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, 1] 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.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: **Comparisons of the Federated/Non-federated bilevel optimization algorithms for finding an ϵ-stationary point of** (1). $Gc(f,\epsilon)$ and $Gc(g,\epsilon)$ denote the number of gradient evaluations *w.r.t.* $f^{(m)}(x,y)$ and $g^{(m)}(x,y)$; $JV(g,\epsilon)$ denotes the number of Jacobian-vector products; $HV(g,\epsilon)$ is the number of Hessian-vector products; $\kappa = L/\mu$ is the condition number, $p(\kappa)$ is used when no dependence is provided. Sample complexities are measured by client.

Setting	Algorithm	Communication	$Gc(f,\epsilon)$	$Gc(g,\epsilon)$	$JV(g,\epsilon)$	$HV(g,\epsilon)$	Heterogeneity
Non-Fed	StocBiO [25]		$O(\kappa^5\epsilon^{-2})$	$O(\kappa^9\epsilon^{-2})$	$O(\kappa^5\epsilon^{-2})$	$O(\kappa^6\epsilon^{-2})$	
	MRBO [59]		$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	$O(p(\kappa)\epsilon^{-1.5})$	
Federated	CommFedBiO [35]	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	✓
	FedNest [51]	$O(\kappa^9\epsilon^{-2})$	$O(\kappa^5\epsilon^{-2})$	$O(\kappa^9\epsilon^{-2})$	$O(\kappa^5\epsilon^{-2})$	$O(\kappa^9\epsilon^{-2})$	✓
	AggITD [56]	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	$O(p(\kappa)\epsilon^{-2})$	✓
	FedMBO [22]	$O(M^{-1}p(\kappa)\epsilon^{-2})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	✓
	SimFBO [61]	$O(p(\kappa)\epsilon^{-1})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	$O(M^{-1}p(\kappa)\epsilon^{-2})$	✓
	Local-BSGVR [13]	$O(p(\kappa)\epsilon^{-1})$	$O(M^{-1}p(\kappa)\epsilon^{-1.5})$	$O(M^{-1}p(\kappa)\epsilon^{-1.5})$	$O(M^{-1}p(\kappa)\epsilon^{-1.5})$	$O(M^{-1}p(\kappa)\epsilon^{-1.5})$	✗
	FedBiOAcc (Ours)	$\boldsymbol{O(\kappa^{19/3}\epsilon^{-1})}$	$\boldsymbol{O(M^{-1}\kappa^8\epsilon^{-1.5})}$	$\boldsymbol{O(M^{-1}\kappa^8\epsilon^{-1.5})}$	$\boldsymbol{O(M^{-1}\kappa^8\epsilon^{-1.5})}$	$\boldsymbol{O(M^{-1}\kappa^8\epsilon^{-1.5})}$	✓

following research question: *Is it possible to develop communication-efficient federated algorithms tailored for bilevel optimization problems that also ensure a rapid convergence rate?*

More specifically, a general Federated Bilevel Optimization problem has the following form:

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

A federated bilevel optimization problem consists of an upper and a lower level problem, the upper problem $f(x,y) \coloneqq \frac{1}{M}\sum_{m=1}^{M} f^{(m)}(x,y)$ relies on the solution y_x of the lower problem, and $g(x,y) \coloneqq \frac{1}{M}\sum_{m=1}^{M} g^{(m)}(x,y)$. Meanwhile, both the upper and the lower level problems are federated: In Eq.(1), we have M clients, and each client has a local upper problem $f^{(m)}(x,y)$ and a lower level problem $g^{(m)}(x,y)$. Compared to single-level federated optimization problems, the estimation of the hyper-gradient in federated bilevel optimization problems is much more challenging. In Eq.(1), the hyper-gradient is not linear *w.r.t* the local hyper-gradients of clients, whereas the gradient of a single-level Federated Optimization problem is the average of local gradients. Consequently, directly applying the vanilla local-sgd method [42] to federated bilevel problems results in a large bias. In the literature [51, 35, 22, 56], researchers evaluate the hyper-gradient through multiple rounds 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 $O(\epsilon^{-1.5})$, 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 ∇ denotes full gradient, ∇_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, 2, 8]. 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 [35, 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:

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

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 μ-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**)

Input: Constants c_ω, c_ν, c_u, γ, η, τ, r; learning rate schedule $\{\alpha_t\}$, $t \in [T]$, initial state (x_1, y_1, u_1);
Initialization: Set $y_1^{(m)} = y_1$, $x_1^{(m)} = x_1$, $u_1^{(m)} = u_1$, $\omega_1^{(m)} = \nabla_y g^{(m)}(x_1, y_1, \mathcal{B}_y)$, $\nu_1^{(m)} = \nabla_x f^{(m)}(x_1, y_1; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_1, y_1; \mathcal{B}_{g,1}) u_1$ and $q_1 = \nabla_{y^2} g^{(m)}(x_1^{(m)}, y_1^{(m)}; \mathcal{B}_{g,2}) u_1 - \nabla_y f^{(m)}(x_1^{(m)}, y_1^{(m)}; \mathcal{B}_{f,2})$ for $m \in [M]$
for $t = 1$ **to** T **do**
  $\hat{y}_{t+1}^{(m)} = y_t^{(m)} - \gamma \alpha_t \omega_t^{(m)}$, $\hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta \alpha_t \nu_t^{(m)}$, $\hat{u}_{t+1}^{(m)} = \mathcal{P}_r(u_t^{(m)} - \tau \alpha_t q_t^{(m)})$
  if t mod I $= 0$ **then**
    $y_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M \hat{y}_{t+1}^{(j)}$; $x_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M \hat{x}_{t+1}^{(j)}$, $u_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M \hat{u}_{t+1}^{(j)}$
  else
    $y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}$, $x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)}$, $u_{t+1}^{(m)} = \hat{u}_{t+1}^{(m)}$
  end if
  Get $\hat{\omega}_{t+1}^{(m)}$, $\hat{\nu}_{t+1}^{(m)}$ and $\hat{q}_{t+1}^{(m)}$ following Eq. (7)
  if t mod I $= 0$ **then**
    $\omega_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M \hat{\omega}_{t+1}^{(j)}$, $\nu_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M \hat{\nu}_{t+1}^{(j)}$, $q_{t+1}^{(m)} = \frac{1}{M} \sum_{j=1}^M \hat{q}_{t+1}^{(j)}$,
  else
    $\omega_{t+1}^{(m)} = \hat{\omega}_{t+1}^{(m)}$, $\nu_{t+1}^{(m)} = \hat{\nu}_{t+1}^{(m)}$, $q_{t+1}^{(m)} = \hat{q}_{t+1}^{(m)}$
  end if
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: $\|\nabla f^{(m)}(z) - \nabla f^{(j)}(z)\| \leq \zeta_f$, $\|\nabla g^{(m)}(z) - \nabla g^{(j)}(z)\| \leq \zeta_g$, $\|\nabla_{xy} g^{(m)}(z) - \nabla_{xy} g^{(j)}(z)\| \leq \zeta_{g,xy}$, $\|\nabla_{y^2} g^{(m)}(z) - \nabla_{y^2} g^{(j)}(z)\| \leq \zeta_{g,yy}$, where ζ_f, ζ_g, $\zeta_{g,xy}$, $\zeta_{g,yy}$, 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, 15]. 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 $\Phi(x, y)$ is denoted as:

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

Based on Assumption 3.1$\sim$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:

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

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 $l(u)$:

$$\min_{u \in \mathbb{R}^d} l(u) = \frac{1}{M} \sum_{m=1}^M u^T (\nabla_{y^2} g^{(m)}(x, y)) u - \langle \nabla_y f^{(m)}(x, y), u \rangle \tag{4}$$

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

$$\nabla h(x) = \frac{1}{M}\sum_{m=1}^{M}\big(\nabla_x f^{(m)}(x, y_x) - \nabla_{xy} g^{(m)}(x, y_x)u^*\big) \tag{5}$$

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 [42], suppose that each client maintains a variable $u_t^{(m)}$, and performs the following update:

$$u_{t+1}^{(m)} = \mathcal{P}_r(u_t^{(m)} - \tau_t \nabla l^{(m)}(u_t^{(m)}; \mathcal{B}))$$
$$\nabla l^{(m)}(u_t^{(m)}; \mathcal{B}) = \nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,2}))u_t^{(m)} - \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,2})$$

where $\nabla l^{(m)}(u_t^{(m)}; \mathcal{B})$ 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, the $\mathcal{P}_r(\cdot)$ denotes the projection to a bounded ball of radius-r. Note that Clients perform multiple local updates of $u_t^{(m)}$ before averaging. As for Eq. (5), each client evaluates $\nabla h^{(m)}(x)$ locally: $\nabla h^{(m)}(x) = \nabla_x f^{(m)}(x, y_x) - \nabla_{xy} g^{(m)}(x, y_x)u^*$ 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 $x_t^{(m)}$, the lower level variable $y_t^{(m)}$ and hyper-gradient computation variable $u_t^{(m)}$. For example, for each client $m \in [M]$, we perform the following local updates:

$$y_{t+1}^{(m)} = y_t^{(m)} - \gamma_t \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y),\ u_{t+1}^{(m)} = \mathcal{P}_r(u_t^{(m)} - \tau_t \nabla l^{(m)}(u_t^{(m)}; \mathcal{B}))$$
$$x_{t+1}^{(m)} = x_t^{(m)} - \eta_t\big(\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,1})u_t^{(m)}\big) \tag{6}$$

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 $x_t^{(m)}$, 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 FedIHGP 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 ϵ-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 $q_t^{(m)}$ to be the momentum estimator for $x_t^{(m)}$, $y_t^{(m)}$ and $u_t^{(m)}$ respectively, and we update them following the rule of STORM [6]:

$$\begin{aligned}
\hat{\omega}_{t+1}^{(m)} &= \nabla_y g^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}, \mathcal{B}_y) + (1 - c_\omega \alpha_t^2)(\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y)) \\
\hat{\nu}_{t+1}^{(m)} &= \big(\nabla_x f^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; \mathcal{B}_{g,1})u_{t+1}^{(m)}\big) \\
&\quad + (1 - c_\nu \alpha_t^2)\big(\nu_t^{(m)} - \big(\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,1})u_t^{(m)}\big)\big) \\
\hat{q}_{t+1}^{(m)} &= \big(\nabla_{y^2} g^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; \mathcal{B}_{g,2})u_{t+1}^{(m)} - \nabla_y f^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; \mathcal{B}_{f,2})\big) \\
&\quad + (1 - c_u \alpha_t^2)\big(q_t^{(m)} - \big(\nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,2})u_t^{(m)} - \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,2})\big)\big)
\end{aligned} \tag{7}$$

where c_ω, c_ν and c_u are constants, α_t is the learning rate. Then we update the $x_t^{(m)}$, $y_t^{(m)}$ and $u_t^{(m)}$ as follows:

$$\hat{y}_{t+1}^{(m)} = y_t^{(m)} - \gamma\alpha_t\omega_t^{(m)}, \hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta\alpha_t\nu_t^{(m)}, \hat{u}_{t+1}^{(m)} = \mathcal{P}_r(u_t^{(m)} - \tau\alpha_t q_t^{(m)}) \tag{8}$$

where γ, η, τ are constants and α_t is the learning rate. The FedBiOAcc algorithm following Eq. (8) 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:

