2213701, CCF 2217003, DBI 2225775.

A federated bilevel optimization problem consists of an upper and a lower level problem, the upper problem \( f(x, y) \) of client-server communication, however, this approach leads to high communication overhead. In contrast, we view the hyper-gradient estimation as solving a quadratic federated problem and solving it with the local-sgd method. More specifically, we formulate the solution of the federated bilevel optimization as three intertwined federated problems: the upper problem, the lower problem and the quadratic problem for the hyper-gradient estimation. Then we address the three problems using alternating gradient descent steps, furthermore, to manage the noise of the stochastic gradient and obtain the fast convergence rate, we employ a momentum-based variance reduction technique.

Beyond the standard federated bilevel optimization problem as defined in Eq. 1, another variant of Federated Bilevel Optimization problem, which entails locally managed lower-level problems, is also frequently utilized in practical applications. For this type of problem, we can get an unbiased estimate of the global hyper-gradient using local hyper-gradient, thus we can solve it with a local-SGD like algorithm, named FedBiOAcc-Local. However, it is challenging to analyze the convergence of the algorithm. In particular, we need to bound the intertwined client drift error, which is intrinsic to FL and the bilevel-related errors e.g. the lower level solution bias. In fact, we prove that the FedBiOAcc-Local algorithm attains the same fast rate as FedBiO algorithm.

Finally, we highlight the main contributions of our paper as follows:

1. We propose FedBiOAcc to solve Federated Bilevel Optimization problems, the algorithm evaluates the hypergradient of federated bilevel optimization problems efficiently and achieves optimal convergence rate through momentum-based variance reduction. FedBiOAcc has sample complexity of \( O(\epsilon^{-1}p) \), communication complexity of \( O(\epsilon^{-1}) \) and achieves linear speed-up \( w.r.t \) the number of clients.

2. We study Federated Bilevel Optimization problem with local lower level problem for the first time, where we show the convergence of a modified version of FedBiOAcc, named FedBiOAcc-Local for this type of problems.

3. We validate the efficacy of the proposed FedBiOAcc algorithm through two real-world tasks: Federated Data Cleaning and Federated Hyper-representation Learning.
Notations \( \nabla \) denotes full gradient, \( \nabla_x \) denotes partial derivative for variable \( x \), higher order derivatives follow similar rules. \([K]\) represents the sequence of integers from 1 to \( K \), \( \bar{x} \) represents average of the sequence of variables \( \{x^{(m)}\}_{m=1}^{M} \). \( \bar{t}_s \) represents the global communication timestamp \( s \).

2 Related Works

Bilevel optimization dates back to at least the 1960s when [54] proposed a regularization method, and then followed by many research works [10, 50, 58, 47], while in machine learning community, similar ideas in the name of implicit differentiation were also used in Hyper-parameter Optimization [32, 3, 28]. Early algorithms for Bilevel Optimization solved the accurate solution of the lower problem for each upper variable. Recently, researchers developed algorithms that solve the lower problem with a fixed number of steps, and use the ‘back-propagation through time’ technique to compute the hyper-gradient [9, 41, 12, 45, 49]. Very Recently, it witnessed a surge of interest in using implicit differentiation to derive single loop algorithms [15, 18, 24, 30, 4, 59, 20, 34, 7, 21, 19]. In particular, [34, 7] proposes a way to iteratively evaluate the hyper-gradients to save computation. In this work, we view the hyper-gradient estimation of Federated Bilevel Optimization as solving a quadratic federated optimization problem and use a similar iterative evaluation rule as [34, 7] in local update.

The bilevel optimization problem is also considered in the more general settings. For example, bilevel optimization with multiple lower tasks is considered in [16]. Furthermore, [5, 60, 40, 14] studies the bilevel optimization problem in the decentralized setting, [26] studies the bilevel optimization problem in the asynchronous setting. In contrast, we study bilevel optimization problems under Federated Learning [42] setting. Federated learning is a promising privacy-preserving learning paradigm for distributed data. Compared to 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, various methods [28, 37, 48, 63, 43, 36] are proposed. However, bilevel optimization problems are less investigated in the federated learning setting. [57] considered the distributed bilevel formulation, but it needs to communicate the Hessian matrix for every iteration, which is computationally infeasible. More recently, FedNest [51] has been proposed to tackle the general federated nest problems, including federated bilevel problems. However, this method evaluates the full hyper-gradient at every iteration; this leads to high communication overhead; furthermore, FedNest also uses SVRG to accelerate the training. Similar works that evaluate the hyper-gradient with multiple rounds of client-server communication are [33, 22, 56, 61]. Finally, there is a concurrent work [13] that investigates the possibility of local gradients on Federated Bilevel Optimization, however, it only considers the homogeneous case, this setting is quite constrained and much simpler than the more general heterogeneous case we considered. Furthermore, [13] only considers the case where both the upper and the lower problem are federated, and omit the equally important case where the lower level problem is not federated.

3 Federated Bilevel Optimization

3.1 Some Mild Assumptions

Note that the formulation of Eq.(1) is very general, and we consider the stochastic heterogeneous case in this work. More specifically, we assume:

\[
 f^{(m)}(x, y) := \mathbb{E}_{\xi \sim \mathcal{D}_f^{(m)}}[f^{(m)}(x, y, \xi)],
 g^{(m)}(x, y) := \mathbb{E}_{\xi \sim \mathcal{D}_g^{(m)}}[g^{(m)}(x, y; \xi)]
\]

where \( \mathcal{D}_f^{(m)} \) and \( \mathcal{D}_g^{(m)} \) are some probability distributions. Furthermore, we assume the local objectives could be potentially different: \( f^{(m)}(x, y) \neq f^{(k)}(x, y) \) or \( g^{(m)}(x, y) \neq g^{(k)}(x, y) \) for \( m \neq k, m, k \in [M] \). Furthermore, we assume the following assumptions in our subsequent discussion:

**Assumption 3.1.** Function \( f^{(m)}(x, y) \) is possibly non-convex and \( g^{(m)}(x, y) \) is \( \mu \)-strongly convex w.r.t \( y \) for any given \( x \).

**Assumption 3.2.** Function \( f^{(m)}(x, y) \) is \( L \)-smooth and has \( C_f \)-bounded gradient;

**Assumption 3.3.** Function \( g^{(m)}(x, y) \) is \( L \)-smooth, and \( \nabla_{xy} g^{(m)}(x, y) \) and \( \nabla_{y^2} g^{(m)}(x, y) \) are Lipschitz continuous with constants \( L_{xy} \) and \( L_{y^2} \) respectively;
Algorithm 1: Accelerated Federated Bilevel Optimization (FedBiOAcc)

1: **Input:** Constants $c_0, c_1, c_u, \gamma, \eta, \tau, \rho$; learning rate schedule $\{\alpha_t\}, t \in [T]$, initial state $(x_1, y_1, u_1)$.

2: **Initialization:** Set $y_1^{(m)} = y_1$, $x_1^{(m)} = x_1$, $u_1^{(m)} = u_1$, $\omega_1^{(m)} = \nabla_y f^{(m)}(x_1, y_1, B_y)$, $\nu_1^{(m)} = \nabla_x f^{(m)}(x_1, y_1, B_{f,1}) - \nabla_x y_g^{(m)}(x_1, y_1, B_{f,1})u_1$ and $q_1 = \nabla_y g^{(m)}(x_1^{(m)}, y_1^{(m)}; B_{f,2}) u_1 - \nabla_y f^{(m)}(x_1^{(m)}, y_1^{(m)}; B_{f,2})$ for $m \in [M]$.

3: for $t = 1$ to $T$ do

4: for $t \mod 1 = 0$ then

5: $y_t^{(m)} = y_t$ and $x_t^{(m)} = x_t$ and $u_t^{(m)} = u_t$

6: else

7: $y_t^{(m)} = y_t^{(m)} - \gamma \alpha_t \nu_t^{(m)}$, $x_t^{(m)} = x_t^{(m)}$, $u_t^{(m)} = u_t^{(m)} + \lambda_t$

8: if $t \mod 1 = 0$ then

9: $\omega_t^{(m)} = \frac{1}{M} \sum_{j=1}^{M} \omega_t^{(j)}$, $\nu_t^{(m)} = \frac{1}{M} \sum_{j=1}^{M} \nu_t^{(j)}$

10: else

11: $\omega_t^{(m)} = \frac{1}{M} \sum_{j=1}^{M} \omega_t^{(j)}$, $\nu_t^{(m)} = \frac{1}{M} \sum_{j=1}^{M} \nu_t^{(j)}$

12: end if

13: end for

14: end if

15: end for

16: end for

Assumption 3.4. We have unbiased stochastic first-order and second-order gradient oracle with bounded variance.

Assumption 3.5. For any $m, j \in [M]$ and $z = (x, y)$, we have: $\|\nabla f^{(m)}(z) - \nabla f^{(j)}(z)\| \leq \zeta_f$, $\|\nabla g^{(m)}(z) - \nabla g^{(j)}(z)\| \leq \zeta_g$, $\|\nabla x y g^{(m)}(z) - \nabla x y g^{(j)}(z)\| \leq \zeta_{g,xy}$, $\|\nabla y g^{(m)}(z) - \nabla y g^{(j)}(z)\| \leq \zeta_{g,yg}$, where $\zeta_f, \zeta_g, \zeta_{g,xy}, \zeta_{g,yg}$, are constants.

As stated in The assumption 3.1 we study the non-convex-strongly-convex bilevel optimization problems, this class of problems is widely studied in the non-distributed bilevel literature [23,25]. Furthermore, Assumption 3.2 and Assumption 3.3 are also standard assumptions made in the non-distributed bilevel literature. Assumption 3.4 is widely used in the study of stochastic optimization problems. For Assumption 3.5 gradient difference is widely used in single level Federated Learning literature as a measure of client heterogeneity [30,55]. Please refer to the full version of Assumptions in Appendix.

3.2 The FedBiOAcc Algorithm

A major difficulty in solving a Federated Bilevel Optimization problem Eq. (1) is **evaluating the hyper-gradient $\nabla h(x)$**. For the function class (non-convex-strongly-convex) we consider, the explicit form of hypergradient $h(x)$ exists as $\nabla h(x) = \Phi(x, y, x)$, where $\Phi(x, y, x)$ is denoted as:

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

(2)

Based on Assumption 3.1-3.3, we can verify $\Phi(x, y, x)$ is the hyper-gradient [15]. But since the clients only have access to their local data, for $\forall m \in [M]$, the client evaluates:

$$\bar{\Phi}^{(m)}(x, y) = \nabla_x f^{(m)}(x, y) - \nabla_x y g^{(m)}(x, y) \times [\nabla y^2 g^{(m)}(x, y)]^{-1} \nabla_y f^{(m)}(x, y),$$

(3)

It is straightforward to verify that $\Phi^{(m)}(x, y)$ is not an unbiased estimate of the full hyper-gradient, i.e. $\Phi(x, y, x) \neq \frac{1}{M} \sum_{m=1}^{M} \Phi^{(m)}(x, y, x)$. To address this difficulty, we can view the **Hyper-gradient computation as the process of solving a federated optimization problem**.

In fact, Evaluating Eq. (2) is equivalent to the following two steps: first, we solve the quadratic federated optimization problem $l(u)$:

$$\min_{u \in \mathbb{R}^d} l(u) = \frac{1}{M} \sum_{m=1}^{M} u^T (\nabla y^2 g^{(m)}(x, y)) u - (\nabla y f^{(m)}(x, y), u)$$

(4)
Suppose we denote the solution of the above problem as $u^*$, then we have the following linear operation to get the hypergradient:

$$
\nabla h(x) = \frac{1}{M} \sum_{m=1}^{M} \left( \nabla_x f^{(m)}(x, y) - \nabla_{xy} g^{(m)}(x, y)u^* \right)
$$

(5)

Compared to the formulation Eq. (2), Eq. (4) and Eq. (5) are more suitable for the distributed setting. In fact, both Eq. (4) and Eq. (5) have a linear structure. Eq. (4) is a (single-level) quadratic federated optimization problem, and we could solve Eq. (4) through local-sgd [22], suppose that each client maintains a variable $u_t^{(m)}$, and performs the following update:

$$
u_t^{(m)} = \mathcal{P}_t(u_t^{(m)} - \tau_t \nabla l^{(m)}(u_t^{(m)}; B))
$$

$$
\nabla l^{(m)}(u_t^{(m)}; B) = \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_g, 2)u_t^{(m)} - \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_f, 2)
$$

where $\nabla l^{(m)}(u_t^{(m)}; B)$ is client $m$'s the stochastic gradient of Eq. (4), and $(x_t^{(m)}, y_t^{(m)})$ denotes the upper and lower variable state at the timestamp $t$. Note that Clients perform multiple updates of $u_t^{(m)}$ before averaging. As for Eq. (5), each client evaluates $\nabla h^{(m)}(x)$ locally: $\nabla h^{(m)}(x) = \nabla_x f^{(m)}(x, y) - \nabla_{xy} g^{(m)}(x, y)u^*$ and the server averages $\nabla h^{(m)}(x)$ to get $\nabla h(x)$. In summary, the linear structure of Eq. (4) and Eq. (5) makes it suitable for local updates, therefore, reduce the communication cost.

More specifically, we perform alternative update of upper level variable $x_{t+1}^{(m)}$, the lower level variable $y_{t+1}^{(m)}$ and hyper-gradient computation variable $u_t^{(m)}$. For example, for each client $m \in [M]$, we perform the following local updates:

$$
\begin{align*}
\dot{y}_{t+1}^{(m)} &= y_t^{(m)} - \gamma_t \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_g), \\
\dot{x}_{t+1}^{(m)} &= x_t^{(m)} - \eta_t (\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; B_f) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_g, 1))u_t^{(m)}
\end{align*}
$$

(6)

Every $I$ steps, the server averages clients’ local states, this resembles the local-sgd method for single level federated optimization problems. Note that in the update of the upper variable $x_{t+1}^{(m)}$, we use $u_t^{(m)}$ as an estimation of $u^*$ in Eq. (5). An algorithm follows Eq. (6) is shown in Algorithm 2 of Appendix and we refer to it as FedBio.

**Comparison with FedNest.** The update rule of Eq. (6) is very different from that of FedNest [51] and its follow-ups [22, 56]. In FedNest, a sub-routine named FedHGDP is used to evaluate Eq. (2) at every global epoch. This involves multiple rounds of client-server communication and leads to higher communication overhead. In contrast, Eq. (6) formulates the hyper-gradient estimation as an quadratic federated optimization problem, and then solves three intertwined federated problems through alternative updates of $x, y$ and $u$.

Note that Eq. (6) updates the related variables through vanilla gradient descent steps. In the non-federated setting, gradient-based methods such as stocBiO [24] requires large batch size $(O(\epsilon^{-1}))$ to reach an $\epsilon$-stationary point, and we also analyze Algorithm 2 in Appendix to show the same dependence. To control the noise and remove the dependence over large batch size, we apply the momentum-based variance-reduction technique STORM [6]. In fact, Eq. (6) solves three intertwined optimization problems: the bilevel problem $h(x)$, the lower level problem $g(x, y)$ and the hyper-gradient computation problem Eq. (4). So we control the noise in the process of solving each of the three problems. More specifically, we have $\omega_t^{(m)}, \nu_t^{(m)}$ and $\tilde{q}_t^{(m)}$ to be the momentum estimator for $x_{t+1}^{(m)}, y_{t+1}^{(m)}$ and $u_t^{(m)}$, respectively, and we update them following the rule of STORM [6]:

$$
\dot{\omega}_{t+1}^{(m)} = \nabla_y g^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}, B_g) + (1 - c_w \alpha_t^2) (\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_g))
$$

$$
\dot{\nu}_{t+1}^{(m)} = (\nabla_x f^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; B_f) - \nabla_{xy} g^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; B_g, 1))u_t^{(m)} + (1 - c_w \alpha_t^2) (\nu_t^{(m)} - (\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; B_f) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_g, 1))u_t^{(m)})
$$

$$
\dot{\tilde{q}}_{t+1}^{(m)} = (\nabla y g^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; B_g, 2))\nu_{t+1}^{(m)} - \nabla f^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; B_f) + (1 - c_w \alpha_t^2) (\tilde{q}_t^{(m)} - (\nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_g, 2))\nu_t^{(m)} - \nabla f^{(m)}(x_t^{(m)}, y_t^{(m)}; B_f, 2))
$$

(7)
where \( c_\omega, c_\nu, c_u \) are constants, \( \alpha_t \) is the learning rate. Then we update the \( x_t^{(m)}, y_t^{(m)} \) and \( u_t^{(m)} \) as follows:

\[
\begin{align*}
\bar{y}_{t+1}^{(m)} &= y_t^{(m)} - \gamma \alpha_t \omega_t^{(m)}, \\
\bar{x}_{t+1}^{(m)} &= x_t^{(m)} - \eta \alpha_t \nu_t^{(m)}, \\
\bar{u}_{t+1} &= P_{\Omega}(u_t^{(m)} - \tau \alpha_t \omega_t^{(m)})
\end{align*}
\]

where \( \gamma, \eta, \tau \) are constants and \( \alpha_t \) is the learning rate. The FedBiOAcc algorithm following Eq. (6) is summarized in Algorithm 1. As shown in line 6 and 12 of Algorithm 1, every \( I \) iterations, we average both variables and the momentum.

### 3.3 Convergence Analysis

In this section, we study the convergence property for the FedBiOAcc algorithm. For any \( t \in [T] \), we define the following virtual sequence:

\[
\begin{align*}
\bar{x}_t &= \frac{1}{M} \sum_{m=1}^{M} x_t^{(m)}, \\
\bar{y}_t &= \frac{1}{M} \sum_{m=1}^{M} y_t^{(m)}, \\
\bar{u}_t &= \frac{1}{M} \sum_{m=1}^{M} u_t^{(m)}
\end{align*}
\]

de note the average of the momentum similarly as \( \bar{\omega}_t, \bar{\nu}_t \) and \( \bar{\eta}_t \). Then we consider the following Lyapunov function \( G_t \):

\[
\begin{align*}
G_t &= h(\bar{x}_t) + \frac{18 \eta L^2}{\mu \gamma} \left( \| \bar{y}_t - \bar{x}_t \|^2 + \| \bar{u}_t - \bar{y}_t \|^2 \right) + \frac{9 b M \eta}{64 \alpha_t} \| \bar{\omega}_t - \frac{1}{M} \sum_{m=1}^{M} \nabla_y g(m) (x_t^{(m)}, y_t^{(m)}) \|^2 \\
&\quad + \frac{9 b M \eta}{64 \alpha_t} \| \bar{\nu}_t - \frac{1}{M} \sum_{m=1}^{M} \left( \nabla_x f(m) (x_t^{(m)}, y_t^{(m)}) - \nabla_y g(m) (x_t^{(m)}, y_t^{(m)}) \right) u_t^{(m)} \|^2 \\
&\quad + \frac{9 b M \eta}{64 \alpha_t} \| \bar{\eta}_t - \frac{1}{M} \sum_{m=1}^{M} \nabla_x f(m) (x_t^{(m)}, y_t^{(m)}) \|^2
\end{align*}
\]

where \( \bar{y}_t, \bar{u}_t, \bar{y}_t \) denotes the solution of the lower level problem \( g(\bar{x}_t, \cdot), \bar{u}_t, \bar{y}_t = [\nabla_y g(\bar{x}_t, \bar{y}_t)]^{-1} \nabla_y f(\bar{x}_t, \bar{y}_t) \) denotes the solution of Eq. (4) at state \( \bar{x}_t \). Besides, \( \gamma, \eta, \tau \) are learning rates and \( L, \bar{L} \) are constants. Note that the first three terms of \( G_t \): \( h(\bar{x}_t), \| \bar{y}_t - \bar{x}_t \|^2, \| \bar{u}_t - \bar{y}_t \|^2 \) measures the errors of three federated problems: the upper level problem, the lower level problem and the hyper-gradient estimation. Then the last three terms measure the estimation error of the momentum variables: \( \bar{\omega}_t, \bar{\nu}_t \) and \( \bar{\eta}_t \). The convergence proof primarily concentrates on bounding these errors, please see Lemma C.2 - C.6 in the Appendix for more details. Meanwhile, as in the single level federated optimization problems, local updates lead to client-drift error. More specifically, we need to bound \( \| x_t^{(m)} - \bar{x}_t \|^2, \| y_t^{(m)} - \bar{y}_t \|^2 \) and \( \| u_t^{(m)} - \bar{u}_t \|^2 \), please see Lemma C.7 - C.11 for more details. Finally, we have the following convergence theorem:

**Theorem 3.6.** Suppose in Algorithm 1, we choose learning rate \( \alpha_t = \frac{\delta}{(u_t + 1)^{3/2}}, t \in [T], \) for some constant \( \delta \) and \( u \), and let \( c_\omega, c_\nu, c_u \) choose some value, \( \eta, \gamma, \nu, \tau \) be some small values decided by the Lipschitz constants of \( h(x) \), we choose the minibatch size to be \( b_x = b_y = b \) and the first batch to be \( b_1 = O(1b) \), then we have:

\[
\frac{1}{T} \sum_{t=1}^{T-1} E[\| \nabla h(\bar{x}_t) \|^2] = O\left(\frac{\kappa^{19/3} I}{T} + \frac{\kappa^{16/3}}{(bMT)^{2/3}}\right)
\]

To reach an \( \epsilon \)-stationary point, we need \( T = O(\kappa^8 (bM)^{-1} \epsilon^{-1.5}, I = O(\kappa^{5/3} (bM)^{-1} \epsilon^{-0.5}) \)

As stated in the Theorem, to reach an \( \epsilon \)-stationary point, we need \( T = O(\kappa^8 (bM)^{-1} \epsilon^{-1.5}) \), then the sample complexity for each client is \( Gc(f, e) = O(M^{-1}\kappa^8 \epsilon^{-1.5}) \), \( Gc(g, e) = O(M^{-1}\kappa^8 \epsilon^{-1.5}) \), \( Jv(g, e) = O(M^{-1}\kappa^8 \epsilon^{-1.5}) \), \( Hv(g, e) = O(M^{-1}\kappa^8 \epsilon^{-1.5}) \). So FedBiOAcc achieves the linear speedup w.r.t. to the number of clients \( M \). Next, suppose we choose \( I = O(\kappa^{5/3} (bM)^{-1} \epsilon^{-0.5}) \), then the number of communication round \( E = O(\kappa^{19/3} \epsilon^{-1}) \). This matches the optimal communication complexity of the single level optimization problems as in the STEM [29]. Furthermore, compared to FedNest and its variants, FedBiOAcc has improved both the communication complexity and the iteration complexity. As for LocalBSCVR [13], FedBiOAcc obtains same rate, but incorporates the heterogeneous case. Note that it is much more challenging to analyze the heterogeneous case. In fact, if we assume homogeneous clients, we have local hyper-gradient (Eq. (3)) equals the global hyper-gradient (Eq. (3)), then we do not need to use the quadratic federated optimization problem view in Section 3.2, while the theoretical analysis is also simplified significantly.
4 Federated Bilevel Optimization with Local Lower Level Problems

In this section, we consider an alternative formulation of the Federated Bilevel Optimization problems as follows:

$$\min_{x \in \mathbb{R}^p} h(x) := \frac{1}{M} \sum_{m=1}^{M} f^{(m)}(x, y_{x}^{(m)}), \text{ s.t. } y_{x}^{(m)} = \arg \min_{y \in \mathbb{R}^d} g^{(m)}(x, y)$$  \hspace{1cm} (10)

Same as Eq. (1), Eq. (10) has a federated upper level problem, however, Eq. (10) has a unique lower level problem for each client, which is different from Eq. (1). In fact, federated bilevel optimization problem Eq (10) can be viewed as a special type of standard federated learning problems. If we denote $h^{(m)}(x) = f^{(m)}(x, y_{x}^{(m)})$, then Eq. (10) can be written as $\min_{x \in \mathbb{R}^p} h(x) := \frac{1}{M} \sum_{m=1}^{M} h^{(m)}(x)$.

But due to the bilevel structure of $h^{(m)}(x)$, Eq. (10) is more challenging than the standard Federated Learning problems.

Hyper-gradient Estimation. Assume Assumption 3.1–Assumption 3.3 hold, then the hyper-gradient is $\Phi(x, y_{x}) = \frac{1}{M} \sum_{m=1}^{M} \Phi^{(m)}(x, y_{x})$, where $\Phi^{(m)}(x, y)$ is defined in Eq. (3). In other words, the local hyper-gradient $\Phi^{(m)}(x, y)$ is an unbiased estimate of the full hyper-gradient. This fact makes it possible to solve Eq. (10) with local-sgd like methods. More specifically, we solve the local bilevel problem $h^{(m)}(x)$ multiple steps on each client and then the server averages the local states from clients. Please refer to Algorithm 3 and the variance-reduction acceleration Algorithm 4 in the Appendix. For ease of reference, we name them FedBiO-Local and FedBiOAcc-Local, respectively.

Several challenges exist in analyzing FedBiO-Local and FedBiOAcc-Local. First, Eq. (1) involves Hessian inverse, so we only evaluate it approximately through the Neumann series [39] as:

$$\Phi^{(m)}(x, y; \xi_{x}) = \nabla_x f^{(m)}(x, y; \xi_{f}) - \tau \nabla_{xy} g^{(m)}(x, y; \xi_{g})$$

$$\times \sum_{q=1}^{Q-1} \prod_{j=q}^{Q} (I - \tau \nabla_{y} g^{(m)}(x, y_{j}; \xi_{j})) \nabla_y f^{(m)}(x, y; \xi_{f})$$  \hspace{1cm} (11)

where $\xi_{x} = \{\xi_{j}(j = 1, \ldots, Q), \xi_{f}, \xi_{g}\}$, and we assume its elements are mutually independent. $\Phi^{(m)}(x, y; \xi_{x})$ is a biased estimate of $\Phi^{(m)}(x, y)$, but with bounded bias and variance (Please see Proposition D.2 for more details.) Furthermore, to reduce the computation cost, each client solves the local lower level problem approximately and we update the upper and lower level variable alternatively.

The idea of alternative update is widely used in the non-distributed bilevel optimization [24, 59]. However, in the federated setting, client variables drift away when performing multiple local steps. As a result, the variable drift error and the bias caused by inexact solution of the lower level problem intertwined with each other. For example, in the local update, clients optimize the lower level variable $y_{x}^{(m)}$ towards the minimizer $y_{x}^{(m)}$, but after the communication step, $x^{(m)}$ is smoothed among clients, as a result, the target of $y_{x}^{(m)}$ changes which causes a huge bias.

In the appendix, we show the FedBiOAcc-Local algorithm achieves the same optimal convergence rate as FedBiOAcc, which has iteration complexity $O(\epsilon^{-1.5})$ and communication complexity $O(\epsilon^{-1})$. However, since the lower level problem in Eq. (10) is unique for each client, FedBiOAcc-Local does not have the property of linear speed-up with the number of clients as FedBiOAcc does.

5 Numerical Experiments

In this section, we assess the performance of the proposed FedBiOAcc algorithm through two federated bilevel tasks: Federated Data Cleaning and Federated Hyper-representation Learning. The Federated Data Cleaning task involves global lower level problems, while the Hyper-representation Learning task involves local lower level problems. The implementation is carried out using PyTorch, and the Federated Learning environment is simulated using the PyTorch.Distributed package. Our experiments were conducted on servers equipped with an AMD EPYC 7763 64-core CPU and 8 NVIDIA V100 GPUs.
5.1 Federated Data Cleaning

In this section, we consider the Federated Data Cleaning task. In this task, we are given a noisy training dataset whose labels are corrupted by noise and a clean validation set. Then we aim to find weights for training samples such that a model that is learned over the weighted training set performs well on the validation set. This is a federated bilevel problem when the noisy training set is distributed over multiple clients. The formulation of the task is included in Appendix B.1. This task is a specialization of Eq. (1).

Dataset and Baselines. We create 10 clients and construct datasets based on MNIST [33]. For the training set, each client randomly samples 4500 images (no overlap among clients) from 10 classes and then randomly uniformly perturb the labels of $\rho$ ($0 \leq \rho \leq 1$) percent samples. For the validation set, each client randomly selects 50 clean images from a different class. In other words, the $m_{th}$ client only has validation samples from the $m_{th}$ class. This single-class validation setting introduces a high level of heterogeneity, such that individual clients are unable to conduct local cleaning due to they only have clean samples from one class. In our experiments, we test our FedBiOAcc algorithm, including the FedBiO algorithm (Algorithm 2 in Appendix) which does not use variance reduction; additionally, we also consider some baseline methods: a baseline that directly performs FedAvg [42] on the noisy dataset, this helps to verify the usefulness of data cleaning; Local-BSGVR [13], FedNest [51], CommFedBiO [55], AggITD [56] and FedMBO [22]. Note that Local-BSGVR is designed for the homogeneous setting, and the last four baselines all need multiple rounds of client-server communication to evaluate the hyper-gradient at each global epoch. We perform grid search to find the best hyper-parameters for each method and report the best results. Specific choices are included in Appendix B.1.

In figure 1, we compare the performance of different methods at various noise levels $\rho$. Note that the larger the $\rho$ value, the more noisy the training data are. The noise level can be illustrated by the performance of the FedAvg algorithm, which learns over the noisy data directly. As shown in the figure, FedAvg learns almost nothing when $\rho = 0.95$. Next, our algorithms are robust under various heterogeneity levels. When the noise level in the training set increases as the value of $\rho$ increases, learning relies more on the signal from the heterogeneous validation set, and our algorithms consistently outperform other baselines. Finally, in figure 2, we vary the number of clients sampled per epoch, and the experimental results show that our FedBiOAcc converges faster with more clients in the training per epoch; in figure 3, we vary the number of local steps under different noisy levels. Interestingly, the algorithm benefit more from the local training under larger noise.
5.2 Federated Hyper-Representation Learning

In the Hyper-representation learning task, we learn a hyper-representation of the data such that a linear classifier can be learned quickly with a small number of data samples. A mathematical formulation of the task is included in Appendix B.2. Note that this task is an instantiation of Eq. (10), due to the fact that each client has its own tasks, and thus only the upper level problem is federated. We consider the Omniglot [31] and MiniImageNet [46] data sets. As in the non-distributed setting, we perform \( N \)-way-\( K \)-shot classification.

In this experiment, we compare FedBiOAcc-Local (Algorithm 4 in the Appendix) with three baselines FedBiO-Local (Algorithm 3 in the Appendix), DistBiO and DistBiOAcc. Note that DistBiO and DistBiOAcc are the distributed version of FedBiO-Local and FedBiOAcc-Local, respectively. In the experiments, we implement DistBiO and DistBiOAcc by setting the local steps as 1 for FedBiO-Local and FedBiOAcc-Local. We perform grid search for the hyper-parameter selection for both methods and choose the best ones, the specific choices of hyper-parameters are deferred to Appendix B.2. The results are summarized in Figure 4 (full results are included in Figure 5 and Figure 6 of Appendix. As shown by the results, FedBiOAcc converges faster than the baselines on both datasets and on all four types of classification tasks, which demonstrates the effectiveness of variance reduction and multiple steps of local training.

6 Conclusion

In this paper, we study the Federated Bilevel Optimization problems and introduce FedBiOAcc. In particular, FedBiOAcc evaluates the hyper-gradient by solving a federated quadratic problem, and mitigates the noise through momentum-based variance reduction technique. We provide a rigorous convergence analysis for our proposed method and show that FedBiOAcc has the optimal iteration complexity \( O(\epsilon^{-1.5}) \) and communication complexity \( O(\epsilon^{-1}) \), and it also achieves linear speed-up \( w.r.t \) the number of clients. Besides, we study a type of novel Federated Bilevel Optimization problems with local lower level problems. We modify FedBiO for this type of problems and propose
FedBiOAcc-Local. FedBiOAcc-Local achieves the same optimal convergence rate as FedBiOAcc. Finally, we validate our algorithms with real-world tasks.

References


A Assumptions

In this section, we restate all assumptions needed in our proof below:

**Assumption A.1** (Assumption 1). The function $f^{(m)}(x, y)$ is possibly non-convex and $g^{(m)}(x, y)$ is $\mu$-strongly convex w.r.t $y$ for any given $x$, i.e. for any $y_1, y_2 \in \mathbb{R}^d$, we have:

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

**Assumption A.2** (Assumption 2). Function $f^{(m)}(x, y)$ is $L$-Lipschitz, i.e. for any $x_1, x_2 \in \mathcal{X}$ and for any $y_1, y_2 \in \mathbb{R}^d$, we denote $z_1 = (x_1, y_1), z_2 = (x_2, y_2)$, then we have:

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

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

**Assumption A.3** (Assumption 3). Function $g^{(m)}(x, y)$ is $L$-Lipschitz. i.e. for any $x_1, x_2 \in \mathcal{X}$ and for any $y_1, y_2 \in \mathbb{R}^d$, we denote $z_1 = (x_1, y_1), z_2 = (x_2, y_2)$, then we have:

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

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

- a) $\nabla_{xy} g^{(m)}(x, y)$ and $\nabla_{y^2} g^{(m)}(x, y)$ are Lipschitz continuous with constant $L_{xy}$ and $L_{y^2}$ respectively, i.e. 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: $\|\nabla_{xy} g^{(m)}(z_1) - \nabla_{xy} g^{(m)}(z_2)\| \leq L_{xy} \|z_1 - z_2\|$ and $\|\nabla_{y^2} g^{(m)}(z_1) - \nabla_{y^2} g^{(m)}(z_2)\| \leq L_{y^2} \|z_1 - z_2\|$.

**Assumption A.4** (Assumption 4). We have an unbiased stochastic first order and second order derivative oracle with bounded variance, more specifically, denote $z = (x, y)$, we have:

- a) we have $\nabla f^{(m)}(z; \xi)$, such that: $E[\nabla f^{(m)}(z; \xi)] = \nabla f^{(m)}(z)$ and $\text{var}(\nabla f^{(m)}(z; \xi)) \leq \sigma^2$.
- b) we have $\nabla g^{(m)}(z; \xi)$, such that: $E[\nabla g^{(m)}(z; \xi)] = \nabla g^{(m)}(z)$ and $\text{var}(\nabla g^{(m)}(z; \xi)) \leq \sigma^2$.
- c) we have $\nabla y^2 g^{(m)}(z; \xi)$, such that: $E[\nabla y^2 g^{(m)}(z; \xi)] = \nabla y^2 g^{(m)}(z)$ and $\text{var}(\nabla y^2 g^{(m)}(z; \xi)) \leq \sigma^2$.
- d) we have $\nabla_{xy} g^{(m)}(z; \xi)$, such that: $E[\nabla_{xy} g^{(m)}(z; \xi)] = \nabla_{xy} g^{(m)}(z)$ and $\text{var}(\nabla_{xy} g^{(m)}(z; x, y; \xi)) \leq \sigma^2$.

**Assumption A.5** (Assumption 5). For any $m, j \in [M]$ and $z = (x, y)$, we have: $\|\nabla f^{(m)}(z) - \nabla f^{(j)}(z)\| \leq \zeta_f$, $\|\nabla g^{(m)}(z) - \nabla g^{(j)}(z)\| \leq \zeta_g$, $\|\nabla_{xy} g^{(m)}(z) - \nabla_{xy} g^{(j)}(z)\| \leq \zeta_{g, xy}$, $\|\nabla y^2 g^{(m)}(z) - \nabla y^2 g^{(j)}(z)\| \leq \zeta_{g, y^2}$, $\|\nabla_{xy} g^{(m)}(z) - \nabla_{xy} g^{(j)}(z)\| \leq \zeta_{g, y^2}$, where $\zeta_f, \zeta_g, \zeta_{g, xy}, \zeta_{g, y^2}$, are constants.

**Assumption A.6** (Assumption 6). For any $m, j \in [M]$ and $z = (x, y)$, we have: $\|\nabla f^{(m)}(z) - \nabla f^{(j)}(z)\| \leq \zeta_f$, $\|\nabla_{xy} g^{(m)}(z) - \nabla_{xy} g^{(j)}(z)\| \leq \zeta_{g, xy}$, $\|\nabla y^2 g^{(m)}(z) - \nabla y^2 g^{(j)}(z)\| \leq \zeta_{g, y^2}$, $\|\nabla_{xy} g^{(m)}(z) - \nabla_{xy} g^{(j)}(z)\| \leq \zeta_{g, y^2}$, where $\zeta_f, \zeta_{g, xy}, \zeta_{g, y^2}$, are constants.
B More Experimental Details and Results

In this section, we introduce more details of the experiments.

B.1 Federated Data Cleaning

The formulation of the problem is as follows:

\[
\min_{x \in \mathbb{R}^p} h(x) := \frac{1}{M} \sum_{m=1}^{M} f^{(m)}(x, y^{(m)}_x) = \frac{1}{M} \sum_{m=1}^{M} \left( \frac{1}{N_m^{(val)}} \sum_{n=1}^{N_m^{(val)}} \Theta(y_{x}^{(m,n)}; \xi_{m,n}^{(val)}) \right)
\]

s.t. \( y_x = \arg \min_{y \in \mathbb{R}^d} g(x, y) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N_m^{(tr)}} x_{m,n} \Theta(y; \xi_{m,n}^{(tr)}) \)

In the above formulation, we have \( M \) clients, each client \( m \in [M] \) has a pair of (noisy) training set \( \{ \xi_{m,n}^{(tr)} \}_{n=1}^{N_m^{(tr)}} \) and validation set \( \{ \xi_{m,n}^{(val)} \}_{n=1}^{N_m^{(val)}} \), and \( x_{m,n} \in [N_m^{(tr)}] \) are weights for training samples, \( y \) is the parameter of a model, and we denote the model by \( \Theta \). Note that \( y_x \) is the model learned over the weighted training set. We fit a model with 3 fully connected layers for the MNIST dataset. We also use \( L_2 \) regularization with coefficient \( 10^{-3} \) to satisfy the strong convexity condition.

In the Experiments, for FedNest and CommFedBiO, we choose learning rate 1 and hyper-learning rate 10000, for FedBiO, we choose learning rate 0.5, hyper learning rate 0.1, \( \gamma \) as 0.5, for FedBiOAcc, we choose learning rate 0.1, hyper learning rate 1000, for FedBiOAcc, we choose learning rate 0.05, hyper learning rate 0.01, \( \eta \) as 200 and \( \gamma \) as 1.

B.2 Federated Hyper-Representation Learning

\[
\min_{x \in \mathbb{R}^p} h(x) := \frac{1}{M} \sum_{m=1}^{M} f^{(m)}(x, y^{(m)}_x) = \frac{1}{M} \sum_{m=1}^{M} \left( \frac{1}{N_m^{(tr)}} \sum_{n=1}^{N_m^{(tr)}} \left( \frac{1}{N_m^{(val)}} \sum_{i=1}^{N_m^{(val)}} \Theta(x, y_{x}^{(m,n)}, \xi_{i}^{(val)}) \right) \right)
\]

s.t. \( y_x^{(T_m,n)} = \arg \min_{y \in \mathbb{R}^d} g^{(T_m,n)}(x, y) = \frac{1}{N_{T_m,n}} \sum_{i=1}^{N_{T_m,n}} \Theta(x, y; \xi_{i}^{(tr)}) \)

In the above formulation, we have \( M \) clients, each client \( m \in [M] \) has \( N_m \) tasks and each task \( T_{m,n} \) is defined by a pair of training set \( \{ \xi_{i}^{(tr)} \}_{i=1}^{N_{T_m,n}} \) and validation set \( \{ \xi_{i}^{(val)} \}_{i=1}^{N_{T_m,n}} \). \( \Theta \) defines the model, \( x \) is the parameter of the backbone model and \( y \) is the parameter of the linear classifier. In summary, the lower level problem is to learn the optimal linear classifier \( y \) given the backbone \( x \), and the upper level problem is to learn the optimal backbone parameter \( x \).

The Omniglot dataset includes 1623 characters from 50 different alphabets and each character consists of 20 samples. We create the Federated version of the Omniglot dataset. Firstly, we follow the experimental protocols of [52] to divide the alphabets to train/validation/test with 33/5/12, respectively. Then we distribute three alphabets to a client, in other words, we consider 11 clients in experiments. As in the non-distributed setting, we perform \( N \)-way-\( K \)-shot classification, more specifically, for each task, we randomly sample \( N \) characters from the alphabet over that client and for each character, we sample \( K \) samples for training and 15 samples for validation. We augment the characters by performing rotation operations (multipliers of 90 degrees). We use a 4-layer convolutional neural network where each convolutional layer has 64 filters of \( 3 \times 3 \) [11]. For the MiniImageNet, it has 64 training classes and 16 validation classes. We distribute the training classes into four clients, similar to Omniglot, we also perform the \( N \)-way-\( K \)-shot classification. We use a 4-layer convolutional neural network where each convolutional layer has 64 filters of \( 3 \times 3 \) [11] for experiments.

In the Experiments for Omniglot, for FedBiO, we choose learning rate 0.4, hyper learning rate 1, \( \tau \) 0.5, for FedBiOAcc, we choose \( \delta \) as 2, \( u \) as 10000, \( C_\eta \) as 100, \( \tau \) as 0.5, \( \epsilon_{ta} \) as 1 and \( \gamma \) as 0.4. For MiniImageNet, for FedBiO, we choose learning rate 0.05, hyper learning rate 0.1, \( \tau \) 0.01, for FedBiOAcc, we choose \( \delta \) as 2, \( u \) as 10000, \( C_\eta \) as 100, \( \tau \) as 0.01, \( \epsilon_{ta} \) as 1 and \( \gamma \) as 0.05.
C Proof for Global Lower Level Problem

This section includes proofs related to the Federated Bilevel Optimization problems with global lower level problems (Eq. 1). First, we have the global and local hyper-gradient \( \nabla h(x) = \Phi(x, y_x) \), \( \nabla h(m)(x) = \Phi(m)(x, y_x) \) as defined in Eq. 2 and Eq. 3, and the following proposition:

**Proposition C.1.** Suppose Assumptions 3.2 and 3.3 hold, the following statements hold:

a) \( y_x \) is Lipschitz continuous in \( x \) with constant \( \rho = \kappa \), where \( \kappa = \frac{L}{\mu} \) is the condition number of \( g(x, y_x) \).

b) \( \| \Phi(x_1; y_1) - \Phi(x_2; y_2) \|^2 \leq \hat{L}^2 (\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2) \), where \( \hat{L} = O(\kappa^2) \).

c) \( h(x) \) is Lipschitz continuous in \( x \) with constant \( \hat{L} \) i.e., for any given \( x_1, x_2 \in X \), we have \( \| \nabla h(x_2) - \nabla h(x_1) \| \leq \hat{L} \|x_2 - x_1\| \) where \( \hat{L} = O(\kappa^3) \).

This is a standard results in bilevel optimization and we omit the proof here.

C.1 Proof for the FedBiOAcc Algorithm

In this section, we prove the convergence of the FedBiOAcc Algorithm. To simplify the notation, we denote

\[
\begin{align*}
\mu^{(m)}_{t, \xi} &= \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{f, 1}) - \nabla_x g^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{g, 1}) u_t^{(m)}, \\
\mathbb{E}_\xi[\mu^{(m)}_{t, \xi}] &= \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_x g^{(m)}(x_t^{(m)}, y_t^{(m)}) u_t^{(m)}
\end{align*}
\]

Figure 5: Results for the Omniglot Dataset. From Left to Right: 5-way-1-shot, 5-way-5-shot, 20-way-1-shot, 20-way-5-shot.
Figure 6: Results for the MiniImageNet Dataset. From Left to Right: 5-way-1-shot, 5-way-5-shot, 20-way-1-shot, 20-way-5-shot.

where the expectation is \( \text{w.r.t} \{\xi_{f,1}, \xi_{g,1}\} \) at iteration \( t \), we denote \( \mu^{(m)}_t = \mathbb{E}_{\xi} [\mu^{(m)}_{t,\xi}] \) for short. Similarly, we denote

\[
p^{(m)}_{t,\xi} = \nabla y \cdot g^{(m)}(x^{(m)}_t, y^{(m)}_t, \xi_{g,2}) u^{(m)}_t + \nabla y \cdot f^{(m)}(x^{(m)}_t, y^{(m)}_t, \xi_{f,2}),
\]

and we have:

\[
\mathbb{E}_{\xi} [p^{(m)}_{t,\xi}] = \nabla y \cdot g^{(m)}(x^{(m)}_t, y^{(m)}_t) u^{(m)}_t + \nabla y \cdot f^{(m)}(x^{(m)}_t, y^{(m)}_t).
\]

where the expectation is \( \text{w.r.t} \{\xi_{f,2}, \xi_{g,2}\} \) at iteration \( t \), we denote \( \bar{p}^{(m)}_t = \mathbb{E}_{\xi} \bar{p}^{(m)}_{t,\xi} \) for short.

### C.1.1 Hyper-Gradient Bias and Inner-Gradient Bias

**Lemma C.2.** Suppose we have \( c_u \alpha_t^2 < 1 \), then we have:

\[
\mathbb{E} \left[ \| \bar{q}_t - \bar{p}_t \|^2 \right] \leq (1 - c_u \alpha_{t-1}^2) \mathbb{E} \left[ \| \bar{q}_{t-1} - \bar{p}_{t-1} \|^2 \right] + \frac{2(c_u \alpha_{t-1}^2)^2}{b_x M} \sigma^2 \\
+ \frac{4\tilde{L}_2^2}{b_x M^2} \sum_{m=1}^{M} \mathbb{E} \left[ \| x^{(m)}_t - x^{(m)}_{t-1} \|^2 + \| y^{(m)}_t - y^{(m)}_{t-1} \|^2 \right] \\
+ \frac{8L^2}{b_x M^2} \sum_{m=1}^{M} \mathbb{E} \left[ \| u^{(m)}_t - u^{(m)}_{t-1} \|^2 \right]
\]

where \( \tilde{L}_2^2 = \left( L^2 + \frac{2L^2 C^2}{\mu^2} \right) \) and the expectation outside is \( \text{w.r.t} \) all the stochasity of the algorithm.
Proof. First, we have:

\[
\mathbb{E}[\|\tilde{v}_t - \tilde{\mu}_t\|^2] = \mathbb{E}[\|\tilde{p}_{t,B_x} + (1 - c_v\alpha_{t-1}^2)(\tilde{p}_{t-1,B_x} - \tilde{\mu}_t)\|^2] \\
= \mathbb{E}[(1 - c_v\alpha_{t-1}^2)(\tilde{p}_{t-1} - \tilde{\mu}_t) + (\tilde{p}_{t,B_x} - \tilde{\mu}_t + (1 - c_v\alpha_{t-1}^2)(\tilde{p}_{t-1,B_x} - \tilde{\mu}_t,B_x))]^2 \\
\leq (1 - c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{p}_{t-1} - \tilde{\mu}_t\|^2] + \mathbb{E}[\|\tilde{p}_{t,B_x} - \tilde{\mu}_t + (1 - c_v\alpha_{t-1}^2)(\tilde{p}_{t-1,B_x} - \tilde{\mu}_t,B_x)\|^2] \\
\leq (1 - c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{p}_{t-1} - \tilde{\mu}_t\|^2] + \frac{1}{b_x^2 M^2} \sum_{m=1}^M \sum_{\xi_x \in B_x} \mathbb{E}[\|\tilde{p}_{t,\xi_x} - \tilde{p}_t + (1 - c_v\alpha_{t-1}^2)(\tilde{p}_{t-1,\xi_x} - \tilde{p}_t)\|^2]
\]

where the first inequality uses the fact that the cross product term is zero in expectation, the condition that \(c_v\alpha_t^2 < 1\) and the second inequality follows that samples are independent among clients. We denote the second term of above as \(T_1\), then we have:

\[
T_1 \leq 2(c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{p}_{t,\xi_x} - \tilde{p}_t\|^2] + 2(1 - c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{p}_{t-1,\xi_x} - \tilde{p}_{t-1}\|^2] \\
\leq 2(c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{p}_{t,\xi_x} - \tilde{p}_t\|^2] + 2\mathbb{E}[\|\tilde{p}_{t-1,\xi_x} - \tilde{p}_{t-1}\|^2]
\]

where inequality (a) follows the generalized triangle inequality; (b) and the bounded variance assumption. We denote the second term above as \(T_1,2\), we have:

\[
T_{1,2} = 2\mathbb{E}[\|\nabla_y g(m)(x_t,\xi_{\mathcal{S}_2})u_t + \nabla_y f(m)(x_t,\xi_{\mathcal{S}_2})u_t - (\nabla_y g(m)(x_{t-1},\xi_{\mathcal{S}_2})u_{t-1} + \nabla_y f(m)(x_{t-1},\xi_{\mathcal{S}_2}))\|^2] \\
\leq 4\mathbb{E}[\|\nabla_y g(m)(x_t,\xi_{\mathcal{S}_2}) - \nabla_y g(m)(x_{t-1},\xi_{\mathcal{S}_2})\|^2] + 4\mathbb{E}[\|\nabla_y f(m)(x_t,\xi_{\mathcal{S}_2}) - \nabla_y f(m)(x_{t-1},\xi_{\mathcal{S}_2})\|^2] \\
\leq 4(L^2 + \frac{2L^2\alpha_{t-1}^2}{\mu^2})\mathbb{E}[\|x_t - x_{t-1}\|^2 + \|y_t - y_{t-1}\|^2] + 8L^2\mathbb{E}[\|u_t - u_{t-1}\|^2]
\]

Combine everything together finishes the proof. \(\square\)

Lemma C.3. Suppose we have \(c_v\alpha_t^2 < 1\), then we have:

\[
\mathbb{E}[\|\tilde{v}_t - \tilde{\mu}_t\|^2] \leq (1 - c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{\mu}_{t-1} - \tilde{\mu}_{t-1}\|^2] + \frac{2(c_v\alpha_{t-1}^2)^2}{b_x M} \sigma^2 \\
+ \frac{4\tilde{L}_1^2}{b_x^2 M^2} \sum_{m=1}^M \mathbb{E}[\|x_t - x_{t-1}\|^2 + \|y_t - y_{t-1}\|^2] \\
+ \frac{8L^2}{b_x^2 M^2} \sum_{m=1}^M \mathbb{E}[\|u_t - u_{t-1}\|^2]
\]

where \(\tilde{L}_1^2 = (L^2 + \frac{2L^2\alpha_{t-1}^2}{\mu^2})\) and the expectation outside is w.r.t all the stochasticity of the algorithm.

Proof. First, we have:

\[
\mathbb{E}[\|\tilde{v}_t - \tilde{\mu}_t\|^2] = \mathbb{E}[\|\tilde{v}_{t,B_x} + (1 - c_v\alpha_{t-1}^2)(\tilde{v}_{t-1,B_x} - \tilde{\mu}_t)\|^2] \\
= \mathbb{E}[(1 - c_v\alpha_{t-1}^2)(\tilde{v}_{t-1} - \tilde{\mu}_t) + (\tilde{v}_{t,B_x} - \tilde{\mu}_t + (1 - c_v\alpha_{t-1}^2)(\tilde{v}_{t-1,B_x} - \tilde{\mu}_t,B_x))]^2 \\
\leq (1 - c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{v}_{t-1} - \tilde{\mu}_t\|^2] + \mathbb{E}[\|\tilde{v}_{t,B_x} - \tilde{\mu}_t + (1 - c_v\alpha_{t-1}^2)(\tilde{v}_{t-1,B_x} - \tilde{\mu}_t,B_x)\|^2] \\
\leq (1 - c_v\alpha_{t-1}^2)\mathbb{E}[\|\tilde{v}_{t-1} - \tilde{\mu}_t\|^2] + \frac{1}{b_x^2 M^2} \sum_{m=1}^M \sum_{\xi_x \in B_x} \mathbb{E}[\|\tilde{v}_{t,\xi_x} - \tilde{v}_t + (1 - c_v\alpha_{t-1}^2)(\tilde{v}_{t-1,\xi_x} - \tilde{v}_t)\|^2]
\]

where the first inequality uses the fact that the cross product term is zero in expectation, the condition that \(c_v\alpha_t^2 < 1\) and the second inequality follows that samples are independent among clients. We
denote the second term of above as $T_1$, then we have:

$$T_1 \leq (a) 2(c_\nu \alpha^2_{t-1})^2 \mathbb{E} \left[ \| \mu^{(m)}_t - \mu^{(m)}_{t-1} \|^2 \right] + (b) 2(1 - c_\nu \alpha^2_{t-1})^2 \mathbb{E} \left[ \| \mu^{(m)}_t - \mu^{(m)}_{t-1} \|^2 - (\mu^{(m)}_t - \mu^{(m)}_{t-1}) \right]$$

where inequality (a) follows the generalized triangle inequality; (b) and the bounded variance assumption. We denote the second term of above as $T_{1,2}$, we have:

$$T_{1,2} = 2 \mathbb{E} \left[ \| \nabla_x f(x_t^{(m)}, y_t^{(m)}; B_{f,1}) - \nabla_x g(x_t^{(m)}, y_t^{(m)}; B_{g,1}) u_{t}^{(m)} \right]$$

$$= \left[ (1 - c_\nu \alpha^2_{t-1}) \mathbb{E} \left[ \| \omega_{t-1} - \frac{1}{M} \sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right] + \frac{2(c_\omega \alpha^2_{t-1})^2 \sigma^2}{b_y M^2} 
+ 2L^2 \sum_{m=1}^M \mathbb{E} \left[ \| x_t^{(m)} - x_{t-1}^{(m)} \|^2 + \| y_t^{(m)} - y_{t-1}^{(m)} \|^2 \right] \right]$$

Combine everything together finishes the proof.

**Lemma C.4.** Suppose we have $c_\omega \alpha^2_{t-1} < 1$, then for $t \neq \bar{t}$, with $s \in [S]$, we have:

$$\mathbb{E} \left[ \| \bar{\omega}_t - \frac{1}{M} \sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right]$$

$$\leq (1 - c_\omega \alpha^2_{t-1}) \mathbb{E} \left[ \| \omega_{t-1} - \frac{1}{M} \sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right] + \frac{2(c_\omega \alpha^2_{t-1})^2 \sigma^2}{b_y M^2} 
+ 2L^2 \sum_{m=1}^M \mathbb{E} \left[ \| x_t^{(m)} - x_{t-1}^{(m)} \|^2 + \| y_t^{(m)} - y_{t-1}^{(m)} \|^2 \right]$$

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

**Proof.** First, we have:

$$\mathbb{E} \left[ \| \bar{\omega}_t - \frac{1}{M} \sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right]$$

$$= \mathbb{E} \left[ \frac{1}{M} \sum_{m=1}^M \left( \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_y) + (1 - c_\omega \alpha^2_{t-1})(\omega_{t-1} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_y)) \right) - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \right]^2 \right]$$

$$= \mathbb{E} \left[ \| (1 - c_\omega \alpha^2_{t-1}) \omega_{t-1} - \frac{1}{M} \sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right]$$

$$+ \frac{1}{M} \sum_{m=1}^M \left( \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}; B_y) - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \right)$$

$$+ (1 - c_\omega \alpha^2_{t-1}) \left( \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \right) \right]^2 \right]$$

$$\leq (1 - c_\omega \alpha^2_{t-1}) \mathbb{E} \left[ \| \omega_{t-1} - \frac{1}{M} \sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right]$$

$$+ \frac{1}{b_y M^2} \sum_{m=1}^M \mathbb{E} \left[ \| \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right]$$

$$+ (1 - c_\omega \alpha^2_{t-1}) \left( \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \right) \right]^2 \right]$$

$$+ \frac{1}{b_y M^2} \sum_{m=1}^M \mathbb{E} \left[ \| \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2 \right]$$

$$+ (1 - c_\omega \alpha^2_{t-1}) \left( \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \right) \right]^2 \right]$$

19
where inequality (a) uses the fact that the cross product term is zero in expectation and the condition that $c_w \alpha^2_t < 1$, $t \in [T]$, furthermore, the samples are sampled independently on clients.

We denote the second term in the above inequality as $T_1$, we have:

$$
T_1^{(b)} \leq 2(c_w \alpha^2_t - 1)^2 \mathbb{E} \left[ \left\| \nabla \gamma (m) (x_t(m), y_t(m), \xi_y) - \nabla \gamma (m) (x_t(m), y_t(m)) \right\|^2 \right] \\
+ 2 \left( 1 - c_w \alpha^2_t - 1 \right)^2 \mathbb{E} \left[ \left\| - \nabla \gamma (m) (x_t(m), y_t(m)) \right\| \right]
$$

$$
+ \nabla \gamma (m) (x_t(m), y_t(m), \xi_y) + \nabla \gamma (m) (x_{t-1}(m), y_{t-1}(m)) - \nabla \gamma (m) (x_t(m), y_t(m), \xi_y) \right\| \right]^2
$$

$$
\leq 2(c_w \alpha^2_t - 1)^2 \sigma^2 + 2 \mathbb{E} \left[ \left\| \nabla \gamma (m) (x_t(m), y_t(m), \xi_y) - \nabla \gamma (m) (x_{t-1}(m), y_{t-1}(m), \xi_y) \right\|^2 \right]
$$

inequality (b) uses the generalized triangle inequality; inequality (c) follows the bounded variance assumption [3.4], Proposition [E.2], inequality (d) uses the smoothness assumption [3.3].

C.1.2 Lower Problem Solution Error

**Lemma C.5.** Suppose we choose $\gamma \leq \frac{1}{\sqrt{2L}}$ and $\alpha_t < 1$. Then for $t \in [T]$, we have:

$$\left\| \bar{y}_{t+1} - y_{\bar{x}_{t+1}} \right\|^2 \leq \left( 1 - \frac{\mu^2 \gamma \alpha_t}{4} \right) \left\| \bar{y}_t - y_{\bar{x}_t} \right\|^2 - \frac{\gamma^2 \alpha_t}{4} \left\| \bar{\omega}_t \right\|^2 + \frac{9 \gamma^2 \eta^2 \alpha_t \mu}{2 \gamma \alpha_t} \left\| \bar{v}_t \right\|^2 \\
+ \frac{9 \gamma \alpha_t L^2}{\mu M} \sum_{m=1}^{M} \left( \left\| x_t(m) - \bar{x}_t \right\|^2 + \left\| y_t(m) - \bar{y}_t \right\|^2 \right) + \frac{9 \gamma \alpha_t}{\mu M} \sum_{m=1}^{M} \left( \left\| \nabla \gamma (m) (x_t(m), y_t(m)) - \bar{w}_t \right\|^2 \right)
$$

**Proof.** First, we exploit Proposition [E.5] and choose the function $g(\bar{x}_t, \cdot)$, by assumption it is $L$ smooth and $\mu$ strongly convex, and we choose $\gamma \leq \frac{1}{\sqrt{2L}}$ and $\alpha_t < 1$, thus:

$$\left\| \bar{y}_{t+1} - y_{\bar{x}_t} \right\|^2 \leq \left( 1 - \frac{\mu^2 \gamma \alpha_t}{4} \right) \left\| \bar{y}_t - y_{\bar{x}_t} \right\|^2 - \frac{\gamma^2 \alpha_t}{4} \left\| \bar{\omega}_t \right\|^2 + \frac{4 \gamma \alpha_t L^2}{\mu M} \sum_{m=1}^{M} \left( \left\| x_t(m) - \bar{x}_t \right\|^2 + \left\| y_t(m) - \bar{y}_t \right\|^2 \right) + \frac{9 \gamma \alpha_t}{\mu M} \sum_{m=1}^{M} \left( \left\| \nabla \gamma (m) (x_t(m), y_t(m)) - \bar{w}_t \right\|^2 \right). \quad (12)
$$

Next, we decompose the term $\left\| \bar{y}_{t+1} - y_{\bar{x}_{t+1}} \right\|^2$ as follows:

$$\left\| \bar{y}_{t+1} - y_{\bar{x}_{t+1}} \right\|^2 \leq \left( 1 + \frac{\mu^2 \gamma \alpha_t}{4} \right) \left\| \bar{y}_{t+1} - y_{\bar{x}_{t+1}} \right\|^2 + \left( 1 + \frac{4 \gamma \alpha_t}{\mu M} \right) \left\| y_{\bar{x}_t} - y_{\bar{x}_{t+1}} \right\|^2 \\
\leq \left( 1 + \frac{\mu^2 \gamma \alpha_t}{4} \right) \left\| \bar{y}_{t+1} - y_{\bar{x}_t} \right\|^2 + \left( 1 + \frac{4 \gamma \alpha_t}{\mu M} \right) \left\| y_{\bar{x}_t} - y_{\bar{x}_{t+1}} \right\|^2 \quad (13)
$$

where the second inequality is due to case a) of Proposition 3.9. Combining the above inequalities [12] and [13] we have

$$\left\| \bar{y}_{t+1} - y_{\bar{x}_{t+1}} \right\|^2 \leq \left( 1 + \frac{\mu^2 \gamma \alpha_t}{4} \right) \left( 1 - \frac{\mu^2 \gamma \alpha_t}{2} \right) \left\| \bar{y}_t - y_{\bar{x}_t} \right\|^2 + \left( 1 + \frac{\mu^2 \gamma \alpha_t}{4} \right) \left\| \bar{\omega}_t \right\|^2 \\
+ \left( 1 + \frac{\mu^2 \gamma \alpha_t}{4} \right) \frac{4 \gamma \alpha_t}{\mu M} \left\| \nabla \gamma (m) (x_t(m), y_t(m)) - \bar{w}_t \right\|^2 + \left( 1 + \frac{4 \gamma \alpha_t}{\mu M} \right) \left\| y_{\bar{x}_t} - y_{\bar{x}_{t+1}} \right\|^2
$$

Since we choose $\gamma \leq \frac{1}{\sqrt{2L}}$, $\alpha_t < 1$, we have:

$$\left( 1 + \frac{\mu^2 \gamma \alpha_t}{4} \right) \left( 1 - \frac{\mu^2 \gamma \alpha_t}{2} \right) = 1 - \mu^2 \gamma^2 \alpha_t^2 \frac{8}{4} - 1 \leq -1 \frac{\mu \gamma \alpha_t}{4}
$$

and $-1 \left( 1 + \frac{\mu \gamma \alpha_t}{4} \right) \leq -1, \left( 1 + \frac{\mu \gamma \alpha_t}{4} \right) \leq \frac{9}{8}, \mu \gamma \alpha_t < \frac{1}{2}$. Thus, we have

$$\left\| \bar{y}_{t+1} - y_{\bar{x}_{t+1}} \right\|^2 \leq \left( 1 - \frac{\mu^2 \gamma \alpha_t}{4} \right) \left\| \bar{y}_t - y_{\bar{x}_t} \right\|^2 - \frac{\gamma^2 \alpha_t}{4} \left\| \bar{\omega}_t \right\|^2 + \frac{9 \gamma \alpha_t}{2 \mu} \left( \frac{\nabla \gamma (m) (x_t(m), y_t(m)) - \bar{w}_t}{T_i} \right) + \frac{9 \gamma^2 \eta^2 \alpha_t}{2 \mu} \left\| \bar{v}_t \right\|^2
$$

20
For the term $T_1$ in the inequality above, we have:
\[
\|\nabla_y g(\tilde{x}_t, \tilde{y}_t) - \bar{u}_t\|^2 \leq 2\|\nabla_y g(\tilde{x}_t, \tilde{y}_t) - \frac{1}{M} \sum_{m=1}^{M} \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\|^2 \\
+ 2\|\frac{1}{M} \sum_{m=1}^{M} \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \bar{w}_t\|^2 \\
\leq 2L^2 \frac{M}{M} \sum_{m=1}^{M} [\|x_t^{(m)} - \tilde{x}_t\|^2 + \|y_t^{(m)} - \tilde{y}_t\|^2] \\
+ 2\|\frac{1}{M} \sum_{m=1}^{M} \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \bar{w}_t\|^2
\]
This completes the proof.

\[\square\]

**Lemma C.6.** Suppose we choose $\tau \leq \frac{1}{2L}$ and $\alpha_t < 1$, $r = \frac{C_t}{\mu}$. Then for $t \in [T]$, we have:
\[
\|\bar{u}_{t+1} - u_{\tilde{x}_{t+1}}\|^2 \leq (1 - \frac{\mu T \alpha_t}{4})\|\bar{u}_t - u_{\tilde{x}_t}\|^2 - \frac{\tau^2 \alpha_t}{4} \|\tilde{q}_t\|^2 + \frac{9\alpha_t^2 \eta^2 \alpha_t}{2\mu} \|\tilde{\nu}_t\|^2 + \frac{9\alpha_t}{\mu} \|\bar{p} - \tilde{q}_t\|^2 \\
+ \frac{18\alpha_t L^2}{\mu M} \sum_{m=1}^{M} [\|x_t^{(m)} - \tilde{x}_t\|^2 + \|y_t^{(m)} - \tilde{y}_t\|^2] + \frac{18\alpha_t L^2}{M} \sum_{m=1}^{M} \|u_t^{(m)} - \bar{u}_t\|^2
\]
where $\tilde{L}_2^2 = (L^2 + \frac{2L^2 \eta^2}{\mu})$ is a constant.

**Proof.** First, we exploit Proposition E.5 and choose the function $\frac{1}{2} x^T \nabla_y g(x, y) x - \nabla_y f(\tilde{x}, \tilde{y})^T x$, by assumption it is $L$ smooth and $\mu$ strongly convex, and we choose $\tau < \frac{1}{2L}$ and $\alpha_t < 1$, thus:
\[
\|\bar{u}_{t+1} - u_{\tilde{x}_t}\|^2 \leq (1 - \frac{\mu T \alpha_t}{2})\|\bar{u}_t - u_{\tilde{x}_t}\|^2 - \frac{\tau^2 \alpha_t}{4} \|\tilde{q}_t\|^2 + \frac{4\alpha_t}{\mu} \|\nabla_y g(\tilde{x}, \tilde{y}) \bar{u}_t - \nabla_y f(\tilde{x}, \tilde{y}) - \tilde{q}_t\|^2.
\]

where we also use the fact that
\[
\|\bar{u}_{t+1} - u_{\tilde{x}_t}\|^2 \leq \|\tilde{u}_t - \tau \alpha_t \tilde{q}_t - u_{\tilde{x}_t}\|^2
\]
for $r = \frac{C_t}{\mu} \geq \|u_{\tilde{x}_t}\|$. Next, we decompose the term $\|\bar{u}_{t+1} - u_{\tilde{x}_{t+1}}\|^2$ as follows:
\[
\|\bar{u}_{t+1} - u_{\tilde{x}_{t+1}}\|^2 \leq (1 + \frac{\mu T \alpha_t}{4})\|\bar{u}_{t+1} - u_{\tilde{x}_t}\|^2 + (1 + \frac{4}{\mu T \alpha_t})\|u_{\tilde{x}_t} - u_{\tilde{x}_{t+1}}\|^2 \\
\leq (1 + \frac{\mu T \alpha_t}{4})\|\bar{u}_{t+1} - u_{\tilde{x}_t}\|^2 + (1 + \frac{4}{\mu T \alpha_t})\|\tilde{L}_2\|\|\tilde{x}_t - \tilde{x}_{t+1}\|^2
\]
where the second inequality is due to case a) of Proposition 3.9. Combining the above inequalities [12] and [13] we have:
\[
\|\bar{u}_{t+1} - u_{\tilde{x}_{t+1}}\|^2 \leq (1 - \frac{\mu T \alpha_t}{4})\|\bar{u}_t - u_{\tilde{x}_t}\|^2 - \frac{\tau^2 \alpha_t}{4} \|\tilde{q}_t\|^2 \\
+ \frac{9\alpha_t}{2\mu} \|\nabla_y g(\tilde{x}, \tilde{y}) \bar{u}_t - \nabla_y f(\tilde{x}, \tilde{y}) - \tilde{q}_t\|^2 + \frac{9\alpha_t^2 \eta^2 \alpha_t}{2\mu} \|\tilde{\nu}_t\|^2
\]
where we use the fact that $\tau \leq \frac{1}{2L}$, $\alpha_t < 1$. For the term $T_1$ in the inequality above, we have:
\[
T_1 \leq 2\|\nabla_y g(\tilde{x}, \tilde{y}) \bar{u}_t - \nabla_y f(\tilde{x}, \tilde{y}) - \tilde{p}_t\|^2 + 2\|\tilde{p}_t - \tilde{q}_t\|^2 \\
\leq 2\|\nabla_y g(\tilde{x}, \tilde{y}) \bar{u}_t - \nabla_y f(\tilde{x}, \tilde{y}) \|
\]
We denote the first term of the above inequality as $T_{1,1}$, we have:

\[ T_{1,1} \leq 4 \left\| \nabla_y f(\bar{x}, y) - \frac{1}{M} \sum_{m=1}^{M} \left( \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}) \right) \right\|^2 + 4 \left\| \nabla_y g(\bar{x}, y) \bar{u}_t - \frac{1}{M} \sum_{m=1}^{M} \left( \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) u_t^{(m)} \right) \right\|^2 \]

\[ \leq \left( \frac{4L^2}{M} + \frac{8L^2 C_f^2}{\mu^2 M} \right) \sum_{m=1}^{M} \left[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 + \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right] + \frac{8L^2}{M} \sum_{m=1}^{M} \left\| u_t^{(m)} - \bar{u}_t \right\|^2 \]

Combine everything completes the proof. \hfill \Box

### C.1.3 Upper Variable Drift

**Lemma C.7.** For any $t \neq \bar{t}$, $s \in [S]$, we have:

\[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 \leq I\eta^2 \sum_{\ell = \bar{t} - 1}^{t-1} \alpha_\ell^2 \left\| \nu_\ell^{(m)} - \nu_\ell \right\|^2 \]

\[ \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \leq I\gamma^2 \sum_{\ell = \bar{t} - 1}^{t-1} \alpha_\ell^2 \left\| \omega_\ell^{(m)} - \omega_\ell \right\|^2 \]

\[ \left\| u_t^{(m)} - \bar{u}_t \right\|^2 \leq I\tau^2 \sum_{\ell = \bar{t} - 1}^{t-1} \alpha_\ell^2 \left\| q_\ell^{(m)} - q_\ell \right\|^2 \]

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

\[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 = \left\| x_{t-1}^{(m)} - \bar{x}_{t-1} - \left( \sum_{\ell = \bar{t}_s - 1}^{t-1} \eta \alpha_\ell \nu_\ell^{(m)} - \sum_{\ell = \bar{t}_s - 1}^{t-1} \eta \alpha_\ell \nu_\ell \right) \right\|^2 = \left\| \sum_{\ell = \bar{t}_s - 1}^{t-1} \eta \alpha_\ell \left( \nu_\ell^{(m)} - \nu_\ell \right) \right\|^2 \]

\[ \leq I\eta^2 \sum_{\ell = \bar{t}_s - 1}^{t-1} \alpha_\ell^2 \left\| \nu_\ell^{(m)} - \nu_\ell \right\|^2 \]

We can derive the bound for $\left\| y_t^{(m)} - \bar{y}_t \right\|^2$ and $\left\| u_t^{(m)} - \bar{u}_t \right\|^2$ similarly. This completes the proof. \hfill \Box

**Lemma C.8.** Suppose $\eta \alpha_t < \frac{1}{16L_M}$, then for $t \neq \bar{t}_s$, $s \in [S]$, we have:

\[ \sum_{m=1}^{M} \mathbb{E}\left\| \nu_t^{(m)} - \nu_t \right\|^2 \]

\[ \leq \left( 1 + \frac{17}{16I} \right) \sum_{m=1}^{M} \mathbb{E}\left\| \nu_{t-1}^{(m)} - \nu_{t-1} \right\|^2 + 8I\bar{L}_t^2 \alpha_t^2 \sum_{m=1}^{M} \mathbb{E}\left[ 2\left\| \nu_t \nu_{t-1} \right\|^2 + \left\| \omega_{t-1}^{(m)} \right\|^2 \right] + 16IL^2 \alpha_t^2 \sum_{m=1}^{M} \mathbb{E}\left[ \tau q_{t-1}^{(m)} \right] \]

\[ + 128I(c_v \alpha_t^2)^2 \bar{L}^2 \sum_{m=1}^{M} \mathbb{E}\left[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 + \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right] + 32I(c_v \alpha_t^2)^2 \bar{L}^2 \sum_{m=1}^{M} \mathbb{E}\left\| u_t^{(m)} - \bar{u}_t \right\|^2 \]

\[ + 8IM(c_v \alpha_t^2)^2 \bar{L}^2 \sum_{m=1}^{M} \mathbb{E}\left[ \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right] + 64I(c_v \alpha_t^2)^2 M \frac{C_f^2 \bar{L}^2 \bar{y}}{\mu^2} + \frac{C_f^2 \bar{L}^2 \bar{y}}{\mu^2} \]

where the expectation is w.r.t the stochasticity of the algorithm.
Proof. For $t \neq \bar{t}$, we have:

$$
\mathbb{E}[\mu_{t,B}^{(m)} - \hat{\mu}_{t,B}]^2 = \mathbb{E}[(1 - c_0\alpha_t^2)(\mu_{t-1,B}^{(m)} - \mu_{t-1,B}\bar{\mu}_{t-1,B}) - \mu_{t,B} - (1 - c_0\alpha_t^2)(\hat{\mu}_{t-1,B} - \bar{\mu}_{t-1,B})]^2
$$

$$
= \mathbb{E}[(1 - c_0\alpha_t^2)(\mu_{t-1,B}^{(m)} - \hat{\mu}_{t,B}) + \mu_{t,B} - (1 - c_0\alpha_t^2)(\mu_{t,B} - \bar{\mu}_{t-1,B})]^2
$$

$$
\leq (1 + \frac{1}{\varpi})((1 - c_0\alpha_t^2)^2 \mathbb{E}[\mu_{t-1,B}^{(m)} - \hat{\mu}_{t,B}]^2 + (1 + \varpi)\mathbb{E}[(\mu_{t,B}^{(m)} - \mu_{t,B}) - (1 - c_0\alpha_t^2)(\mu_{t,B} - \bar{\mu}_{t-1,B})]^2
$$

$$
\leq \left(1 + \frac{1}{\varpi}\right)\mathbb{E}[\mu_{t-1,B}^{(m)} - \hat{\mu}_{t,B}]^2 + (1 + \varpi)\mathbb{E}[(\mu_{t,B}^{(m)} - \mu_{t,B}) - (1 - c_0\alpha_t^2)(\mu_{t,B} - \bar{\mu}_{t-1,B})]^2
$$

(14)

where (a) follows from the the generalized triangle inequality.