$$\bar{x}_t = \frac{1}{M}\sum_{m=1}^{M} x_t^{(m)}, \bar{y}_t = \frac{1}{M}\sum_{m=1}^{M} y_t^{(m)}, \bar{u}_t = \frac{1}{M}\sum_{m=1}^{M} u_t^{(m)}$$

we denote the average of the momentum similarly as $\bar{\omega}_t$, $\bar{\nu}_t$ and $\bar{q}_t$. Then we consider the following Lyapunov function $\mathcal{G}_t$:

$$\begin{aligned}\mathcal{G}_t = h(\bar{x}_t) + \frac{18\eta\tilde{L}^2}{\mu\gamma}(\|\bar{y}_t - y_{\bar{x}_t}\|^2 + \|\bar{u}_t - u_{\bar{x}_t}\|^2) + \frac{9bM\eta}{64\alpha_t}\|\bar{\omega}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\|^2 \\ + \frac{9bM\eta}{64\alpha_t}\|\bar{q}_t - \frac{1}{M}\sum_{m=1}^{M}(\nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)})u_t^{(m)} - \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}))\|^2 \\ + \frac{9bM\eta}{64\alpha_t}\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}(\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)})u_t^{(m)})\|^2\end{aligned} \tag{9}$$

where $y_{\bar{x}_t}$ denotes the solution of the lower level problem $g(\bar{x}_t, \cdot)$, $u_{\bar{x}_t} = [\nabla_{y^2} g(\bar{x}, y_{\bar{x}})]^{-1}\nabla_y f(\bar{x}, y_{\bar{x}})$ denotes the solution of Eq (4) at state $\bar{x}_t$. Besides, γ, η, τ are learning rates and $L, \tilde{L}$ are constants. Note that the first three terms of $\mathcal{G}_t$: $h(\bar{x}_t)$, $\|\bar{y}_t - y_{\bar{x}_t}\|^2$, $\|\bar{u}_t - u_{\bar{x}_t}\|^2$ 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 $\bar{q}_t$. 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 $\alpha_t = \frac{\delta}{(u+t)^{1/3}}, t \in [T]$, for some constant δ and u, and let c_ν, c_ω, c_u choose some value, η, γ and τ, r 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 $b_1 = O(Ib)$, then we have:*

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

To reach an ϵ-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 ϵ-stationary point, we need $T = O(\kappa^8(bM)^{-1}\epsilon^{-1.5})$, then the sample complexity for each client is $Gc(f,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$, $Gc(g,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$, $Jv(g,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$, $Hv(g,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$. So FedBiOAcc achieves the linear speed up *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. (2)), 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:

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

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 $\min_{x\in\mathbb{R}^p} h(x) \coloneqq \frac{1}{M}\sum_{m=1}^{M} h^{(m)}(x)$. 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$\sim$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. (3) involves Hessian inverse, so we only evaluate it approximately through the Neumann series [39] as:

$$\begin{aligned}\Phi^{(m)}(x,y;\xi_x) = \nabla_x f^{(m)}(x,y;\xi_f) - \tau\nabla_{xy} g^{(m)}(x,y;\xi_g)\\ \times \sum_{q=-1}^{Q-1}\prod_{j=Q-q}^{Q}(I-\tau\nabla_{y^2} g^{(m)}(x,y;\xi_j))\nabla_y f^{(m)}(x,y;\xi_f)\end{aligned} \tag{11}$$

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 $y^{(m)}$ towards the minimizer $y_{x^{(m)}}^{(m)}$, but after the communication step, $x^{(m)}$ is smoothed among clients, as a result, the target of $y_t^{(m)}$ 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 *w.r.t* 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.

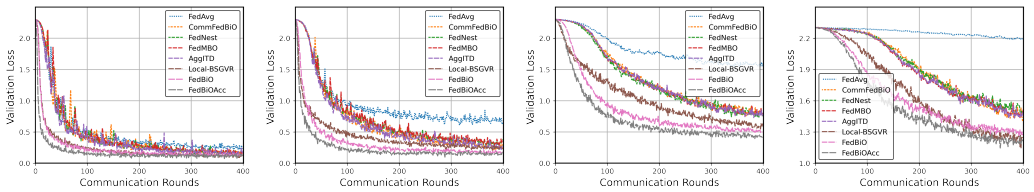

Figure 1: Validation Error vs Communication Rounds. From Left to Right: $\rho = 0.1, 0.4, 0.8, 0.95$. The local step I is set as 5 for FedBiO, FedBiOAcc and FedAvg.

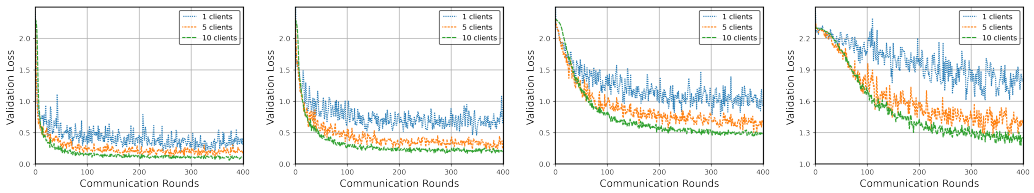

Figure 2: Validation Error vs Communication Rounds with different number of clients per epoch. From Left to Right: $\rho = 0.1, 0.4, 0.8, 0.95$. The local step I is set as 5.

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 ρ ($0 \leq \rho \leq 1$) 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 [35], 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 ρ. Note that the larger the ρ 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 ρ 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.

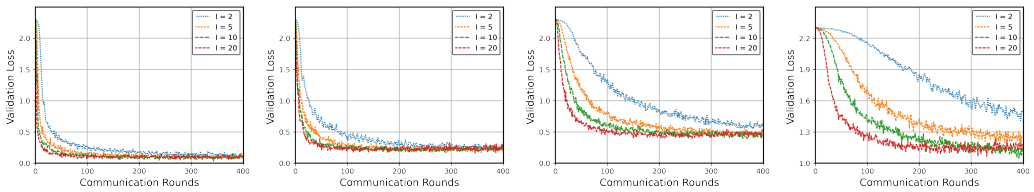

Figure 3: Validation Error vs Communication Rounds with different number of local steps I. From Left to Right: $\rho = 0.1, 0.4, 0.8, 0.95$.

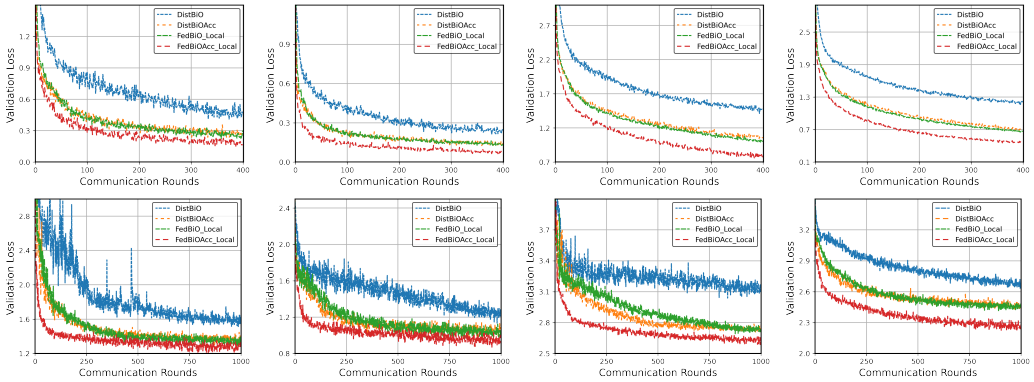

Figure 4: Validation Error vs Communication Rounds. The top row shows the result for the Omniglot Dataset and the bottom row shows MiniImageNet. From Left to Right: 5-way-1-shot, 5-way-5-shot, 20-way-1-shot, 20-way-5-shot. The local step I is set to 5.

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

[1] A. K. R. Bayoumi, K. Mishchenko, and P. Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529, 2020.

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

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

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

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

[6] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pages 15236–15245, 2019.

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

[8] C. B. Do, C.-S. Foo, and A. Y. Ng. Efficient multiple hyperparameter learning for log-linear models. In *NIPS*, volume 2007, pages 377–384. Citeseer, 2007.

[9] J. Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

[10] M. C. Ferris and O. L. Mangasarian. Finite perturbation of convex programs. *Applied Mathematics and Optimization*, 23(1):263–273, 1991.

[11] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[12] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1165–1173. JMLR. org, 2017.

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

[14] H. Gao, B. Gu, and M. T. Thai. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *International Conference on Artificial Intelligence and Statistics*, pages 9238–9281. PMLR, 2023.

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

[16] Z. Guo, Q. Hu, L. Zhang, and T. Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

[17] F. Haddadpour and M. Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

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

[19] F. Huang. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023.

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

[21] F. Huang, J. Li, S. Gao, and H. Huang. Enhanced bilevel optimization via bregman distance. *Advances in Neural Information Processing Systems*, 35:28928–28939, 2022.

[22] M. Huang, D. Zhang, and K. Ji. Achieving linear speedup in non-iid federated bilevel learning. *arXiv preprint arXiv:2302.05412*, 2023.

[23] K. Ji and Y. Liang. Lower bounds and accelerated algorithms for bilevel optimization. *arXiv preprint arXiv:2102.03926*, 2021.

[24] K. Ji, J. Yang, and Y. Liang. Provably faster algorithms for bilevel optimization and applications to meta-learning. *arXiv preprint arXiv:2010.07962*, 2020.

[25] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021.

[26] Y. Jiao, K. Yang, T. Wu, D. Song, and C. Jian. Asynchronous distributed bilevel optimization. *arXiv preprint arXiv:2212.10048*, 2022.

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

[28] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

[29] P. Khanduri, P. Sharma, H. Yang, M. Hong, J. Liu, K. Rajawat, and P. Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34:6050–6061, 2021.

[30] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *arXiv preprint arXiv:2102.07367*, 2021.

[31] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[32] J. Larsen, L. K. Hansen, C. Svarer, and M. Ohlsson. Design and regularization of neural networks: the optimal use of a validation set. In *Neural Networks for Signal Processing VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop*, pages 62–71. IEEE, 1996.

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

[34] J. Li, B. Gu, and H. Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7426–7434, 2022.

[35] J. Li, J. Pei, and H. Huang. Communication-efficient robust federated learning with noisy labels. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 914–924, 2022.

[36] T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

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

[38] H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[39] J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.

[40] S. Lu, S. Zeng, X. Cui, M. Squillante, L. Horesh, B. Kingsbury, J. Liu, and M. Hong. A stochastic linearized augmented lagrangian method for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 35:30638–30650, 2022.

[41] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122, 2015.

[42] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[43] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. *arXiv preprint arXiv:1902.00146*, 2019.

[44] T. Okuno, A. Takeda, and A. Kawana. Hyperparameter learning via bilevel nonsmooth optimization. *arXiv preprint arXiv:1806.01520*, 2018.

[45] F. Pedregosa. Hyperparameter optimization with approximate gradient. *arXiv preprint arXiv:1602.02355*, 2016.

[46] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017.

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

[48] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3, 2018.

[49] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. *arXiv preprint arXiv:1810.10667*, 2018.

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

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

[52] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[53] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

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

[55] B. Woodworth. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021.

[56] P. Xiao and K. Ji. Communication-efficient federated hypergradient computation via aggregated iterative differentiation. *arXiv preprint arXiv:2302.04969*, 2023.

[57] P. Xing, S. Lu, L. Wu, and H. Yu. Big-fed: Bilevel optimization enhanced graph-aided federated learning.

[58] I. Yamada, M. Yukawa, and M. Yamagishi. Minimizing the moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 345–390. Springer, 2011.