Next we bound the second term of the above inequality:

$$
\sum_{m=1}^{M} \mathbb{E}[\mu_{t,B}^{(m)} - \mu_{t,B} - (1 - c_0\alpha_t^2)(\mu_{t-1,B}^{(m)} - \mu_{t-1,B})]^2
$$

$$
\leq 2 \sum_{m=1}^{M} \mathbb{E}[\mu_{t,B}^{(m)} - \mu_{t,B} - (1 - c_0\alpha_t^2)(\mu_{t-1,B}^{(m)} - \mu_{t-1,B})]^2 + 2(c_0\alpha_t^2)^2 \sum_{m=1}^{M} \mathbb{E}[\mu_{t-1,B}^{(m)} - \mu_{t-1,B}]^2
$$

where the inequality follows the triangle inequality. We bound the two terms separately, for the first term, we have:

$$
\sum_{m=1}^{M} \mathbb{E}[\mu_{t,B}^{(m)} - \mu_{t,B} - (1 - c_0\alpha_t^2)(\mu_{t-1,B}^{(m)} - \mu_{t-1,B})]^2 \leq \sum_{m=1}^{M} \mathbb{E}[\mu_{t,B}^{(m)} - \mu_{t,B}]^2
$$

$$
\leq \sum_{m=1}^{M} \mathbb{E}[\nabla_x f^{(m)}(x_{t-1,B}^{(m)}, y_{t-1,B}^{(m)}; \xi_{t-1,B}) - \nabla_x g^{(m)}(x_{t-1,B}^{(m)}, y_{t-1,B}^{(m)}; \xi_{t-1,B})]u_t^{(m)}
$$

$$
- \nabla_x f^{(m)}(x_{t-1,B}^{(m)}, y_{t-1,B}^{(m)}; \xi_{t-1,B}) - \nabla_x g^{(m)}(x_{t-1,B}^{(m)}, y_{t-1,B}^{(m)}; \xi_{t-1,B})]u_t^{(m)}]^2
$$

$$
\leq 2(L^2 + \frac{2L^2 \varpi^2}{\mu_t^2}) \sum_{m=1}^{M} \mathbb{E}[\|x_t^{(m)} - x_{t-1}^{(m)}\|^2 + \|y_t^{(m)} - y_{t-1}^{(m)}\|^2] + 4L^2 \sum_{m=1}^{M} \mathbb{E}[u_t^{(m)} - u_{t-1}^{(m)}]^2
$$

$$
\leq 2L^2 \alpha_t^2 \sum_{m=1}^{M} \mathbb{E}[\|\eta_t^{(m)}\|^2 + \|\gamma_t^{(m)}\|^2] + 4L^2 \alpha_t^2 \sum_{m=1}^{M} \mathbb{E}[\gamma_t^{(m)}]\|\|^2
$$

(15)

where (a) follows Proposition E.2, (b) follows Proposition C.1 and the fact that $x_t^{(m)} = x_{t-1}^{(m)}$ when $t \neq \bar{t}$; Next for the second term, we have:

$$
\sum_{m=1}^{M} \mathbb{E}[\mu_{t-1,B}^{(m)} - \mu_{t-1,B}]^2 = \sum_{m=1}^{M} \mathbb{E}[\mu_{t-1,B}^{(m)} - \mu_{t-1,B} - (\mu_{t-1,B}^{(m)} - \mu_{t-1,B})] + \mu_{t-1,B}^{(m)} - \mu_{t-1,B}]^2
$$

$$
\leq 2 \sum_{m=1}^{M} \mathbb{E}[\mu_{t-1,B}^{(m)} - \mu_{t-1,B} - (\mu_{t-1,B}^{(m)} - \mu_{t-1,B})] + \mu_{t-1,B}^{(m)} - \mu_{t-1,B}]^2 + 2 \sum_{m=1}^{M} \mathbb{E}[\mu_{t-1,B}^{(m)} - \mu_{t-1,B}]^2
$$

$$
\leq 2 \sum_{m=1}^{M} \mathbb{E}[\mu_{t-1,B}^{(m)} - \mu_{t-1,B}]^2 + 2 \sum_{m=1}^{M} \mathbb{E}[\mu_{t-1,B}^{(m)} - \mu_{t-1,B}]^2
$$

(16)
Note for the term $T_1$ of Eq. 16, we have $\mathbb{E}\|\mu_{t-1,b_s}^{(m)} - \mu_{t-1}^{(m)}\|^2 \leq \frac{\epsilon^2}{b_s}$ by the bounded variance assumption; Next for the term $T_2$, we have:

$$T_2 = \sum_{m=1}^{M} \left\| \nabla_x f^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_x g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) u_{t-1}^{(m)} \right\|
$$

$$- \frac{1}{M} \sum_{j=1}^{M} \left( \nabla_x f^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}) - \nabla_x g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}) u_{t-1}^{(j)} \right)^2$$

$$\leq 16(L^2 + \frac{2L^2 C_f^2}{\mu^2}) \sum_{m=1}^{M} \mathbb{E}\|x_t^{(m)} - x_t\|^2 + \|y_t^{(m)} - y_t\|^2 + 4L^2 \sum_{m=1}^{M} \mathbb{E}\|u_t^{(m)} - \bar{u}_t\|^2 + 4M \zeta_f^2 + \frac{8MC_g^2 \zeta_g^2 \epsilon}{\mu^2}$$

Finally, combine Eq. 15, Eq. 16 with Eq. 14 and use the fact that $I \geq 1$, we have:

$$\sum_{m=1}^{M} \mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$$

$$\leq \left(1 + \frac{1}{I} + 16I \tilde{L}_2^2 \eta^2 \alpha_{t-1}^2 \right) \sum_{m=1}^{M} \mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$$

$$+ 8I \tilde{L}_1^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\|\eta \nu_t^{(m)}\|^2 + \|\gamma \omega_t^{(m)}\|^2 + 16IL^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\|\tau q_t^{(m)}\|$$

$$+ 128I(c_r \alpha_{t-1}^2) \tilde{L}_1^2 \sum_{m=1}^{M} \mathbb{E}\|x_t^{(m)} - x_t\|^2 + \|y_t^{(m)} - y_t\|^2 + 32I(c_r \alpha_{t-1}^2)^2 L^2 \sum_{m=1}^{M} \mathbb{E}\|u_t^{(m)} - \bar{u}_t\|^2$$

$$+ 8IM(c_r \alpha_{t-1}^2) \frac{\sigma^2}{b_x} + 32IM(c_r \alpha_{t-1}^2)^2 \zeta_f^2 + 64I(c_r \alpha_{t-1}^2)^2 M \frac{C_g^2 \zeta_g^2 \epsilon}{\mu^2}$$

We separate the term $T_1$ with triangle inequality to get:

$$\sum_{m=1}^{M} \mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$$

$$\leq \left(1 + \frac{1}{I} + 16I \tilde{L}_2^2 \eta^2 \alpha_{t-1}^2 \right) \sum_{m=1}^{M} \mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$$

$$+ 8I \tilde{L}_1^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\|\eta \nu_t^{(m)}\|^2 + \|\gamma \omega_t^{(m)}\|^2 + 16IL^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\|\tau q_t^{(m)}\|$$

$$+ 128I(c_r \alpha_{t-1}^2) \tilde{L}_1^2 \sum_{m=1}^{M} \mathbb{E}\|x_t^{(m)} - x_t\|^2 + \|y_t^{(m)} - y_t\|^2 + 32I(c_r \alpha_{t-1}^2)^2 L^2 \sum_{m=1}^{M} \mathbb{E}\|u_t^{(m)} - \bar{u}_t\|^2$$

$$+ 8IM(c_r \alpha_{t-1}^2)^2 \frac{\sigma^2}{b_x} + 32IM(c_r \alpha_{t-1}^2)^2 \zeta_f^2 + 64I(c_r \alpha_{t-1}^2)^2 M \frac{C_g^2 \zeta_g^2 \epsilon}{\mu^2}$$

This completes the proof.

**Lemma C.9.** Suppose $\gamma \alpha_t < \frac{1}{16IL}$, then for $t \neq \bar{t}, s \in [S]$, we have:

$$\sum_{m=1}^{M} \mathbb{E}\|\omega_t^{(m)} - \bar{\omega}_t\|^2 \leq \left(1 + \frac{33}{32I} \right) \sum_{m=1}^{M} \mathbb{E}\|\omega_t^{(m)} - \bar{\omega}_t\|^2 + 4IL^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\|\eta \nu_t^{(m)}\|^2$$

$$+ 8IM(c_r \alpha_{t-1}^2)^2 \frac{\sigma^2}{b_y} + 16IM(c_r \alpha_{t-1}^2)^2 \zeta_g^2 + 16IL^2 (c_r \alpha_{t-1}^2)^2 \sum_{m=1}^{M} \mathbb{E}\|x_t^{(m)} - x_{t-1}\|^2$$

$$+ 16IL^2 (c_r \alpha_{t-1}^2)^2 \sum_{m=1}^{M} \mathbb{E}\|y_t^{(m)} - y_{t-1}\|^2$$

where the expectation is w.r.t the stochasticity of the algorithm.
Proof. By the update step in Line 7 of Algorithm 1 for $t \neq \tilde{t}_s$, we have:

$$
\mathbb{E}||\tilde{\omega}_t^{(m)} - \bar{\omega}_t||^2 = \mathbb{E}||1 - c_\omega \alpha_{t-1}^2 \omega_{t-1}^{(m)}(\omega_{t-1}^{(m)} - \bar{\omega}_{t-1}) + \nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, B_y) - \frac{1}{M} \sum_{j=1}^M \nabla g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, B_y) - (1 - c_\omega \alpha_{t-1}^2) \nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, B_y) - \frac{1}{M} \sum_{j=1}^M \nabla g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, B_y)||^2
\leq (1 + \frac{1}{I}) (1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}||\omega_{t-1}^{(m)} - \bar{\omega}_{t-1}||^2 + (1 + I) \mathbb{E}||\nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, B_y) - \frac{1}{M} \sum_{j=1}^M \nabla g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, B_y) - (1 - c_\omega \alpha_{t-1}^2) \nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, B_y) - \frac{1}{M} \sum_{j=1}^M \nabla g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, B_y)||^2
$$

(17)

where the inequality follows from the generalized triangle inequality and the condition that $c_\omega \alpha_{t-1}^2 < 1$.

Next we denote the second term in Eq. (17) as $T_1$, then we have:

$$
T_1 \leq 2 \sum_{m=1}^M \mathbb{E}||\nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, B_y) - \frac{1}{M} \sum_{j=1}^M \nabla g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, B_y)||^2
+ 2(c_\omega \alpha_{t-1}^2)^2 \sum_{m=1}^M \mathbb{E}||\nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, B_y) - \frac{1}{M} \sum_{j=1}^M \nabla g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, B_y)||^2
$$

We bound the two terms separately, we denote them as $T_{1,1}$ and $T_{1,2}$ separately, then we have:

$$
T_{1,1} \overset{(a)}{=} \sum_{m=1}^M \mathbb{E}||\nabla g^{(m)}(x_{t}^{(m)}, y_{t}^{(m)}, B_y) - \nabla g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, B_y)||^2
\leq L^2 \sum_{m=1}^M \mathbb{E}[||x_{t}^{(m)} - x_{t-1}^{(m)}||^2 + ||y_{t}^{(m)} - y_{t-1}^{(m)}||^2] \leq L^2 \alpha_{t-1} \sum_{m=1}^M \mathbb{E}[||y_{t}^{(m)}||^2 + ||\omega_{t-1}^{(m)}||^2]
$$

(18)
where \((a)\) follows Proposition \ref{eq:E9} \((b)\) follows Proposition \ref{eq:11} \((b)\) and the fact that \(\hat{x}_t^{(m)} = x_t^{(m)}\) and \(\hat{y}_t^{(m)} = y_t^{(m)}\) when \(t \neq \hat{t}_s\); Next for the second term, we have:

\[
T_{1,2} = \sum_{m=1}^{M} \mathbb{E} \left[ \nabla_y g^{(m)}(x_{t-1}, y_{t-1}, B_y) - \nabla_y g^{(m)}(x_{t-1}, y_{t-1}) \right]
- \frac{1}{M} \sum_{j=1}^{M} \left( \nabla_y g^{(j)}(x_{t-1}, y_{t-1}, B_y) - \nabla_y g^{(j)}(x_{t-1}, y_{t-1}) \right)
+ \nabla_y g^{(m)}(x_{t-1}, y_{t-1}) - \frac{1}{M} \sum_{j=1}^{M} \nabla_y g^{(j)}(x_{t-1}, y_{t-1}) \right|^2
\]

\( \leq 2 \sum_{m=1}^{M} \mathbb{E} \left[ \nabla_y g^{(m)}(x_{t-1}, y_{t-1}, B_y) - \nabla_y g^{(m)}(x_{t-1}, y_{t-1}) \right]^2
+ 4 \sum_{m=1}^{M} \frac{1}{M} \sum_{j=1}^{M} \mathbb{E} \left[ \nabla_y g^{(m)}(\bar{x}_{t-1}, \bar{y}_{t-1}) - \nabla_y g^{(j)}(\bar{x}_{t-1}, \bar{y}_{t-1}) \right]^2
+ 4 \sum_{m=1}^{M} \mathbb{E} \left[ \nabla_y g^{(m)}(x_{t-1}, y_{t-1}) - \nabla_y g^{(m)}(\bar{x}_{t-1}, \bar{y}_{t-1}) \right]^2
+ \frac{1}{M} \sum_{j=1}^{M} \nabla_y g^{(j)}(\bar{x}_{t-1}, \bar{y}_{t-1}) - \nabla_y g^{(j)}(\bar{x}_{t-1}, \bar{y}_{t-1}) \right|^2 \]

Finally, combine Eq. \ref{eq:17} Eq. \ref{eq:18} with Eq. \ref{eq:19} and use the fact that \(I \geq 1\), we have:

\[
\sum_{m=1}^{M} \mathbb{E} \left[ \omega_t^{(m)} - \bar{\omega}_t \right]^2 \leq (1 + \frac{1}{I}) \sum_{m=1}^{M} \mathbb{E} \left[ \omega_t^{(m)} - \bar{\omega}_t \right]^2 + 4I^2 \sigma^2 \sum_{m=1}^{M} \mathbb{E} \left[ \gamma_t^{(m)} \right] \sum_{m=1}^{M} \mathbb{E} \left[ \eta_t^{(m)} \right]^2
+ 8IM(c_x \alpha_t^{(m)})^2 \sigma^2 \frac{\sigma^2}{b_y}
+ 16IM(c_x \alpha_t^{(m)})^2 \sigma^2 + 16I^2 \sum_{m=1}^{M} \mathbb{E} \left[ \left| x_t^{(m)} - \bar{x}_t \right| \right] ^2
+ 16I^2 \sum_{m=1}^{M} \mathbb{E} \left[ \left| y_t^{(m)} - \bar{y}_t \right| \right] ^2
\]
We separate the term $T_1$ with triangle inequality to get:

$$
\sum_{m=1}^{M} \mathbb{E}||\tilde{\omega}_t^{(m)} - \tilde{\omega}_t||^2 \leq \left( 1 + \frac{1}{T} + 8IL^2\gamma^2\alpha_t^{-1} \right) \sum_{m=1}^{M} \mathbb{E}||\omega_{t-1}^{(m)} - \omega_{t-1}||^2 \\
+ 4IL^2\alpha_t^{-1} \sum_{m=1}^{M} \mathbb{E}[2||\gamma\tilde{\omega}_{t-1}||^2 + ||\eta\nu_t^{(m)}||^2] \\
+ 8IM(c_u\alpha_t^{-1})^2\sigma_y^2 \\
+ 16IM(c_u\alpha_t^{-1})^2\zeta_y^2 + 16IL^2(c_u\alpha_t^{-1})^2 \sum_{m=1}^{M} \mathbb{E}[||x_t^{(m)} - \bar{x}_{t-1}||^2] \\
+ 16IL^2(c_u\alpha_t^{-1})^2 \sum_{m=1}^{M} \mathbb{E}[||y_t^{(m)} - \bar{y}_{t-1}||^2]
$$

This completes the proof. $\square$

**Lemma C.10.** Suppose $\tau \alpha_t < \frac{1}{32IL}$, then for $t \neq \bar{t}_s$, $s \in [S]$, we have:

$$
\sum_{m=1}^{M} \mathbb{E}||\tilde{q}_t^{(m)} - \tilde{q}_t||^2 \leq \left( 1 + \frac{33}{32I} \right) \sum_{m=1}^{M} \mathbb{E}||\tilde{q}_t^{(m)} - \tilde{q}_t||^2 + 8IL^2\alpha_t^{-1} \sum_{m=1}^{M} \mathbb{E}[||\omega_t^{(m)}||^2 + ||\nu_t^{(m)}||^2] \\
+ 32IL^2\alpha_t^{-1} \sum_{m=1}^{M} \mathbb{E}[||\tau^2\tilde{q}_t||^2] + 8IM(c_u\alpha_t^{-1})^2\frac{\sigma_y^2}{b_x} \\
+ 16IM(c_u\alpha_t^{-1})^2\zeta_f^2 + 32IM(c_u\alpha_t^{-1})^2\frac{C_f\zeta_g\eta\nu}{\mu^2} \\
+ 64I(c_u\alpha_t^{-1})^2L^2 \sum_{m=1}^{M} \mathbb{E}[||x_t^{(m)} - \bar{x}_t||^2 + ||y_t^{(m)} - \bar{y}_t||^2] \\
+ 16I(c_u\alpha_t^{-1})^2L^2 \sum_{m=1}^{M} \mathbb{E}[||u_t^{(m)} - \bar{u}_t||^2]
$$

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

**Proof.** For $t \neq \bar{t}_s$, we have:

$$
\mathbb{E}||\tilde{q}_t^{(m)} - \tilde{q}_t||^2 = \mathbb{E}||\left(1 - c_u\alpha_t^{-1}\right)\tilde{q}_t^{(m)} - \tilde{q}_t - (1 - c_u\alpha_t^{-1})\left(p_t^{(m)} - \bar{p}_t.B_s - (1 - c_u\alpha_t^{-1})(p_t^{(m)} - \bar{p}_t.B_s)\right)||^2 \\
\leq (1 + \frac{1}{I})(1 - c_u\alpha_t^{-1})^2\mathbb{E}||\tilde{q}_t^{(m)} - \tilde{q}_t||^2 \\
+ (1 + I)\mathbb{E}||p_t^{(m)} - \bar{p}_t.B_s - (1 - c_u\alpha_t^{-1})\left(p_t^{(m)} - \bar{p}_t.B_s\right)||^2
$$

where the inequality follows from the the generalized triangle inequality and the condition that $c_u\alpha_t^2 < 1$.

Next we sum over $M$ for the second term in Eq. [17] and denote it as $T_1$, then we have:

$$
T_1 \leq 2 \sum_{m=1}^{M} \mathbb{E}[p_t^{(m)} - \bar{p}_t.B_s - \left(p_t^{(m)} - \bar{p}_t.B_s - \bar{p}_t.B_s\right)||^2 + 2(c_u\alpha_t^{-1})^2 \sum_{m=1}^{M} \mathbb{E}[p_t^{(m)} - \bar{p}_t.B_s - \bar{p}_t.B_s||^2
$$

27
We bound the two terms separately, we denote them as $T_{1,1}$ and $T_{1,2}$ separately, then we have:

\[
T_{1,1} \leq \sum_{m=1}^{M} \mathbb{E}\left\| p_{1,B_x}^{(m)} - p_{1-1,B_x}^{(m)} \right\|^2 \\
\leq 2(L^2 + \frac{2L^2C_f^2}{\mu^2}) \sum_{m=1}^{M} \mathbb{E}\left[\| \tilde{x}_t^{(m)} - x_t^{(m)} \|_2^2 + \| \tilde{y}_t^{(m)} - y_t^{(m)} \|_2^2 \right] + 4L^2 \sum_{m=1}^{M} \mathbb{E}\left[ \| u_t^{(m)} - u_{t-1}^{(m)} \|_2^2 \right] \\
\leq 2L^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\left[\| \eta \tilde{u}_t^{(m)} \|_2^2 + \| \gamma \omega_{t-1}^{(m)} \|_2^2 \right] + 4L^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\left[ \| r^2 \tilde{q}_{t-1}^{(m)} \|_2^2 \right]
\]

where (a) follows Proposition E.2, (b) follows Proposition C.1 and the fact that $\tilde{x}_t^{(m)} = x_t^{(m)}$ and $\tilde{y}_t^{(m)} = y_t^{(m)}$ when $t \neq t_s$. Next for the second term, we have:

\[
T_{1,2} = \sum_{m=1}^{M} \mathbb{E}\left\| p_{1-1,B_x}^{(m)} - p_{1-1}^{(m)} \right\|^2 \\
\leq 2 \sum_{m=1}^{M} \mathbb{E}\left[ \| p_{1-1,B_x}^{(m)} - p_{1-1}^{(m)} \|_2^2 \right] + 2 \sum_{m=1}^{M} \mathbb{E}\left[ \| p_{1-1}^{(m)} - \bar{p}_{1-1} \|_2^2 \right]
\]

We denote the two terms above as $T_{1,2,1}, T_{1,2,2}$ respectively. For the term $T_{1,2,1}$ of Eq. 19 we have $T_{1,2,1} \leq 2M\sigma^2/b_x$ by the bounded variance assumption; For the term $T_{1,2,2}$ of Eq. 19 we have

\[
T_{1,2,2} \leq 16\left( L^2 + \frac{2L^2C_f^2}{\mu^2} \right) \sum_{m=1}^{M} \mathbb{E}\left[\| \bar{x}_t^{(m)} \|_2^2 + \| \bar{y}_t^{(m)} \|_2^2 \right] \\
+ 4L^2 \sum_{m=1}^{M} \mathbb{E}\left[ u_t^{(m)} - \bar{u}_t \|_2^2 \right] + 4MC^2\sigma^2 \zeta^2 + \frac{8MC^2\sigma^2 \zeta^2}{\mu^2}
\]

Finally, combine everything together and use the fact that $I \geq 1$, we have:

\[
\sum_{m=1}^{M} \mathbb{E}\left\| \tilde{q}_t^{(m)} - \bar{q}_t \right\|^2 \leq \left(1 + \frac{1}{7}\right) \sum_{m=1}^{M} \mathbb{E}\left\| \tilde{q}_{t-1}^{(m)} - \bar{q}_t^{(m)} \right\|^2 + 8IL^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\left[\| \gamma \omega_{t-1}^{(m)} \|_2^2 + \| \eta \tilde{u}_{t-1}^{(m)} \|_2^2 \right] \\
+ 16IL^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\left[ \| \tau^2 \tilde{q}_{t-1}^{(m)} \|_2^2 \right] \\
+ 8IM \sigma_x^2 \frac{\tau^2}{b_x^2} + 16IM \sigma_x^2 \frac{\tilde{L}_2^2}{\mu^2} + 32IM \sigma_x^2 \frac{\tilde{L}_2^2}{\mu^2} + 64I \sigma_x^2 \frac{\tilde{L}_2^2}{\mu^2}
\]

28
We separate the term $T_1$ with triangle inequality to get:
\[
\sum_{m=1}^{M} \mathbb{E} \| q_t^{(m)} - \tilde{q}_t \|^2 \leq \left( 1 + \frac{1}{T} + 32IL^2\gamma^2\alpha_t^{-1} \right) \sum_{m=1}^{M} \mathbb{E} \| q_t^{(m)} - \tilde{q}_t \|^2 \\
+ 8IL^2\alpha_t^{-1} \sum_{m=1}^{M} \mathbb{E} \| \gamma \omega_t^{(m)} \|^2 + \| \eta \mu_t^{(m)} \|^2 \\
+ 32IL^2\alpha_t^{-1} \sum_{m=1}^{M} \mathbb{E} \| \tau^2 \tilde{q}_t \|^2 + 8IM(c_u\alpha_t^{-1})^2 \sigma_b^2 \\
+ 16IM(c_u\alpha_t^{-1})^2 \zeta_f^2 + 32IM(c_u\alpha_t^{-1})^2 \frac{C^2\zeta_f^2 \mu_t}{\mu^2} \\
+ 64I(c_u\alpha_t^{-1})^2 L^2 \sum_{m=1}^{M} \mathbb{E} \| x_t^{(m)} - \bar{x}_t \|^2 + \| y_t^{(m)} - \bar{y}_t \|^2 \\
+ 16I(c_u\alpha_t^{-1})^2 L^2 \sum_{m=1}^{M} \mathbb{E} \| y_t^{(m)} - \bar{y}_t \|^2
\]

This completes the proof. \(\square\)

Next, to simply the notation, we denote $A_t = \mathbb{E} \| \bar{\nu}_t - \bar{\mu}_t \|^2$, $B_t = \mathbb{E} \| \bar{\gamma}_t - y_{z,t} \|^2$, $C_t = \mathbb{E} \| \bar{\omega}_t - \frac{1}{M} \sum_{m=1}^{M} \nabla_y g_t^{(m)}(x_t^{(m)}, y_t^{(m)}) \|^2$, $D_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \| \nu_t^{(m)} - \bar{v}_t \|^2$, $E_t = \mathbb{E} \| \bar{\nu}_t \|^2$, $F_t = \mathbb{E} \| \bar{\omega}_t \|^2$, $G_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \| \omega_t^{(m)} - \bar{\omega}_t \|^2$, $H_t = \mathbb{E} \| \bar{q}_t - \bar{p}_t \|^2$, $I_t = \mathbb{E} \| \bar{u}_t - \bar{u}_t \|^2$, $J_t = \mathbb{E} \| \bar{q}_t^{(m)} - \bar{q}_t \|^2$, $Q_t = \mathbb{E} \| \bar{q}_t \|^2$.

**Lemma C.11.** For $\eta < \min\left( \frac{L^2}{\epsilon_c^{2}}, \frac{L^2}{\epsilon_c^{2}}, \frac{L^2}{\epsilon_c^{2}} \right)$, $\gamma < \min\left( \frac{L^2}{\epsilon_c^{2}}, \frac{L^2}{\epsilon_c^{2}}, \frac{L^2}{\epsilon_c^{2}} \right)$, $\tau < \min\left( \frac{L^2}{\epsilon_c^{2}}, \frac{L^2}{\epsilon_c^{2}}, \frac{1}{2} \right)$ and $\alpha_t < \frac{1}{16L^4}$, where $L = \max(L_1, L_2)$, we have:
\[
\sum_{t=\ell_s}^{\ell_s+1} \alpha_t D_t \leq \sum_{t=\ell_s}^{\ell_s+1} \left( \alpha_t E_t + \alpha_t F_t + \alpha_t Q_t + \frac{c_u^2 \alpha_t^3 \sigma_b^2}{L^2} + \frac{c_u^2 \alpha_t^3 \zeta_f^2}{L^2} \right) \\
+ \frac{c_u^2 \alpha_t^3 \sigma_b^2}{L^2} + \frac{c_u^2 \alpha_t^3 \zeta_f^2}{L^2} + \frac{c_u^2 \alpha_t^3 \zeta_f^2}{L^2} + \frac{c_u^2 \alpha_t^3 \zeta_f^2}{L^2} + \frac{2c_u^2 \alpha_t^3 \zeta_f^2}{L^2} + \left( \frac{4c_u^2 \alpha_t^3 \zeta_f^2}{L^2} \right)
\]

**Proof.** Based on Lemma C.8 for $t \neq \ell_s$, we have:
\[
D_t \leq \left( 1 + \frac{17}{16L} \right) D_{t-1} + 16IL_1^2\alpha_t^{-1}\eta^2 E_{t-1} + 16IL_1^2\alpha_t^{-1}\gamma^2 F_{t-1} + 16IL_1^2\alpha_t^{-1}\gamma^2 G_{t-1} + 32IL^2\tau^2\alpha_t^{-1}J_{t-1} \\
+ 32IL^2\tau^2\alpha_t^{-1}Q_{t-1} + 8Ic_u^2\alpha_t^{-1}\sigma_b^2 + 32Ic_u^2\alpha_t^{-1}\zeta_f^2 + 64Ic_u^2\alpha_t^{-1}\frac{C^2\zeta_f^2 \mu_t}{\mu^2} \\
+ 128I^2L_1^2\gamma^2c_u^2\alpha_t^{-1} \sum_{t=\ell_s}^{\ell_s+1} \alpha_t^2 D_t + 128I^2L_1^2\gamma^2c_u^2\alpha_t^{-1} \sum_{t=\ell_s}^{\ell_s+1} \alpha_t^2 G_t + 32I^2L^2\tau^2c_u^2\alpha_t^{-1} \sum_{t=\ell_s}^{\ell_s+1} \alpha_t^2 J_t
\]
while for \( t = \tilde{t}_s \), we have \( D_{t_s} = 1/M \sum_{m=1}^M \mathbb{E} ||\nu_t^{(m)} - \bar{v}_t||^2 = 0 \). Apply the above equation recursively from \( \tilde{t}_{s-1} + 1 \) to \( t \), so we have:

\[
D_t \leq \sum_{\ell = \tilde{t}_{s-1}}^t (1 + \frac{17}{16t})^{t-\ell} \left( 16I_2^2 \alpha_t^2 \eta_t^2 E_{\ell} + 16I_2^2 \alpha_t^2 \gamma_t^2 F_{\ell} + 16I_2^2 \alpha_t^2 \gamma_t^2 G_{\ell} + 32I_2^2 \tau^2 \alpha_t^2 J_{\ell} \\
+ 32I_2^2 \tau^2 \alpha_t^2 Q_{\ell} + 8I_2^2 \alpha_t^2 \sigma_t^2 \frac{\sigma_t^2}{b_x} + 32I_2^2 \alpha_t^4 \zeta_t^2 + 64I_2^2 \alpha_t^4 \frac{C_{f_{g,x,y}}^2}{\mu^2} \right) \left( \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 D_{\ell} + \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 G_{\ell} + 32I_2^2 \tau^2 \alpha_t^2 \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 J_{\ell} \right)
\]

\[
\leq \sum_{\ell = \tilde{t}_{s-1}}^{t-1} (48I_2^2 \alpha_t^2 \eta_t^2 E_{\ell} + 48I_2^2 \alpha_t^2 \gamma_t^2 F_{\ell} + 48I_2^2 \alpha_t^2 \gamma_t^2 G_{\ell} + 96I_2^2 \tau^2 \alpha_t^2 J_{\ell} \\
+ 96I_2^2 \tau^2 \alpha_t^2 Q_{\ell} + 24I_2^2 \alpha_t^4 \sigma_t^2 \frac{\sigma_t^2}{b_x} + 96I_2^2 \alpha_t^4 \zeta_t^2 + 192I_2^2 \alpha_t^4 \frac{C_{f_{g,x,y}}^2}{\mu^2} \right) \left( \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 D_{\ell} + \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 G_{\ell} + 96I_2^2 \tau^2 \alpha_t^2 \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 J_{\ell} \right)
\]

The second inequality uses the fact that \( t - \ell \leq I \) and the inequality \( \log(1 + a/x) \leq a/x \) for \( x > -a \), so we have \( (1 + a/x)^x \leq e^a \). Then we choose \( a = 17/16 \) and \( x = I \). Finally, we use the fact that \( e^{17/16} \leq 3 \).