[59] J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *arXiv preprint arXiv:2106.04692*, 2021.

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

[61] Y. Yang, P. Xiao, and K. Ji. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. *arXiv preprint arXiv:2305.19442*, 2023.

[62] H. Yu, S. Yang, and S. Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.

[63] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[64] L. Zintgraf, K. Shiarli, V. Kurin, K. Hofmann, and S. Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019.

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 μ-strongly convex *w.r.t* y for any given x, *i.e.* for any $y_1, y_2 \in \mathbb{R}^d$, we have:

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

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

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

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 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 for any $x_1, x_2 \in \mathcal{X}$ and for any $y_1, y_2 \in \mathbb{R}^d$, and we denote $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, then we have:

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

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 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 $var(\nabla f^{(m)}(z; \xi)) \leq \sigma^2$.

b) we have $\nabla g^{(m)}(z; \xi)$, such that: $E[\nabla g^{(m)}(z; \xi)] = \nabla g^{(m)}(z)$ and $var(\nabla g^{(m)}(z; \xi)) \leq \sigma^2$.

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 $var(\nabla_{y^2} g^{(m)}(z; \xi)) \leq \sigma^2$;

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

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

Assumption A.6 (Assumption 6)**.** For any $m, j \in [M]$ and $z = (x, y)$, we have: $\|\nabla f^{(m)}(z) - \nabla f^{(j)}(z)\| \leq \zeta_f$, $\|\nabla_{xy} g^{(m)}(z) - \nabla_{xy} g^{(j)}(z)\| \leq \zeta_{g,xy}$, $\|\nabla_{y^2} g^{(m)}(z) - \nabla_{y^2} g^{(j)}(z)\| \leq \zeta_{g,yy}$, $\|y_x^{(m)} - y_x^{(j)}\| \leq \zeta_{g^*}$, where ζ_f, $\zeta_{g,xy}$, $\zeta_{g,yy}$, ζ_{g^*} 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:

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

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 $x_{m,n}, n \in [N_m^{(tr)}]$ are weights for training samples, y is the parameter of a model, and we denote the model by Θ. 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 1000, for FedBiOAcc, we choose δ as 30, u as 10000, c_η as 0.2, C_γ as 0.2, τ as 0.01, η as 200 and γ as 1.

B.2 Federated Hyper-Representation Learning

$$\min_{x\in\mathbb{R}^p} h(x) \coloneqq \frac{1}{M}\sum_{m=1}^{M} f^{(m)}(x, y_x^{(m)}) = \frac{1}{M}\sum_{m=1}^{M}\big(\frac{1}{N_m}\sum_{n=1}^{N_m}\big(\frac{1}{N_{m,n}^{val}}\sum_{i=1}^{N_{m,n}^{val}}\Theta(x, y_x^{(\mathcal{T}_{m,n})};\xi_i^{val})\big)\big)$$
$$\text{s.t. } y_x^{(\mathcal{T}_{m,n})} = \arg\min_{y\in\mathbb{R}^d} g^{(\mathcal{T}_{m,n})}(x,y) = \frac{1}{N_{m,n}^{tr}}\sum_{i=1}^{N_{m,n}^{tr}}\Theta(x,y;\xi_i^{tr})$$

In the above formulation, we have M clients, each client $m \in [M]$ has N_m tasks and each task $\mathcal{T}_{m,n}$ is defined by a pair of training set $\{\xi_i^{tr}\}_{i=1}^{N_{m,n}^{tr}}$ and validation set $\{\xi_i^{val}\}_{i=1}^{N_{m,n}^{val}}$. Θ 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 3×3 [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 3×3 [11] for experiments.

In the Experiments for Omniglot, for FedBiO, we choose learning rate 0.4, hyper learning rate 1, τ 0.5, for FedBiOAcc, we choose δ as 2, u as 10000, C_η as 100, τ as 0.5, eta as 1 and γ as 0.4. For MiniImageNet, for FedBiO, we choose learning rate 0.05, hyper learning rate 0.1, τ 0.01, for FedBiOAcc, we choose δ as 2, u as 10000, C_η as 100, τ as 0.01, eta as 1 and γ as 0.05.

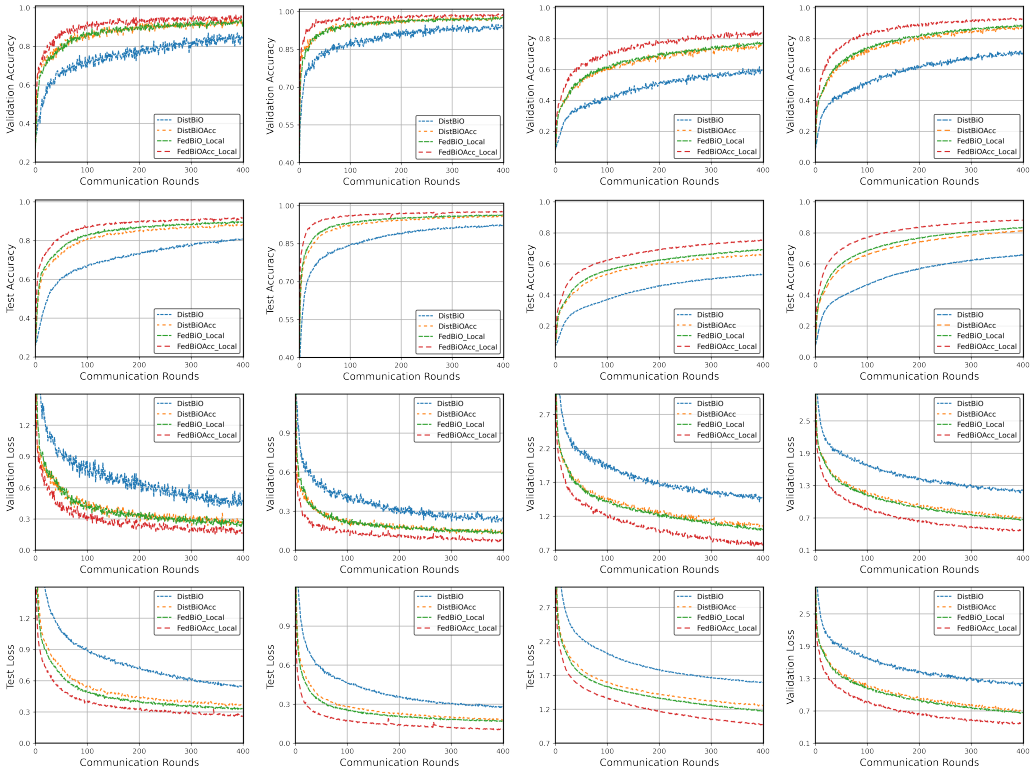

Figure 5: Results for the Omniglot Dataset. From Left to Right: 5-way-1-shot, 5-way-5-shot, 20-way-1-shot, 20-way-5-shot.

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* $g(x, y)$*.*

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* $\bar{L}$ *i.e., for any given* $x_1, x_2 \in X$*, we have* $\|\nabla h(x_2) - \nabla h(x_1)\| \leq \bar{L}\|x_2 - x_1\|$ *where* $\bar{L} = O(\kappa^3)$*.*

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

$$\mu_{t,\xi}^{(m)} = \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{f,1}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{g,1}) u_t^{(m)},$$

and we have:

$$\mathbb{E}_\xi[\mu_{t,\xi}^{(m)}] = \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}) u_t^{(m)}$$

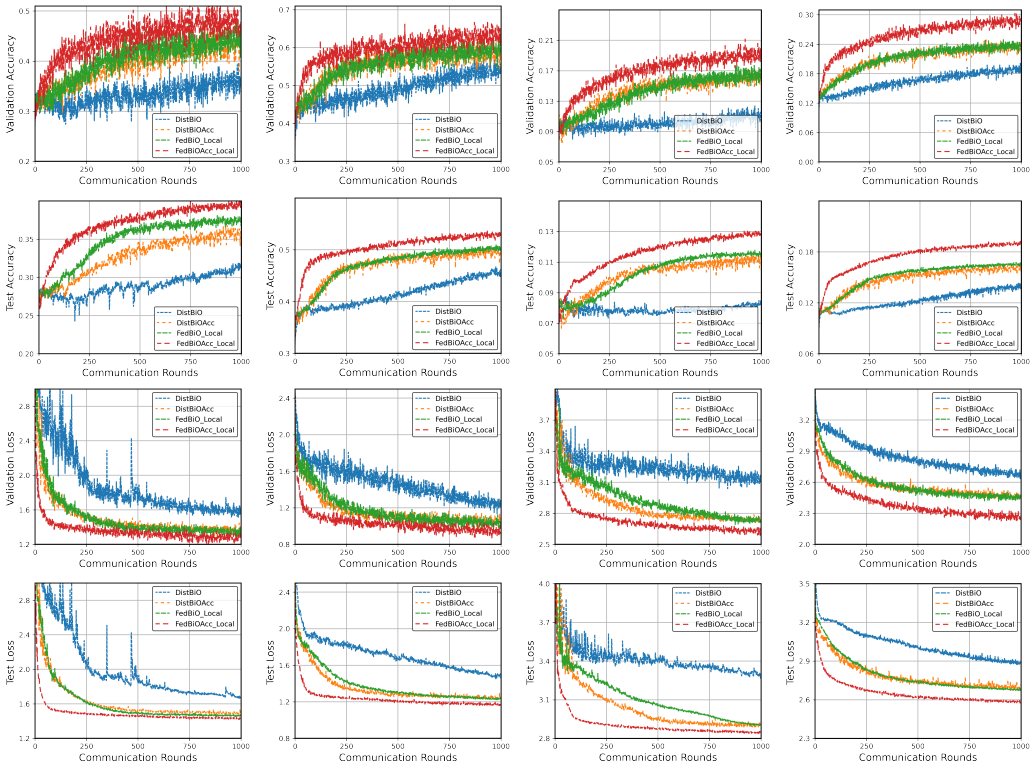

Figure 6: Results for the MiniImageNet Dataset. From Left to Right: 5-way-1-shot, 5-way-5-shot, 20-way-1-shot, 20-way-5-shot.

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

$$p_{t,\xi}^{(m)} = \nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{g,2}) u_t^{(m)} + \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{f,2}),$$

and we have:

$$\mathbb{E}_\xi[p_{t,\xi}^{(m)}] = \nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}) u_t^{(m)} + \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}).$$

where the expectation is *w.r.t* $\{\xi_{f,2}, \xi_{g,2}\}$ at iteration t, we denote $p_t^{(m)} = \mathbb{E}_\xi[p_{t,\xi}^{(m)}]$ 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:*

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{q}_t - \bar{p}_t\right\|^2\right] \leq (1 - c_u\alpha_{t-1}^2)\mathbb{E}\left[\left\|\bar{q}_{t-1} - \bar{p}_{t-1}\right\|^2\right] + \frac{2(c_u\alpha_{t-1}^2)^2}{b_x M}\sigma^2 \\
+ \frac{4\tilde{L}_2^2}{b_x M^2}\sum_{m=1}^{M}\mathbb{E}\left[\left\|x_t^{(m)} - x_{t-1}^{(m)}\right\|^2 + \left\|y_t^{(m)} - y_{t-1}^{(m)}\right\|^2\right] \\
+ \frac{8L^2}{b_x M^2}\sum_{m=1}^{M}\mathbb{E}\left[\left\|u_t^{(m)} - u_{t-1}^{(m)}\right\|^2\right]
\end{aligned}$$

where $\tilde{L}_2^2 = \left(L^2 + \frac{2L_{y^2}^2 C_f^2}{\mu^2}\right)$ *and the expectation outside is* w.r.t *all the stochasity of the algorithm.*

Proof. First, we have:

$$\begin{aligned}
&\mathbb{E}\big[\big\|\bar{q}_t - \bar{p}_t\big]\big\|^2\big] = \mathbb{E}\big[\big\|\bar{p}_{t,\mathcal{B}_x} + (1 - c_u\alpha_{t-1}^2)(\bar{q}_{t-1} - \bar{p}_{t-1,\mathcal{B}_x}) - \bar{p}_t\big\|^2\big] \\
&= \mathbb{E}\big[\big\|(1 - c_u\alpha_{t-1}^2)\big(\bar{q}_{t-1} - \bar{p}_{t-1}\big) + \big(\bar{p}_{t,\mathcal{B}_x} - \bar{p}_t + (1 - c_u\alpha_{t-1}^2)(\bar{p}_{t-1} - \bar{p}_{t-1,\mathcal{B}_x})\big)\big\|^2\big] \\
&\leq (1 - c_u\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{q}_{t-1} - \bar{p}_{t-1}\big\|^2\big] + \mathbb{E}\big[\big\|\bar{p}_{t,\mathcal{B}_x} - \bar{p}_t + (1 - c_u\alpha_{t-1}^2)(\bar{p}_{t-1} - \bar{p}_{t-1,\mathcal{B}_x})\big\|^2\big] \\
&\leq (1 - c_u\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{q}_{t-1} - \bar{p}_{t-1}\big\|^2\big] \\
&\quad + \frac{1}{b_x^2 M^2}\sum_{m=1}^{M}\sum_{\xi_x \in \mathcal{B}_x}\mathbb{E}\big[\big\|p_{t,\xi_x}^{(m)} - p_t^{(m)} + (1 - c_u\alpha_{t-1}^2)(p_{t-1}^{(m)} - p_{t-1,\xi_x}^{(m)})\big\|^2\big]
\end{aligned}$$

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

$$\begin{aligned}
T_1 &\overset{(a)}{\leq} 2(c_u\alpha_{t-1}^2)^2\mathbb{E}\big[\big\|p_{t,\xi_x}^{(m)} - p_t^{(m)}\big\|^2\big] + 2(1 - c_u\alpha_{t-1}^2)^2\mathbb{E}\big[\big\|p_{t,\xi_x}^{(m)} - p_{t-1,\xi_x}^{(m)} - (p_t^{(m)} - p_{t-1}^{(m)})\big\|^2\big] \\
&\overset{(b)}{\leq} 2(c_u\alpha_{t-1}^2)^2\sigma^2 + 2\mathbb{E}\big[\big\|p_{t,\xi_x}^{(m)} - p_{t-1,\xi_x}^{(m)}\big\|^2\big]
\end{aligned}$$

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:

$$\begin{aligned}
T_{1,2} &= 2\mathbb{E}\big\|\nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{g,2})u_t^{(m)} + \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{f,2}) \\
&\qquad - \big(\nabla_{y^2} g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \xi_{g,2})u_{t-1}^{(m)} + \nabla_y f^{(m)}(x_t^{(m)}, y_{t-1}^{(m)}; \xi_{f,2})\big)\big\|^2 \\
&\leq 4\mathbb{E}\big\|\nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,1}) - \nabla_y f^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_{f,1})\big\|^2 \\
&\qquad + 4\mathbb{E}\big\|\nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,1})u_t^{(m)} - \nabla_{y^2} g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_{g,1})u_{t-1}^{(m)}\big\|^2 \\
&\leq 4\big(L^2 + \frac{2L_{y^2}^2 C_f^2}{\mu^2}\big)\mathbb{E}\big[\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \big\|y_t^{(m)} - y_{t-1}^{(m)}\big\|^2\big] + 8L^2\mathbb{E}\big[\big\|u_t^{(m)} - u_{t-1}^{(m)}\big\|^2\big]
\end{aligned}$$

Combine everything together finishes the proof. □

Lemma C.3. *Suppose we have $c_\nu\alpha_t^2 < 1$, then we have:*

$$\begin{aligned}
\mathbb{E}\big[\big\|\bar{\nu}_t - \bar{\mu}_t\big\|^2\big] \leq (1 - c_\nu\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{\nu}_{t-1} - \bar{\mu}_{t-1}\big\|^2\big] + \frac{2(c_\nu\alpha_{t-1}^2)^2}{b_x M}\sigma^2 \\
+ \frac{4\tilde{L}_1^2}{b_x M^2}\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \big\|y_t^{(m)} - y_{t-1}^{(m)}\big\|^2\big] \\
+ \frac{8L^2}{b_x M^2}\sum_{m=1}^{M}\mathbb{E}\big[\big\|u_t^{(m)} - u_{t-1}^{(m)}\big\|^2\big]
\end{aligned}$$

where $\tilde{L}_1^2 = \big(L^2 + \frac{2L_{xy}^2 C_f^2}{\mu^2}\big)$ and the expectation outside is w.r.t *all the stochasity of the algorithm.*

Proof. First, we have:

$$\begin{aligned}
&\mathbb{E}\big[\big\|\bar{\nu}_t - \bar{\mu}_t\big]\big\|^2\big] = \mathbb{E}\big[\big\|\bar{\mu}_{t,\mathcal{B}_x} + (1 - c_\nu\alpha_{t-1}^2)(\bar{\nu}_{t-1} - \bar{\mu}_{t-1,\mathcal{B}_x}) - \bar{\mu}_t\big\|^2\big] \\
&= \mathbb{E}\big[\big\|(1 - c_\nu\alpha_{t-1}^2)\big(\bar{\nu}_{t-1} - \bar{\mu}_{t-1}\big) + \big(\bar{\mu}_{t,\mathcal{B}_x} - \bar{\mu}_t + (1 - c_\nu\alpha_{t-1}^2)(\bar{\mu}_{t-1} - \bar{\mu}_{t-1,\mathcal{B}_x})\big)\big\|^2\big] \\
&\leq (1 - c_\nu\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{\nu}_{t-1} - \bar{\mu}_{t-1}\big\|^2\big] + \mathbb{E}\big[\big\|\bar{\mu}_{t,\mathcal{B}_x} - \bar{\mu}_t + (1 - c_\nu\alpha_{t-1}^2)(\bar{\mu}_{t-1} - \bar{\mu}_{t-1,\mathcal{B}_x})\big\|^2\big] \\
&\leq (1 - c_\nu\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{\nu}_{t-1} - \bar{\mu}_{t-1}\big\|^2\big] + \frac{1}{b_x^2 M^2}\sum_{m=1}^{M}\sum_{\xi_x \in \mathcal{B}_x}\mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mu_t^{(m)} + (1 - c_\nu\alpha_{t-1}^2)(\mu_{t-1}^{(m)} - \mu_{t-1,\xi_x}^{(m)})\big\|^2\big]
\end{aligned}$$

where the first inequality uses the fact that the cross product term is zero in expectation, the condition that $c_\nu\alpha_t^2 < 1$ and the second inequality follows that samples are independent among clients. We

denote the second term of above as T_1, then we have:

$$T_1 \overset{(a)}{\leq} 2(c_\nu\alpha_{t-1}^2)^2\mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mu_t^{(m)}\big\|^2\big] + 2(1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mu_{t-1,\xi_x}^{(m)} - (\mu_t^{(m)} - \mu_{t-1}^{(m)})\big\|^2\big]$$
$$\overset{(b)}{\leq} 2(c_\nu\alpha_{t-1}^2)^2\sigma^2 + 2\mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mu_{t-1,\xi_x}^{(m)}\big\|^2\big]$$

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:

$$\begin{aligned}
T_{1,2} &= 2\mathbb{E}\big\|\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,1})u_t^{(m)} \\
&\quad - \big(\nabla_x f^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_{g,1})u_{t-1}^{(m)}\big)\big\|^2 \\
&\leq 4\mathbb{E}\big\|\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,1}) - \nabla_x f^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_{f,1})\big\|^2 \\
&\quad + 4\mathbb{E}\big\|\nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,1})u_t^{(m)} - \nabla_{xy} g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \mathcal{B}_{g,1})u_{t-1}^{(m)}\big\|^2 \\
&\leq 4\big(L^2 + \frac{2L_{xy}^2 C_f^2}{\mu^2}\big)\mathbb{E}\big[\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \big\|y_t^{(m)} - y_{t-1}^{(m)}\big\|^2\big] + 8L^2\mathbb{E}\big[\big\|u_t^{(m)} - u_{t-1}^{(m)}\big\|^2\big]
\end{aligned}$$

Combine everything together finishes the proof. □

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

$$\begin{aligned}
&\mathbb{E}\big[\big\|\bar{\omega}_t - \frac{1}{M}\sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\big\|^2\big] \\
&\leq (1-c_\omega\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{\omega}_{t-1} - \frac{1}{M}\sum_{m=1}^M \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\big\|^2\big] + \frac{2(c_\omega\alpha_{t-1}^2)^2\sigma^2}{b_y M} \\
&\quad + \frac{2L^2}{b_y M^2}\sum_{m=1}^M \mathbb{E}\big[\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \big\|y_t^{(m)} - y_{t-1}^{(m)}\big\|^2\big]
\end{aligned}$$

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

Proof. First, we have:

$$\begin{aligned}
&\mathbb{E}\big[\big\|\bar{\omega}_t - \frac{1}{M}\sum_{m=1}^M \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\big\|^2\big] \\
&= \mathbb{E}\big[\big\|\frac{1}{M}\sum_{m=1}^M \big(\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) \\
&\quad + (1-c_\omega\alpha_{t-1}^2)(\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y)) - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\big)\big\|^2\big] \\
&= \mathbb{E}\big[\big\|(1-c_\omega\alpha_{t-1}^2)(\bar{\omega}_{t-1} - \frac{1}{M}\sum_{m=1}^M \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) \\
&\quad + \frac{1}{M}\sum_{m=1}^M \big(\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \\
&\quad\quad + (1-c_\omega\alpha_{t-1}^2)(\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y))\big)\big\|^2\big] \\
&\overset{(a)}{\leq} (1-c_\omega\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{\omega}_{t-1} - \frac{1}{M}\sum_{m=1}^M \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\big\|^2\big] \\
&\quad + \frac{1}{b_y^2 M^2}\sum_{m=1}^M \mathbb{E}\sum_{\xi_y \in \mathcal{B}_y}\big[\big\|\big(\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \xi_y) - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \\
&\quad + (1-c_\omega\alpha_{t-1}^2)(\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \xi_y))\big)\big\|^2\big]
\end{aligned}$$

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

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

$$\begin{aligned}
T_1 &\overset{(b)}{\leq} 2(c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\big[\big\|\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \xi_y) - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\big\|^2\big] \\
&\quad + 2(1 - c_\omega \alpha_{t-1}^2)^2 \mathbb{E}\big[\big\| - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) \\
&\quad + \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \xi_y) + \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \xi_y)\big\|^2\big] \\
&\overset{(c)}{\leq} 2(c_\omega \alpha_{t-1}^2)^2 \sigma^2 + 2\mathbb{E}\big[\big\|\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \xi_y) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \xi_y)\big\|^2\big] \\
&\overset{(d)}{\leq} 2(c_\omega \alpha_{t-1}^2)^2 \sigma^2 + 2L^2 \mathbb{E}\big[\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \big\|y_t^{(m)} - y_{t-1}^{(m)}\big\|^2\big]
\end{aligned}$$