Next we multiply \( \alpha_t \) over both sides and take sum from \( \tilde{t}_{s-1} + 1 \) to \( \tilde{t}_s \), we have:

\[
\sum_{\ell = \tilde{t}_{s-1} + 1}^{\tilde{t}_s} \alpha_t D_{\ell} \leq \sum_{\ell = \tilde{t}_{s-1}}^{\tilde{t}_s - 1} \alpha_t \left( 48I_2^2 \alpha_t^2 \eta_t^2 E_{\ell} + 48I_2^2 \alpha_t^2 \gamma_t^2 F_{\ell} + 48I_2^2 \alpha_t^2 \gamma_t^2 G_{\ell} + 96I_2^2 \tau^2 \alpha_t^2 J_{\ell} \\
+ 96I_2^2 \tau^2 \alpha_t^2 Q_{\ell} + 24I_2^2 \alpha_t^4 \sigma_t^2 \frac{\sigma_t^2}{b_x} + 96I_2^2 \alpha_t^4 \zeta_t^2 + 192I_2^2 \alpha_t^4 \frac{C_{f_{g,x,y}}^2}{\mu^2} \right) \left( \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 D_{\ell} + \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 G_{\ell} + 96I_2^2 \tau^2 \alpha_t^2 \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 J_{\ell} \right)
\]

\[
= \sum_{\ell = \tilde{t}_{s-1}}^{\tilde{t}_s - 1} \left( 3I_1 \alpha_t^2 \eta_t^2 E_{\ell} + 3I_1 \alpha_t^2 \gamma_t^2 F_{\ell} + 3I_1 \alpha_t^2 \gamma_t^2 G_{\ell} + 6I_1 \tau^2 \alpha_t^2 J_{\ell} \\
+ 6I_1 \tau^2 \alpha_t^2 Q_{\ell} + \frac{3I_2^2 \alpha_t^4 \sigma_t^2}{2L} + \frac{6I_2^2 \alpha_t^4 \zeta_t^2}{L} + \frac{12I_2^2 \alpha_t^4 \frac{C_{f_{g,x,y}}^2}{\mu^2}}{L^2} \right) \left( \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 D_{\ell} + \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 G_{\ell} + 6I_2^2 \tau^2 \alpha_t^2 \sum_{\ell = \tilde{t}_{s-1}}^{\ell-1} \alpha_t^2 J_{\ell} \right)
\]

\[
\leq \sum_{\ell = \tilde{t}_{s-1}}^{\tilde{t}_s - 1} \left( \frac{3\eta_t^2}{16} \alpha_t E_{\ell} + \frac{3\gamma_t^2}{16} \alpha_t F_{\ell} + \frac{3\gamma_t^2}{16} \alpha_t G_{\ell} + \frac{3\tau^2}{8} \alpha_t J_{\ell} + \frac{3\tau^2}{8} \alpha_t Q_{\ell} \\
+ \frac{3c_2 \alpha_t^2 \sigma_t^2}{32L^2} \frac{\sigma_t^2}{b_x} + \frac{3c_2 \alpha_t^2 \zeta_t^2}{8L^2} + \frac{3c_2 \alpha_t^2 \frac{C_{f_{g,x,y}}^2}{\mu^2}}{4L^2} + \frac{\eta_t^2 \zeta_t^2}{8 \times 16^3 I_2^2 L^4} \alpha_t D_{\ell} \\
+ \frac{\gamma_t^2 \zeta_t^2}{8 \times 16^3 I_2^2 L^4} \alpha_t G_{\ell} + \frac{3\tau^2 \zeta_t^2}{8 \times 16^4 I_2^2 L^4} \alpha_t J_{\ell} \right)
\]
In inequalities (a) and (b), we use \( \alpha_t < \frac{1}{16L} \) and \( \eta < \frac{l^2}{c_y} \). Note that \( \sum_{t=t_{s-1}+1}^{t_{s-1}} \alpha_tD_t = \sum_{t=t_{s-1}+1}^{t_{s-1}} \alpha_tD_t = D_{t_{s-1}} = 0 \).

Then if we choose \( \eta < \frac{l^2}{c_y} \) and \( \gamma < \frac{l^2}{c_y} \), we have

\[
\sum_{t=t_{s-1}+1}^{t_{s-1}} \alpha_tD_t \leq \sum_{t=t_{s-1}+1}^{t_{s-1}} (\frac{\eta^2}{4} \alpha_tE_t + \frac{\gamma^2}{4} \alpha_tF_t + \frac{\gamma^2}{2} \alpha_tG_t + \tau^2 \alpha_tJ_t + \frac{\gamma^2}{2} \alpha_tQ_t
\]

\[
+ \frac{c_2^2\alpha_t^2 \sigma^2}{8L^2 b_x} + \frac{c_2^2\alpha_t^2 \sigma^2}{2L^2} + \frac{c_2^2\alpha_t^2 C^2_{f q, y}}{\mu^2})
\]

(20)

Based on Lemma[9] for \( t \neq t_s \), we have:

\[
G_t \leq \left( 1 + \frac{33}{32L} \right) G_{t-1} + 8IL^2 \eta^2 \alpha_t^2 D_{t-1} + 8IL^2 \eta^2 \alpha_t^2 E_t + 8IL^2 \gamma^2 \alpha_t^2 F_{t-1}
\]

\[
+ 8IC_2^2 \alpha_t^2 \sigma^2 + 16IC_2^2 \alpha_t^2 \sigma_g^2 + 16I^2 L^2 \eta^2 \alpha_t^2 \alpha_t^2 \sum_{t=t_{s-1}}^{t_{s-1}-2} \alpha_t^2 D_t + 16I^2 L^2 \gamma^2 \alpha_t^2 \alpha_t^2 \sum_{t=t_{s-1}}^{t_{s-1}-2} \alpha_t^2 G_t
\]

Follow similar derivation, by recursively applying the above inequality, we have:

\[
G_t \leq \sum_{t=t_{s-1}}^{t_{s-1}+1} (24IL^2 \eta^2 \alpha_t^2 D_t + 24IL^2 \eta^2 \alpha_t^2 E_t + 24IL^2 \gamma^2 \alpha_t^2 F_t
\]

\[
+ 24IC_2^2 \alpha_t^2 \sigma^2 + 48IC_2^2 \alpha_t^2 \sigma_g^2 + 48I^2 L^2 \eta^2 \alpha_t^2 \alpha_t^2 \sum_{t=t_{s-1}}^{t_{s-1}-1} \alpha_t^2 D_t + 48I^2 L^2 \gamma^2 \alpha_t^2 \alpha_t^2 \sum_{t=t_{s-1}+1}^{t_{s-1}} \alpha_t^2 G_t
\]

Next we multiply \( \alpha_t \) over both sides and take sum from \( t_{s-1}+1 \) to \( t_s \), use the condition that \( \alpha_t < \frac{1}{16L} < \frac{1}{16L} \), \( \eta < \frac{l^2}{c_y} \) and \( \gamma < \frac{l^2}{c_y} \), we have:

\[
\sum_{t=t_{s-1}+1}^{t_{s-1}} \alpha_tD_t \leq \sum_{t=t_{s-1}+1}^{t_{s-1}} (\frac{1}{2} \eta^2 \alpha_tE_t + \frac{1}{8} \eta^2 \alpha_tE_t + \frac{1}{8} \gamma^2 \alpha_tF_t + \alpha_t^2 \frac{c_2^2 \sigma^2}{8L^2 b_x} + \frac{c_2^2 \sigma_g^2}{8L^2} \alpha_t^2)
\]

(21)

Based on Lemma[10] we have:

\[
J_t \leq \left( 1 + \frac{33L}{32L} \right) J_{t-1} + 16IL^2 \eta^2 \alpha_t^2 G_{t-1} + 16IL^2 \eta^2 \alpha_t^2 F_{t-1} + 16IL^2 \gamma^2 \alpha_t^2 D_{t-1} + 16IL^2 \gamma^2 \alpha_t^2 E_{t-1}
\]

\[
+ 16IL^2 \gamma^2 \alpha_t^2 \sigma_g^2 + 8IC_2^2 \alpha_t^2 \sigma_g^2 + 32IC_2^2 \alpha_t^2 \alpha_t^2 \]

\[
+ 64IC_2^2 \gamma^2 \alpha_t^2 \alpha_t^2 \sum_{t=t_{s-1}}^{t_{s-1}-1} \alpha_t^2 D_t + 64IC_2^2 \gamma^2 \alpha_t^2 \alpha_t^2 \sum_{t=t_{s-1}+1}^{t_{s-1}} \alpha_t^2 G_t + 64IC_2^2 \gamma^2 \alpha_t^2 \alpha_t^2 \sum_{t=t_{s-1}}^{t_{s-1}-1} \alpha_t^2 J_t
\]

Suppose we have \( \alpha_t < \frac{1}{16L} \), \( \eta < \frac{l^2}{c_y} \), \( \gamma < \frac{l^2}{c_y} \), \( \tau < \frac{l^2}{c_y} \),

\[
\sum_{t=t_{s-1}+1}^{t_{s-1}} \alpha_tJ_t \leq \sum_{t=t_{s-1}+1}^{t_{s-1}} (\frac{\gamma^2}{2} \alpha_tG_t + \frac{\gamma^2}{4} \alpha_tF_t + \frac{\eta^2}{4} \alpha_tD_t + \frac{\eta^2}{4} \alpha_tE_t
\]

\[
+ \frac{\gamma^2}{4} \alpha_tQ_t + \alpha_t^2 \frac{c_2^2 \sigma^2}{8L^2 b_x} + \frac{c_2^2 \sigma_g^2}{4L^2} + \frac{3c_2^2 \alpha_t^2 C^2_{f q, y}}{\mu^2})
\]

(22)

Next, we combine Eq. 20, Eq. 21 and Eq. 22 to have the result in the lemma.
C.1.4 Descent Lemma

Lemma C.12. For all $t \in [\bar{t}_s-1, \bar{t}_s-1]$, the iterates generated satisfy:

$$
\mathbb{E}\|\nabla h(\bar{x}_t) - \bar{\mu}_t\|^2 \leq \frac{2\bar{L}^2}{M} \sum_{m=1}^{M} \mathbb{E}\left[\|\bar{x}_t - x_t^{(m)}\|^2 + 2\|\bar{y}_t - y_t^{(m)}\|^2 + 2\|y_{\bar{x}_t} - \bar{y}_t\|^2 + 4L^2\mathbb{E}\|u_{\bar{x}_t} - \bar{u}_t\|^2 \right]
$$

where we denote $u_{\bar{x}_t} = \left[\nabla_y g(\bar{x}_t, y_{\bar{x}_t})\right]^{-1}\nabla_y f(\bar{x}_t, y_{\bar{x}_t})$ and $\bar{L}_t^2 = (L^2 + \frac{2L^2C^2}{\mu^2})$ is a constant.

Proof. This lemma follows the same derivation as Lemma C.22.

Lemma C.13. Suppose $\eta t < \frac{1}{2L}$, for all $t \in [\bar{t}_s-1, \bar{t}_s-1]$ and $s \in [S]$, the iterates generated satisfy:

$$
\mathbb{E}\left[h(\bar{x}_{t+1})\right] \leq \mathbb{E}\left[h(\bar{x}_t) + \langle \nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \right]
$$

(a) $\mathbb{E}\left[h(\bar{x}_t) - \eta t \langle \nabla h(\bar{x}_t), \bar{\nu}_t \rangle + \frac{\eta^2 t^2 L^2}{2}\|\bar{\nu}_t\|^2 \right]$

(b) $\mathbb{E}\left[h(\bar{x}_t) - \eta t \|\bar{\nu}_t\|^2 + \frac{\eta^2 t^2 L^2}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta^2 t^2 L^2}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta^2 t^2 L^2}{2}\|\bar{\nu}_t\|^2 \right]$

where equality (a) follows from the iterate update given in Algorithm I (b) uses $(a, b) = \frac{1}{2}\|a\|^2 + \|b\|^2 - \|a - b\|^2$ and $\eta t < \frac{1}{2L}$. For the term $T_1$, we have:

$$
\mathbb{E}\left[\|\nabla h(\bar{x}_t) - \bar{\nu}_t\|^2 \right] \leq 2\mathbb{E}\left[\|\nabla h(\bar{x}_t) - \bar{\nu}_t\|^2 \right] + 2\mathbb{E}\left[\|\bar{\nu}_t - \bar{\nu}_t\|^2 \right]
$$

Use Lemma C.3 for the first term and combine everything together finishes the proof. □

C.1.5 Proof of Convergence Theorem

We first denote the following potential function $G(t)$:

$$
G_t = h(\bar{x}_t) + \frac{9bM\eta}{64\alpha_t} \|\bar{\nu}_t - \bar{\mu}_t\|^2 + \frac{18\eta L^2}{\mu\gamma}\|\bar{y}_t - y_{x_t}\|^2 + \frac{9bM\eta}{64\alpha_t} \|\bar{\nu}_t - \bar{\mu}_t\|^2 + \frac{9bM\eta}{64\alpha_t} \|\bar{\nu}_t - \bar{\mu}_t\|^2
$$

Furthermore, we have constants $\bar{L}_t^2 = \left(\bar{L}^2 + \frac{2L^2C^2}{\mu^2}\right)$ and $\bar{L}_t^2 = \left(\bar{L}^2 + \frac{2L^2C^2}{\mu^2}\right)$, to ease the writing, without loss of generality, we assume the second order Lipschitz constants $L_{xy} = L_{by}$, as a result $\bar{L}_t^2 = \bar{L}_t^2$, we denote it as $\bar{L}$ in the subsequent proof.

Theorem C.14. Suppose we choose $c_\nu = \frac{9bM\eta}{64\alpha_t} + \frac{2L^2C^2}{\mu^2}$, $c_\omega = \frac{48^2}{48\alpha_t} + \frac{2L^2C^2}{\mu^2}$, $c_u = \frac{48^2}{48\alpha_t} + \frac{2L^2C^2}{\mu^2}$, $u = (b\bar{M})^{\bar{u}}$, where $\bar{u} = \max\left(2, \frac{16\bar{M}^2 \bar{L}^2}{L_{xy}}, \frac{3}{\sqrt{2}}\right)$, $\delta = \frac{(\bar{M})^{2/3}}{(16\bar{L})^{3/2}}$, $\alpha_t = \frac{\delta}{(u + t)\bar{u}^{3/2}}$, $t \in [T]$, $\gamma < \min\left(\frac{1}{8C_1^2}, \frac{\bar{L}}{4C_1^2}, \frac{L_t^2}{C_1^2}, \frac{x_{\bar{x}_t}^2}{C_1^2}, \frac{L_t^2}{C_1^2}, \frac{L_t}{2C_1}, 1\right)$, $\eta < \min\left(\frac{\mu \gamma}{36\bar{M}^2}, \frac{\bar{L}^2}{4C_1^2}, \frac{L_t^2}{C_1^2}, \frac{L_t^2}{C_1^2}, \frac{1}{2C_1}, 1\right)$, $r < \frac{1}{8C_1^2}$.
min \left( \frac{1}{4c_1^2}, \frac{L}{4c_1^2}, \frac{L^2}{c_1}, \frac{1}{c_1}, \frac{1}{2} \right) \) where \( C_1 \) is a constant, we set the mini-batch size \( b_x = b_y = b \) and the first batch with size \( b_1 = O(Ib) \), \( r = \frac{C_1}{b} \), then we have:

\[
\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}[\| \nabla h(\bar{x}_t) \|^2] = O\left( \frac{\kappa^{16/3}}{T} + \left( \frac{bMT}{1.5} \right)^{2/3} \right)
\]

To reach an \( \epsilon \)-stationary point, we need \( T = O(\kappa^{8}(bM)^{-1}\epsilon^{-1.5}) \), \( I = O(\kappa^{5/3}(bM)^{-1}\epsilon^{-0.5}) \).

**Proof.** By the condition that \( u \geq \epsilon^{3/2} \delta^3 \), it is straightforward to verify that \( c_\nu \alpha^2 < 1 \). By Lemma C.3 we have:

\[
\frac{A_t}{\alpha_t-1} - \frac{A_{t-1}}{\alpha_{t-2}} \leq \left( \alpha_{t-1} - \alpha_t-2 - c_\nu \alpha_{t-1} \right) \frac{2\bar{c}_2 \alpha_{t-1} \alpha_t^2}{bM} + \frac{16L^2 \tau \alpha_{t-1}}{bM} (J_{t-1} + Q_{t-1})
\]

where we choose \( b_x = b_y = b \). For \( \alpha_{t-1} - \alpha_t-2 \), we have:

\[
\alpha_{t-1} - \alpha_t-2 = \frac{(u + \sigma^t - t - 1)^{1/3}}{\delta} \leq \frac{\sigma^2}{3\delta(u + \sigma^t(t - 1))^{2/3}} \leq \frac{2\alpha^2}{3\delta u^2 (t - 1)^{2/3}} \leq \frac{2}{3b^2 M^2} \alpha_t \leq \frac{2}{3b^2 M^2} \alpha_t
\]

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 \( \alpha_t \). By choosing \( c_\nu = \frac{64}{9bM^2} + \frac{2}{3b^2 M^2} \), we have:

\[
\frac{A_t}{\alpha_t-1} - \frac{A_{t-1}}{\alpha_{t-2}} \leq \frac{64}{9bM} \alpha_{t-1} A_t + \frac{2\epsilon^2 \alpha_{t-1} \alpha_t^2}{bM} + \frac{16L^2 \tau \alpha_{t-1}}{bM} (J_{t-1} + Q_{t-1})
\]

Next, we telescope from \( \bar{t}_{n-1} + 1 \) to \( \bar{t}_n \):

\[
\left( \frac{A_{\bar{t}_n}}{\alpha_{\bar{t}_n-1}} - \frac{A_{\bar{t}_{n-1}}}{\alpha_{\bar{t}_{n-1}-1}} \right) \leq \frac{64}{9bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t A_t + \frac{2\epsilon^2 \alpha_{t-1} \alpha_t^2}{bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t^3 + \frac{16L^2 \tau \alpha_{t-1}}{bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t D_t
\]

Next, we follow similar derivation as \( A_t/\alpha_t-1 - A_{t-1}/\alpha_{t-2} \). By Lemma C.4 we choose \( c_\nu = \frac{48^2}{bM^2} + \frac{2}{3b^2 M^2} \), to obtain:

\[
\frac{C_t}{\alpha_{t-1}} - \frac{C_{t-1}}{\alpha_{t-2}} \leq -\frac{48^2 \alpha_{t-1} C_{t-1}}{bM^2} + \frac{2\epsilon^2 \alpha_{t-1} \alpha_t^2}{bM} + \frac{4L^2 \eta^2 \alpha_{t-1}}{bM} (J_{t-1} + E_{t-1}) + \frac{4L^2 \eta^2 \alpha_{t-1}}{bM} (F_{t-1} + G_{t-1})
\]

Then telescope from \( \bar{t}_{n-1} + 1 \) to \( \bar{t}_n \), we have:

\[
\left( \frac{C_{\bar{t}_n}}{\alpha_{\bar{t}_n-1}} - \frac{C_{\bar{t}_{n-1}}}{\alpha_{\bar{t}_{n-1}-1}} \right) \leq -\frac{48^2}{bM^2} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t C_t + \frac{2\epsilon^2 \alpha_{t-1} \alpha_t^2}{bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t^3 + \frac{16L^2 \eta^2 \alpha_{t-1}}{bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t D_t
\]

\[
+ \frac{8L^2 \eta^2}{bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t E_t + \frac{8L^2 \eta^2}{bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t F_t + \frac{16L^2 \eta^2 \alpha_{t-1}}{bM} \sum_{t=\bar{t}_{n-1}}^{\bar{t}_n} \alpha_t G_t
\]
Next from Lemma [C.2] we choose $c_u = \frac{48^2}{bM\mu^2} + \frac{2}{3\gamma M}$, to obtain:

\[
\frac{H_t}{\alpha_{t-1}} - \frac{H_{t-1}}{\alpha_{t-2}} \leq -\frac{48^2\alpha_{t-1}}{bM\mu^2} H_{t-1} + \frac{2c_u^2\alpha_{t-1}}{bM} + \frac{8\eta^2\alpha_{t-1}\tilde{L}^2}{bM} (D_{t-1} + E_{t-1}) \\
+ \frac{8\gamma^2\alpha_{t-1}\tilde{L}^2}{bM} (F_{t-1} + G_{t-1}) + \frac{8\tau^2\alpha_{t-1}L^2}{bM} (J_{t-1} + Q_{t-1})
\]

Then telescope from $\tilde{t}_{s-1} + 1$ to $\tilde{t}_s$, we have:

\[
\frac{H_{\tilde{t}_s}}{\alpha_{\tilde{t}_{s-1}}} - \frac{H_{\tilde{t}_{s-1}}}{\alpha_{\tilde{t}_{s-2}}} \leq -\frac{48^2}{bM\mu^2} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t H_t + \frac{2c_u^2}{bM} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t^3 \sigma^2 + \frac{8\eta^2\tilde{L}^2}{bM} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_{t-1} (D_t + E_t) \\
+ \frac{8\gamma^2\tilde{L}^2}{bM} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_{t-1} (F_t + G_t) + \frac{8\tau^2\alpha_{t-1}L^2}{bM} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t (J_t + Q_{t-1})
\]

Next from Lemma [C.5] for $t \neq \tilde{t}_s$, we have:

\[
B_{t+1} - B_t \leq -\frac{\mu\gamma\alpha_t B_t}{4} - \frac{\gamma^2\alpha_t F_t}{4} + \frac{9\gamma\alpha_t C_t}{\mu} + \frac{9\kappa^2\eta^2\alpha_t E_t}{2\mu\gamma} \\
+ \frac{9\gamma\alpha_t L^2}{\mu} \sum_{\ell=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} I_{\ell}^2 \gamma^2 \alpha_{\ell}^2 D_{\ell} + \frac{9\gamma\alpha_t L^2}{\mu} \sum_{\ell=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} I_{\ell}^2 \gamma^2 \alpha_{\ell}^2 G_{\ell}
\]

When $t = \tilde{t}_s$, we do not have the last two terms in the above inequality. Next, we telescope from $\tilde{t}_{s-1} + 1$ to $\tilde{t}_s$ and have:

\[
B_{\tilde{t}_s} - B_{\tilde{t}_{s-1}} \leq -\frac{\mu\gamma}{4} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t B_t - \frac{\gamma^2}{4} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t F_t + \frac{9\gamma}{\mu} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t C_t + \frac{9\kappa^2\eta^2}{2\mu\gamma} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t E_t \\
+ \frac{9I\eta^2\gamma L^2}{\mu} \sum_{t=\tilde{t}_{s-1}+1}^{\tilde{t}_{s-1}} \alpha_t \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t^2 D_{\ell} + \frac{9I\gamma^2 L^2}{\mu} \sum_{t=\tilde{t}_{s-1}+1}^{\tilde{t}_{s-1}} \alpha_t \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t^2 G_{\ell} \\
\leq -\frac{\mu\gamma}{4} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t B_t - \frac{\gamma^2}{4} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t F_t + \frac{9\gamma}{\mu} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t C_t + \frac{9\kappa^2\eta^2}{2\mu\gamma} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t E_t \\
+ \frac{9L^2\eta^2\gamma^2}{16L^2\mu} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t D_{\ell} + \frac{9L^2\gamma^3}{16L^2\mu} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t G_{\ell}
\]

where we use the fact that $\alpha_t < \frac{1}{16L^2}$. Next, from Lemma [C.6] we have:

\[
I_{t+1} - I_t \leq -\frac{\mu\tau\alpha_t I_t}{4} - \frac{\tau^2\alpha_t}{4} Q_t + \frac{9\kappa^2\eta^2\alpha_t E_t}{2\mu\tau} + \frac{9\tau\alpha_t}{\mu} H_t \\
+ \frac{18I\eta^2\tau\alpha_t L^2}{\mu} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t^2 D_{\ell} + \frac{18I\gamma^2\tau\alpha_t L^2}{\mu} \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t^2 G_{\ell} + 18I\tau^3\alpha_t L^2 \sum_{t=\tilde{t}_{s-1}}^{\tilde{t}_{s-1}} \alpha_t^2 J_{\ell}
\]
when $t = \bar{t}_s$, we do not have the last three terms in the above inequality. Next, we telescope from $t_{s-1} + 1$ to $t_s$ and have:

\[
I_{t_s} - I_{t_{s-1}} \leq -\frac{\mu \tau}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t I_t - \frac{\tau^2}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t Q_t + \frac{9\kappa^2 \eta^2}{2\mu \tau} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t E_t + \frac{9\tau}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t H_t \\
+ \frac{18I\eta^2 \tau^2 \bar{L}^2}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 D_t + \frac{18I\gamma^2 \tau^2 \bar{L}^2}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 G_t \\
+ \frac{18I\gamma^3 \bar{L}^2}{16^2 \mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 J_t \\
\leq -\frac{\mu \tau}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t I_t - \frac{\tau^2}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t Q_t + \frac{9\kappa^2 \eta^2}{2\mu \tau} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t E_t + \frac{9\tau}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t H_t \\
+ \frac{18I\eta^2 \tau^2 \bar{L}^2}{16^2 \mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 D_t + \frac{18I\gamma^2 \tau^2 \bar{L}^2}{16^2 \mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 G_t \\
+ \frac{18I\gamma^3 \bar{L}^2}{16^2 \mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 J_t 
\]  

(27)

Next, by Lemma C.13 when $t + 1 \neq \bar{t}_s$, we have:

\[
\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta \alpha_t}{4} E_t - \frac{\eta \alpha_t}{2} \mathbb{E} [\| \nabla h(\bar{x}_t) \|^2] + \eta \alpha_t A_t + 4\bar{L}^2 \eta \alpha_t B_t + 4\bar{L}^2 \eta \alpha_t I_t \\
+ 2\bar{L}^2 I \eta^3 \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 D_t + 4\bar{L}^2 I \eta^2 \gamma \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 G_t 
\]

When $t = \bar{t}_s$, we do not have the last two terms. Next, we telescope from $\bar{t}_{s-1}$ to $\bar{t}_s - 1$ to have:

\[
\mathbb{E}[h(\bar{x}_{\bar{t}_s}) - h(\bar{x}_{\bar{t}_{s-1}})] \\
\leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \frac{\eta \alpha_t}{4} E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \frac{\eta \alpha_t}{2} \mathbb{E} [\| \nabla h(\bar{x}_t) \|^2] + 4\bar{L}^2 \eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t I_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \eta \alpha_t A_t \\
+ \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} 4\bar{L}^2 \eta \alpha_t B_t + 2\bar{L}^2 I \eta^3 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 D_t + 4\bar{L}^2 I \eta^2 \gamma \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t \sum_{t=\bar{t}_{s-1}}^{t-1} \alpha_t^2 G_t \\
\leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \frac{\eta \alpha_t}{4} E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \frac{\eta \alpha_t}{2} \mathbb{E} [\| \nabla h(\bar{x}_t) \|^2] + 4\bar{L}^2 \eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t I_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \eta \alpha_t A_t \\
+ \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} 4\bar{L}^2 \eta \alpha_t B_t + \frac{\eta^3}{128} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t D_t + \frac{2 \eta}{64} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s} \alpha_t G_t 
\]

(28)

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

35
Combine Eq. (23), Eq. (24), Eq. (26) and Eq. (28) and we have:

$$E[G_{t_i}] - E[G_{t_{i-1}}] \leq - \sum_{t=t_{i-1}}^{t_{i-1}} \frac{\eta \alpha}{2} E[\|\nabla h(\bar{x}_t)\|^2] + \left( \frac{9\eta^2 \sigma^2}{32} + \frac{9\eta^2 \sigma^2}{32} + \frac{9\eta^2 \sigma^2}{32} \right) \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t^3$$

$$- \frac{L^2 \eta}{2} \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t B_t - \frac{L^2 \eta}{2} \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t I_t - \frac{9\eta \gamma L^2}{2} \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t F_t$$

$$- \sum_{t=t_{i-1}}^{t_{i-1}} \left( \frac{9\eta \gamma^2 \eta^2}{4} - \frac{9\eta \gamma^2 \eta^2}{4} \right) \eta \alpha_t E_t$$

$$- \sum_{t=t_{i-1}}^{t_{i-1}} \left( \frac{9\eta \gamma^2 \eta^2}{4} - \frac{9\eta \gamma^2 \eta^2}{4} \right) \eta \alpha_t D_t$$

By the condition that $\eta < \frac{\sqrt{C_{11}}}{90k^2L}$ and $\gamma < \frac{1}{32} < \frac{1}{2\mu}$. Next, we denote:

$$C_1 = \frac{1}{64} + \frac{81\kappa^2}{32} + 9L^2 = O(k^2)$$

Then, we have:

$$E[G_{t_i}] - E[G_{t_{i-1}}] \leq - \sum_{t=t_{i-1}}^{t_{i-1}} \frac{\eta \alpha}{2} E[\|\nabla h(\bar{x}_t)\|^2] + \left( \frac{9\eta^2 \sigma^2}{32} + \frac{9\eta^2 \sigma^2}{32} + \frac{9\eta^2 \sigma^2}{32} \right) \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t^3$$

$$- \frac{9\eta \gamma L^2}{8} \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t F_t - \frac{L^2 \eta}{8} \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t E_t - \frac{L^2 \eta}{2} \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t B_t$$

$$- \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t Q_t + C_1 \eta \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t D_t + C_1 \eta \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t G_t$$

$$- \frac{9\eta \gamma^2 \eta^2}{4} \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t E_t$$

Combine Eq. (29) with Lemma C.11 and use the condition that $\eta < \min \left( \frac{1}{8C_{11}^2}, \frac{L}{4\kappa_2^2}, 1 \right)$, $\gamma < \min \left( \frac{1}{8C_{11}^2}, \frac{L}{4\kappa_2^2}, 1 \right)$ and $\tau < \min \left( \frac{1}{8C_{11}^2}, \frac{L}{4\kappa_2^2}, 1 \right)$ we have:

$$E[G_{t_i}] - E[G_{t_{i-1}}] \leq - \sum_{t=t_{i-1}}^{t_{i-1}} \frac{\eta \alpha}{2} E[\|\nabla h(\bar{x}_t)\|^2] + C_{\sigma, \zeta} \eta \sum_{t=t_{i-1}}^{t_{i-1}} \alpha_t^3$$

For ease of notation, we denote

$$C_{\sigma, \zeta} = \left( 4C_2^2 \sigma^2 + 4C_2^2 \sigma^2 + 3C_2^2 \sigma^2 + 3C_2^2 \sigma^2 + 3C_2^2 \sigma^2 + \frac{3C_2^2 \sigma^2}{\mu^2} + \frac{12C_2^2 \sigma^2}{\mu^2} \right)$$

Next, sum over all $s \in [S]$ (assume $T = SI + 1$ without loss of generality), we have:

$$E[G_T] - E[G_1] \leq - \sum_{t=1}^{T-1} \frac{\eta \alpha}{2} E[\|\nabla h(\bar{x}_t)\|^2] + \eta C_{\sigma, \zeta} \sum_{t=1}^{T-1} \alpha_t^3$$
Rearranging the terms and use the fact that $\alpha_t$ is non-increasing, we have:

$$
\frac{\eta \alpha^T}{2} \sum_{t=1}^{T-1} \mathbb{E} \left[ \left\| \nabla h(\bar{x}_t) \right\|^2 \right] \leq \mathbb{E} [G_1] - \mathbb{E} [G_T] + \eta C_{\sigma, \zeta} \sum_{t=1}^{T-1} \alpha_t^3
$$

$$
\leq h(x_1) - h^* + \frac{9bM \eta A_1}{64 \alpha_1} + \frac{18 \eta \tilde{L}^2 B_1}{\mu \gamma} + \frac{9bM \eta C_1}{64 \alpha_1} + \frac{9bM \eta H_1}{64 \alpha_1} + \frac{18 \eta \tilde{L}^2 I_1}{\mu \tau} + \eta C_{\sigma, \zeta} \sum_{t=1}^{T-1} \alpha_t^3
$$

where we use $G_T \geq h^*$ ($h^*$ is the optimal value of $h$), and for the last term, we use the following fact:

$$
\sum_{t=1}^{T} \alpha_t^3 = \sum_{t=1}^{T} \frac{\delta^3}{u + \sigma^2 \bar{t}} \leq \sum_{t=1}^{T} \frac{\delta^3}{\sigma^2 + \sigma^2 \bar{t}} = \frac{\delta^3}{\sigma^2} \sum_{t=1}^{T} \frac{1}{1 + t} \leq \frac{\delta^3}{\sigma^2} \ln(T+1) = \frac{b^2 M^2 \ln(T+1)}{16 L}
$$

the first inequality follows $u_t > \sigma^2$, the last inequality follows Proposition E.2

Next, we denote the initial sub-optimality as $\Delta = h(\bar{x}_1) - h^*$, initial inner variable estimation error i.e. $B_1 = \|y_1 - y_{x_1}\|^2 \leq \Delta_y$, and the initial hyper-gradient computation error $I_1 = \|u_1 - [\nabla g(x_1, y_{x_1})]^{-1} \nabla f(x_1, y_{x_1})\|^2 \leq \Delta_u$.

Furthermore, we have $A_1 = \leq \frac{\sigma^2}{b_1 M}, C_1 \leq \frac{\sigma^2}{b_1 M}, H_1 \leq \frac{\sigma^2}{b_1 M}$ where $b_1$ be the size of the first batch. Then, we divide both sides by $\eta \alpha T/2$ to have:

$$
\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E} \left[ \left\| \nabla h(\bar{x}_t) \right\|^2 \right] \leq \left( \frac{2 \Delta}{\eta} + \frac{27 b^2 \sigma^2}{32 b_1 \alpha_1} + \frac{36 \tilde{L}^2 \Delta_y}{\mu \gamma} + \frac{36 \tilde{L}^2 \Delta_u}{\mu \tau} + \frac{b^2 \mu^2 C_{\sigma, \zeta} \ln(T)}{8 \tilde{L}} \right) \frac{1}{T \alpha T}
$$

Note that we have:

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

where the inequality uses the fact that $(x + y)^{1/3} \leq x^{1/3} + y^{1/3}$. In particular, when $t = 1$, we have

$$
\frac{1}{\alpha_1} \leq \frac{u^{1/3} + \sigma^{2/3}}{\delta} = \frac{(16 \tilde{L})^{1/3}((b_1 M)^{2/3} \eta^{1/3} + 1)}{(b_1 M)^{2/3}} \quad (30)
$$

when $t = T$, we have:

$$
\frac{1}{\alpha_T} \leq \frac{u^{1/3}}{\delta T} + \frac{\sigma^{2/3}}{\delta T^{2/3}} = \left(16 \tilde{L} \bar{u}^{1/3} \frac{T}{(b_1 M)^{2/3}}ight) \quad (31)
$$

In summary, we have:

$$
\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E} \left[ \left\| \nabla h(\bar{x}_t) \right\|^2 \right] \leq \left( \frac{2 \Delta}{\eta} + \frac{27 b^2 \sigma^2}{32 b_1 \alpha_1} + \frac{36 \tilde{L}^2 \Delta_y}{\mu \gamma} + \frac{36 \tilde{L}^2 \Delta_u}{\mu \tau} + \frac{b^2 \mu^2 C_{\sigma, \zeta} \ln(T)}{8 \tilde{L}} \right) \left( \frac{(16 \tilde{L} \bar{u})^{1/3}}{T} + \frac{(16 \tilde{L})^{1/3}}{(b_1 M)^{2/3}} \right)
$$

Recall that $\tilde{L} = O(\kappa), \bar{L} = O(\kappa^3)$, therefore we have $c_{\nu} = \Theta((b_1 M)^{-1}), c_{\omega} = \Theta(\kappa^2 (b_1 M)^{-1})$, $c_u = \Theta(\kappa^2 (b_1 M)^{-1}) \bar{u} = \Theta(I_3 \kappa^3)$, then for $\eta, \gamma, \tau$, we have:

$$
\gamma < \min \left( \frac{1}{8c_1^{1/2}}, \frac{\tilde{L}}{4c_1^{1/2}}, \frac{\tilde{L}^2}{c_{\nu}}, \frac{\tilde{L}^2}{c_{\omega}}, \frac{\tilde{L}^2}{c_u}, \frac{1}{2\tilde{L}}, 1 \right)
$$

$$
\tau < \min \left( \frac{1}{8c_1^{1/2}}, \frac{\tilde{L}}{4c_1^{1/2}}, \frac{\tilde{L}^2}{c_{\nu}}, \frac{\tilde{L}^2}{c_{\omega}}, \frac{1}{2\tilde{L}}, 1 \right)
$$

37
\[ \eta < \min \left( \frac{\mu \gamma}{36\kappa \bar{L}}, \frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, \frac{\tilde{L}^2}{c_v}, \frac{\tilde{L}^2}{c_v}, \frac{\tilde{L}^2}{c_v}, \frac{1}{2\bar{L}}, 1 \right) \]

where \( C_1 = O(\kappa^2) \), so we have \( \gamma^{-1} = O(\kappa) \), \( \eta^{-1} = O(\kappa^3) \), \( \tau^{-1} = O(\kappa) \), furthermore, \( \alpha_1^{-1} = O(I\kappa^{4/3}) \), \( C_{\sigma, \zeta} = O(\kappa^5(bM)^{-2}) \), assume we choose the size of the first batch to be \( b_1 = Ib \).

Combine everything together, we have:

\[ \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E} \left[ \| \nabla h(\bar{x}_t) \|^2 \right] = O \left( \frac{\kappa^{19/3}I}{T} + \frac{\kappa^{16/3}}{(bMT)^{2/3}} \right) \]

To reach an \( \epsilon \)-stationary point, we need \( T = O(\kappa^8(bM)^{-1}\epsilon^{-1.5}) \), \( I = O(\kappa^{5/3}(bM)^{-1}\epsilon^{-0.5}) \). The communication cost is \( E = T/I \geq \kappa^{19/3}\epsilon^{-1} \), the sample complexity is \( G_c(f, \epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5}) \), \( G_c(g, \epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5}) \), \( J_v(g, \epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5}) \), \( H_v(g, \epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5}) \).
Algorithm 2 Federated Bilevel Optimization (FedBiO)

1: Input: Initial states $x_1, y_1$ and $u_1$; learning rates $\{\gamma_t, \eta_t, \tau_t\}_{t=1}^T$
2: Initialization: Set $x_1^{(m)} = x_1$, $y_1^{(m)} = y_1$, $u_1^{(m)} = u_1$;
3: for $t = 1$ to $T$ do
4:   Randomly sample mutually independent minibatch of samples $B_y$ and $B_x = \{B_y, B_x\}$ of size $b$;
5:   $\omega_t^{(m)} = \nabla_y g_t^{(m)}(x_t^{(m)}, y_t^{(m)}, B_y)$
6:   $\nu_t^{(m)} = \nabla_x f_t^{(m)}(x_t^{(m)}, y_t^{(m)}, B_x) - \nabla_{xy} g_t^{(m)}(x_t^{(m)}, y_t^{(m)}, B_y)u_t^{(m)}$;
7:   $\hat{y}_{t+1} = y_t - \gamma_t \omega_t^{(m)}$, $\hat{x}_{t+1} = x_t - \eta_t \nu_t^{(m)}$;
8:   if $t \mod 1 = 0$ then
9:     $\gamma_t^{(m)} = \frac{1}{M} \sum_{j=1}^M \gamma_t(j)$, $x_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M \hat{x}_t(j)$
10:   else
11:     $\gamma_t^{(m)} = \gamma_t^{(m)}$, $x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)}$
12:   end if
13:   $u_{t+1}^{(m)} = \mathcal{P}_y(\tau_t \nabla_y f_t^{(m)}(x_t^{(m)}, y_t^{(m)}, B_{t+1})) + (I - \tau_t \nabla y g_t^{(m)}(x_t^{(m)}, y_t^{(m)}, B_{t+1}))(t_{t+1}^{(m)})$;
14:   if $t \mod 1 = 0$ then
15:     $u_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M u_t(j)$
16:   else
17:     $u_{t+1}^{(m)} = u_t^{(m)}$
18:   end if
19: end for

C.2 Proof for the FedBiO Algorithm

Algorithm 2 follows Eq. 5 and we discuss its convergence property in this subsection.

C.2.1 Lower Problem Solution Error and Hyper-gradient Estimation Error

Lemma C.15. When $\gamma < \frac{1}{2L}$, we have:

$$
\mathbb{E}\|\bar{y}_t - y_{\bar{x}_t}\|^2 \leq (1 - \frac{\mu \gamma}{4})\mathbb{E}\|\bar{y}_{t-1} - y_{\bar{x}_{t-1}}\|^2 + \frac{9\kappa^2 \eta^2}{2\mu \gamma} \mathbb{E}\|\bar{y}_{t-1}\|^2 + \frac{\gamma^2}{4} \mathbb{E}\|\bar{w}_{t-1}\|^2 \leq \frac{9L^2 \gamma}{2\mu M} \sum_{m=1}^M \mathbb{E}\|x_t^{(m)} - \bar{x}_{t-1}\|^2 + \frac{5\gamma^2 \sigma^2}{4b_M} + \frac{5\eta^2 \tilde{L}^2}{\mu \gamma} \mathbb{E}\|\bar{y}_{t-1}\| \leq \frac{9L^2 \gamma}{2\mu M} \sum_{m=1}^M \mathbb{E}\|x_t^{(m)} - \bar{x}_{t-1}\|^2 + \frac{5\gamma^2 \sigma^2}{4b_M} + \frac{5\eta^2 \tilde{L}^2}{\mu \gamma} \mathbb{E}\|\bar{y}_{t-1}\|$$

Lemma C.16. Suppose we choose $\tau < \frac{1}{L}$, then we have:

$$
\mathbb{E}\|\bar{u}_{t+1} - u_{\bar{x}_{t+1}}\|^2 \leq (1 - \frac{\mu \tau}{4})\mathbb{E}\|\bar{u}_{t} - u_{\bar{x}_{t}}\|^2 + \frac{5\gamma^2 \sigma^2}{4b_M} + \frac{5\eta^2 \tilde{L}^2}{\mu \gamma} \mathbb{E}\|\bar{y}_{t}\| \leq \frac{3\gamma^2 \sigma^2}{4b_M} + \frac{\tau L^2 \gamma^2 C^2}{2\mu^2 M} \sum_{m=1}^M \mathbb{E}\|x_t^{(m)} - \bar{x}_{t-1}\|^2 + 2\|\bar{y}_{t} - y_t^{(m)}\|^2 + 2\|y_{\bar{x}_t} - \bar{y}_{t}\|^2
$$

We provide the proof for Lemma C.16 here and Lemma C.15 can be derived similarly.
Proof. First, by proposition 5 (set \( \alpha = 1 \)) and choose \( \gamma < \frac{1}{2L} \), we have:

\[
\mathbb{E}\|\tilde{y}_t - \tilde{x}_{t-1}\|^2 \leq (1 - \frac{\mu \gamma}{2}) \mathbb{E}\|\tilde{y}_t - y_{t-1}\|^2 - \frac{\gamma^2}{4} \mathbb{E}\|\tilde{x}_{t-1}\|^2 + \frac{4 \gamma \sigma^2}{\mu b} \\
+ \frac{4 \gamma}{\mu} \mathbb{E}\|\nabla g(x_{t-1}, y_{t-1}) - \nabla g(x_{t-1}, y_{t-1})\|^2 + \frac{4 \gamma \sigma^2}{\mu b} M \\
\leq (1 - \frac{\mu \gamma}{2}) \mathbb{E}\|\tilde{y}_t - y_{t-1}\|^2 - \frac{\gamma^2}{4} \mathbb{E}\|\tilde{x}_{t-1}\|^2 \\
+ \frac{4 \gamma}{\mu} M \sum_{m=1}^{M} \mathbb{E}\|\nabla g(x_{t-1}, y_{t-1}) - \nabla g(x_{t-1}, y_{t-1})\|^2 + \frac{4 \gamma \sigma^2}{\mu b} M \\
\leq (1 - \frac{\mu \gamma}{4}) \mathbb{E}\|\tilde{y}_t - y_{t-1}\|^2 - \frac{\gamma^2}{4} \mathbb{E}\|\tilde{x}_{t-1}\|^2 \\
+ \frac{4 L^2 \gamma}{\mu M} \sum_{m=1}^{M} \mathbb{E}\|\nabla g(x_{t-1}, y_{t-1}) - \nabla g(x_{t-1}, y_{t-1})\|^2 + \frac{4 \gamma \sigma^2}{\mu b} M \\
\leq (1 - \frac{\mu \gamma}{4}) \mathbb{E}\|\tilde{y}_t - y_{t-1}\|^2 - \frac{9 \mu \gamma^2}{2\mu^2} \mathbb{E}\|\tilde{y}_{t-1}\|^2 - \frac{\gamma^2}{4} \mathbb{E}\|\tilde{x}_{t-1}\|^2 \\
+ \frac{9 L^2 \gamma}{2\mu M} \sum_{m=1}^{M} \mathbb{E}\|\nabla g(x_{t-1}, y_{t-1}) - \nabla g(x_{t-1}, y_{t-1})\|^2 + \frac{4 \gamma \sigma^2}{\mu b} M \\
\]
Lemma C.18. For any $t \in [T]$, we have:

$$
\frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| (\nu_t^{(m)} - \bar{\nu}_t) \right\|^2 \right] \leq \frac{4L^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| u_t^{(m)} - \bar{u}_t \right\|^2 \right] + 4\zeta_f^2 + \frac{8C_2^2 \zeta_{g,xy}}{\mu^2} + \frac{2\sigma^2}{b_x}
$$

$$
+ \left( \frac{16L^2}{M} + \frac{32L^2 \mu C_f^2}{\mu^2 M^2} \right) \sum_{m=1}^{M} \mathbb{E} \left[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 + \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right]
$$

Lemma C.19. For $t \in T$, we have:

$$
\frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| (\omega_t^{(m)} - \bar{\omega}_t) \right\|^2 \right] \leq \frac{2L^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 \right] + \frac{2L^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right] + \frac{2\sigma^2}{b_y} + 2\zeta_f^2
$$

Lemma C.20. For $t \in [T]$, we have:

$$
\frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| (u_{t+1}^{(m)} - \bar{u}_{t+1}) \right\|^2 \right] \leq \left( 1 + \frac{1}{T} \right) \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| u_t^{(m)} - \bar{u}_t \right\|^2 \right] + \frac{64I^2 \tau^2 C_f^2 \zeta_{g,yy}}{\mu^2} + 32I \tau^2 \zeta_f^2 + \frac{2\tau^2 \sigma^2}{b_x}
$$

$$
+ \left( \frac{128I^2 \tau^2}{M} + \frac{256I^2 \tau^2 \mu^2 C_f^2}{M^2} \right) \sum_{m=1}^{M} \mathbb{E} \left[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 + \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right]
$$

Lemma C.18, Lemma C.20 bounds the local drift of $x_t^{(m)}$, $\omega_t^{(m)}$ and $u_{t+1}^{(m)}$. We provide the proof for Lemma C.18 here and the other two bounds can be derived similarly.

**Proof.** We have:

$$
\frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| (\nu_t^{(m)} - \bar{\nu}_t) \right\|^2 \right] \leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}) u_t^{(m)} \right\|^2 \right]
$$

$$
- \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}) u_t^{(m)} \right\|^2 \right] + \frac{2\sigma^2}{b_x}
$$

$$
\leq \frac{2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) \right\|^2 \right] - \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}) u_t^{(m)} \right\|^2 \right]
$$

For the term $T_1$, we have:

$$
T_1 \leq \frac{16}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_x f^{(m)}(\bar{x}_t, \bar{y}_t) \right\|^2 \right] + \frac{4}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| \nabla_x f^{(m)}(\bar{x}_t, \bar{y}_t) - \nabla_x f(\bar{x}_t, \bar{y}_t) \right\|^2 \right]
$$

$$
\leq \frac{16L^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 + \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right] + 4\zeta_f^2
$$

Next for the term $T_2$, we have:

$$
T_2 \leq \frac{4L^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| u_t^{(m)} - \bar{u}_t \right\|^2 \right] + \frac{4C_f^2}{\mu^2 M^2} \sum_{m=1}^{M} \sum_{j=1}^{M} \mathbb{E} \left[ \left\| \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}) \right\|^2 \right]
$$

$$
\leq \frac{4L^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| u_t^{(m)} - \bar{u}_t \right\|^2 \right] + \frac{32L^2 \mu C_f^2}{\mu^2 M} \sum_{m=1}^{M} \mathbb{E} \left[ \left\| x_t^{(m)} - \bar{x}_t \right\|^2 + \left\| y_t^{(m)} - \bar{y}_t \right\|^2 \right] + \frac{8C_2^2 \zeta_{g,xy}}{\mu^2}
$$

Combine everything together, we get the claim in the lemma. \(\square\)
Lemma C.18 and Lemma C.20 have recursive dependence of each other. Next, we provide an un-intertwined bound for each of them. For ease of notation, we denote $D_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \|v_t^{(m)} - \bar{v}_t\|^2$, $B_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \|u_t^{(m)} - \bar{u}_t\|^2$ and $C_t = \mathbb{E} \|y_t - y_{x_t}\|^2$.

Lemma C.21. For $\gamma \leq \frac{1}{8L_1}$ and $\eta < \frac{1}{8L_2}$, $\tau < \frac{1}{128L_2}$, where $\bar{L}_1 = (L^2 + \frac{2L^2C^2}{\mu^2})$ and $\bar{L}_2 = \frac{2L^2C^2}{\mu^2}$ are constants, then we have:

$$\sum_{t=t_a-1}^{t_a-1} D_t \leq 96I\zeta_f^2 + 16IC^2 + \frac{16C^2C^2}{\mu^2} + \frac{32IC^2}{\mu^2} + \frac{16I\sigma^2}{b_y} + \frac{20I\sigma^2}{b_x}$$

$$\sum_{t=t_a-1}^{t_a-1} B_t \leq 24I\zeta_f^2 + 8I\zeta_f^2 + \frac{4IC^2C^2}{\mu^2} + \frac{8IC^2}{\mu^2} + \frac{8I\sigma^2}{b_y} + \frac{5I\sigma^2}{b_x}$$

Proof. Based on Lemma C.18 and sum from $t_a - 1$ to $t_a - 1$, we have:

$$\sum_{t=t_a-1}^{t_a-1} D_t \leq 16I\eta^2\bar{L}_1^2 + \frac{16I\gamma^2\bar{L}_1^2}{\mu^2} + \frac{8(I-1)C^2}{\mu^2} + \frac{2(I-1)\sigma^2}{b_x}$$

$$\sum_{t=t_a-1}^{t_a-1} B_t \leq 16I\eta^2\bar{L}_1^2 + \frac{8(I-1)C^2}{\mu^2} + \frac{2(I-1)\sigma^2}{b_x}$$

where we denote $\bar{L}_1^2 = (L^2 + \frac{2L^2C^2}{\mu^2})$. Combine with the case of $t = t_a$ in Lemma C.18, we have:

$$\sum_{t=t_a-1}^{t_a-1} D_t \leq 16I\eta^2\bar{L}_1^2 + \frac{16I\gamma^2\bar{L}_1^2}{\mu^2} + \frac{8(I-1)C^2}{\mu^2} + \frac{2I\sigma^2}{b_x}$$

(32)

Based on Lemma C.19 and sum from $t_a - 1$ to $t_a - 1$, we have:

$$\sum_{t=t_a-1}^{t_a-1} B_t \leq 8I\eta^2L^2 + \frac{8(I-1)C^2}{\mu^2} + \frac{2(I-1)\sigma^2}{b_x}$$

Combine with the case of $t = t_a$ in Lemma C.19, we have:

$$\sum_{t=t_a-1}^{t_a-1} B_t \leq 8I\eta^2L^2 + \frac{8(I-1)C^2}{\mu^2} + \frac{2I\sigma^2}{b_x}$$

(33)

Apply Lemma C.20 recursively, we have:

$$A_t \leq \sum_{t=t_a-1}^{t_a-1} \left( \frac{1}{L} \right)^t \left( \frac{64I\gamma^2\bar{L}_1^2}{\mu^2} + \frac{32\gamma^2\zeta_f^2}{b_x} + \frac{2\gamma^2\sigma^2}{b_x} + 128I^2\gamma^2\bar{L}_2^2 + \frac{128I^2\gamma^2\bar{L}_2^2}{\mu^2} + \frac{128I^2\gamma^2\bar{L}_2^2}{\mu^2} B_t \right)$$

$$\leq \sum_{t=t_a-1}^{t_a-1} \left( \frac{192I^2\gamma^2\bar{L}_2^2}{\mu^2} + 96I^2\zeta_f^2 + \frac{6\gamma^2\sigma^2}{b_x} + 384I^2\gamma^2\bar{L}_2^2 B_t \right)$$

42
where we denote \( \tilde{L}_2^2 = (L^2 + \frac{2L^2C^2}{\mu^2}) \), and 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 \((1 + a/x)^2 \leq e^a\). Then we choose \( a = 1 \) and \( x = I \). Finally, we use the fact that \( e^1 \leq 3 \). Next, we sum from \( t_{s-1} \) to \( t_s - 1 \) to have:

\[
\sum_{t = t_{s-1}}^{t_s - 1} A_t \leq \sum_{t = t_{s-1}}^{t_s - 1} \left( \frac{48I^3\tau^2C^2\gamma^2}{\mu^2} + 96I^3\tau^2\gamma^2 + \frac{6I^3\tau^2\sigma^2}{b_x} + 24I^3\eta^2\tau^2\tilde{L}_2^2 \sum_{t = t_{s-1}}^{t_s - 1} D_t + 24I^3\gamma^2\tau^2\tilde{L}_2^2 \sum_{t = t_{s-1}}^{t_s - 1} B_t \right) \leq \frac{192I^3\tau^2C^2\gamma^2}{\mu^2} + 96I^3\tau^2\gamma^2 + \frac{6I^3\tau^2\sigma^2}{b_x} + 384I^4\eta^2\tau^2\tilde{L}_2^4 \sum_{t = t_{s-1}}^{t_s - 1} D_t + 384I^4\gamma^2\tau^2\tilde{L}_2^4 \sum_{t = t_{s-1}}^{t_s - 1} B_t \]

(34)

Combine Eq. 32, Eq. 33 and Eq. 34 and we choose \( \eta, \gamma \) and \( \tau \) such that \( I^2\gamma^2L^2 < \frac{1}{4}, I^2\tau^2\tilde{L}_1^2 < \frac{1}{64}, I^2\eta^2\tilde{L}_2^2 < \frac{1}{16} \). \( I^2\gamma^2L^2 < \frac{1}{4} \), \( I^2\eta^2\tilde{L}_2^2 < \frac{1}{16} \). we get the claim in the lemma, by using the fact that \( \tilde{L}_1 > L \) and \( \tilde{L}_2 > L \), we get the simplified condition in the lemma.

\[\square\]

C.2.3 Descend Lemma

**Lemma C.22.** For all \( t \in [t_{s-1}, t_s - 1] \), the iterates generated satisfy:

\[
\mathbb{E}\| \nabla h(\bar{x}_t) - \nabla h(\bar{x}_{t_s}) \|^2 \leq \frac{2L^2}{M} \sum_{m=1}^{M} \mathbb{E}[\| x_t - x^{(m)}_t \|^2 + 2\| y_t - y^{(m)}_t \|^2 + 2\| y_t - y^{(m)}_{t_s} \|^2 + 2\| y_{t_s} - y^{(m)}_{t_s} \|^2] + 4L^2\mathbb{E}\| u_{\bar{x}_t} - u_{\bar{x}_{t_s}} \|^2
\]

where we denote \( u_{\bar{x}_t} = [\nabla_y g(\bar{x}_t, y_{\bar{x}_t})]^{-1} \nabla_y f(\bar{x}_t, y_{\bar{x}_t}) \) and \( \tilde{L}_2^2 = (L^2 + \frac{2L^2C^2}{\mu^2}) \) is a constant.

**Proof.** By \( \nabla h(\bar{x}_t) = \Phi(\bar{x}_t, y_{\bar{x}_t}) \), we have:

\[
\mathbb{E}\| \nabla h(\bar{x}_t) - \nabla h(\bar{x}_{t_s}) \|^2 \leq \mathbb{E}\| \nabla x f(\bar{x}_t, y_{\bar{x}_t}) - \nabla x g(\bar{x}_t, y_{\bar{x}_t}) \| \mathbb{E}\| \nabla y g(\bar{x}_t, y_{\bar{x}_t}) \|^{-1} \nabla y f(\bar{x}_t, y_{\bar{x}_t})
\]

\[-\frac{1}{M} \sum_{m=1}^{M} \left( \left\| \nabla x f(\bar{x}_t, y_{\bar{x}_t}) - \nabla x g(\bar{x}_t, y_{\bar{x}_t}) \right\| \left\| \nabla y g(\bar{x}_t, y_{\bar{x}_t}) \right\|^{-1} \left\| \nabla y f(\bar{x}_t, y_{\bar{x}_t}) \right\| \right)^2
\]

\[\leq 2\mathbb{E}\| \nabla x f(\bar{x}_t, y_{\bar{x}_t}) - \frac{1}{M} \sum_{m=1}^{M} \nabla x f(\bar{x}_t, y_{\bar{x}_t}) \| ^2 + 2\mathbb{E}\| \nabla x g(\bar{x}_t, y_{\bar{x}_t}) - \frac{1}{M} \sum_{m=1}^{M} \nabla x g(\bar{x}_t, y_{\bar{x}_t}) \| ^2
\]

\[-\frac{1}{M} \sum_{m=1}^{M} \| \nabla y g(\bar{x}_t, y_{\bar{x}_t}) \|^{-1} \| \nabla y f(\bar{x}_t, y_{\bar{x}_t}) \| ^2
\]

We denote the two terms above as \( T_1, T_2 \) respectively. For the first term \( T_1 \), we have:

\[
T_1 \leq \frac{2L^2}{M} \sum_{m=1}^{M} \mathbb{E}[\| x_t - x^{(m)}_t \|^2 + 2\| y_t - y^{(m)}_t \|^2] \leq \frac{2L^2}{M} \sum_{m=1}^{M} \mathbb{E}[\| x_t - x^{(m)}_t \|^2 + 2\| y_t - y^{(m)}_t \|^2] + 2\| y_{t_s} - y^{(m)}_{t_s} \|^2]
\]

For the second term \( T_2 \), we have:

\[
T_2 \leq \frac{4C^2}{\mu^2M} \sum_{m=1}^{M} \mathbb{E}\| \nabla y g(\bar{x}_t, y_{\bar{x}_t}) - \nabla y g(\bar{x}_t, y_{\bar{x}_t}) \| ^2 + 4L^2\mathbb{E}\| \nabla y g(\bar{x}_t, y_{\bar{x}_t}) \| ^2
\]

We denote the first term above as \( T_{2,1} \). For the term \( T_{2,1} \), we have:

\[
T_{2,1} \leq \frac{4L^2C^2}{\mu^2M} \sum_{m=1}^{M} \mathbb{E}[\| x_t - x^{(m)}_t \|^2 + 2\| y_t - y^{(m)}_t \|^2] + 2\| y_{t_s} - y^{(m)}_{t_s} \|^2]
\]

Combine everything together, we get the claim in the lemma.

\[\square\]
Lemma C.23. For all \( t \in [\bar{t}_s - 1, \bar{t}_s - 1] \) and \( s \in [S] \), suppose \( \eta < \frac{1}{2\bar{L}} \), the iterates generated satisfy:

\[
\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4} \mathbb{E}\|
abla \mathbb{E}[\nu_t^{(m)}]\|^2 + \frac{\eta^2 \bar{L} \sigma^2}{2b_x M} + 2\eta L^2 \mathbb{E}\|u_{\bar{x}_t} - \bar{u}_t\|^2
\]

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

Proof. Using the smoothness of \( f \) we have:

\[
\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] + \mathbb{E} \langle \nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{\bar{L}}{2} \mathbb{E}\|\bar{x}_{t+1} - \bar{x}_t\|^2
\]

\[
= \mathbb{E}[h(\bar{x}_t)] - \eta \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 + \eta \mathbb{E}\|\nabla h(\bar{x}_t) - \nabla \mathbb{E}[\nu_t]\|^2 - \left( \frac{\eta}{2} - \frac{\eta^2 \bar{L}}{2} \right) \mathbb{E}\|\nabla \mathbb{E}[\nu_t]\|^2 + \frac{\eta^2 \bar{L} \sigma^2}{2b_x M}
\]

\[
\leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4} \mathbb{E}\|
abla \mathbb{E}[\nu_t^{(m)}]\|^2 + \frac{\eta^2 \bar{L} \sigma^2}{2b_x M}
\]

\[
+ \frac{\eta \bar{L}^2}{M} \sum_{m=1}^{M} \mathbb{E}\|\bar{x}_t - x^{(m)}_t\|^2 + 2\mathbb{E}\|\bar{y}_t - y^{(m)}_t\|^2 + 2\mathbb{E}\|u_{\bar{x}_t} - \bar{u}_t\|^2
\]

where equality (a) uses \( \langle a, b \rangle = \frac{1}{2}\|a\|^2 + \|b\|^2 - \|a - b\|^2 \); (b) follows the assumption that \( \eta < 1/2 \bar{L} \) and Lemma C.23.

C.2.4 Proof of Convergence Theorem

We first denote the following potential function \( G(t) \):

\[
G_t = \mathbb{E}[h(\bar{x}_t)] + \frac{9\eta \bar{L}^2}{\mu T} \mathbb{E}\|\bar{y}_t - y\|^2 + \frac{9\eta \bar{L}^2}{\mu T} \mathbb{E}\|\bar{u}_t - u\|^2
\]

Theorem C.24. Suppose we choose \( \tau = \min\left(\frac{1}{128T \bar{L}^2}, \frac{1}{144\gamma T}\right) \), then denote \( \tilde{\gamma} = \min\left(\frac{1}{8\tilde{L}^2}, \frac{7}{36\bar{L}^2}, \frac{1}{144\gamma T}\right) \), if we choose \( \gamma = \frac{\bar{L}^2}{36\mu T \bar{L}^1} \), and \( \gamma = \min\left(\tilde{\gamma}, \left(\frac{\Delta'}{C'_r} T\right)^{1/3}\left(\frac{\Delta' \bar{L}}{C'_r T}\right)^{1/3}\right) \) and \( r = \frac{C'}{\mu} \) where \( \Delta' \) and \( C'_r \) are constants denoted in Eq. 36 then we have:

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 = O\left(\frac{\kappa^8}{T} + \left(\frac{\kappa^{12}}{T^2}\right)^{1/3} + \frac{\kappa^4 \sigma^2}{b_y M} + \frac{\sigma^2}{b_x M}\right)
\]

and to reach an \( \epsilon \) stationary point, we choose the inner batch size \( b_y = O(M^{-1} \kappa^4 \epsilon^{-1}) \), upper batch size \( b_x = O(M^{-1} \epsilon^{-1}) \), and \( T = O(\kappa^8 \epsilon^{-1.5}) \) number of iterations.

Proof. Similar to Lemma C.21, we denote \( D_t = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nu_t^{(m)} - \nu_t\|^2 \), \( B_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\omega_t^{(m)} - \omega_t\|^2 \), \( C_t = \mathbb{E}\|\bar{y}_t - y_{\bar{x}_t}\|^2 \) and \( A_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\nu_t^{(m)} - \bar{u}_t\|^2 \), additionally, we denote \( E_t = \mathbb{E}\|\nabla \mathbb{E}[\nu_t]\|^2 \) and \( F_t = \mathbb{E}\|u_{\bar{x}_t} - \bar{u}_t\|^2 \). Combine Lemma C.15 Lemma C.16 and the definition of the
potential function we have:

\[
\mathcal{G}_{t_s} - \mathcal{G}_{t_{s-1}} \leq -\frac{\eta}{2} \sum_{t = t_{s-1}}^{t_s-1} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta L^2}{4} \sum_{t = t_{s-1}}^{t_s-1} F_t - \frac{\eta L^2}{4} \sum_{t = t_{s-1}}^{t_s-1} C_t \\
- \frac{\eta}{4} \left( 1 - \frac{162\kappa^2 \eta^2 L^2_\ell}{\mu^2 \gamma^2} - \frac{180\kappa^2 \eta^2 L^2_\ell}{\tau^2} \right) \sum_{t = t_{s-1}}^{t_s-1} E_t + \left( \tilde{L}^2_\ell + \frac{81\kappa^2 \tilde{L}^2_\ell}{2} + \frac{45\kappa^2 \tilde{L}^2_\ell}{16} \right) t^2 \eta^3 \sum_{t = t_{s-1}}^{t_s-1} D_t \\
+ (2 \tilde{L}^2_\ell + \frac{81\kappa^2 \tilde{L}^2_\ell}{2} + \frac{45\kappa^2 \tilde{L}^2_\ell}{16}) t^2 \gamma^2 \eta \sum_{t = t_{s-1}}^{t_s-1} B_t + \frac{81\kappa^2 \tilde{L}^2_\ell \eta^3 \sigma^2}{2\mu^2 \gamma^2 b_\times M} + \frac{36 \tilde{I}_\ell \eta \gamma \eta^2}{\mu^2 b_\times M} \\
+ \frac{45\kappa^2 \tilde{L}^2_\ell \eta^3 \sigma^2}{\tau^2 b_\times M} + \frac{45\kappa^2 \tilde{L}^2_\ell \eta \gamma \sigma^2}{4 b_\times M} + \frac{\eta^2 \tilde{I} \sigma^2}{2 b_\times M}
\]

to bound the coefficients above, we choose \( \eta \leq \min \left( \frac{\mu^2}{36\kappa \tilde{L}_\ell}, \frac{\tau}{36\kappa \tilde{L}_\ell}, \frac{1}{\tilde{t}_\ell} \right) \), \( \tau < \frac{1}{144 \kappa \tilde{L}_\ell} \) and we denote \( C_1 = (2 \tilde{L}^2_\ell + \frac{81\kappa^2 \tilde{L}^2_\ell}{2} + \frac{45\kappa^2 \tilde{L}^2_\ell}{16}) t^2 \). Then we have:

\[
\mathcal{G}_{t_s} - \mathcal{G}_{t_{s-1}} \leq -\frac{\eta}{2} \sum_{t = t_{s-1}}^{t_s-1} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta L^2}{4} \sum_{t = t_{s-1}}^{t_s-1} F_t - \frac{\eta L^2}{4} \sum_{t = t_{s-1}}^{t_s-1} C_t - \frac{\eta}{8} \sum_{t = t_{s-1}}^{t_s-1} E_t \\
+ C_1 \eta^3 \sum_{t = t_{s-1}}^{t_s-1} D_t + C_1 \gamma^2 \eta \sum_{t = t_{s-1}}^{t_s-1} B_t + \frac{I \eta \sigma^2}{2 b_\times M} + \frac{36 \tilde{I}_\ell \eta \gamma \eta^2}{\mu^2 b_\times M}
\]

Next, we combine with lemma [C.21] to have:

\[
\mathcal{G}_{t_s} - \mathcal{G}_{t_{s-1}} \leq -\frac{\eta}{2} \sum_{t = t_{s-1}}^{t_s-1} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta L^2}{4} \sum_{t = t_{s-1}}^{t_s-1} F_t - \frac{\eta L^2}{4} \sum_{t = t_{s-1}}^{t_s-1} C_t - \frac{\eta}{8} \sum_{t = t_{s-1}}^{t_s-1} E_t + \frac{I \eta \sigma^2}{2 b_\times M} + \frac{36 \tilde{I}_\ell \eta \gamma \eta^2}{\mu^2 b_\times M} \\
+ C_1 \eta^3 (96 I \zeta^2 + 16 I \zeta^2_{g,yy} + \frac{16 I \mathcal{C}^2_{g,yy}}{\mu^2} + \frac{32 I \mathcal{C}^2_{g,xy}}{\mu^2} + \frac{16 I \sigma^2}{b_\times} + \frac{20 I \sigma^2}{b_\times}) \\
+ C_1 \gamma^2 \eta (24 I \zeta^2 + 8 I \zeta^2_{g,yy} + \frac{4 I \mathcal{C}^2_{g,yy}}{\mu^2} + \frac{8 I \mathcal{C}^2_{g,xy}}{\mu^2} + \frac{8 I \sigma^2}{b_\times} + \frac{5 I \sigma^2}{b_\times})
\]

Sum over all \( s \in [S] \) (assume \( T = SI + 1 \) without loss of generality) to obtain:

\[
\frac{\eta}{2} \sum_{t = 1}^{T} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq \mathcal{G}_1 - \mathcal{G}_T + \frac{T \eta \sigma^2}{2 b_\times M} + \frac{36 T \tilde{I} \eta \gamma \eta^2}{\mu^2 b_\times M} \\
+ C_1 \eta^3 (96 I \zeta^2_f + 16 I \zeta^2_g + \frac{16 T \mathcal{C}^2_{g,yy}}{\mu^2} + \frac{32 T \mathcal{C}^2_{g,xy}}{\mu^2} + \frac{16 T \sigma^2}{b_\times} + \frac{20 T \sigma^2}{b_\times}) \\
+ C_1 \gamma^2 \eta (24 T \zeta^2_f + 8 T \zeta^2_g + \frac{4 T \mathcal{C}^2_{g,yy}}{\mu^2} + \frac{8 T \mathcal{C}^2_{g,xy}}{\mu^2} + \frac{8 T \sigma^2}{b_\times} + \frac{5 T \sigma^2}{b_\times})
\]

\[
\leq \Delta + \frac{9 \eta \tilde{L}^2 \Delta_u}{\mu^2} + \frac{9 \eta L^2 \Delta_u}{\mu^2 T} + \frac{T \eta \sigma^2}{2 b_\times M} + \frac{36 \tilde{I} \eta \gamma \eta^2}{\mu^2 b_\times M} \\
+ C_1 \eta^3 (96 I \zeta^2_f + 16 I \zeta^2_g + \frac{16 T \mathcal{C}^2_{g,yy}}{\mu^2} + \frac{32 T \mathcal{C}^2_{g,xy}}{\mu^2} + \frac{16 T \sigma^2}{b_\times} + \frac{20 T \sigma^2}{b_\times}) \\
+ C_1 \gamma^2 \eta (24 T \zeta^2_f + 8 T \zeta^2_g + \frac{4 T \mathcal{C}^2_{g,yy}}{\mu^2} + \frac{8 T \mathcal{C}^2_{g,xy}}{\mu^2} + \frac{8 T \sigma^2}{b_\times} + \frac{5 T \sigma^2}{b_\times})
\]

we define \( \Delta = h(x_1) - h^* \) as the initial sub-optimality of the function, \( \Delta_u = \|y_1 - y_{x_1}\|^2 \) as the initial sub-optimality of the inner variable estimation, \( \Delta_u = \|u_1 - u_{x_1}\|^2 \) as the initial sub-optimality
of the hyper-gradient estimation. Then we divide by $\eta T/2$ on both sides and have:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 \leq \frac{2\Delta}{\eta T} + \frac{18\tilde{L}_1^2 \Delta_y}{\mu \gamma T} + \frac{18L^2 \Delta_u}{\mu T} + \frac{\sigma^2}{b_x M} + \frac{72\tilde{L}_2^2 \sigma^2}{\mu^2 b_y M} + 2C_1 \eta^2 \left( 96\zeta_f^2 + 16\zeta_g^2 + \frac{16C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{32C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{16\sigma^2}{b_y} + \frac{20\sigma^2}{b_x} \right) + 2C_1 \gamma^2 \left( 24\zeta_f^2 + 8\zeta_g^2 + \frac{4C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{8C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{8\sigma^2}{b_y} + \frac{5\sigma^2}{b_x} \right)
$$

For ease of notation, we denote constants $C_\eta = 2C_1 \left( 96\zeta_f^2 + 16\zeta_g^2 + \frac{16C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{32C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{16\sigma^2}{b_y} + \frac{20\sigma^2}{b_x} \right)$ and $C_\gamma = 2C_1 \left( 24\zeta_f^2 + 8\zeta_g^2 + \frac{4C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{8C_1^2 \zeta_f \zeta_g}{\mu^2} + \frac{8\sigma^2}{b_y} + \frac{5\sigma^2}{b_x} \right)$, we have:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 \leq \frac{2\Delta}{\eta T} + \frac{18\tilde{L}_1^2 \Delta_y}{\mu \gamma T} + \frac{18L^2 \Delta_u}{\mu T} + \frac{\sigma^2}{b_x M} + \frac{72\tilde{L}_2^2 \sigma^2}{\mu^2 b_y M} + C_\eta \eta^2 + C_\gamma \gamma^2 \quad (35)
$$

Recall that, we have the condition that $\eta \leq \min \left( \frac{1}{8\tilde{L}_2}, \frac{\mu \gamma}{36\kappa L}, \frac{1}{128\tilde{L}_2}, \frac{1}{4L}, \frac{1}{8\tilde{L}_1} \right)$, $\gamma \leq \min \left( \frac{1}{8\tilde{L}_2}, \frac{\tau}{36\kappa L}, \frac{1}{4L}, \frac{1}{8\tilde{L}_1} \right)$, then denote

$$
\bar{\gamma} = \min \left( \frac{1}{8\tilde{L}_2}, \frac{\tau}{36\kappa L}, \frac{1}{4L}, \frac{1}{8\tilde{L}_1} \right),
$$

and let $\gamma \leq \bar{\gamma}$, and $\eta = \frac{\mu \gamma}{36\kappa L}$, then we have:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 \leq \frac{72\kappa \tilde{L}_1 \Delta + 18\tilde{L}_1^2 \Delta_y}{\mu \gamma T} + \left( \frac{C_\eta \mu^2}{36\kappa^2 \tilde{L}_1^2} + C_\gamma \right) \gamma^2 + \frac{18L^2 \Delta_u}{\mu T} + \frac{\sigma^2}{b_x M} + \frac{72\tilde{L}_2^2 \sigma^2}{\mu^2 b_y M}
$$

We denote

$$
C'_\gamma = \left( \frac{C_\eta \mu^2}{36\kappa^2 \tilde{L}_1^2} + C_\gamma \right), \quad C'_\delta = \frac{72\kappa \tilde{L}_1 \Delta + 18\tilde{L}_1^2 \Delta_y}{\mu}
$$

then we choose $\gamma$ as:

$$
\gamma = \min \left( \bar{\gamma}, \left( \frac{C'_\gamma}{C'_\delta \tau} \right)^{1/3} \right)
$$

and obtain:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 \leq \Delta' + \left( \frac{C'_\gamma \Delta' \gamma^2}{T^2} \right)^{1/3} + \frac{18L^2 \Delta_u}{\mu T} + \frac{\sigma^2}{b_x M} + \frac{72\tilde{L}_2^2 \sigma^2}{\mu^2 b_y M}
$$

Finally, since $\tilde{L}_1 = O(\kappa), L = O(\kappa^3)$, suppose we choose $I = O(1)$, then we have $\tau^{-1} = O(\kappa)$, $\bar{\gamma}^{-1} = O(\kappa^5), C'_\gamma = O(\kappa^3), C_\gamma = O(\kappa^4), C_\eta = O(\kappa^6)$, then we have:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 = O \left( \frac{\kappa^8}{T} + \left( \frac{\kappa^{12}}{T^2} \right)^{1/3} + \frac{\kappa^4 \sigma^2}{b_y M} + \frac{\sigma^2}{b_x M} \right)
$$

and to reach an $\epsilon$ stationary point, we choose the inner batch size $b_y = O(M^{-1} \kappa^4 \epsilon^{-1})$, upper batch size $b_x = O(M^{-1} \epsilon^{-1})$, and $T = O(\kappa^8 \epsilon^{-1.5})$ number of iterations.
Algorithm 3 FedBiO- Local Lower Level Problem

1: **Input:** Initial states \( x_1, y_1 \); learning rates \( \{\gamma_t, \eta_t\}_{t=1}^{T} \)
2: **Initialization:** Set \( x_1^{(m)} = x_1, y_1^{(m)} = y_1 \);
3: for \( t = 1 \) to \( T \) do
4: Randomly sample mutually independent minibatch of samples \( B_y \) and \( B_x \) of size \( b \);
5: \( \omega_t^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, B_y) \) and \( \nu_t^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}, B_x) \);
6: \( \hat{y}_{t+1}^{(m)} = y_t^{(m)} - \gamma_t \omega_t^{(m)}, \hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta_t \nu_t^{(m)} \);
7: if \( t \mod I = 0 \) then
8: \( y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}, x_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^{M} \hat{y}_j^{(m)} \)
9: else
10: \( y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}, x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)} \)
11: end if
12: end for

Algorithm 4 FedBiOAcc- Local Lower Level Problem

1: **Input:** Constants \( c_o, c_r, \gamma, \eta \); learning rate schedule \( \{\alpha_t\}, t \in [T] \); initial state \( (x_1, y_1) \);
2: **Initialization:** Set \( \hat{y}_1^{(m)} = y_1, x_1^{(m)} = x_1, \omega_1^{(m)} = \nabla_y g^{(m)}(x_1, y_1, B_y), \nu_1^{(m)} = \Phi^{(m)}(x_1, y_1, B_x) \) for \( m \in [M] \)
3: for \( t = 1 \) to \( T \) do
4: \( \omega_{t+1}^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, B_y) + (1 - c_o \alpha_t^2) \omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, B_y) \)
5: \( \nu_{t+1}^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}, B_x) + (1 - c_r \alpha_t^2) \nu_t^{(m)} - \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}, B_x) \)
6: if \( t \mod I = 0 \) then
7: \( \omega_{t+1}^{(m)} = \omega_{t+1}^{(m)}, \nu_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^{M} \nu_j^{(m)} \)
8: else
9: \( \omega_{t+1}^{(m)} = \omega_{t+1}^{(m)}, \nu_{t+1}^{(m)} = \nu_{t+1}^{(m)} \)
10: end if
11: end for

D Proof for Local Lower Level Problem

The FedBiOAcc-Local and FedBiO-Local are presented in Algorithm 4 and Algorithm 3 respectively. Then in this section, we discuss the convergence rate of the two algorithms. Please see Theorem D.12 and Theorem D.19 for the convergence rates.