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

$$\begin{aligned}
\|\bar{y}_{t+1} - y_{\bar{x}_{t+1}}\|^2 \leq \big(1 - \frac{\mu\gamma\alpha_t}{4}\big)\|\bar{y}_t - y_{\bar{x}_t}\|^2 - \frac{\gamma^2\alpha_t}{4}\|\bar{\omega}_t\|^2 + \frac{9\kappa^2\eta^2\alpha_t}{2\mu\gamma}\|\bar{\nu}_t\|^2 \\
+ \frac{9\gamma\alpha_t L^2}{\mu M}\sum_{m=1}^{M}\big[\|x_t^{(m)} - \bar{x}_t\|^2 + \|y_t^{(m)} - \bar{y}_t\|^2\big] + \frac{9\gamma\alpha_t}{\mu}\|\frac{1}{M}\sum_{m=1}^{M}\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}) - \bar{w}_t\|^2
\end{aligned}$$

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

$$\|\bar{y}_{t+1} - y_{\bar{x}_t}\|^2 \leq (1 - \frac{\mu\gamma\alpha_t}{2})\|\bar{y}_t - y_{\bar{x}_t}\|^2 - \frac{\gamma^2\alpha_t}{4}\|\bar{\omega}_t\|^2 + \frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(\bar{x}_t, \bar{y}_t) - \bar{w}_t\|^2. \tag{12}$$

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

$$\begin{aligned}
\|\bar{y}_{t+1} - y_{\bar{x}_{t+1}}\|^2 &\leq (1 + \frac{\mu\gamma\alpha_t}{4})\|\bar{y}_{t+1} - y_{\bar{x}_t}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\|y_{\bar{x}_t} - y_{\bar{x}_{t+1}}\|^2 \\
&\leq (1 + \frac{\mu\gamma\alpha_t}{4})\|\bar{y}_{t+1} - y_{\bar{x}_t}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\kappa^2\|\bar{x}_t - \bar{x}_{t+1}\|^2
\end{aligned} \tag{13}$$

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

$$\begin{aligned}
\|\bar{y}_{t+1} - y_{\bar{x}_{t+1}}\|^2 \leq (1 + \frac{\mu\gamma\alpha_t}{4})(1 - \frac{\mu\gamma\alpha_t}{2})\|\bar{y}_t - y_{\bar{x}_t}\|^2 - (1 + \frac{\mu\gamma\alpha_t}{4})\frac{\gamma^2\alpha_t}{4}\|\bar{\omega}_t\|^2 \\
+ (1 + \frac{\mu\gamma\alpha_t}{4})\frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(\bar{x}_t, \bar{y}_t) - \bar{w}_t\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\kappa^2\eta^2\alpha_t^2\|\bar{\nu}_t\|^2
\end{aligned}$$

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

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

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

$$\|\bar{y}_{t+1} - y_{\bar{x}_{t+1}}\|^2 \leq \big(1 - \frac{\mu\gamma\alpha_t}{4}\big)\|\bar{y}_t - y_{\bar{x}_t}\|^2 - \frac{\gamma^2\alpha_t}{4}\|\bar{\omega}_t\|^2 + \frac{9\gamma\alpha_t}{2\mu}\underbrace{\|\nabla_y g(\bar{x}_t, \bar{y}_t) - \bar{w}_t\|^2}_{T_1} + \frac{9\kappa^2\eta^2\alpha_t}{2\mu\gamma}\|\bar{\nu}_t\|^2$$

For the term T_1 in the inequality above, we have:

$$\begin{aligned}\|\nabla_y g(\bar{x}_t,\bar{y}_t)-\bar{w}_t\|^2 &\le 2\|\nabla_y g(\bar{x}_t,\bar{y}_t)-\frac{1}{M}\sum_{m=1}^{M}\nabla_y g^{(m)}(x_t^{(m)},y_t^{(m)})\|^2\\ &\quad+2\|\frac{1}{M}\sum_{m=1}^{M}\nabla_y g^{(m)}(x_t^{(m)},y_t^{(m)})-\bar{w}_t\|^2\\ &\le \frac{2L^2}{M}\sum_{m=1}^{M}\big[\|x_t^{(m)}-\bar{x}_t\|^2+\|y_t^{(m)}-\bar{y}_t\|^2\big]\\ &\quad+2\|\frac{1}{M}\sum_{m=1}^{M}\nabla_y g^{(m)}(x_t^{(m)},y_t^{(m)})-\bar{w}_t\|^2\end{aligned}$$

This completes the proof. □

Lemma C.6. *Suppose we choose $\tau \le \frac{1}{2L}$ and $\alpha_t < 1$, $r = \frac{C_f}{\mu}$. Then for $t \in [T]$, we have:*

$$\begin{aligned}\|\bar{u}_{t+1}-u_{\bar{x}_{t+1}}\|^2 &\le \big(1-\frac{\mu\tau\alpha_t}{4}\big)\|\bar{u}_t-u_{\bar{x}_t}\|^2-\frac{\tau^2\alpha_t}{4}\|\bar{q}_t\|^2+\frac{9\kappa^2\eta^2\alpha_t}{2\mu\tau}\|\bar{\nu}_t\|^2+\frac{9\tau\alpha_t}{\mu}\|\bar{p}-\bar{q}_t\|^2\\ &\quad+\frac{18\tau\alpha_t\tilde{L}_2^2}{\mu M}\sum_{m=1}^{M}\big[\|x_t^{(m)}-\bar{x}_t\|^2+\|y_t^{(m)}-\bar{y}_t\|^2\big]+\frac{18\tau\alpha_t L^2}{M}\sum_{m=1}^{M}\big\|u_t^{(m)}-\bar{u}_t\big\|^2\end{aligned}$$

where $\tilde{L}_2^2 = (L^2+\frac{2L_{y^2}^2C_f^2}{\mu^2})$ is a constant.

Proof. First, we exploit Proposition E.5, and choose the function $\frac{1}{2}x^T\nabla_{y^2}g(\bar{x},y_{\bar{x}})x-\nabla_y f(\bar{x},y_{\bar{x}})^T x$, by assumption it is L smooth and μ strongly convex, and we choose $\tau < \frac{1}{2L}$ and $\alpha_t < 1$, thus:

$$\|\bar{u}_{t+1}-u_{\bar{x}_t}\|^2 \le (1-\frac{\mu\tau\alpha_t}{2})\|\bar{u}_t-u_{\bar{x}_t}\|^2-\frac{\tau^2\alpha_t}{4}\|\bar{q}_t\|^2+\frac{4\tau\alpha_t}{\mu}\|\nabla_{y^2}g(\bar{x},y_{\bar{x}})\bar{u}_t-\nabla_y f(\bar{x},y_{\bar{x}})-\bar{q}_t\|^2.$$

where we also use the fact that

$$\|\bar{u}_{t+1}-u_{\bar{x}_t}\|^2 \le \|\bar{u}_t-\tau\alpha_t\bar{q}_t-u_{\bar{x}_t}\|^2$$

for $r = \frac{C_f}{\mu} \ge \|u_{\bar{x}_t}\|$. Next, we decompose the term $\|\bar{u}_{t+1}-u_{\bar{x}_{t+1}}\|^2$ as follows:

$$\begin{aligned}\|\bar{u}_{t+1}-u_{\bar{x}_{t+1}}\|^2 &\le (1+\frac{\mu\tau\alpha_t}{4})\|\bar{u}_{t+1}-u_{\bar{x}_t}\|^2+(1+\frac{4}{\mu\tau\alpha_t})\|u_{\bar{x}_t}-u_{\bar{x}_{t+1}}\|^2\\ &\le (1+\frac{\mu\tau\alpha_t}{4})\|\bar{u}_{t+1}-u_{\bar{x}_t}\|^2+(1+\frac{4}{\mu\tau\alpha_t})\bar{L}^2\|\bar{x}_t-\bar{x}_{t+1}\|^2\end{aligned}$$

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

$$\begin{aligned}\|\bar{u}_{t+1}-u_{\bar{x}_{t+1}}\|^2 &\le (1-\frac{\mu\tau\alpha_t}{4})\|\bar{u}_t-u_{\bar{x}_t}\|^2-\frac{\tau^2\alpha_t}{4}\|\bar{q}_t\|^2\\ &\quad+\frac{9\tau\alpha_t}{2\mu}\underbrace{\|\nabla_{y^2}g(\bar{x},y_{\bar{x}})\bar{u}_t-\nabla_y f(\bar{x},y_{\bar{x}})-\bar{q}_t\|^2}_{T_1}+\frac{9\bar{L}^2\eta^2\alpha_t}{2\mu\tau}\|\bar{\nu}_t\|^2\end{aligned}$$

where we use the fact that $\tau \le \frac{1}{2L}$, $\alpha_t < 1$ For the term T_1 in the inequality above, we have:

$$\begin{aligned}T_1 &\le 2\|\nabla_{y^2}g(\bar{x},y_{\bar{x}})\bar{u}_t-\nabla_y f(\bar{x},y_{\bar{x}})-\bar{p}_t\|^2+2\|\bar{p}_t-\bar{q}_t\|^2\\ &\le 2\big\|\nabla_{y^2}g(\bar{x},y_{\bar{x}})\bar{u}_t-\nabla_y f(\bar{x},y_{\bar{x}})\\ &\quad-\frac{1}{M}\sum_{m=1}^{M}\big(\nabla_{y^2}g^{(m)}(x_t^{(m)},y_t^{(m)})u_t^{(m)}+\nabla_y f^{(m)}(x_t^{(m)},y_t^{(m)})\big)\big\|^2+2\|\bar{p}_t-\bar{q}_t\|^2\end{aligned}$$

We denote the first term of the above inequality as $T_{1,1}$, we have:

$$\begin{aligned}
T_{1,1} &\leq 4\big\|\nabla_y f(\bar{x}, y_{\bar{x}}) - \frac{1}{M}\sum_{m=1}^{M}\big(\nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)})\big)\big\|^2 \\
&\quad + 4\big\|\nabla_{y^2} g(\bar{x}, y_{\bar{x}})\bar{u}_t - \frac{1}{M}\sum_{m=1}^{M}\big(\nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)})u_t^{(m)}\big)\big\|^2 \\
&\leq \big(\frac{4L^2}{M} + \frac{8L_{y^2}^2 C_f^2}{\mu^2 M}\big)\sum_{m=1}^{M}\big[\big\|x_t^{(m)} - \bar{x}_t\big\|^2 + \big\|y_t^{(m)} - \bar{u}_t\big\|^2\big] + \frac{4L^2}{M}\sum_{m=1}^{M}\big\|u_t^{(m)} - \bar{u}_t\big\|^2
\end{aligned}$$

Combine everything completes the proof. □

C.1.3 Upper Variable Drift

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

$$\begin{aligned}
\|x_t^{(m)} - \bar{x}_t\|^2 &\leq I\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 \big\|\nu_\ell^{(m)} - \bar{\nu}_\ell\big\|^2 \\
\|y_t^{(m)} - \bar{y}_t\|^2 &\leq I\gamma^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 \big\|\omega_\ell^{(m)} - \bar{\omega}_\ell\big\|^2 \\
\|u_t^{(m)} - \bar{u}_t\|^2 &\leq I\tau^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 \big\|q_\ell^{(m)} - \bar{q}_\ell\big\|^2
\end{aligned}$$