For Eq. (10), we also assume Assumptions 3.1, 3.4 with a slightly different assumption to the heterogeneity as follows:

**Assumption D.1.** For any \( m, j \in [M] \) and \( z = (x, y) \), we have:
\[
\| \nabla f^{(m)}(z) - \nabla f^{(j)}(z) \| \leq \zeta_f, \\
\| \nabla_x g^{(m)}(z) - \nabla_x g^{(j)}(z) \| \leq \zeta_g, \\
\| \nabla_y g^{(m)}(z) - \nabla_y g^{(j)}(z) \| \leq \zeta_g, \\
\| y_x^{(m)} - y_x^{(j)} \| \leq \zeta_g, \\
\| y_y^{(m)} - y_y^{(j)} \| \leq \zeta_g, \\
\| x_x^{(m)} - x_x^{(j)} \| \leq \zeta_g, \\
\| x_y^{(m)} - x_y^{(j)} \| \leq \zeta_g,
\]
where \( \zeta_f, \zeta_g, \zeta_{g,x}, \zeta_{g,y}, \zeta_{g,x,y} \) are constants.

Note that we remove the requirement of gradient dissimilarity \( \zeta_g \) in Assumption 3.5 and add the dissimilarity bound \( \zeta_g \) for the minimizer of the lower level problem. Note that Assumption D.1 is a sufficient condition such that the dissimilarity of local hyper-gradient is bounded by some constant \( \zeta \).

**Proposition D.2.** (Lemma 4 and 7 in [59]) Suppose Assumptions 3.2, 3.3, and 3.4 hold and \( \tau < \frac{1}{T} \), the hypergradient estimator \( \Phi(x, y; B_x) \) w.r.t. \( x \) based on a minibatch \( B_x \) has bounded variance and bias:

\( a) \| \mathbb{E}[\Phi^{(m)}(x, y; B_x)] - \Phi^{(m)}(x, y) \| \leq G_1, \) where \( G_1 = R(1 - \tau \mu)^{Q+1} C_f \)
Proposition D.3. Suppose Assumptions 3.2 and 3.3 hold, the following statements hold:

a) \( y'(m) \) is Lipschitz continuous in \( x \) with constant \( \rho = \kappa \), where \( \kappa = \frac{\mu}{\lambda} \) is the condition number of \( g'(m)(x, y) \).

b) \( \| \Phi^m(x_1; y_1) - \Phi^m(x_2; y_2) \|^2 \leq \hat{L}^2 (\| x_1 - x_2 \|^2 + \| y_1 - y_2 \|^2) \), where \( \hat{L} = O(\kappa^2) \).

d) \( h'(m)(x) \) is Lipschitz continuous in \( x \) with constant \( \tilde{L} \) i.e., for any given \( x_1, x_2 \in X \), we have \( \| \nabla h'(m)(x_2) - \nabla h'(m)(x_1) \| \leq \tilde{L} \| x_2 - x_1 \| \) where \( \tilde{L} = O(\kappa^3) \).

This is a standard result in bilevel optimization and we omit the proof here.

Proposition D.4. In Eq. 10 suppose Assumption 3.1, 3.2, 3.3, 3.4 hold, we have:

\[
\| \nabla h'(m)(x) - \nabla h'(j)(x) \| \leq (1 + \kappa) \zeta_f + \frac{C_f}{\mu} \zeta_{g,xy} + \frac{\kappa C_f}{\mu} \zeta_{g,yy} + \left( (1 + \kappa) L + \frac{C_f L_{xy}}{\mu} + \frac{\kappa C_f L_{yy}}{\mu} \right) \zeta_g^* := \zeta
\]

Proof. For \( h'(m)(x) = f'(m)(x, y'(m)) \), \( m \in [M] \) in Eq. 10 we have:

\[
\| \nabla h'(m)(x) - \nabla h'(j)(x) \| = \| \nabla_x f'(m)(x, y'(m)) - \nabla_x g'(m)(x, y'(m)) \| \nabla_y g'(m)(x, y'(m))^{-1} \| \nabla_y f'(m)(x, y'(m)) \|
\]

\[
\leq \| \nabla_x f'(m)(x, y'(m)) - \nabla_x f'(j)(x, y'(j)) \| + \| \nabla_x g'(m)(x, y'(m)) - \nabla_x g'(j)(x, y'(j)) \| \cdot \| \nabla_y g'(m)(x, y'(m))^{-1} \| \nabla_y f'(j)(x, y'(j)) \|
\]

Next we bound the three terms separately. For the first term:

\[
\| \nabla_x f'(m)(x, y'(m)) - \nabla_x f'(j)(x, y'(j)) \| \leq \| \nabla_x f'(m)(x, y'(m)) - \nabla_x f'(j)(x, y'(m)) \|
\]

\[
+ \| \nabla_x f'(j)(x, y'(j)) - \nabla_x f'(j)(x, y'(m)) \| \leq \zeta_f + L \| y'(m) - y'(j) \| \leq \zeta_f + L \zeta_g^* \quad (37)
\]

where the second inequality is due to Assumption 3.2, and Assumption D.1. The last inequality also follows the Assumption D.1. Next, for the second term, we have:

\[
\| \nabla_x g'(m)(x, y'(m)) - \nabla_x g'(j)(x, y'(j)) \| \leq C_f \| \nabla_x g'(m)(x, y'(m)) - \nabla_x g'(j)(x, y'(j)) \|
\]

\[
\leq C_f \| \nabla_y g'(m)(x, y'(m))^{-1} \| \| \nabla_y f'(j)(x, y'(j)) \|
\]

\[
\leq C_f \| \nabla_y g'(m)(x, y'(m))^{-1} \| \| \nabla_y g'(m)(x, y'(m))^{-1} \| \| \nabla_y f'(j)(x, y'(j)) \|
\]

\[
\leq C_f \| \nabla y_{g,xy} \| \| \nabla y_{g,yy} \| \| \nabla y_{g,yy} \| \| \nabla y_{g,yy} \| \| \nabla y_{g,yy} \|
\]

\[
\leq C_f \| \nabla y_{g,xy} \| \| \nabla y_{g,yy} \| \| \nabla y_{g,yy} \| \| \nabla y_{g,yy} \| \| \nabla y_{g,yy} \|
\]

where the first inequality follows from the Assumption 3.1, 3.2, the third inequality follows from Assumption D.1, 3.3, the last inequality follows from Assumption D.1. Next, for the third term, we
have:
\[
\|\nabla_{xy} g(x, y) \| (\nabla_{yy} g^t(x, y^m) - \nabla_y f^m(x, y^m)) - (\nabla_{yy} g^j(x, y^j))^{-1} \nabla_y f^j(x, y^j) \|
\leq L \| (\nabla_{yy} g^m(x, y^m))^{-1} \nabla_y f^m(x, y^m) - (\nabla_{yy} g^j(x, y^j))^{-1} \nabla_y f^j(x, y^j) \|
\leq L \| (\nabla_{yy} g^m(x, y^m))^{-1} \| \| \nabla_y f^m(x, y^m) - \nabla_y f^j(x, y^j) \|
+ L \| (\nabla_{yy} g^j(x, y^j))^{-1} \|
\leq \frac{L \| (\nabla_{yy} g^m(x, y^m))^{-1} \| + C_f L \| (\nabla_{yy} g^j(x, y^j))^{-1} \|}{\mu} 
\leq \frac{L(\zeta_f + L\zeta_g)}{\mu} + \frac{C_f L(\zeta_g, y^m) + C_f L_{xy} g^m}{\mu} + \frac{L(\zeta_f + L\zeta_g)}{\mu}
\]
where the first inequality is by Assumption 3.3; the third inequality is by Assumption 3.2; the fourth inequality is by Cauchy Schwartz inequality; the last inequality is by Assumption 3.1 and the result in Eq. 3.7. Combine everything together, we have:
\[
\|\nabla_z f^m(x, y^m) - \nabla_z f^j(x, y^j)\| \leq \zeta_f + L\zeta_g + \frac{C_f \zeta_g, x^m}{\mu} + \frac{C_f L_{xy} g^m}{\mu} + \frac{L(\zeta_f + L\zeta_g)}{\mu}
\]
which completes the proof.

\[\square\]

D.1 Proof for the FedBioAcc-Local Algorithm

D.1.1 Hyper-Gradient Bias and Inner-Gradient Bias

**Lemma D.5.** Suppose we have \( c_\alpha^2 < 1 \), then:
\[
\mathbb{E}\left[\|\hat{\nu}_t - \mathbb{E}_\xi[\hat{\mu}_t, B]\|^2\right] \leq (1 - c_\alpha^2 - 1)\mathbb{E}\left[\|\hat{\nu}_{t-1} - \mathbb{E}_\xi[\hat{\mu}_{t-1}, B]\|^2\right] + \frac{2c_\alpha^2}{\alpha^2 - 1} \mathbb{E}\left[\|\hat{\mu}_{t-1} - \mathbb{E}_\xi[\hat{\mu}_{t-1}, B]\|^2\right] + \frac{2L^2}{\alpha^2 - 1} \sum_{m=1}^{M} \mathbb{E}\left[\|x_{m} - \bar{x}_{m}\|^2 + \|y_{m} - \bar{y}_{m}\|^2\right]
\]
where \( \mu_{t, \xi}^{(m)} = \Phi^{(m)}(x^{(m)}; \xi, B) \) and the expectation outside is w.r.t all the stochasticity of the algorithm.

**Proof.** For ease of notation, we denote \( \mu_{t, \xi}^{(m)} = \Phi^{(m)}(x^{(m)}; \xi, B) \), and \( \mu_{t}^{(m)} = \Phi^{(m)}(x^{(m)}; \xi, B) \), then by the definition of \( \hat{\nu}_t \) we have:
\[
\mathbb{E}\left[\|\hat{\nu}_t - \mathbb{E}_\xi[\hat{\mu}_t, B]\|^2\right] = \mathbb{E}\left[\|\mu_{t, \xi}^{(m)} - \mathbb{E}_\xi[\mu_{t, B}]\|^2\right]
\]
\[
= \mathbb{E}\left[\|\mu_{t, \xi}^{(m)} - (1 - c_\alpha^2 - 1)\mathbb{E}_\xi[\hat{\mu}_{t-1}, B] - \mathbb{E}_\xi[\hat{\mu}_{t-1}, B] - (1 - c_\alpha^2 - 1)\mathbb{E}_\xi[\hat{\mu}_{t-1}, B] + (1 - c_\alpha^2 - 1)\mathbb{E}_\xi[\hat{\mu}_{t-1}, B] + (1 - c_\alpha^2 - 1)\mathbb{E}_\xi[\hat{\mu}_{t-1}, B]\|\|^2\right]
\]
\[
\leq (1 - c_\alpha^2 - 1)\mathbb{E}\left[\|\hat{\nu}_{t-1} - \mathbb{E}_\xi[\hat{\mu}_{t-1}, B]\|^2\right] + \frac{1}{\alpha^2 - 1} \sum_{m=1}^{M} \mathbb{E}\left[\|\mu_{t, \xi}^{(m)} - \mathbb{E}_\xi[\mu_{t, \xi}^{(m)}] + (1 - c_\alpha^2 - 1)\mathbb{E}_\xi[\mu_{t, \xi}^{(m)}] - \mathbb{E}_\xi[\mu_{t, \xi}^{(m)}] - (1 - c_\alpha^2 - 1)\mathbb{E}_\xi[\mu_{t, \xi}^{(m)}] - \mathbb{E}_\xi[\mu_{t, \xi}^{(m)}]\|^2\right]
\]
where inequality (a) uses the fact that the cross product term is zero in expectation, the condition that \( c_v \alpha_1^2 < 1 \) and the fact that clients independently choose samples.

We denote the second term above as \( T_1 \), then we have:

\[
T_1^{(a)} \leq 2(c_v \alpha_1^2 - 1)E_f E_x [\| \mu_0 - \mu_1 \|^2] + 2(1 - c_v \alpha_1^2 - 1)E_f E_x [\| \mu_0 - \mu_1 \|^2]
\]

(b) follows Proposition E.2 due to the definition of \( \nu \), (b) follows Propse Proposition E.2 due to the generalized triangle inequality; (b) follows Proposition E.2 due to the definition of \( \nu \), (b) follows Proposition E.2 due to the bounded variance assumption; This completes the proof.

**Lemma D.6.** Suppose we have \( c_v \alpha_1^2 < 1 \), then we have:

\[
\frac{1}{M} \sum_{m=1}^M E_{\nu} \left[ \| \omega_t - \nabla g(m)(x_t^{(m)}, y_t^{(m)}) \|^2 \right] 
\]

\[
\leq \frac{1}{M} \sum_{m=1}^M E_{\nu} \left[ \| \omega_{t-1} - \nabla g(m)(x_{t-1}^{(m)}, y_{t-1}^{(m)}) \|^2 \right] + \frac{2(c_v \alpha_1^2 - 1)^2 \sigma^2}{b_y} 
\]

\[
+ \frac{2L^2}{b_m^2} \sum_{m=1}^M E_{\nu} \left[ \| x_t^{(m)} - x_{t-1}^{(m)} \|^2 + \| y_t^{(m)} - y_{t-1}^{(m)} \|^2 \right]
\]

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

The proof of Lemma D.6 can be derived similar as Lemma D.5.

**D.1.2 Lower Problem Solution Error**

**Lemma D.7.** Suppose we choose \( \gamma \leq \frac{1}{2L} \) and \( \alpha_t < 1 \). Then for \( t \neq t_s \), we have:

\[
E \left[ \| y_t^{(m)} - y_t^{(m)} \|^2 \right] \leq (1 - \frac{\mu \gamma \alpha_t - 1}{4})E \left[ \| y_{t-1}^{(m)} - y_{t-1}^{(m)} \|^2 \right] - \frac{\gamma^2 \alpha_t - 1}{4}E \left[ \| \omega_{t-1}^{(m)} \|^2 \right] 
\]

\[
+ \frac{9 \gamma^2 \alpha_t - 1}{2\mu}E \left[ \| \omega_{t-1}^{(m)} - \nabla g(m)(x_{t-1}^{(m)}, y_{t-1}^{(m)}) \|^2 \right] + \frac{9 \kappa^2 \gamma^2 \alpha_t - 1}{3\mu}E \left[ \| v_{t-1}^{(m)} \|^2 \right]
\]

for \( t = t_s \), we have:

\[
E \left[ \| y_t^{(m)} - y_t^{(m)} \|^2 \right] \leq (1 - \frac{\mu \gamma \alpha_t - 1}{4})E \left[ \| y_{t-1}^{(m)} - y_{t-1}^{(m)} \|^2 \right] - \frac{\gamma^2 \alpha_t - 1}{4}E \left[ \| \omega_{t-1}^{(m)} \|^2 \right] 
\]

\[
+ \frac{9 \gamma^2 \alpha_t - 1}{2\mu}E \left[ \| \omega_{t-1}^{(m)} - \nabla g(m)(x_{t-1}^{(m)}, y_{t-1}^{(m)}) \|^2 \right] 
\]

\[
+ \frac{9 \kappa^2 \gamma^2 \alpha_t - 1}{3\mu}E \left[ \| v_{t-1}^{(m)} \|^2 \right] + \frac{9 \kappa^2 \gamma^2 \alpha_t - 1}{3\mu}E \left[ \| v_t^{(m)} - \xi_t \|^2 \right]
\]

**Proof.** First, we exploit Proposition E.5 and choose the function \( g(m)(x_t^{(m)}, \cdot) \), by assumption it is \( L \) smooth and \( \mu \) strongly convex, and we choose \( \gamma < \frac{1}{2L} \) and \( \alpha_t < 1 \), thus:

\[
\| y_{t+1}^{(m)} - y_{t+1}^{(m)} \|^2 \leq (1 - \frac{\mu \gamma \alpha_t}{2})\| y_t^{(m)} - y_{t+1}^{(m)} \|^2 + \frac{\gamma^2 \alpha_t}{4} \| \omega_t^{(m)} \|^2 + \frac{\gamma^2 \alpha_t}{\mu} \| \nabla g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)} \|^2.
\]

(38)

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

\[
\| y_{t+1}^{(m)} - y_{t+1}^{(m)} \|^2 \leq (1 + \frac{\mu \gamma \alpha_t}{4})\| y_{t+1}^{(m)} - y_{t+1}^{(m)} \|^2 + (1 + \frac{\gamma^2 \alpha_t}{\mu})\| y_t^{(m)} - y_t^{(m)} \|^2 
\]

\[
\leq (1 + \frac{\mu \gamma \alpha_t}{4})\| y_{t+1}^{(m)} - y_{t+1}^{(m)} \|^2 + (1 + \frac{\gamma^2 \alpha_t}{\mu})\| x_t^{(m)} - x_{t+1}^{(m)} \|^2
\]

(39)
where the first inequality holds by the generalized triangle inequality, and the second inequality is due to case a) of Proposition 3.9. Combining the above inequalities [38] and [39] we have
\[
\|y_{t+1}^{(m)} - y_{x_{t+1}}^{(m)}\|^2 \leq (1 + \frac{\mu \gamma}{4})(1 - \frac{\mu \gamma}{2})\|y_t^{(m)} - y_{x_t}^{(m)}\|^2 - (1 + \frac{\mu \gamma}{4})\frac{\gamma^2}{4}\|\omega_t^{(m)}\|^2
\]
\[\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\ Quad
Proof. By the update step in Line 7 of Algorithm 1 for $t \neq \bar{t}_s$, we have:

$$
\mathbb{E}\|\mu_t^{(m)} - \bar{\nu}_t\|^2 = \mathbb{E}\|\mu_{t, B_s}^{(m)} + (1 - c_o \alpha_t^2) (\nu_t^{(m)} - \mu_{t, 1, B_s}^{(m)}) - (\bar{\mu}_{t, B_s} + (1 - c_o \alpha_t^2) (\bar{\nu}_{t-1} - \bar{\mu}_{t-1, B_s}))\|^2
$$

$$
= \mathbb{E}\|(1 - c_o \alpha_t^2) (\nu_t^{(m)} - \bar{\nu}_{t-1}) + \mu_{t, B_s}^{(m)} - \bar{\mu}_{t, B_s} - (1 - c_o \alpha_t^2) (\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s})\|^2
$$

$$(a) \leq (1 + \frac{1}{T}) (1 - c_o \alpha_t^2)^2 \mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_{t-1}\|^2
$$

$$
+ (1 + I) \mathbb{E}\|\mu_{t, B_s}^{(m)} - \bar{\mu}_{t, B_s} - (1 - c_o \alpha_t^2) (\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s})\|^2
$$

$$
\leq \left(1 + \frac{1}{T}\right) \mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 + (1 + I) \mathbb{E}\|\mu_{t, B_s}^{(m)} - \bar{\mu}_{t, B_s} - (1 - c_o \alpha_t^2) (\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s})\|^2
$$

(40)

where $(a)$ follows from the generalized triangle inequality.

Next we bound the second term of the above inequality (denoted as $T_1$):

$$
\sum_{m=1}^{M} \mathbb{E}\|\mu_{t, B_s}^{(m)} - \bar{\mu}_{t-1, B_s} - (1 - c_o \alpha_t^2) (\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s})\|^2
$$

$$
\leq 2 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t, B_s}^{(m)} - \bar{\mu}_{t-1, B_s} - (\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s})\|^2 + 2(c_o \alpha_t^2)^2 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s}\|^2
$$

$$
\leq 2 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t, B_s}^{(m)} - \mu_{t-1, B_s}^{(m)}\|^2 + 2(c_o \alpha_t^2)^2 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s}\|^2
$$

where the second inequality follows Proposition 2. We bound the two terms separately, for the first term, we have:

$$
\sum_{m=1}^{M} \mathbb{E}\|\mu_{t, B_s}^{(m)} - \mu_{t-1, B_s}^{(m)}\|^2 \leq \bar{L}^2 \sum_{m=1}^{M} \mathbb{E}\|\hat{x}_{t}^{(m)} - \hat{x}_{t-1}^{(m)}\|^2 + \gamma \omega_{t-1}^{2}\|\gamma \omega_{t-1}^{2}\|^2
$$

(41)

where the inequalities follow Proposition 3(b) and the fact that $\hat{x}_{t}^{(m)} = \hat{x}_{t}^{(m)}$ when $t \neq \bar{t}_s$.

Next for the second term, we have:

$$
\sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s}\|^2 = \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} + (1 - c_o \alpha_t^2) (\nu_{t-1}^{(m)} - \mu_{t-1, B_s}^{(m)}) - (\bar{\mu}_{t-1, B_s} + (1 - c_o \alpha_t^2) (\bar{\nu}_{t-1} - \bar{\mu}_{t-1, B_s}))\|^2
$$

$$(a) \leq 2 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \nu_{t-1}^{(m)} - (\bar{\mu}_{t-1, B_s} - \bar{\nu}_{t-1})\|^2 + 2 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \bar{\mu}_{t-1, B_s}\|^2
$$

$$(b) \leq 2 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \nu_{t-1}^{(m)}\|^2 + 4 \sum_{m=1}^{M} \mathbb{E}\|\nabla h^{(m)}(\bar{x}_{t-1}) - \nabla h(\bar{x}_{t-1})\|^2
$$

$$
+ 4 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \nabla h^{(m)}(\bar{x}_{t-1}) + \nabla h(\bar{x}_{t-1}) - \bar{\mu}_{t-1}\|^2
$$

(42)

Note for the term $T_1$ of Eq. 42 we have $\mathbb{E}\|\mu_{t-1, B_s}^{(m)} - \nu_{t-1}^{(m)}\|^2 \leq G_2^2 + \frac{G_2^2}{b_s}$; For the term $T_2$ of Eq. 42 by the bounded intra-node heterogeneity assumption we have:

$$
T_2 \leq 4 \sum_{m=1}^{M} \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}\|\nabla h^{(m)}(\bar{x}_{t-1}) - \nabla h^{(j)}(\bar{x}_{t-1})\|^2 \leq 4M\varepsilon^2
$$

52
Finally, For the term $T_3$ of Eq. 42

$$
T_3 \leq 8 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1}^{(m)} - \nabla h^{(m)}(\tilde{x}_{t-1})\|^2 + 8 \sum_{m=1}^{M} \mathbb{E}\|\nabla h(\tilde{x}_{t-1}) - \tilde{\mu}_{t-1}\|^2 \leq 16 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1}^{(m)} - \nabla h^{(m)}(\tilde{x}_{t-1})\|^2
$$

$$
\leq 32 \sum_{m=1}^{M} \mathbb{E}\|\mu_{t-1}^{(m)} - \nabla h^{(m)}(x_{t-1}^{(m)})\|^2 + \|\nabla h^{(m)}(x_{t-1}^{(m)}) - \nabla h^{(m)}(\tilde{x}_{t-1})\|^2
$$

$$(a) \leq 32\tilde{L}^2 \sum_{m=1}^{M} \mathbb{E}\|x_{t-1}^{(m)} - \tilde{x}_{t-1}\|^2 + 32\tilde{L}^2 \sum_{m=1}^{M} \mathbb{E}\|y_{t-1}^{(m)} - y_{t-1}^{(m)}\|^2$$

where inequality (b) follows Proposition D.3 (c) and (d).

Combine Eq. 40, Eq. 41 and Eq. 42 use the fact that $I \geq 1$, we have:

$$
\sum_{m=1}^{M} \mathbb{E}\|\nu_{t}^{(m)} - \tilde{\nu}_{t}\|^2
$$

$$\leq \left( 1 + \frac{1}{I} \right) \sum_{m=1}^{M} \mathbb{E}\|\nu_{t-1}^{(m)} - \tilde{\nu}_{t-1}\|^2 + 4I\tilde{L}^2\alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\left[ 2\|\eta\nu_{t-1}^{(m)}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2 \right]
$$

$$+ 8IM(c\nu_{t-1})^2G_1^2 + \frac{8IM(c\nu_{t-1})^2G_2^2}{\beta_X} + 16IM(c\nu_{t-1})^2\zeta^2
$$

$$+ 128IL^2(c\nu_{t-1})^2 \sum_{m=1}^{M} \mathbb{E}\|x_{t-1}^{(m)} - \tilde{x}_{t-1}\|^2 + 128IL^2(c\nu_{t-1})^2 \sum_{m=1}^{M} \mathbb{E}\|y_{t-1}^{(m)} - y_{t-1}^{(m)}\|^2$$

We separate the term $T_1$ with triangle inequality to get:

$$
\sum_{m=1}^{M} \mathbb{E}\|\nu_{t}^{(m)} - \tilde{\nu}_{t}\|^2
$$

$$\leq \left( 1 + \frac{1}{I} \right) \sum_{m=1}^{M} \mathbb{E}\|\nu_{t-1}^{(m)} - \tilde{\nu}_{t-1}\|^2 + 4I\tilde{L}^2\alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\left[ 2\|\eta\nu_{t-1}^{(m)}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2 \right]
$$

$$+ 8IM(c\nu_{t-1})^2G_1^2 + \frac{8IM(c\nu_{t-1})^2G_2^2}{\beta_X} + 16IM(c\nu_{t-1})^2\zeta^2
$$

$$+ 128IL^2(c\nu_{t-1})^2 \sum_{m=1}^{M} \mathbb{E}\|x_{t-1}^{(m)} - \tilde{x}_{t-1}\|^2 + 128IL^2(c\nu_{t-1})^2 \sum_{m=1}^{M} \mathbb{E}\|y_{t-1}^{(m)} - y_{t-1}^{(m)}\|^2$$

Finally, choose $\eta\alpha_t < \frac{1}{16IL^4}$ and combine with Lemma D.8 to bound the term $T_1$, we get the bound in the lemma. This completes the proof. \qed

Next, to simply the notation, we denote $A_t = \mathbb{E}\|\nu_t - E[c\tilde{\nu}_{t}]\|^2$, $B_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|y_{t-1}^{(m)} - y_{t-1}^{(m)}\|^2$, $C_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\omega_{t}^{(m)} - \nabla g^{(m)}(x_{t}^{(m)}, y_{t}^{(m)})\|^2$, $D_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\nu_{t}^{(m)} - \tilde{\nu}_{t}\|^2$, $E_t = \mathbb{E}\|\tilde{\nu}_{t}\|^2$, $F_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\omega_{t}^{(m)}\|^2$.

**Lemma D.10.** For $\alpha_t < \frac{1}{16IL^4}$, we have:

$$(1 - \frac{3\nu_{t-1}^2\alpha_{t-1}^2}{4 + 16^3IL^4}) \sum_{t=t_{t-1}}^{t-1} \alpha_t D_t \leq \frac{3\nu_{t-1}^2}{2 + 16^2IL^2} \sum_{t=t_{t-1}}^{t-1} \alpha_t B_t + \frac{3\beta}{32I} \sum_{t=t_{t-1}}^{t-1} \alpha_t E_t + \frac{3\gamma^2}{64I} \sum_{t=t_{t-1}}^{t-1} \alpha_t F_t$$

$$+ \left( \frac{3\nu_{t-1}^2G_1^2}{32IL^2} + \frac{3\nu_{t-1}^2G_2^2}{32I\beta_XL^2} + \frac{3\nu_{t-1}^2\alpha_{t-1}^2}{16IL^2} \right) \sum_{t=t_{t-1}}^{t-1} \alpha_t^3$$

where the terms $D_t$, $E_t$ and $F_t$ are denoted above.
Proof. Based on Lemma D.9, for \( t \neq \bar{t}_s \), we have:

\[
D_t \leq (1 + \frac{33}{32}I)D_{t-1} + 128\hat{L}^2c_\nu^2\alpha_t^4B_{t-1} + 8\hat{L}^2D_{t-1}^2E_{t-1} + 4I\hat{L}^2\alpha_t^2F_{t-1} \\
+ 8Ic_\nu^2\alpha_t^2G_{t-1}^2 + \frac{8Ic_\nu^2\alpha_t^2G_{t-1}^2}{b_x} + 16Ic_\nu^2\alpha_t^2c_\nu\alpha_t^4 \sum_{\ell=\bar{t}_s-1}^{t-2} \alpha_t^2 D_{\ell}
\]

while for \( t = \bar{t}_s \), we have \( D_{\bar{t}_s} = 1/M \sum_{m=1}^{M} E[\|v(m) - \bar{v}_{\bar{t}_s}\|^2] = 0 \). Apply the above equation recursively from \( \bar{t}_s-1 \) to \( t \) so we have:

\[
D_t \leq \sum_{\ell=\bar{t}_s-1}^{t-1} (1 + \frac{33}{32}I)^{t-\ell} (128\hat{L}^2c_\nu^2\alpha_t^4B_\ell + 8\hat{L}^2D_{\ell}^2E_\ell + 4I\hat{L}^2\gamma^2\alpha_t^2F_\ell + 8Ic_\nu^2G_{\ell}^2 + \frac{8Ic_\nu^2G_{\ell}^2}{b_x}) \sum_{\ell=\bar{t}_s-1}^{t-1} \alpha_t^2 D_{\ell}
\]

\[
\leq \sum_{\ell=\bar{t}_s-1}^{t-1} (384\hat{L}^2c_\nu^2\alpha_t^4B_\ell + 24I\hat{L}^2\gamma^2\alpha_t^2E_\ell + 12I\hat{L}^2\gamma^2\alpha_t^2F_\ell + 24Ic_\nu^2G_{\ell}^2 + \frac{24Ic_\nu^2G_{\ell}^2}{b_x}) \sum_{\ell=\bar{t}_s-1}^{t-1} \alpha_t^2 D_{\ell}
\]

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 \( (1 + a/x)^x \leq e^a \). Then we choose \( a = 33/32 \) and \( x = I \). Finally, we use the fact that \( e^{33/32} \leq 3 \).

Next we multiply \( \alpha_t \) over both sides and take sum from \( \bar{t}_s-1 \) to \( t \), we have:

\[
\sum_{t=\bar{t}_s-1+1}^{\bar{t}_s} \alpha_t D_t \leq \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} \alpha_t \sum_{t=\bar{t}_s-1}^{t-1} (384\hat{L}^2c_\nu^2\alpha_t^4B_\ell + 24I\hat{L}^2\gamma^2\alpha_t^2E_\ell + 12I\hat{L}^2\gamma^2\alpha_t^2F_\ell \\
+ (24Ic_\nu^2G_{\ell}^2 + \frac{24Ic_\nu^2G_{\ell}^2}{b_x} + 48Ic_\nu^2c_\nu^2) \sum_{\ell=\bar{t}_s-1}^{t-1} \alpha_t^2 D_{\ell})
\]

\[
\leq \sum_{t=\bar{t}_s-1}^{t-1} (24I^{1/2}c_\nu G_{\ell}^2 + \frac{3I^{1/2}\hat{L}c_\nu^2}{2} + \frac{3I^{1/2}\hat{L}c_\nu^2}{4} \alpha_t^2 E_\ell + \frac{3I^{1/2}\hat{L}\gamma^2}{4} \alpha_t^2 F_\ell \\
+ \left(3I^{1/2}\nu^2c_\nu^2 c_{1}^2 \frac{G_{\ell}^2}{2L} + \frac{3I^{1/2}\nu^2c_\nu^2}{2b_x L} + \frac{3I^{1/2}\nu^2c_\nu^2}{4L} \alpha_t^2 \sum_{\ell=\bar{t}_s-1}^{t-1} \alpha_t^2 D_{\ell})
\]

\[
\leq \sum_{t=\bar{t}_s-1}^{t-1} \frac{3c_\nu^2}{2} \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t B_\ell + \frac{3c_\nu^2}{2b_x} \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t E_\ell + \frac{3c_\nu^2}{4b_x} \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t F_\ell \\
+ \left(3c_\nu^2G_{\ell}^2 + \frac{3c_\nu^2G_{\ell}^2}{32b_x L^2} + \frac{3c_\nu^2\gamma^2}{16L^2} \right) \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t^3 + \frac{3c_\nu^2\gamma^2c_\nu^2}{4} \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t D_{\ell}
\]

In inequalities (a) and (b), we use \( \alpha_t < \frac{1}{16t\pi^2} \). Note that \( \sum_{t=\bar{t}_s-1+1}^{t} \alpha_t D_t = \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t D_{\ell} \) as \( D_{\bar{t}_s} = D_{\bar{t}_s-1} = 0 \), so we have:

\[
(1 - \frac{3c_\nu^2\gamma^2c_\nu^2}{4 + 163t^5 L^4}) \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t D_t \leq \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t B_\ell + \frac{3c_\nu^2}{2} \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t E_\ell + \frac{3c_\nu^2}{4b_x} \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t F_\ell \\
+ \left(3c_\nu^2G_{\ell}^2 + \frac{3c_\nu^2G_{\ell}^2}{32b_x L^2} + \frac{3c_\nu^2\gamma^2}{16L^2} \right) \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t^3 + \sum_{t=\bar{t}_s-1}^{t-1} \alpha_t D_{\ell}
\]

This completes the proof. \( \square \)
D.1.4 Descent Lemma

**Lemma D.11.** Suppose \( \eta < \frac{1}{2L}, \alpha_t < 1 \), for all \( t \in [\bar{t}_{s-1}, \bar{t}_s - 1] \) and \( s \in [S] \), the iterates generated satisfy:

\[
E[h(\bar{x}_{t+1})] \leq E[h(\bar{x}_t)] - \frac{\eta \alpha_t}{4} E[\|\bar{\nu}_t\|^2] - \frac{\eta \alpha_t}{2} E[\|\nabla h(\bar{x}_t)\|^2] + 2\eta \alpha_t E[\|\mu_t - \bar{\nu}_t\|^2] + 4\eta \alpha_t G_1^2 \\
+ \frac{\bar{L}^2 \eta^3 \alpha_t}{M} \sum_{\ell = \ell_{t-1}}^{t-1} \sum_{m=1}^{M} E[\|\nu_{t}\| - \bar{\nu}_{t}\|^2] + \frac{4\bar{L}^2 \eta \alpha_t}{M} \sum_{m=1}^{M} E[\|y_{t} - y_{\bar{x}_t}\|^2]
\]

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

**Proof.** By the smoothness of \( h(x) \) we have:

\[
E[h(\bar{x}_{t+1})] \leq E[h(\bar{x}_t)] + \langle \nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{\bar{L}}{2} \|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
\overset{(a)}{=} E[h(\bar{x}_t)] - \eta \alpha_t \langle \nabla h(\bar{x}_t), \bar{\nu}_t \rangle + \frac{\eta^2 \alpha_t^2 \bar{L}}{2} \|\bar{\nu}_t\|^2 \\
\overset{(b)}{=} E[h(\bar{x}_t)] - \frac{\eta \alpha_t}{2} \|\bar{\nu}_t\|^2 - \frac{\eta \alpha_t}{2} \|\nabla h(\bar{x}_t)\|^2 + \frac{\eta \alpha_t}{2} \|\nabla h(\bar{x}_t) - \bar{\nu}_t\|^2 + \frac{\eta^2 \alpha_t^2 \bar{L}}{2} \|\bar{\nu}_t\|^2 \\
= E[h(\bar{x}_t)] - \frac{\eta \alpha_t}{4} \|\bar{\nu}_t\|^2 - \frac{\eta \alpha_t}{2} \|\nabla h(\bar{x}_t)\|^2 + \frac{\eta \alpha_t}{2} \|\nabla h(\bar{x}_t) - \bar{\nu}_t\|^2 \\
\leq E[h(\bar{x}_t)] - \frac{\eta \alpha_t}{4} \|\bar{\nu}_t\|^2 - \frac{\eta \alpha_t}{2} \|\nabla h(\bar{x}_t)\|^2 + \frac{\eta \alpha_t}{2} \|\nabla h(\bar{x}_t) - \bar{\nu}_t\|^2 \\
\leq \frac{2}{M} \sum_{m=1}^{M} E[\|\nabla h(x^{(m)}_t)\|^2] + 4\eta \alpha_t E[\|\mu_t - \bar{\nu}_t\|^2] \\
\quad + 4E[\|\nabla h(x^{(m)}_t)\|^2]
\]

where equality (a) follows from the iterate update given in Algorithm[1]. (b) uses \( \langle a, b \rangle = \frac{1}{2} \|a\|^2 + \|b\|^2 - \|a - b\|^2 \) and \( \eta \alpha_t < \frac{1}{2M} \).

**For the first term,** we have:

\[
2E[\|\nabla h(\bar{x}_t) - \frac{1}{M} \sum_{m=1}^{M} \nabla h(x^{(m)}_t)\|^2] \leq \frac{2}{M} \sum_{m=1}^{M} E[\|\nabla h(x^{(m)}_t)\|^2] - \frac{2\bar{L}^2}{M} \sum_{m=1}^{M} E[\|\bar{x}_t - x^{(m)}_t\|^2] \\
\leq \frac{2\bar{L}^2 \eta^2}{M} \frac{1}{\sum_{\ell = \ell_{t-1}}^{t-1} \sum_{m=1}^{M} E[\|\nu_{t}\| - \bar{\nu}_{t}\|^2]} \quad \text{where the last inequality uses Lemma D.8}
\]

**For the second term,** we have:

\[
4E[\|\frac{1}{M} \sum_{m=1}^{M} \nabla h(x^{(m)}_t) - E[\mu_t]\|^2] \leq \frac{8}{M} \sum_{m=1}^{M} E[\|\nabla h(x^{(m)}_t) - \mu_t\|^2] + \frac{8}{M} E[\|\mu_t - \mu^{(m)}_t\|^2] \\
\leq \frac{8\bar{L}^2}{M} \sum_{m=1}^{M} E[\|y^{(m)}_x - y^{(m)}_t\|^2] + 8G_1^2
\]

Plug the bound for term \( T_1 \) back gets the claim in the lemma.

**D.1.5 Proof of Convergence Theorem**

We first denote the following potential function \( G(t) \):

\[
G_t = h(\bar{x}_t) + \frac{9bM \eta}{16 \alpha_t} \|\bar{\nu}_t\| - \frac{1}{M} \sum_{m=1}^{M} \nabla h(x^{(m)}_t)\|^2 + \frac{18 \eta \bar{L}^2}{\mu^2} \times \frac{1}{M} \sum_{m=1}^{M} E[\|y^{(m)}_t - y^{(m)}_{x_t}\|^2] \\
+ \frac{9bM \bar{L} \eta^2}{16 \bar{L}^2 \alpha_t} \times \frac{1}{M} \sum_{m=1}^{M} \|\nu^{(m)}_t - \nabla y^{(m)}_t (x^{(m)}_t, y^{(m)}_t)\|^2
\]

55
Theorem D.12. Suppose \( \gamma \leq \frac{1}{\sqrt{T}} \), \( \eta \leq \min \left( \frac{\gamma}{4u^2L^2}, \frac{\sqrt{c_1^2 + c_2^2}/L}{\sqrt{c_1^2 + c_2^2}/L}, \frac{\bar{L}}{\sqrt{c_1^2 + c_2^2}/L} \right) \), \( c_\nu = \frac{32}{9bM} + \frac{L}{24bM^2} \), \( c_\omega = \frac{144L^2}{bM\sigma^2} + \frac{L}{24bM^2} \), \( u = (bM\sigma)^2 \bar{u} \), where \( \bar{u} = \max \left( 2, 16^3 I/9^2 \hat{L}, c_\nu^2, c_\omega^2 \right) \), \( \delta = \frac{1}{T^{1-\epsilon}} \), then we have:

\[
\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E} \left[ \| \nabla \mathbb{h}(\bar{x}_t) \|^2 \right] = O \left( \frac{\kappa^{19/3} I^{3/2}}{T} + \frac{\kappa^{16/3}}{(bM\tau)^2/3} + \kappa^3 b^2 M^2 I^{9/2} G_1^2 \right)
\]

To reach an \( \epsilon \)-stationary point, we need \( T = O(\kappa^8 (bM)^{-1} \epsilon^{-1.5}) \), \( l = O(\kappa^{10/9} (bM)^{-2/3} \epsilon^{-1/3}) \) and \( Q = O(\kappa \log \left( \frac{\kappa}{bM} \right)) \).

Proof. By the condition that \( u \geq \bar{u} ^{3/2} \delta^3 \), it is straightforward to verify that \( c_\nu \alpha_t^2 < 1 \). By Lemma D.5 (in new notation), when \( t \neq \bar{t}_s \), we have:

\[
\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \leq \left( \alpha_{t-1} - \alpha_{t-2} - c_\nu \alpha_t \right) \frac{A_{t-1} + \frac{2c_\nu^2 \alpha_t^3 G_2}{bM} + \frac{4L^2 \eta^2 \alpha_t}{bM} (D_{t-1} + E_{t-1}) + \frac{2L^2 \gamma^2 \alpha_{t-1} F_{t-1}}{bM}}{\alpha_{t-1} - \alpha_{t-2}}
\]

Note we choose \( b_x = b \). For \( \alpha_{t-1} - \alpha_{t-2} \), we have:

\[
\alpha_{t-1} - \alpha_{t-2} = \frac{u + \sqrt{u^2 + 4x^2}}{\delta} - \frac{u + \sqrt{u^2 + 4(x-1)^2}}{\delta} \leq \frac{\delta^2}{3\delta(u + \sqrt{u^2 + 4x^2})}
\]

\[
\leq \frac{2^{3/2} \delta^3 u^2}{(u + \sqrt{u^2 + 4x^2})} \leq \frac{2L^2}{3M^2} \alpha_t^2 \leq \frac{L}{24bM^2} \alpha_t
\]

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 \( \alpha_t \). By choosing \( c_\nu = \frac{32}{9bM} + \frac{L}{24bM^2} \),

\[
\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \leq \frac{32}{9bM} \alpha_{t-1} A_{t-1} + \frac{2c_\nu^2 \alpha_t^3 G_2}{bM} + \frac{4L^2 \eta^2 \alpha_t}{bM} (D_{t-1} + E_{t-1}) + \frac{2L^2 \gamma^2 \alpha_{t-1} F_{t-1}}{bM}
\]

When \( t = \bar{t}_s \), by Lemma D.5 and Lemma D.6, we have:

\[
\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \leq \frac{32}{9bM} \alpha_{t-1} A_{t-1} + \frac{2c_\nu^2 \alpha_t^3 G_2}{bM} + \frac{4L^2 \eta^2 \alpha_t}{bM} (D_{t-1} + E_{t-1}) + \frac{2L^2 \gamma^2 \alpha_{t-1} F_{t-1}}{bM}
\]

Note we use the fact \( \alpha_t/\alpha_{t-1} < 2 \) in the last term, which is due to:

\[
\frac{\alpha_t}{\alpha_{t-1}} = \frac{(u_{t-1} + \sqrt{u_{t-1}^2 + 4x^2})^{1/3}}{(u_{t} + \sqrt{u_{t}^2 + 4x^2})^{1/3}} = \left( 1 + \frac{u_{t-1} - u_t + \sigma^2(I_t - 1 - t)}{u_t + \sigma^2t} \right)^{1/3}
\]

\[
\leq \left( 1 + \frac{(I - 1) \sigma^2}{u_t + \sigma^2t} \right)^{1/3} \leq 1 + \frac{(I - 1)}{3(I + 1)} \leq 2
\]

where we use the condition \( u_t \geq (I + 1) \sigma^2 \). Next, we telescope from \( \bar{t}_{s-1} + 1 \) to \( \bar{t}_s \):

\[
\left( \frac{A_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}} - \frac{A_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}} \right) \leq - \frac{32}{9bM} \sum_{t=1}^{\bar{t}_s} \alpha_t A_t + \frac{2c_\nu^2 G_2}{bM} \sum_{t=1}^{\bar{t}_s} \alpha_t^{-1} + \frac{16L^2 \eta^2}{bM} \sum_{t=1}^{\bar{t}_s} \alpha_t^{-1} D_t
\]

\[
+ \frac{8L^2 \eta^2}{bM} \sum_{t=1}^{\bar{t}_s} \alpha_t^{-1} E_t + \frac{2L^2 \gamma^2}{bM} \sum_{t=1}^{\bar{t}_s} \alpha_t^{-1} F_t \quad (43)
\]

Next, we follow similar derivation as \( A_t/\alpha_{t-1} - A_{t-1}/\alpha_{t-2} \). By Lemma D.6, for \( t \neq \bar{t}_s \), we choose \( c_\omega = \frac{144L^2}{bM\sigma^2} + \frac{L}{24bM^2} \), to obtain:

\[
\frac{C_t}{\alpha_{t-1}} - \frac{C_{t-1}}{\alpha_{t-2}} \leq - \frac{144L^2 \alpha_t}{bM\mu^2} C_t + \frac{2c_\omega G_2}{bM} \sum_{t=1}^{\bar{t}_s} \alpha_t^{-1} \sigma^2 + \frac{4L^2 \eta^2 \alpha_t}{bM} (D_{t-1} + E_{t-1}) + \frac{2L^2 \gamma^2}{bM} \alpha_{t-1} F_{t-1}
\]

56
Note we choose $b_y = bM$. When $t = \tilde{t}_s$, by Lemma D.5 and Lemma D.8 we have:

$$
\frac{C_t}{\alpha_{t-1}} - \frac{C_{t-1}}{\alpha_{t-2}} \leq -\frac{144L^2 \alpha_t}{bM} C_t + \frac{2c_2 \alpha_t}{bM} \alpha_t \sigma^2 + \frac{8L^2 \eta^2 \alpha_t}{bM} (D_t + E_t)
$$

$$
+ \frac{2L^2 \gamma \alpha_t}{bM} F_t + \frac{4L^2}{bM} \sum_{t=\tilde{t}_s}^{t-1} \alpha_t D_t + \frac{4L^2}{bM} \sum_{t=\tilde{t}_s}^{t-1} \alpha_t F_t
$$

Divide $\tilde{c}_t$ for both sides and then telescope from $\tilde{t}_s + 1$ to $\tilde{t}_s$, we have:

$$
\left( \frac{C_{\tilde{t}_s}}{\alpha_{\tilde{t}_s-1}} - \frac{C_{\tilde{t}_s-1}}{\alpha_{\tilde{t}_s-2}} \right) \leq -\frac{144L^2}{2bM} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t C_t + \frac{2c_2 \alpha_t}{bM} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t \sigma^2 + \frac{16L^2 \eta^2}{bM} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t D_t
$$

$$
+ \frac{8L^2 \eta^2}{bM} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t E_t + \frac{2L^2 \gamma \alpha_t}{bM} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t F_t.
$$

(44)

Next from Lemma D.7 for $t \neq \tilde{t}_s$, we have:

$$
B_t - B_{t-1} \leq \frac{\mu \gamma}{4} \alpha_t B_t - \frac{\gamma^2}{4} \alpha_t F_t + \frac{9\gamma}{2\mu} \alpha_t C_t + \frac{9\kappa^2 \eta^2}{\mu} \alpha_t D_t
$$

When $t = \tilde{t}_s$, we have:

$$
B_{\tilde{t}_s} - B_{\tilde{t}_s-1} \leq \frac{\mu \gamma}{4} \alpha_{\tilde{t}_s} B_{\tilde{t}_s} - \frac{\gamma^2}{4} \alpha_{\tilde{t}_s} F_{\tilde{t}_s} + \frac{9\gamma}{2\mu} \alpha_{\tilde{t}_s} C_{\tilde{t}_s} + \frac{9\kappa^2 \eta^2}{\mu} \alpha_{\tilde{t}_s} D_{\tilde{t}_s}
$$

For the coefficient of the last term, we use $\alpha_t/\alpha_{\tilde{t}_s} < 2$. We telescope from $\tilde{t}_s + 1$ to $\tilde{t}_s$ and have:

$$
B_{\tilde{t}_s} - B_{\tilde{t}_s-1} \leq -\frac{\mu \gamma}{4} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t B_t - \frac{\gamma^2}{4} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t F_t + \frac{9\gamma}{2\mu} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t C_t
$$

$$
+ \frac{36L^2 \kappa^2 \eta^2}{\mu^2} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t D_t + \frac{18\kappa^2 \eta^2}{\mu^2} \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t F_t
$$

(45)

Next, by Lemma D.11 we have:

$$
E[h(\tilde{x}_{t+1})] \leq E[h(\tilde{x}_t)] - \frac{\eta \alpha_t}{4} E_t - \frac{\eta \alpha_t}{2} E[\|\nabla h(\tilde{x}_t)\|^2] + \tilde{L}^2 \eta^3 \alpha_t \sum_{t=\tilde{t}_s}^{\tilde{t}_s-1} \alpha_t^2 D_t + 2\eta \alpha_t A_t + 4\tilde{L}^2 \eta \alpha_t B_t + 4\eta \alpha_t G_1^2
$$

We telescope from $\tilde{t}_s - 1$ to $\tilde{t}_s$ to have:

$$
E[h(\tilde{x}_{t_s}) - h(\tilde{x}_{t_s - 1})] \leq -\sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \frac{\eta \alpha_t}{4} E_t - \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \frac{\eta \alpha_t}{2} E[\|\nabla h(\tilde{x}_t)\|^2] + 4\eta \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \alpha_t G_1^2
$$

$$
+ \tilde{L}^2 \eta^3 \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \alpha_t \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \alpha_t^2 D_t + \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} 2\eta \alpha_t A_t + \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} 4\tilde{L}^2 \eta \alpha_t B_t
$$

$$
\leq -\sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \frac{\eta \alpha_t}{4} E_t - \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \frac{\eta \alpha_t}{2} E[\|\nabla h(\tilde{x}_t)\|^2] + 4\eta \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \alpha_t G_1^2
$$

$$
+ \frac{\kappa^2 \eta^2}{64} \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} \alpha_t D_t + \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} 2\eta \alpha_t A_t + \sum_{t=\tilde{t}_s - 1}^{\tilde{t}_s - 1} 4\tilde{L}^2 \eta \alpha_t B_t
$$

(46)
In the last inequality, we use the fact that $\bar{t}_s - \bar{t}_{s-1} \leq I$, $\alpha_t < \frac{1}{16L}\bar{L}$ and $\bar{L}/\bar{L} = \kappa + 1 \leq 2\kappa$

Combine Eq. (43), Eq. (44), Eq. (45) and Eq. (46) and we have:

$$
\mathbb{E}[G_{t_s}] - \mathbb{E}[G_{t_{s-1}}] \leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \frac{\eta_0}{2} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + \left( \frac{9\bar{L}^2\eta\bar{c}^2\sigma^2}{8L^2} + \frac{9\eta_0^2 G_2^2}{8} \right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t^3 - \frac{\bar{L}^2\eta}{2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t B_t \\
- \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \left( \frac{\eta_0^2 \bar{L}^2}{2} - \frac{9\eta_0^2 \gamma \bar{L}^2}{8} - \frac{9\eta_0^2 \gamma \bar{L}^2}{8\mu} \right) \alpha_t F_t \\
- \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \left( \eta_0 \gamma \bar{L}^2 \eta \eta - \frac{9\bar{L}^2 \eta \eta}{2} - \frac{9\bar{L}^2 \eta \eta}{2} \right) \eta \alpha_t E_t \\
+ \left( \frac{\kappa^2}{64} + \frac{648\kappa^2 \bar{L}^2}{\mu^2 \gamma^2} + 9I \bar{L}^2 + 9I \bar{L}^2 \right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t D_t + 4\eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t G_1^2
$$

By the condition that $\eta < \frac{\mu^2 \gamma}{144\kappa L}$, So we have:

$$
\mathbb{E}[G_{t_s}] - \mathbb{E}[G_{t_{s-1}}] \leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \frac{\eta_0}{2} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + \left( \frac{9\eta_0^2 \bar{c}^2 \sigma^2}{8\mu^2} + \frac{9\eta_0^2 G_2^2}{8\mu^2} \right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t^3 + 4\eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t G_1^2 \\
- \frac{9\eta_0^2 \bar{L}^2}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t F_t - \frac{3\eta}{16} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t E_t - \frac{\bar{L}^2 \eta}{2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t E_t + C_1 I \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t D_t
$$

(47)

where we denote $C_1 = \left( \frac{\kappa^2}{64} + \frac{648\kappa^2 \bar{L}^2}{\mu^2 \gamma^2} + 9I \bar{L}^2 + 9I \bar{L}^2 \right)$. By Lemma D.10 and choose $\eta < \frac{\bar{L}^2}{\kappa c_{\bar{c}}}$, we have:

$$
\sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t D_t \leq \frac{c_2^2}{128I^2 \bar{L}^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t B_t + \frac{\eta_0^2}{8I} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t E_t + \frac{\eta_0^2}{16I} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t F_t \\
+ \left( \frac{c_2^2 G_1^2}{8I^2} + \frac{c_2^2 G_2^2}{8I b \bar{L}^2} + \frac{c_2^2 c_{\bar{c}}^2}{4I^2} \right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t^3
$$

(48)

Combine Eq. (47) and Eq. (48), and use the condition that $\eta < \min \left( \frac{\bar{L}^2}{c_1 c_{\bar{c}}}, \frac{\bar{L}^2}{c_1}, \frac{\bar{L}^2}{c_1^2}, 1 \right)$, the fact that $I \geq 1$, we have:

$$
\mathbb{E}[G_{t_s}] - \mathbb{E}[G_{t_{s-1}}] \leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \frac{\eta_0}{2} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + 4\eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t G_1^2 \\
+ \eta \left( \frac{9\bar{L}^2 \bar{c}^2 \sigma^2}{8L^2} + \frac{9\eta_0^2 G_2^2}{8} + \frac{c_2^2 G_1^2}{8} + \frac{c_2^2 G_2^2}{8} + \frac{c_2^2 c_{\bar{c}}^2}{4} \right) \sum_{t=\bar{t}_{s-1}}^{\bar{t}_{s-1}} \alpha_t^3
$$

Sum over all $s \in [S]$ (assume $T = SI + 1$ without loss of generality), we have:

$$
\mathbb{E}[G_T] - \mathbb{E}[G_1] \leq - \sum_{t=1}^{T-1} \frac{\eta_0}{2} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + \left( \eta C_{\sigma, \bar{c}} + \frac{4\eta G_1^2}{\alpha_T T} \right) \sum_{t=1}^{T-1} \alpha_t^3
$$

For ease of notation, we denote $C_{\sigma, \bar{c}} = \left( \frac{9\bar{L}^2 \bar{c}^2 \sigma^2}{8L^2} + \frac{9\eta_0^2 G_2^2}{8} + \frac{c_2^2 G_1^2}{8} + \frac{c_2^2 G_2^2}{8} + \frac{c_2^2 c_{\bar{c}}^2}{4} \right)$. Rearranging the terms and use the fact that $\alpha_t$ is non-increasing, we have:

$$
\frac{\eta_0 T}{2} \sum_{t=1}^{T-1} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] \leq \mathbb{E}[G_1] - \mathbb{E}[G_T] + \left( \eta C_{\sigma, \bar{c}} + \frac{4\eta G_1^2}{\alpha_T T} \right) \sum_{t=1}^{T-1} \alpha_t^3 \\
\leq h(x_1) - h^* + \frac{9bM\eta A_1}{16\alpha_1} + \frac{18\eta \bar{L}^2 B_1}{\mu \gamma} + \frac{9bM\eta C_1}{16\alpha_1} + \left( \eta C_{\sigma, \bar{c}} + \frac{4\eta G_1^2}{\alpha_T T} \right) \sum_{t=1}^{T-1} \alpha_t^3
$$

58
where we use \( G_T \geq h^* \) (\( h^* \) is the optimal value of \( h \)), and for the last term, we use the following fact:

\[
\sum_{t=1}^{T} \sigma^3_t = \sum_{t=1}^{T} \frac{\delta^3}{u + \sigma^2 t} \leq \sum_{t=1}^{T} \frac{\delta^3}{\sigma^2} = \frac{\delta^3}{\sigma^2} \sum_{t=1}^{T} \frac{1}{1 + t} \leq \frac{\delta^3}{\sigma^2} \ln(T + 1) = \frac{b^2 M^2}{L^2} \ln(T + 1)
\]

the first inequality follows \( u_t > \sigma^2 \), the last inequality follows Proposition [E.3]. Next, we denote the initial sub-optimality as \( \Delta = h(\bar{x}_1) - h^* \), and inner variable estimation error \( \hat{b}_1 = \frac{1}{M} \sum_{m=1}^{M} \|y^{(m)}_1 - y^{(m)}_{\hat{x}_1}\|^2 \leq \Delta_y \), and we assume \( \Delta_y = O(\kappa^{-1}) \), furthermore, we have:

\[
A_1 = \mathbb{E} \left[ \frac{1}{M} \sum_{m=1}^{M} \left( \Phi^{(m)}(x_1^{(m)}, y_1^{(m)}; \mathcal{B}_x) - \Phi^{(m)}(x_1^{(m)}, y_1^{(m)}) \right)^2 \right] \leq \frac{\sigma^2}{b_1 M}
\]

and

\[
C_1 = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \| \omega^{(m)}_1 - \nabla_y g^{(m)}(x_1^{(m)}, y_1^{(m)}) \|^2 \leq \frac{\sigma^2}{b_1}
\]

where we choose the size of the first minibatch to be \( b_x = b_1 \) and \( b_y = b_1 M \). Then, we divide both sides by \( \eta \alpha T^2 / 2 \) to have:

\[
\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E} \left[ \| \nabla h(\bar{x}_t) \|^2 \right] \leq \frac{2 \Delta}{\eta \alpha T} + \frac{9 b_1^2}{8 b_1 T \alpha \alpha T} + \frac{36 L^2 \Delta_y}{\kappa \mu \gamma \alpha T} + \frac{9 \beta_1}{8 b_1 T \alpha \alpha T} + \frac{2 b^2 M^2 C_{\sigma, \zeta} \ln(T)}{L^2 T \alpha T} + \frac{8 b^2 M^2 G^2_1 \ln(T)}{L^2 T \alpha T^2}
\]

Note that we have:

\[
\frac{1}{\alpha_t} = \frac{u(t) + \sigma^2 t}{\delta t} \leq \frac{u_1^{1/3} + \sigma^{2/3}}{\delta t^{2/3}}
\]

where the inequality uses the fact that \( (x + y)^{1/3} \leq x^{1/3} + y^{1/3} \). In particular, when \( t = 1 \), we have

\[
\frac{1}{\alpha_1} \leq \frac{u_1^{1/3} + \sigma^{2/3}}{\delta} = \frac{\hat{L}^2 / (\bar{b}_1^3 / (bM)^{2/3} \bar{\kappa} + 1)}{(bM)^{2/3}}
\]

when \( t = T \), we have:

\[
\frac{1}{\alpha_T} \leq \frac{u_1^{1/3}}{\delta T} + \frac{\sigma^{2/3}}{\delta T^{2/3}} = \frac{\hat{L}^2 / (b MT)^{2/3}}{T} + \frac{\hat{L}^2 / (b MT)^{2/3}}{T}
\]

In summary, we have:

\[
\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E} \left[ \| \nabla h(\bar{x}_t) \|^2 \right] \leq \frac{2 \Delta}{\eta} + \frac{9 \beta_1}{8 b_1 T \alpha \alpha T} + \frac{36 L^2 \Delta_y}{\kappa \mu \gamma \alpha T} + \frac{9 \beta_1}{8 b_1 T \alpha \alpha T} + \frac{2 \ln(T)}{(bM)^{2/3}} \left( \frac{b^2 M^2 C_{\sigma, \zeta}}{L^2} + \frac{4 b^2 M^2 G^2_1}{L^2 \alpha T^2} \right)
\]

Note that \( \hat{L} = O(\kappa^2), \bar{L} = O(\kappa^3), \bar{\kappa}_\nu = O((bM)^{-1} \kappa^2) \) and \( \bar{\kappa}_\omega = O((bM)^{-1} \kappa^2) \), \( \bar{u} = O((bM)^{-1/2} \kappa^2) \), \( \bar{\kappa}_\nu = O(I^{3/2}) \), \( \bar{\kappa}_\omega = O((bM)^{-1/2} \kappa^{7/3}) \), then for \( \eta \), we have:

\[
\eta \leq \min \left( \frac{\hat{L}^2}{144 \kappa \bar{L}}, \frac{\hat{L}^2}{C_1^{1/2} \bar{\kappa}_\nu}, \frac{1}{C_1^{1/2}}, \frac{\hat{L}}{12 L^{1/2}} \right)
\]

Recall that \( C_1 = \left( \frac{\kappa^2}{2^2 + 54 \kappa^2 \bar{L}^2} + b \bar{L}^2 + 9 \bar{L}^2 \right) \), suppose we choose \( \gamma = \frac{1}{2L} \), then \( C_1 = O(\kappa^8) \) and \( \eta^{-1} = O(\kappa^4), \mu \gamma = O(\kappa^{-1}) \). recall that \( C_{\sigma, \zeta} = \left( \frac{9 \bar{L}^2 \bar{\kappa}_\nu^2}{8 L^2} + \frac{9 \bar{L}^2 \bar{\kappa}_\omega^2}{8 L^2} + \frac{2 \bar{L}^2 \bar{\kappa}_\nu^2}{8 L^2} + \frac{2 \bar{L}^2 \bar{\kappa}_\omega^2}{8 L^2} + \frac{2 \bar{L}^2 \bar{\kappa}_\nu^2}{8 L^2} \right) \), so we have \( C_{\sigma, \zeta} = O((bM)^{-2} \kappa^8) \), suppose we choose \( b_1 = O(I^{3/2}) \). Finally, for the coefficient of the hyper-gradient bias term \( G_1^2 \), we have:

\[
\frac{8 b^2 M^2 \ln(T)}{L^{1/3} \alpha T^2} \left( \frac{\bar{u}^{1/3}}{T} + \frac{1}{(bM)^{2/3}} \right) \leq 16 b^2 M^2 \bar{u} + 16 \left( \frac{b^2 M^2 \bar{u}}{T} \right)^{2/3} + 16 \left( \frac{b^2 M^2 \bar{u}}{T} \right)^{1/3} + 16 = O(\kappa^3 b^2 M^2 T^{9/2})
\]
Then, we have:

$$1 \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] = O\left(\frac{\kappa^{19/3}T^{3/2}}{T} + \frac{\kappa^{16/3}}{(bMT)^{2/3}} + \kappa^3 b^2 M^2 I^{1/2} G^2_1\right)$$

To reach an $\epsilon$-stationary point, we need $T = O(\kappa^8 (bM)^{-1} \epsilon^{-1.5})$, $I = O(\kappa^{10/9} (bM)^{-2/3} \epsilon^{-1/3})$ and $Q = O(\kappa \log(\frac{\kappa}{\epsilon T}))$. The communication cost is $E = T / I \geq \kappa^{62/9} (bM)^{-1/3} \epsilon^{-7/6}$, the sample complexity is $G_c(f, \epsilon) = O(M^{-1} \kappa^8 \epsilon^{-1.5})$, $G_c(g, \epsilon) = O(\kappa^8 \epsilon^{-1.5})$, $J_v(g, \epsilon) = O(\kappa^8 \epsilon^{-1.5})$, $H_v(g, \epsilon) = O(\kappa^8 \epsilon^{-1.5})$.

Suppose we choose $b = O(\epsilon^{-0.5})$, we have $T = O(\kappa^8 (M^{-1} \epsilon^{-1.5}))$, $I = \kappa^{10/9} M^{-2/3}$, $Q = O(\kappa \log(\frac{\kappa}{\epsilon T}))$ and $E = \kappa^{62/9} M^{-1/3} \epsilon^{-1}$. If we instead choose $b = O(1)$, we have $T = O(\kappa^8 M^{-1} \epsilon^{-1.5})$, $I = O(\kappa^{10/9} M^{-2/3} \epsilon^{-1.5})$ and $Q = O(\kappa \log(\frac{\kappa}{\epsilon T}))$. The communication cost is $E = O(\kappa^{62/9} M^{-1/3} \epsilon^{-7/6})$.

\[\Box\]

### D.2 Proof for the FedBiO-Local Algorithm

In this section, we investigate the convergence rate for the FedBiO-Local algorithm (Algorithm 3).