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 $\bar{x}_t = \bar{x}_{\bar{t}_{s-1}} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\alpha_\ell\bar{\nu}_\ell$. So for $t \neq \bar{t}_s$, with $s \in [S]$ we have:

$$\begin{aligned}
\|x_t^{(m)} - \bar{x}_t\|^2 &= \big\|x_{\bar{t}_{s-1}}^{(m)} - \bar{x}_{\bar{t}_{s-1}} - \big(\sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\alpha_\ell\nu_\ell^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\alpha_\ell\bar{\nu}_\ell\big)\big\|^2 = \big\|\sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\alpha_\ell\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\big\|^2 \\
&\leq I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2\big\|\nu_\ell^{(m)} - \bar{\nu}_\ell\big\|^2
\end{aligned}$$

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

Lemma C.8. *Suppose $\eta\alpha_t < \frac{1}{16I\tilde{L}_1}$, then for $t \neq \bar{t}_s, s \in [S]$, we have:*

$$\begin{aligned}
&\sum_{m=1}^{M}\mathbb{E}\|\hat{\nu}_t^{(m)} - \bar{\nu}_t\|^2 \\
&\leq \left(1 + \frac{17}{16I}\right)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 + 8I\tilde{L}_1^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[2\|\eta\bar{\nu}_{t-1}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\big] + 16IL^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\|\tau q_{t-1}^{(m)}\|^2 \\
&\quad + 128I(c_\nu\alpha_{t-1}^2)^2\tilde{L}_1^2\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)} - \bar{x}_t\big\|^2 + \big\|y_t^{(m)} - \bar{y}_t\big\|^2\big] + 32I(c_\nu\alpha_{t-1}^2)^2L^2\sum_{m=1}^{M}\mathbb{E}\big\|u_t^{(m)} - \bar{u}_t\big\|^2 \\
&\quad + 8IM(c_\nu\alpha_{t-1}^2)^2\frac{\sigma^2}{b_x} + 32IM(c_\nu\alpha_{t-1}^2)^2\zeta_f^2 + 64I(c_\nu\alpha_{t-1}^2)^2M\frac{C_f^2\zeta_{g,xy}^2}{\mu^2}
\end{aligned}$$

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

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

$$\begin{aligned}
\mathbb{E}\|\hat{\nu}_t^{(m)} - \bar{\nu}_t\|^2 &= \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} + (1-c_\nu\alpha_{t-1}^2)\big(\nu_{t-1}^{(m)} - \mu_{t-1,\mathcal{B}_x}^{(m)}\big) - \big(\bar{\mu}_{t,\mathcal{B}_x} + (1-c_\nu\alpha_{t-1}^2)\big(\bar{\nu}_{t-1} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big)\big\|^2 \\
&= \mathbb{E}\big\|(1-c_\nu\alpha_{t-1}^2)\big(\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\big) + \mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 \\
&\overset{(a)}{\leq} (1+\frac{1}{I})(1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 \\
&\quad + \big(1+I\big)\mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 \\
&\leq \left(1+\frac{1}{I}\right)\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 + \big(1+I\big)\mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2
\end{aligned} \tag{14}$$

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

Next we bound the second term of the above inequality:

$$\begin{aligned}
&\sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 \\
&\leq 2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - \big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 + 2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big\|^2
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - \big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 \overset{(a)}{\leq} \sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \mu_{t-1,\mathcal{B}_x}^{(m)}\big\|^2 \\
&\leq \sum_{m=1}^M \mathbb{E}\big\|\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{f,1}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_{g,1})u_t^{(m)} \\
&\quad - \big(\nabla_x f^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \xi_{f,1}) - \nabla_{xy} g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}; \xi_{g,1})u_{t-1}^{(m)}\big)\big\|^2 \\
&\overset{(b)}{\leq} 2\big(L^2 + \frac{2L_{xy}^2C_f^2}{\mu^2}\big)\sum_{m=1}^M \mathbb{E}\big[\|x_t^{(m)} - x_{t-1}^{(m)}\|^2 + \|y_t^{(m)} - y_{t-1}^{(m)}\|^2\big] + 4L^2\sum_{m=1}^M \mathbb{E}\|u_t^{(m)} - u_{t-1}^{(m)}\|^2 \\
&\leq 2\tilde{L}_1^2\alpha_{t-1}^2\sum_{m=1}^M \mathbb{E}\big[\|\eta\nu_{t-1}^{(m)}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\big] + 4L^2\alpha_{t-1}^2\sum_{m=1}^M \mathbb{E}\|\tau q_{t-1}^{(m)}\|^2
\end{aligned} \tag{15}$$

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

$$\begin{aligned}
&\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big\|^2 = \sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \mu_{t-1}^{(m)} - \big(\bar{\mu}_{t-1,\mathcal{B}_x} - \bar{\mu}_{t-1}\big) + \mu_{t-1}^{(m)} - \bar{\mu}_{t-1}\big\|^2 \\
&\overset{(a)}{\leq} 2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \mu_{t-1}^{(m)} - \big(\bar{\mu}_{t-1,\mathcal{B}_x} - \bar{\mu}_{t-1}\big)\big\|^2 + 2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1}^{(m)} - \bar{\mu}_{t-1}\big\|^2 \\
&\overset{(b)}{\leq} \underbrace{2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \mu_{t-1}^{(m)}\big\|^2}_{T_1} + \underbrace{2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1}^{(m)} - \bar{\mu}_{t-1}\big\|^2}_{T_2}
\end{aligned} \tag{16}$$

Note for the term T_1 of Eq. 16, we have $\mathbb{E}\big\|\mu^{(m)}_{t-1,\mathcal{B}_x} - \mu^{(m)}_{t-1}\big\|^2 \leq \frac{\sigma^2}{b_x}$ by the bounded variance assumption; Next for the term T_2, we have:

$$
\begin{aligned}
T_2 = &\sum_{m=1}^{M} \big\|\nabla_x f^{(m)}(x^{(m)}_{t-1}, y^{(m)}_{t-1}) - \nabla_{xy} g^{(m)}(x^{(m)}_{t-1}, y^{(m)}_{t-1}) u^{(m)}_{t-1} \\
&- \frac{1}{M}\sum_{j=1}^{M} \big(\nabla_x f^{(j)}(x^{(j)}_{t-1}, y^{(j)}_{t-1}) - \nabla_{xy} g^{(j)}(x^{(j)}_{t-1}, y^{(j)}_{t-1}) u^{(j)}_{t-1}\big)\big\|^2 \\
\leq\ & 16\big(L^2 + \frac{2L^2_{xy}C_f^2}{\mu^2}\big)\sum_{m=1}^{M}\mathbb{E}\big[\big\|x^{(m)}_t - \bar{x}_t\big\|^2 + \big\|y^{(m)}_t - \bar{y}_t\big\|^2\big] + 4L^2\sum_{m=1}^{M}\mathbb{E}\big\|u^{(m)}_t - \bar{u}_t\big\|^2 + 4M\zeta_f^2 + \frac{8MC_f^2\zeta^2_{g,xy}}{\mu^2}
\end{aligned}
$$

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

$$
\begin{aligned}
&\sum_{m=1}^{M}\mathbb{E}\|\hat{\nu}^{(m)}_t - \bar{\nu}_t\|^2 \\
&\leq \big(1 + \frac{1}{I}\big)\sum_{m=1}^{M}\mathbb{E}\|\nu^{(m)}_{t-1} - \bar{\nu}_{t-1}\|^2 + 8I\tilde{L}_1^2\alpha^2_{t-1}\sum_{m=1}^{M}\mathbb{E}\big[\underbrace{\|\eta\nu^{(m)}_{t-1}\|^2}_{T_1} + \|\gamma\omega^{(m)}_{t-1}\|^2\big] + 16IL^2\alpha^2_{t-1}\sum_{m=1}^{M}\mathbb{E}\|\tau q^{(m)}_{t-1}\|^2 \\
&\quad + 128I(c_\nu\alpha^2_{t-1})^2\tilde{L}_1^2\sum_{m=1}^{M}\mathbb{E}\big[\big\|x^{(m)}_t - \bar{x}_t\big\|^2 + \big\|y^{(m)}_t - \bar{y}_t\big\|^2\big] + 32I(c_\nu\alpha^2_{t-1})^2L^2\sum_{m=1}^{M}\mathbb{E}\big\|u^{(m)}_t - \bar{u}_t\big\|^2 \\
&\quad + 8IM(c_\nu\alpha^2_{t-1})^2\frac{\sigma^2}{b_x} + 32IM(c_\nu\alpha^2_{t-1})^2\zeta_f^2 + 64I(c_\nu\alpha^2_{t-1})^2M\frac{C_f^2\zeta^2_{g,xy}}{\mu^2}
\end{aligned}
$$

We separate the term T_1 with triangle inequality to get:

$$
\begin{aligned}
&\sum_{m=1}^{M}\mathbb{E}\|\hat{\nu}^{(m)}_t - \bar{\nu}_t\|^2 \\
&\leq \left(1 + \frac{1}{I} + 16I\tilde{L}_1^2\eta^2\alpha^2_{t-1}\right)\sum_{m=1}^{M}\mathbb{E}\|\nu^{(m)}_{t-1} - \bar{\nu}_{t-1}\|^2 \\
&\quad + 8I\tilde{L}_1^2\alpha^2_{t-1}\sum_{m=1}^{M}\mathbb{E}\big[2\|\eta\bar{\nu}_{t-1}\|^2 + \|\gamma\omega^{(m)}_{t-1}\|^2\big] + 16IL^2\alpha^2_{t-1}\sum_{m=1}^{M}\mathbb{E}\|\tau q^{(m)}_{t-1}\|^2 \\
&\quad + 128I(c_\nu\alpha^2_{t-1})^2\tilde{L}_1^2\sum_{m=1}^{M}\mathbb{E}\big[\big\|x^{(m)}_t - \bar{x}_t\big\|^2 + \big\|y^{(m)}_t - \bar{y}_t\big\|^2\big] + 32I(c_\nu\alpha^2_{t-1})^2L^2\sum_{m=1}^{M}\mathbb{E}\big\|u^{(m)}_t - \bar{u}_t\big\|^2 \\
&\quad + 8IM(c_\nu\alpha^2_{t-1})^2\frac{\sigma^2}{b_x} + 32IM(c_\nu\alpha^2_{t-1})^2\zeta_f^2 + 64I(c_\nu\alpha^2_{t-1})^2M\frac{C_f^2\zeta^2_{g,xy}}{\mu^2}
\end{aligned}
$$

This completes the proof. □

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

$$
\begin{aligned}
\sum_{m=1}^{M}\mathbb{E}\|\omega^{(m)}_t - \bar{\omega}_t\|^2 \leq &\left(1 + \frac{33}{32I}\right)\sum_{m=1}^{M}\mathbb{E}\|\omega^{(m)}_{t-1} - \bar{\omega}_{t-1}\|^2 + 4IL^2\alpha^2_{t-1}\sum_{m=1}^{M}\mathbb{E}\big[2\|\gamma\bar{\omega}_{t-1}\|^2 + \|\eta\nu^{(m)}_{t-1}\|^2\big] \\
&+ 8IM(c_\omega\alpha^2_{t-1})^2\frac{\sigma^2}{b_y} + 16IM(c_\omega\alpha^2_{t-1})^2\zeta_g^2 + 16IL^2(c_\omega\alpha^2_{t-1})^2\sum_{m=1}^{M}\mathbb{E}\big[\|x^{(m)}_{t-1} - \bar{x}_{t-1}\|^2\big] \\
&+ 16IL^2(c_\omega\alpha^2_{t-1})^2\sum_{m=1}^{M}\mathbb{E}\big[\|y^{(m)}_{t-1} - \bar{y}_{t-1}\|^2\big]
\end{aligned}
$$