#### D.2.1 Lower Problem Solution Error

**Lemma D.13.** When $\gamma < \frac{\mu}{2}$, when $t \neq \bar{t}_s$, we have:

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] \leq \left(1 - \frac{\mu \gamma}{2}\right) \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] + \frac{5 \kappa^2 \eta^2}{\mu \gamma M} \sum_{m=1}^{M} \mathbb{E}[\|v_{t-1}^{(m)}\|^2] + \frac{3 \gamma^2 \sigma^2}{b_y}$$

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

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] \leq \left(1 - \frac{\mu \gamma}{2}\right) \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] + \frac{10 \kappa^2 \eta^2}{\mu \gamma M} \sum_{m=1}^{M} \mathbb{E}[\|v_{t-1}^{(m)}\|^2] + \frac{10 \kappa^2 \eta^2}{\mu \gamma M} \sum_{m=1}^{M} \mathbb{E}[\|v_{t-1}^{(m)}\|^2] + \frac{3 \gamma^2 \sigma^2}{b_y}$$

*Proof.* First, we have:

$$\mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] \leq \left(1 + \frac{\mu \gamma}{2}\right) \mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] + \left(1 + \frac{2}{\mu \gamma}\right) \mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2]$$

$$\leq \left(1 + \frac{\mu \gamma}{2}\right) (1 - \mu \gamma) \mathbb{E}[\|y_{t-1}^{(m)} - y_{x_t}^{(m)}\|^2] + \left(1 + \frac{2}{\mu \gamma}\right) \mathbb{E}[\|y_{t-1}^{(m)} - y_{x_t}^{(m)}\|^2] + \left(1 + \frac{\mu \gamma}{2}\right) \frac{2 \gamma^2 \sigma^2}{b_y}$$

where the second inequality is due to Proposition [E.4](#), where we choose $\gamma < 1/L$; in the last inequality, we use $\gamma < 1/(\kappa I)$ and $\mu \leq L$. For the last term, when $t \neq \bar{t}_s$, we have:

$$\mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] \leq \left(1 - \frac{\mu \gamma}{2}\right) \mathbb{E}[\|y_{t-1}^{(m)} - y_{x_t}^{(m)}\|^2] + \frac{5 \kappa^2 \eta^2}{\mu \gamma M} \mathbb{E}[\|v_{t-1}^{(m)}\|^2] + \frac{3 \gamma^2 \sigma^2}{b_y}$$

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

$$\mathbb{E}[\|y_t^{(m)} - y_{x_t}^{(m)}\|^2] \leq \left(1 - \frac{\mu \gamma}{2}\right) \mathbb{E}[\|y_{t-1}^{(m)} - y_{x_t}^{(m)}\|^2] + \frac{10 \kappa^2 \eta^2}{\mu \gamma M} \mathbb{E}[\|v_{t-1}^{(m)}\|^2] + \frac{10 \kappa^2 \eta^2}{\mu \gamma M} \mathbb{E}[\|v_{t-1}^{(m)}\|^2] + \frac{3 \gamma^2 \sigma^2}{b_y}$$

Average over all clients, we get the claim in the lemma.  \[\Box\]

60
D.2.2 Upper Variable Drift

Lemma D.14. For any \( t \neq \bar{t}_s, s \in [S] \), we have:

\[
\| \bar{x}^{(m)}_t - \bar{x}_t \|^2 \leq I\eta^2 \sum_{t = t_{s-1}}^{t-1} \| \nu^{(m)}_t - \bar{\nu}_t \|^2
\]

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

\[
\| x^{(m)}_t - \bar{x}_t \|^2 = \| x^{(m)}_{t_{s-1}} - \bar{x}_{t_{s-1}} - ( \sum_{t = t_{s-1}}^{t-1} \eta_t^{(m)} - \sum_{t = t_{s-1}}^{t-1} \eta \bar{\nu}_t ) \|^2 = ( \sum_{t = t_{s-1}}^{t-1} \eta \bar{\nu}_t )^2
\]

This completes the proof. \( \square \)

Lemma D.15. For \( t \neq \bar{t}_s, s \in [S], \) we have:

\[
\frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \| \nu_t^{(m)} - \bar{\nu}_t \|^2 \right] \leq \frac{4\hat{L}^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \| y_t^{(m)} - y_{x_t^{(m)}}^{(m)} \|^2 \right] + \frac{12\hat{L}^2I\eta^2}{M} \sum_{m=1}^{M} \sum_{t = t_{s-1}}^{t-1} \mathbb{E} \left[ \| \nu_t^{(m)} - \bar{\nu}_t \|^2 \right] + 8\zeta^2 + 4G_1^2 + \frac{4G_2^2}{b_x}
\]

for \( t = \bar{t}_s, s \in [S], \) we have:

\[
\frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \| \nu_t^{(m)} - \bar{\nu}_t \|^2 \right] \leq \frac{4\hat{L}^2}{M} \sum_{m=1}^{M} \mathbb{E} \left[ \| y_t^{(m)} - y_{x_t^{(m)}}^{(m)} \|^2 \right] + 8\zeta^2 + 4G_1^2 + \frac{4G_2^2}{b_x}
\]

Proof. For \( t \in [T], \) we have:

\[
\mathbb{E} \left[ \| \nu_t^{(m)} - \bar{\nu}_t \|^2 \right] \leq 2\mathbb{E} \left[ \| \nu_t^{(m)} - \nabla h^{(m)}(x_t^{(m)}) \|^2 \right] + 2\mathbb{E} \left[ \| \nabla h^{(m)}(x_t^{(m)}) - \frac{1}{M} \sum_{j=1}^{M} \nabla h(j)(x_t^{(j)}) \|^2 \right]
\]

\[
\leq 2\mathbb{E} \left[ \| \nu_t^{(m)} - \nabla h^{(m)}(x_t^{(m)}) \|^2 \right] + 2\mathbb{E} \left[ \| \nabla h^{(m)}(x_t^{(m)}) - \frac{1}{M} \sum_{j=1}^{M} \nabla h(j)(x_t^{(j)}) \|^2 \right]
\]

\[
\leq 4\hat{L}^2 \mathbb{E} \left[ \| y_t^{(m)} - y_{x_t^{(m)}}^{(m)} \|^2 \right] + 4G_1^2 + \frac{4G_2^2}{b_x}
\]

where the equality \((a)\) uses triangle inequality and \((b)\) follows from the application of Proposition E.2.

Next, for the second term of \((49)\), we have:

\[
\sum_{m=1}^{M} \left\| \nabla h^{(m)}(x_t^{(m)}) - \frac{1}{M} \sum_{j=1}^{M} \nabla h(j)(x_t^{(j)}) \right\|^2
\]

\[
\leq 2 \sum_{m=1}^{M} \left\| \nabla h^{(m)}(x_t^{(m)}) - \nabla h^{(m)}(\bar{x}_t) \right\|^2 + 4M \left\| \nabla h(\bar{x}_t) - \frac{1}{M} \sum_{j=1}^{M} \nabla h(j)(x_t^{(j)}) \right\|^2
\]

\[
+ 4 \sum_{m=1}^{M} \left\| \nabla h^{(m)}(\bar{x}_t) - \nabla h(\bar{x}_t) \right\|^2 \leq 6\hat{L}^2 \sum_{m=1}^{M} \left\| x_t^{(m)} - \bar{x}_t \right\|^2 + 4Mc^2 \tag{50}
\]
where \((a)\) follows the generalized triangle inequality; \((b)\) utilizes the heterogeneity Assumption 3.5. Next for the first term, it is 0 when \(t = \bar{t}_s\) and when \(t \neq \bar{t}_s\), we use Lemma D.14. Substituting back to \(\bar{t}_s\) we get the results in the lemma.

**Lemma D.16.** For \(s \in [S]\), we have:

\[
(1 - 12\bar{L}^2 I^2 \eta^2) \sum_{t = \bar{t}_{s-1}}^{\bar{t}_s - 1} D_t \leq 4\bar{L}^2 \sum_{t = \bar{t}_{s-1}}^{\bar{t}_s - 1} B_t + 8I\zeta^2 + 4IG_1^2 + \frac{4IG_2^2}{b_x}
\]

**Proof.** For ease of notation, we denote \(D_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[[\nu_t^{(m)}] - \bar{\nu}_t] \|^2\) and \(B_t = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[[y_t^{(m)}] - y_{\bar{x}_t^{(m)}}] \|^2\). Based on Lemma D.15, we have:

\[
D_t \leq 4\bar{L}^2 B_t + 12\bar{L}^2 I\eta^2 \sum_{t = \bar{t}_{s-1}}^{\bar{t}_s - 1} D_t + 8\zeta^2 + 4G_1^2 + \frac{4G_2^2}{b_x}
\]

Next, we sum from \(\bar{t}_{s-1} + 1\) to \(\bar{t}_s - 1\), we have:

\[
\sum_{t = \bar{t}_{s-1} + 1}^{\bar{t}_s - 1} D_t \leq 4\bar{L}^2 \sum_{t = \bar{t}_{s-1} + 1}^{\bar{t}_s - 1} B_t + 12\bar{L}^2 I\eta^2 \sum_{t = \bar{t}_{s-1} + 1}^{\bar{t}_s - 1} D_t + 8(I - 1)\zeta^2 + 4(I - 1)G_1^2 + \frac{4(I - 1)G_2^2}{b_x}
\]

In the second inequality, we use \(t - 1 \leq \bar{t}_s - 1\), combine with the case when \(t = \bar{t}_{s-1}\) in lemma D.14 we have:

\[
(1 - 12\bar{L}^2 I^2 \eta^2) \sum_{t = \bar{t}_{s-1}}^{\bar{t}_s - 1} D_t \leq 4\bar{L}^2 \sum_{t = \bar{t}_{s-1}}^{\bar{t}_s - 1} B_t + 8I\zeta^2 + 4IG_1^2 + \frac{4IG_2^2}{b_x}
\]

This completes the proof.

**D.2.3 Descent Lemma**

**Lemma D.17.** For all \(t \in [\bar{t}_{s-1}, \bar{t}_s - 1]\), the iterates generated satisfy:

\[
\mathbb{E}\|\triangledown h(\bar{x}_t) - \mathbb{E}[\bar{\nu}_t]\|^2 \leq \frac{2\bar{L}^2}{M} \sum_{m=1}^{M} (4\kappa^2\mathbb{E}\|x_t^{(m)} - \bar{x}_t\|^2 + 2\mathbb{E}\|y_t^{(m)} - y_{\bar{x}_t^{(m)}}\|^2) + 2G_1^2
\]

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

\[
\mathbb{E}\|\triangledown h(\bar{x}_t) - \mathbb{E}[\bar{\nu}_t]\|^2 \leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\mathbb{E}[\bar{\nu}_t^{(m)}] - \nabla h^{(m)}(\bar{x}_t)\|^2
\]

\[
\leq \frac{2}{M} \sum_{m=1}^{M} \left[\mathbb{E}\|\mathbb{E}[\bar{\nu}_t^{(m)}] - \mu_t^{(m)}\|^2 + \|\mu_t^{(m)} - \nabla h^{(m)}(\bar{x}_t)\|^2]\right]
\]

\[
\leq \frac{\bar{L}^2}{M} \sum_{m=1}^{M} (\mathbb{E}\|x_t^{(m)} - \bar{x}_t\|^2 + \mathbb{E}\|y_t^{(m)} - y_{\bar{x}_t^{(m)}}\|^2) + 2G_1^2
\]

\[
\leq \frac{\bar{L}^2}{M} \sum_{m=1}^{M} (\|x_t^{(m)} - \bar{x}_t\|^2 + \mathbb{E}\|y_t^{(m)} - y_{\bar{x}_t^{(m)}} - y_{\bar{x}_t^{(m)}}\|^2) + 2G_1^2
\]

\[
\leq \frac{\bar{L}^2}{M} \sum_{m=1}^{M} (1 + 2\kappa^2) \mathbb{E}\|x_t^{(m)} - \bar{x}_t\|^2 + 2\mathbb{E}\|y_t^{(m)} - y_{\bar{x}_t^{(m)}}\|^2) + 2G_1^2
\]

62
where inequality (a) follows the generalized triangle inequality; inequality (b) follows the Proposition D.3 and Proposition D.2.

Lemma D.18. For $t \neq \bar{t}_s$, the iterates generated satisfy:

$$
\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4} \mathbb{E}\|\nabla \xi[\bar{n}_t]\|^2 + \frac{\eta^2 L G^2}{2b_x M} + \eta G^2_t + \frac{\eta L^2}{M} \left(\mathbb{E}\|\nabla \xi[\bar{n}_t]\|^2 + \mathbb{E}\|\nabla \xi[\bar{n}_t(i)]\|^2\right)
$$

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

$$
\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4} \mathbb{E}\|\nabla \xi[\bar{n}_t]\|^2 + \frac{\eta^2 L G^2}{2b_x M} + \eta G^2_t + \frac{2\eta L^2}{M} \mathbb{E}\|y_t^{(m)} - y_{x_t}^{(m)}\|^2
$$

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

Proof. Using the smoothness of $f$ we have:

$$
\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] + \mathbb{E}\langle \nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{\bar{L}}{2} \mathbb{E}\|\bar{x}_{t+1} - \bar{x}_t\|^2
$$

(a) follows from the iterate update given in Step 6 of Algorithm 2, (b) uses $\langle a, b \rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a - b\|^2)$, (c) follows the assumption that $\eta < 1/2L$. Finally, use Lemma D.17 to bound $T_1$ when $t \neq \bar{t}_s$ finishes the proof.

D.2.4 Proof of Convergence Theorem

We first denote the following potential function $G(t)$:

$$
G_t = \mathbb{E}[h(\bar{x}_t)] + \frac{9\eta L^2}{\mu \gamma} \times \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|y_t^{(m)} - y_{x_t}^{(m)}\|^2
$$

Theorem D.19. Suppose we have constant $\bar{\eta} = \min\left(\frac{1}{2C_{\gamma}^2}, \frac{\mu \gamma}{12L}, \frac{1}{2b_x}, \frac{1}{6L}\right)$, if we choose $\eta = \min\left(\bar{\eta}, \left(\frac{2A}{C_{\gamma}^2}\right)^{1/3}\right)$ and $\gamma = \frac{1}{2T}$, we have:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 = O\left(\frac{\kappa^5}{T} + \left(\frac{\kappa^{16}}{T^2}\right)^{1/3} + \frac{\kappa^5 \sigma^2}{b_y} + \frac{G^2}{b_x M} + G^2_t\right)
$$

To reach an $\epsilon$ stationary point, we choose the inner batch size $b_y = O(\kappa^5 \epsilon^{-1})$, upper batch size $b_x = O(M^{-1} \epsilon^{-1})$ and $Q = O(\kappa \log(\frac{\epsilon}{\delta}))$ in Eq. 17 and $T = O(\kappa^8 \epsilon^{-1.5})$ number of iterations.
Proof. Similar to Lemma D.16 we denote \( D_t = \frac{1}{M} \sum_{m=1}^{M} \| (\nu_t^{(m)} - \bar{\nu}_t^{(m)}) \|^2 \), \( B_t = \frac{1}{M} \sum_{m=1}^{M} \| y_{t,x}^{(m)} - \bar{y}_{t,x}^{(m)} \|^2 \), additionally, we denote \( E_t = \| E_t[\bar{\nu}_t] \|^2 \). First, by Lemma D.13 when \( t \neq \bar{t}_s \), by the triangle inequality, we have:

\[
B_t - B_{t-1} \leq -\frac{\mu \gamma}{2} B_{t-1} + \frac{10 \kappa^2 \eta^2}{\mu \gamma} D_{t-1} + \frac{10 \kappa^2 \eta^2}{\mu \gamma} E_{t-1} + \frac{10 \kappa^2 \eta^2 G_2^2}{\mu \gamma b_z M} + \frac{3 \gamma^2 \sigma^2}{b_y}
\]

When \( t = \bar{t}_s \), we have:

\[
B_t - B_{t-1} \leq -\frac{\mu \gamma}{2} B_{t-1} + \frac{20 \kappa^2 \eta^2}{\mu \gamma} D_{t-1} + \frac{20 \kappa^2 \eta^2}{\mu \gamma} E_{t-1} + \frac{10 \kappa^2 I \eta^2}{\mu \gamma} \sum_{t=1}^{\bar{t}_s-1} D_t + \frac{10 \kappa^2 I \eta^2 G_2^2}{\mu \gamma b_z M} + \frac{3 \gamma^2 \sigma^2}{b_y}
\]

We telescope from \( \bar{t}_s-1 \) to \( \bar{t}_s \) and have:

\[
B_{t_s} - B_{t_s-1} \leq -\frac{\mu \gamma}{2} \sum_{t=1}^{\bar{t}_s-1} B_t + \frac{4 \kappa^2 I \eta^2}{\mu \gamma} \sum_{t=1}^{\bar{t}_s-1} D_t + \frac{20 \kappa^2 \eta^2}{\mu \gamma} \sum_{t=1}^{\bar{t}_s-1} E_t + \frac{10 \kappa^2 I \eta^2 G_2^2}{\mu \gamma b_z M} + \frac{3 \gamma^2 \sigma^2}{b_y}
\]

(51)

Next, by Lemma D.18 when \( t \neq \bar{t}_s \), we have:

\[
\mathbb{E}[h(\bar{x}_{t+1})] - \mathbb{E}[h(\bar{x}_t)] \leq -\frac{\eta}{2} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 - \frac{\eta}{4} E_t + 4 \eta^2 \bar{L}^2 I \eta^3 \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} D_t + 2 \eta \bar{L}^2 B_t + \eta^2 \bar{L} G_2^2 + \eta G_1^2
\]

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

\[
\mathbb{E}[h(\bar{x}_{t+1})] - \mathbb{E}[h(\bar{x}_t)] \leq -\frac{\eta}{2} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 - \frac{\eta}{4} E_t + 2 \eta \bar{L}^2 B_t + \eta^2 \bar{L} G_2^2 + \eta G_1^2
\]

We telescope from \( \bar{t}_s-1 \) to \( \bar{t}_s \) to have:

\[
\mathbb{E}[h(\bar{x}_{t_s})] - \mathbb{E}[h(\bar{x}_{t_s-1})] \leq -\sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} \frac{\eta}{2} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 - \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} \frac{\eta}{4} E_t + 4 \eta^2 \bar{L}^2 I \eta^3 \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} D_t \\
+ \frac{1}{2} L^2 \eta B_t + \frac{I \eta^2 \bar{L} G_2^2}{2 b_z M} + I \eta G_1^2
\]

\[
\leq -\sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} \frac{\eta}{2} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 - \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} \frac{\eta}{4} E_t + 4 \eta^2 \bar{L}^2 I^2 \eta^3 \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} D_t \\
+ 2 \bar{L}^2 \eta \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} B_t + \frac{I \eta^2 \bar{L} G_2^2}{2 b_z M} + I \eta G_1^2
\]

(52)

In the last inequality, we use the fact that \( \bar{t}_s - \bar{t}_s-1 \leq I \).

Next, by the definition of the potential function and combine with Eq. 51 and Eq. 52 we have:

\[
\mathcal{G}_{t_s} - \mathcal{G}_{t_s-1} \leq -\frac{\eta}{2} \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} \mathbb{E} \| \nabla h(\bar{x}_t) \|^2 - \frac{5 \eta \bar{L}^2}{2} \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} B_t - \frac{\eta}{4} \left( 1 - \frac{720 \kappa^2 \eta \bar{L}^2}{\mu^2 \gamma^2} \right) \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} E_t \\
+ \left( 360 \frac{\mu^2 \gamma^2}{\mu^2 \gamma^2} + 4 I \right) \eta^3 \kappa^2 \bar{L}^2 I \sum_{t=\bar{t}_s-1}^{\bar{t}_s-1} D_t + \frac{27 \bar{L}^2 \gamma \eta^2}{b_y \mu} \\
+ 90 I \kappa^2 \bar{L}^2 \eta^2 G_2^2 + \frac{I \eta^2 \bar{L} G_2^2}{2 b_z M} + I \eta G_1^2
\]
to bound the coefficients above, we choose $\eta \leq \frac{\mu^2}{48\kappa L}$. Then we have:

$$G_t - G_{t-1} \leq \frac{\eta}{2} \sum_{t=t_{s-1}}^{t_s} E\|\nabla h(\tilde{x}_t)\|^2 - \frac{5\eta \beta^2}{2} \sum_{t=t_{s-1}}^{t_s} B_t - \frac{\eta}{8} \sum_{t=t_{s-1}}^{t_s} E_t$$

$$+ \left( \frac{360}{\mu^2 \kappa^2} + 4I \right) \eta^3 \kappa^2 \beta^2 I \sum_{t=t_{s-1}}^{t_s} D_t + \frac{27 I \bar{L}^2 \eta \gamma^2}{b_M \mu}$$

$$+ \frac{90 I \kappa^2 \bar{L}^2 \eta^2 G_2^2}{b_M \mu} + \frac{I \eta^2 \bar{L} G_2^2}{2b_M} + I \eta G_1^2$$

By lemma D.16 and choosing $\eta < \frac{1}{6T}$, we have:

$$\sum_{t=t_{s-1}}^{t_s} D_t \leq 6\bar{L}^2 \sum_{t=t_{s-1}}^{t_s} B_t + 18\bar{I} \eta^2 + 6I G_2^2 + \frac{6 I G_2^2}{b_x}$$

Next, we denote $C_1 = \left( \frac{360}{\mu^2 \kappa^2} + 4I \right) \kappa^2 \bar{L}^2$, and choose $\eta < \min \left( \frac{1}{2C_1 \gamma^2}, \frac{\mu^2}{12 \kappa L}, \frac{1}{2T} \right)$ then we have:

$$G_t - G_{t-1} \leq \frac{\eta}{2} \sum_{t=t_{s-1}}^{t_s} E\|\nabla h(\tilde{x}_t)\|^2 - \frac{5\eta \beta^2}{2} \sum_{t=t_{s-1}}^{t_s} B_t - \frac{\eta}{8} \sum_{t=t_{s-1}}^{t_s} E_t$$

$$+ C_1 I \eta^3 \left( 6\bar{L}^2 \sum_{t=t_{s-1}}^{t_s} B_t + 18 \bar{I} \eta^2 + 6I G_2^2 + \frac{6 I G_2^2}{b_x} \right)$$

$$+ \frac{27 I \bar{L}^2 \eta \gamma^2}{b_M \mu} + \frac{I \eta^2 \bar{L} G_2^2}{2b_M} + I \eta G_1^2$$

$$\leq \frac{\eta}{2} \sum_{t=t_{s-1}}^{t_s} E\|\nabla h(\tilde{x}_t)\|^2 + 18C_1 I \bar{L}^2 \eta^3 \bar{I} \eta^2 + 6C_1 I \bar{L}^2 \eta^3 G_2^2 + \frac{6 C_1 I \bar{L}^2 \eta^3 G_2^2}{b_x} + \frac{5 I \eta G_2^2}{8b_M}$$

Sum over all $s \in [S]$ (assume $T = SI + 1$ without loss of generality) to obtain:

$$\frac{\eta}{2} \sum_{t=1}^{T} E\|\nabla h(\tilde{x}_t)\|^2 \leq G_1 - G_T + 18C_1 T I \bar{L}^2 \eta^3 \bar{I} \eta^2 + 6C_1 T I \bar{L}^2 \eta^3 G_2^2 + \frac{6 C_1 T \bar{L}^2 \eta^3 G_2^2}{b_x}$$

$$+ \frac{5 T \eta G_2^2}{8b_M} + \frac{27 T I \bar{L}^2 \eta \gamma^2}{b_M \mu} + \frac{T \eta^2 \bar{L} G_2^2}{2b_M} + T \eta G_1^2$$

$$\leq \Delta + \frac{9 \eta \bar{L}^2 \eta^2}{\mu \gamma^2} + 18C_1 T I \bar{L}^2 \eta^3 \bar{I} \eta^2 + 6C_1 T I \bar{L}^2 \eta^3 G_2^2 + \frac{6 C_1 T \bar{L}^2 \eta^3 G_2^2}{b_x}$$

$$+ \frac{5 T \eta G_2^2}{8b_M} + \frac{27 T I \bar{L}^2 \eta \gamma^2}{b_M \mu} + \frac{T \eta^2 \bar{L} G_2^2}{2b_M} + T \eta G_1^2$$

we define $\Delta = h(x_1) - h^*$ as the initial sub-optimality of the function and $\Delta_y = \frac{1}{M} \sum_{m=1}^{M} \|y_1^{(m)} - \hat{y}_x^{(m)}\|^2$ as the initial sub-optimality of the inner variable estimation, then we divide by $\eta T/2$ on both sides and have:

$$\frac{1}{T} \sum_{t=1}^{T} E\|\nabla h(\tilde{x}_t)\|^2 \leq \frac{2 \Delta}{\eta T} + \frac{\eta^2 L G_2^2}{2b_M} + \left( \frac{36C_1 I \bar{L}^2 \eta^2 + 12C_1 I \bar{L}^2 G_2^2}{b_x} \right) \eta^2$$

$$+ \frac{18 I \bar{L}^2 \Delta_y}{\mu \gamma T} + \frac{54 I \bar{L}^2 \eta \gamma^2}{b_M \mu} + \frac{5 G_2^2}{4b_M} + G_1^2$$

65
As shown in the inequality, we break the bound into three parts. The $T_1$ part has a structure similar to that for the single level federated learning problems. Then the $T_2$ part includes the optimization error of the lower problem, and the statistical error of sampling. Finally, the $T_3$ part includes the bias and variance of the hyper-gradient estimate.

Next, by $\eta < \frac{1}{2L}$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq \frac{2\Delta}{\eta T} + \left(36C_1 \dot{L}^2 \zeta^2 + 12C_1 \dot{L}^2 G_1^2 + \frac{12C_1 \dot{L}^2 G_1^2}{b_x}\right)\eta^2$$

$$+ \frac{18\dot{L}^2 \Delta_y}{\mu \gamma T} + \frac{54\dot{L}^2 \gamma^2 \sigma^2}{b_y \mu} + \frac{G_2^2}{4b_x M} + \frac{5G_2^2}{4b_x M} + G_1^2$$

(53)

Next, we denote constant $\bar{\eta} = \min\left(\frac{1}{2C_1 \dot{L}^2}, \frac{\mu \gamma T}{2\dot{L}^2}, \frac{1}{6L}\right)$ and $C_\eta = \left(36C_1 \dot{L}^2 \zeta^2 + 12C_1 \dot{L}^2 G_1^2 + \frac{12C_1 \dot{L}^2 G_1^2}{b_x}\right)$ we choose

$$\eta = \min \left(\bar{\eta}, \left(\frac{2\Delta}{C_\eta \bar{T}}\right)^{1/3}\right)$$

and $\gamma = \frac{1}{2L}$, and obtain:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq \frac{2\Delta}{\eta T} + \frac{36C_1 \dot{L}^2 \Delta_y}{T} + \left(\frac{4C_1 \dot{L}^2 \Delta_y}{T^2}\right)^{1/3} + \frac{27\dot{L}^2 \sigma^2}{\mu \gamma \mu} + \frac{3G_2^2}{2b_x M} + 2G_1^2$$

Finally, since $\dot{L} = O(\kappa^2)$, $\dot{L} = O(\kappa^3)$ and $\mu \gamma = O(\kappa^{-1})$. Suppose we choose $I = O(1)$, then $\bar{\eta} = O(\kappa^{-4})$ and $C_1 = O(\kappa^3)$, $\zeta = O(\kappa^2)$, $C_\eta = O(\kappa^{16})$, thus, we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla h(\bar{x}_t)\|^2 = O \left(\frac{\kappa^5}{T} + \left(\frac{\kappa^{16}}{T^2}\right)^{1/3} + \frac{\kappa^5 \sigma^2}{b_y} + \frac{G_2^2}{b_x M} + G_1^2\right)$$

and to reach an $\epsilon$ stationary point, we choose the inner batch size $b_y = O(\kappa^5 \epsilon^{-1})$, upper batch size $b_x = O(M^{-1} \epsilon^{-1})$ and $Q = O(\kappa \log(\frac{1}{\epsilon}))$ in Eq. (11) and $T = O(\kappa^8 \epsilon^{-1.5})$ number of iterations.

### E Useful Propositions

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

**Proposition E.1** (Lemma 3 of [27]). (generalized triangle inequality) Let $\{x_k\}, k \in K$ be $K$ vectors. Then the following are true:

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

**Proposition E.2** (Lemma C.1 of [30]). 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 E.3** (Lemma C.2 of [30]). Let $a_0 > 0$ and $a_1, a_2, \ldots, a_T \geq 0$. We have

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

**Proposition E.4.** Suppose we have function $g(y)$, which is $L$-smooth and $\mu$-strongly-convex, then suppose $\gamma < \frac{1}{2}$, the progress made by one step of gradient descent is:

$$\mathbb{E}\|y_{t+1} - y^*\|^2 \leq (1 - \mu \gamma)\|y^* - y_t\|^2 + 2\gamma^2 \sigma^2$$

where $y^*$ is the minimum of $g(y)$ and we have update rule $g(y_{t+1}) = g(y_t) - \gamma \nabla g(y_t, \xi)$, where the error of stochastic gradient estimate is bounded by $\sigma^2$. 

66
Proof. First, by the strong convexity of of function $g(y)$, we have:

$$g(y^*) \geq g(y_t) + \langle \nabla_y g(y_t), y^* - y_t \rangle + \frac{\mu}{2} \| y^* - y_t \|^2$$

$$= g(y_t) + \langle \nabla_y g(y_t), y^* - y_{t+1} \rangle + \langle \nabla_y g(y_t), y_{t+1} - y_t \rangle + \frac{\mu}{2} \| y^* - y_t \|^2$$

Then by $L$-smoothness, we have: $$\frac{L}{2} \| y_{t+1} - y_t \|^2 \geq g(y_{t+1}) - g(y_t) - \langle \nabla_y g(y_t), y_{t+1} - y_t \rangle.$$ Combining above two inequalities and take expectation on both sides, we have

$$g(y^*) \geq \mathbb{E} g(y_{t+1}) + \mathbb{E} \langle \nabla_y g(y_t), y^* - y_{t+1} \rangle + \frac{\mu}{2} \| y^* - y_t \|^2 - \frac{L}{2} \mathbb{E} \| y_{t+1} - y_t \|^2$$

$$\geq \mathbb{E} g(y_{t+1}) + \gamma \| \nabla_y g(y_t) \|^2 + \langle \nabla_y g(y_t), y^* - y_{t+1} \rangle + \frac{\mu}{2} \| y^* - y_t \|^2 - \frac{L\gamma^2}{2} \mathbb{E} \| \nabla_y g(y_t) \|^2$$

Then we choose $\gamma < \frac{1}{L}$, we obtain

$$0 \geq \mathbb{E} \| y_{t+1} - y_t \|^2 = \mathbb{E} \| y_{t+1} - \gamma \nabla_y g(y_t, \xi) - y^* \|^2 = \mathbb{E} \| y_{t+1} - y^* \|^2 - 2\gamma \langle \nabla_y g(y_t), y_{t+1} - y^* \rangle + \gamma^2 \mathbb{E} \| \nabla_y g(y_t, \xi) \|^2$$

Then since we choose $\gamma < \frac{1}{L}$, we obtain:

$$\mathbb{E} \| y_{t+1} - y_t \|^2 \leq (1 - \mu \gamma) \| y^* - y_t \|^2 + 2\gamma^2 \sigma^2$$

This completes the proof. 

\[\square\]

**Proposition E.5.** Suppose we have function $g(y)$, which is $L$-smooth and $\mu$-strongly-convex, then suppose $\gamma < \frac{1}{L}$ and $\alpha_t < 1$, the progress made by one step of gradient descent is:

$$\| y_{t+1} - y^* \|^2 \leq (1 - \frac{\mu \gamma \alpha_t}{2}) \| y_t - y^* \|^2 - \frac{\gamma^2 \alpha_t}{4} \| \omega_t \|^2$$

$$+ \frac{4\gamma \alpha_t}{\mu} \| \nabla_y g(x_t, y_t) \|^2 + \frac{3\gamma^2 \alpha_t}{2} \text{Var}[\omega_t].$$

where $y^*$ is the minimum of $g(y)$ and we have update rule $g(y_{t+1}) = g(y_t) - \gamma \alpha_t \omega_t$.

Proof. First, Suppose we denote $\tilde{y}_{t+1} = y_t - \gamma \omega_t$, then we have $y_{t+1} = y_t + \alpha_t (\tilde{y}_{t+1} - y_t)$. By the strong convexity of of function $g(y_t)$, we have:

$$g(y^*) \geq g(y_{t+1}) + \langle \nabla_y g(y_{t+1}), y^* - y_{t+1} \rangle + \frac{\mu}{2} \| y^* - y_{t+1} \|^2$$

$$= g(y_t) + \mathbb{E} \langle \nabla_y g(y_t), y^* - \tilde{y}_{t+1} \rangle + \mathbb{E} \langle \nabla_y g(y_{t+1}) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle$$

$$- \gamma \langle \nabla_y g(y_t), \mathbb{E}[w_t] \rangle + \frac{\mu}{2} \| y^* - y_{t+1} \|^2$$

(54)

where the expectation is w.r.t the stochasticity of $\omega_t$. Then by $L$-smoothness, we have:

$$\frac{L}{2} \mathbb{E} \| \tilde{y}_{t+1} - y_t \|^2 \geq \mathbb{E} g(y_{t+1}) - g(y_t) + \gamma \langle \nabla_y g(y_t), \mathbb{E}[w_t] \rangle$$

(55)
Combining the (54) with (55), we have
\[
g(y^*) \geq \mathbb{E}g(\tilde{y}_{t+1}) + \mathbb{E}\langle \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle \\
+ \frac{\mu}{2} \|y^* - y_t\|^2 - \frac{L}{2} \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 \\
\geq \mathbb{E}g(\tilde{y}_{t+1}) + \gamma \|\mathbb{E}[w_t]\|^2 + \langle \mathbb{E}[w_t], y^* - y_t \rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle \\
+ \frac{\mu}{2} \|y^* - y_t\|^2 - \frac{L\gamma^2}{2} \mathbb{E}\|\omega_t\|^2 \\
\geq \mathbb{E}g(\tilde{y}_{t+1}) + \langle \mathbb{E}[w_t], y^* - y_t \rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle \\
+ \frac{\mu}{2} \|y^* - y_t\|^2 + \left(\gamma - \frac{L\gamma^2}{2}\right) \|\mathbb{E}[\omega_t]\|^2 - \frac{L\gamma^2}{2} \text{Var} [\omega_t]
\]
where \(\text{Var}\) denotes the variance. By definition of \(y^*\), we have \(g(y^*) \geq g(\tilde{y}_{t+1})\). Thus, we obtain
\[
0 \geq \mathbb{E}[w_t], y^* - y_t \rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle \\
+ \frac{\mu}{2} \|y^* - y_t\|^2 + \left(\gamma - \frac{L\gamma^2}{2}\right) \|\mathbb{E}[\omega_t]\|^2 - \frac{L\gamma^2}{2} \text{Var} [\omega_t]
\]

(56)

Considering the upper bound of the second term \(\langle \nabla_y g(y_t) - w_t, y^* - y_t \rangle\), we have
\[
-\mathbb{E}\langle \nabla_y g(y_t) - w_t, y^* - \tilde{y}_{t+1} \rangle \\
= -\mathbb{E}\langle \nabla_y g(y_t) - w_t, y^* - y_t \rangle + \mathbb{E}\langle \nabla_y g(y_t) - w_t, w_t, \omega_t \rangle \\
\leq \frac{1}{\mu} \|\nabla_y g(y_t) - w_t\|^2 + \frac{\mu}{2} \|y^* - y_t\|^2 + \frac{1}{\mu} \|\nabla_y g(y_t) - w_t\|^2 + \frac{\mu\gamma^2}{4} \|\mathbb{E}[\omega_t]\|^2 \\
= \frac{2}{\mu} \|\nabla_y g(y_t) - w_t\|^2 + \frac{\mu}{4} \|y^* - y_t\|^2 + \frac{\mu\gamma^2}{4} \|\mathbb{E}[\omega_t]\|^2.
\]

Combining with Eq. (56)
\[
0 \geq \mathbb{E}[w_t], y^* - y_t \rangle - \frac{2}{\mu} \|\nabla_y g(y_t) - w_t\|^2 + \left(\gamma - \frac{3L\gamma^2}{4}\right) \|\mathbb{E}[w_t]\|^2 + \frac{\mu}{4} \|y^* - y_t\|^2 - \frac{L\gamma^2}{2} \text{Var} [\omega_t]
\]

By \(y_{t+1} = y_t - \gamma \alpha_t \omega_t\), we have:
\[
\mathbb{E}[y_{t+1} - y^*] = \mathbb{E}[y_t - \gamma \alpha_t \omega_t - y^*] = \|y_t - y^*\|^2 - 2\gamma \alpha_t \mathbb{E}[\omega_t, y_t - y^*] + 2\gamma^2 \alpha_t^2 \mathbb{E}[\|\omega_t\|^2] \\
\leq \left(1 - \frac{\mu\gamma^2}{2}\right) \|y_t - y^*\|^2 - 2\gamma \alpha_t \left(\gamma - \frac{\alpha_t}{2} - \frac{3L\gamma^2}{4}\right) \|\omega_t\|^2 \\
+ \frac{4\gamma^2 \alpha_t}{\mu} \|\nabla_y g(y_t) - w_t\|^2 + (L\gamma^3 \alpha_t + \alpha_t^2) \text{Var} [\omega_t]
\]

Then since we choose \(\gamma < \frac{1}{2L}, \alpha_t < 1\), we obtain:
\[
\mathbb{E}[y_{t+1} - y^*] \leq \left(1 - \frac{\mu\gamma^2}{2}\right) \|y_t - y^*\|^2 - \frac{\gamma^2 \alpha_t^2}{4} \|\mathbb{E}[\omega_t]\|^2 \\
+ \frac{4\gamma \alpha_t}{\mu} \|\nabla_y g(x_t, y_t) - w_t\|^2 + \frac{3\gamma^2 \alpha_t}{2} \text{Var} [\omega_t].
\]

This completes the proof. \(\square\)