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

Proof. By the update step in Line 7 of Algorithm 1, for $t \neq \bar{t}_s$, we have:

$$\begin{aligned}\mathbb{E}\|\hat{\omega}_t^{(m)} - \bar{\omega}_t\|^2 = \mathbb{E}\big\|&(1-c_\omega\alpha_{t-1}^2)\big(\omega_{t-1}^{(m)} - \bar{\omega}_{t-1}\big) + \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(x_t^{(j)}, y_t^{(j)}, \mathcal{B}_y) \\ &- (1-c_\omega\alpha_{t-1}^2)\big(\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, \mathcal{B}_y)\big)\big\|^2 \\ \leq\ & (1+\frac{1}{I})(1-c_\omega\alpha_{t-1}^2)^2\mathbb{E}\|\omega_{t-1}^{(m)} - \bar{\omega}_{t-1}\|^2 \\ &+ (1+I)\mathbb{E}\big\|\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(x_t^{(j)}, y_t^{(j)}, \mathcal{B}_y) \\ &- (1-c_\omega\alpha_{t-1}^2)\big(\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, \mathcal{B}_y)\big)\big\|^2\end{aligned} \tag{17}$$

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

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

$$\begin{aligned}T_1 \leq\ & 2\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(m)}(x_t^{(j)}, y_t^{(j)}, \mathcal{B}_y) \\ &- \big(\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(m)}, \mathcal{B}_y)\big)\big\|^2 \\ &+ 2(c_\omega\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, \mathcal{B}_y)\big\|^2\end{aligned}$$

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

$$\begin{aligned}T_{1,1} &\overset{(a)}{\leq} \sum_{m=1}^{M}\mathbb{E}\big\|\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y)\big\|^2 \\ &\overset{(b)}{\leq} L^2\sum_{m=1}^{M}\mathbb{E}\big[\|x_t^{(m)} - x_{t-1}^{(m)}\|^2 + \|y_t^{(m)} - y_{t-1}^{(m)}\|^2\big] \leq L^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[\|\eta\nu_{t-1}^{(m)}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\big]\end{aligned} \tag{18}$$

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

$$
\begin{aligned}
T_{1,2} = \sum_{m=1}^{M} \mathbb{E}\big\| & \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) \\
& - \frac{1}{M}\sum_{j=1}^{M}\big(\nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)}, \mathcal{B}_y) - \nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)})\big) \\
& + \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)})\big\|^2 \\
\overset{(b)}{\leq} 2\sum_{m=1}^{M}\mathbb{E}\big\| & \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}, \mathcal{B}_y) - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\big\|^2 \\
& + 4\sum_{m=1}^{M}\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}\|\nabla g^{(m)}(\bar{x}_{t-1}, \bar{y}_{t-1}) - \nabla_y g^{(j)}(\bar{x}_{t-1}, \bar{y}_{t-1})\|^2 \\
& + 4\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(\bar{x}_{t-1}, \bar{y}_{t-1}) \\
& + \frac{1}{M}\sum_{j=1}^{M}\nabla_y g^{(j)}(\bar{x}_{t-1}, \bar{y}_{t-1}) - \nabla_y g^{(j)}(x_{t-1}^{(j)}, y_{t-1}^{(j)})\big\|^2 \qquad (19)
\end{aligned}
$$

We denote the three terms above as $T_{1,2,1} - T_{1,2,3}$ respectively. For the term $T_{1,2,1}$ of Eq. 19, we have $T_{1,2,1} \leq 2M\sigma^2/b_y$ by the bounded variance assumption; For the term $T_{1,2,2}$ of Eq. 19, by the bounded intra-node heterogeneity assumption we have $T_{1,2,2} \leq 4M\zeta_g^2$. Finally, For the term $T_{1,2,3}$ of Eq. 19:

$$
\begin{aligned}
T_{1,2,3} &\leq 4\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)}) - \nabla_y g^{(m)}(\bar{x}_{t-1}, \bar{y}_{t-1})\big\|^2 \\
&\leq 4L^2\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2\big] + 4L^2\sum_{m=1}^{M}\mathbb{E}\big[\|y_{t-1}^{(m)} - \bar{y}_{t-1}\|^2\big]
\end{aligned}
$$

Finally, combine Eq. 17, Eq. 18 with Eq. 19 and use the fact that $I \geq 1$, we have:

$$
\begin{aligned}
\sum_{m=1}^{M}\mathbb{E}\|\hat{\omega}_t^{(m)} - \bar{\omega}_t\|^2 \leq \big(1 + \frac{1}{I}\big)\sum_{m=1}^{M}\mathbb{E}\|\omega_{t-1}^{(m)} - \bar{\omega}_{t-1}\|^2 + 4IL^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[\underbrace{\|\gamma\omega_{t-1}^{(m)}\|^2}_{T_1} + \|\eta\nu_{t-1}^{(m)}\|^2\big] \\
+ 8IM(c_\omega\alpha_{t-1}^2)^2\frac{\sigma^2}{b_y} \\
+ 16IM(c_\omega\alpha_{t-1}^2)^2\zeta_g^2 + 16IL^2(c_\omega\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2\big] \\
+ 16IL^2(c_\omega\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|y_{t-1}^{(m)} - \bar{y}_{t-1}\|^2\big]
\end{aligned}
$$

We separate the term T_1 with triangle inequality to get:

$$\begin{aligned}\sum_{m=1}^{M}\mathbb{E}\|\hat{\omega}_t^{(m)}-\bar{\omega}_t\|^2 \le &\left(1+\frac{1}{I}+8IL^2\gamma^2\alpha_{t-1}^2\right)\sum_{m=1}^{M}\mathbb{E}\|\omega_{t-1}^{(m)}-\bar{\omega}_{t-1}\|^2\\ &+4IL^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[2\|\gamma\bar{\omega}_{t-1}\|^2+\|\eta\nu_{t-1}^{(m)}\|^2\big]\\ &+8IM(c_\omega\alpha_{t-1}^2)^2\frac{\sigma^2}{b_y}\\ &+16IM(c_\omega\alpha_{t-1}^2)^2\zeta_g^2+16IL^2(c_\omega\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)}-\bar{x}_{t-1}\|^2\big]\\ &+16IL^2(c_\omega\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|y_{t-1}^{(m)}-\bar{y}_{t-1}\|^2\big]\end{aligned}$$

This completes the proof. □

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

$$\begin{aligned}\sum_{m=1}^{M}\mathbb{E}\|\hat{q}_t^{(m)}-\bar{q}_t\|^2 \le &\left(1+\frac{33}{32I}\right)\sum_{m=1}^{M}\mathbb{E}\|q_{t-1}^{(m)}-\bar{q}_{t-1}\|^2+8I\tilde{L}_2^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[\|\gamma\omega_{t-1}^{(m)}\|^2+\|\eta\nu_{t-1}^{(m)}\|^2\big]\\ &+32IL^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\|\tau^2\bar{q}_{t-1}\|^2+8IM(c_u\alpha_{t-1}^2)^2\frac{\sigma^2}{b_x}\\ &+16IM(c_u\alpha_{t-1}^2)^2\zeta_f^2+32IM(c_u\alpha_{t-1}^2)^2\frac{C_f^2\zeta_{g,yy}^2}{\mu^2}\\ &+64I(c_u\alpha_{t-1}^2)^2\tilde{L}_2^2\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)}-\bar{x}_t\big\|^2+\big\|y_t^{(m)}-\bar{y}_t\big\|^2\big]\\ &+16I(c_u\alpha_{t-1}^2)^2L^2\sum_{m=1}^{M}\mathbb{E}\big\|u_t^{(m)}-\bar{u}_t\big\|^2\end{aligned}$$

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

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

$$\begin{aligned}\mathbb{E}\|\hat{q}_t^{(m)}-\bar{q}_t\|^2 &= \mathbb{E}\big\|(1-c_u\alpha_{t-1}^2)\big(q_{t-1}^{(m)}-\bar{q}_{t-1}\big)+p_{t,\mathcal{B}_x}^{(m)}-\bar{p}_{t,\mathcal{B}_x}-(1-c_u\alpha_{t-1}^2)\big(p_{t-1,\mathcal{B}_x}^{(m)}-\bar{p}_{t-1,\mathcal{B}_x}\big)\big\|^2\\ &\le (1+\frac{1}{I})(1-c_u\alpha_{t-1}^2)^2\mathbb{E}\|q_{t-1}^{(m)}-\bar{q}_{t-1}\|^2\\ &\quad+(1+I)\mathbb{E}\big\|p_{t,\mathcal{B}_x}^{(m)}-\bar{p}_{t,\mathcal{B}_x}-(1-c_u\alpha_{t-1}^2)\big(p_{t-1,\mathcal{B}_x}^{(m)}-\bar{p}_{t-1,\mathcal{B}_x}\big)\big\|^2\end{aligned}$$

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:

$$T_1 \le 2\sum_{m=1}^{M}\mathbb{E}\big\|p_{t,\mathcal{B}_x}^{(m)}-\bar{p}_{t,\mathcal{B}_x}-\big(p_{t-1,\mathcal{B}_x}^{(m)}-\bar{p}_{t-1,\mathcal{B}_x}\big)\big\|^2+2(c_u\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big\|p_{t-1,\mathcal{B}_x}^{(m)}-\bar{p}_{t-1,\mathcal{B}_x}\big\|^2$$

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

$$
\begin{aligned}
T_{1,1} &\overset{(a)}{\leq} \sum_{m=1}^{M} \mathbb{E}\big\|p_{t,\mathcal{B}_x}^{(m)} - p_{t-1,\mathcal{B}_x}^{(m)}\big\|^2 \\
&\overset{(b)}{\leq} 2\big(L^2 + \frac{2L_{y^2}^2 C_f^2}{\mu^2}\big) \sum_{m=1}^{M} \mathbb{E}\big[\|x_t^{(m)} - x_{t-1}^{(m)}\|^2 + \|y_t^{(m)} - y_{t-1}^{(m)}\|^2\big] + 4L^2 \sum_{m=1}^{M} \mathbb{E}\|u_t^{(m)} - u_{t-1}^{(m)}\|^2 \\
&\leq 2\tilde{L}_2^2 \alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\big[\|\eta\nu_{t-1}^{(m)}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\big] + 4L^2\alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\|\tau^2 q_{t-1}^{(m)}\|^2
\end{aligned}
$$

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

$$
\begin{aligned}
T_{1,2} &= \sum_{m=1}^{M} \mathbb{E}\big\|p_{t-1,\mathcal{B}_x}^{(m)} - p_{t-1}^{(m)} - (\bar{p}_{t-1,\mathcal{B}_x} - \bar{p}_{t-1}) + p_{t-1}^{(m)} - \bar{p}_{t-1}\big\|^2 \\
&\overset{(b)}{\leq} 2\sum_{m=1}^{M} \mathbb{E}\big\|p_{t-1,\mathcal{B}_x}^{(m)} - p_{t-1}^{(m)}\big\|^2 + 2\sum_{m=1}^{M} \mathbb{E}\big\|p_{t-1}^{(m)} - \bar{p}_{t-1}\big\|^2
\end{aligned}
$$

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

$$
\begin{aligned}
T_{1,2,2} \leq\ & 16\big(L^2 + \frac{2L_{y^2}^2 C_f^2}{\mu^2}\big) \sum_{m=1}^{M} \mathbb{E}\big[\big\|x_t^{(m)} - \bar{x}_t\big\|^2 + \big\|y_t^{(m)} - \bar{y}_t\big\|^2\big] \\
&+ 4L^2 \sum_{m=1}^{M} \mathbb{E}\big\|u_t^{(m)} - \bar{u}_t\big\|^2 + 4M\zeta_f^2 + \frac{8MC_f^2\zeta_{g,yy}^2}{\mu^2}
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_{m=1}^{M} \mathbb{E}\|\hat{q}_t^{(m)} - \bar{q}_t\|^2 \leq\ & \big(1 + \frac{1}{I}\big) \sum_{m=1}^{M} \mathbb{E}\|q_{t-1}^{(m)} - \bar{q}_{t-1}\|^2 + 8I\tilde{L}_2^2\alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\big[\|\gamma\omega_{t-1}^{(m)}\|^2 + \|\eta\nu_{t-1}^{(m)}\|^2\big] \\
&+ 16IL^2\alpha_{t-1}^2 \sum_{m=1}^{M} \mathbb{E}\underbrace{\|\tau^2 q_{t-1}^{(m)}\|^2}_{T_1} \\
&+ 8IM(c_u\alpha_{t-1}^2)^2\frac{\sigma^2}{b_x} + 16IM(c_u\alpha_{t-1}^2)^2\zeta_f^2 + 32IM(c_u\alpha_{t-1}^2)^2\frac{C_f^2\zeta_{g,yy}^2}{\mu^2} \\
&+ 64I(c_u\alpha_{t-1}^2)^2\tilde{L}_2^2 \sum_{m=1}^{M} \mathbb{E}\big[\big\|x_t^{(m)} - \bar{x}_t\big\|^2 \\
&+ \big\|y_t^{(m)} - \bar{y}_t\big\|^2\big] + 16I(c_u\alpha_{t-1}^2)^2L^2 \sum_{m=1}^{M} \mathbb{E}\big\|u_t^{(m)} - \bar{u}_t\big\|^2
\end{aligned}
$$

We separate the term T_1 with triangle inequality to get:

$$\begin{aligned}\sum_{m=1}^{M}\mathbb{E}\|\hat{q}_t^{(m)}-\bar{q}_t\|^2 \le &\left(1+\frac{1}{I}+32IL^2\tau^2\alpha_{t-1}^2\right)\sum_{m=1}^{M}\mathbb{E}\|q_{t-1}^{(m)}-\bar{q}_{t-1}\|^2\\ &+8I\tilde{L}_2^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[\|\gamma\omega_{t-1}^{(m)}\|^2+\|\eta\nu_{t-1}^{(m)}\|^2\big]\\ &+32IL^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\|\tau^2\bar{q}_{t-1}\|^2+8IM(c_u\alpha_{t-1}^2)^2\frac{\sigma^2}{b_x}\\ &+16IM(c_u\alpha_{t-1}^2)^2\zeta_f^2+32IM(c_u\alpha_{t-1}^2)^2\frac{C_f^2\zeta_{g,yy}^2}{\mu^2}\\ &+64I(c_u\alpha_{t-1}^2)^2\tilde{L}_2^2\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)}-\bar{x}_t\big\|^2+\big\|y_t^{(m)}-\bar{y}_t\big\|^2\big]\\ &+16I(c_u\alpha_{t-1}^2)^2L^2\sum_{m=1}^{M}\mathbb{E}\big\|u_t^{(m)}-\bar{u}_t\big\|^2\end{aligned}$$

This completes the proof. □

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

Lemma C.11. *For $\eta<\min(\frac{\tilde{L}^2}{c_\nu},\frac{\tilde{L}^2}{c_\omega},\frac{\tilde{L}^2}{c_u},1)$, $\gamma<\min(\frac{\tilde{L}^2}{c_\nu},\frac{\tilde{L}^2}{c_\omega},\frac{\tilde{L}^2}{c_u},1)$, $\tau<\min(\frac{\tilde{L}^2}{c_\nu},\frac{\tilde{L}^2}{c_u},\frac{1}{2})$ and $\alpha_t<\frac{1}{16\tilde{L}I}$, where $\tilde{L}=max(\tilde{L}_1,\tilde{L}_2)$, we have:*

$$\begin{aligned}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \le &\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\alpha_tE_t+\alpha_tF_t+\alpha_tQ_t+\frac{c_\omega^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_y}+\frac{c_\omega^2\alpha_t^3}{\tilde{L}^2}\zeta_g^2\\ &+\frac{c_\nu^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_x}+\frac{c_u^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_x}+\frac{c_\nu^2\alpha_t^3\zeta_f^2}{\tilde{L}^2}+\frac{c_u^2\alpha_t^3\zeta_f^2}{\tilde{L}^2}+\frac{2c_\nu^2\alpha_t^3}{\tilde{L}^2}\frac{C_f^2\zeta_{g,xy}^2}{\mu^2}+\frac{4c_u^2\alpha_t^3}{\tilde{L}^2}\frac{C_f^2\zeta_{g,yy}^2}{\mu^2}\big)\\ \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_t \le &\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\alpha_tE_t+\alpha_tF_t+\alpha_tQ_t+\frac{2c_\omega^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_y}+\frac{2c_\omega^2\alpha_t^3}{\tilde{L}^2}\zeta_g^2\\ &+\frac{c_\nu^2\alpha_t^3}{\tilde{L}^2}\frac{2\sigma^2}{b_x}+\frac{c_u^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_x}+\frac{c_\nu^2\alpha_t^3\zeta_f^2}{\tilde{L}^2}+\frac{c_u^2\alpha_t^3\zeta_f^2}{\tilde{L}^2}+\frac{c_\nu^2\alpha_t^3}{\tilde{L}^2}\frac{C_f^2\zeta_{g,xy}^2}{\mu^2}+\frac{2c_u^2\alpha_t^3}{\tilde{L}^2}\frac{C_f^2\zeta_{g,yy}^2}{\mu^2}\big)\\ \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s}\alpha_t J_t \le &\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\alpha_tF_t+\alpha_tE_t+\alpha_tQ_t+\frac{c_\omega^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_y}+\frac{c_\omega^2\alpha_t^3}{\tilde{L}^2}\zeta_g^2\\ &+\frac{c_u^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_x}+\frac{c_\nu^2\alpha_t^3}{\tilde{L}^2}\frac{\sigma^2}{b_x}+\frac{c_u^2\alpha_t^3\zeta_f^2}{\tilde{L}^2}+\frac{c_\nu^2\alpha_t^3\zeta_f^2}{2\tilde{L}^2}+\frac{c_\nu^2\alpha_t^3}{\tilde{L}^2}\frac{C_f^2\zeta_{g,xy}^2}{\mu^2}+\frac{40c_u^2\alpha_t^3}{\tilde{L}^2}\frac{C_f^2\zeta_{g,yy}^2}{\mu^2}\big)\end{aligned}$$

Proof. Based on Lemma C.8, for $t\neq\bar{t}_s$, we have:

$$\begin{aligned}D_t\le &\big(1+\frac{17}{16I}\big)D_{t-1}+16I\tilde{L}_1^2\alpha_{t-1}^2\eta^2E_{t-1}+16I\tilde{L}_1^2\alpha_{t-1}^2\gamma^2F_{t-1}+16I\tilde{L}_1^2\alpha_{t-1}^2\gamma^2G_{t-1}+32IL^2\tau^2\alpha_{t-1}^2J_{t-1}\\ &+32IL^2\tau^2\alpha_{t-1}^2Q_{t-1}+8Ic_\nu^2\alpha_{t-1}^4\frac{\sigma^2}{b_x}+32Ic_\nu^2\alpha_{t-1}^4\zeta_f^2+64Ic_\nu^2\alpha_{t-1}^4\frac{C_f^2\zeta_{g,xy}^2}{\mu^2}\\ &+128I^2\tilde{L}_1^2\eta^2c_\nu^2\alpha_{t-1}^4\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2D_l+128I^2\tilde{L}_1^2\gamma^2c_\nu^2\alpha_{t-1}^4\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2G_l+32I^2L^2\tau^2c_\nu^2\alpha_{t-1}^4\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2J_l\end{aligned}$$

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

$$
\begin{aligned}
D_t \leq & \sum_{\ell=\bar{t}_{s-1}}^{\ell} \big(1+\frac{17}{16I}\big)^{t-\ell}\big(16I\tilde{L}_1^2\alpha_\ell^2\eta^2 E_\ell + 16I\tilde{L}_1^2\alpha_\ell^2\gamma^2 F_\ell + 16I\tilde{L}_1^2\alpha_\ell^2\gamma^2 G_\ell + 32IL^2\tau^2\alpha_\ell^2 J_\ell \\
& + 32IL^2\tau^2\alpha_\ell^2 Q_\ell + 8Ic_\nu^2\alpha_\ell^4\frac{\sigma^2}{b_x} + 32Ic_\nu^2\alpha_\ell^4\zeta_f^2 + 64Ic_\nu^2\alpha_\ell^4\frac{C_f^2\zeta_{g,xy}^2}{\mu^2} \\
& + 128I^2\tilde{L}_1^2\eta^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{l}}^2 D_{\bar{\ell}} + 128I^2\tilde{L}_1^2\gamma^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 G_{\bar{\ell}} + 32I^2L^2\tau^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 J_{\bar{\ell}}\big) \\
\leq & \sum_{\ell=\bar{t}_{s-1}}^{t-1} \big(48I\tilde{L}_1^2\alpha_\ell^2\eta^2 E_\ell + 48I\tilde{L}_1^2\alpha_\ell^2\gamma^2 F_\ell + 48I\tilde{L}_1^2\alpha_\ell^2\gamma^2 G_\ell + 96IL^2\tau^2\alpha_\ell^2 J_\ell \\
& + 96IL^2\tau^2\alpha_\ell^2 Q_\ell + 24Ic_\nu^2\alpha_\ell^4\frac{\sigma^2}{b_x} + 96Ic_\nu^2\alpha_\ell^4\zeta_f^2 + 192Ic_\nu^2\alpha_\ell^4\frac{C_f^2\zeta_{g,xy}^2}{\mu^2} \\
& + 384I^2\tilde{L}_1^2\eta^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{l}}^2 D_{\bar{\ell}} + 384I^2\tilde{L}_1^2\gamma^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 G_{\bar{\ell}} + 96I^2L^2\tau^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 J_{\bar{\ell}}\big)
\end{aligned}
$$

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

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

$$
\begin{aligned}
& \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s} \alpha_t D_t \\
\leq & \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \big(48I\tilde{L}_1^2\alpha_\ell^2\eta^2 E_\ell + 48I\tilde{L}_1^2\alpha_\ell^2\gamma^2 F_\ell + 48I\tilde{L}_1^2\alpha_\ell^2\gamma^2 G_\ell + 96IL^2\tau^2\alpha_\ell^2 J_\ell \\
& + 96IL^2\tau^2\alpha_\ell^2 Q_\ell + 24Ic_\nu^2\alpha_\ell^4\frac{\sigma^2}{b_x} + 96Ic_\nu^2\alpha_\ell^4\zeta_f^2 + 192Ic_\nu^2\alpha_\ell^4\frac{C_f^2\zeta_{g,xy}^2}{\mu^2} \\
& + 384I^2\tilde{L}_1^2\eta^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{l}}^2 D_{\bar{\ell}} + 384I^2\tilde{L}_1^2\gamma^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 G_{\bar{\ell}} + 96I^2L^2\tau^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 J_{\bar{\ell}}\big) \\
\overset{(a)}{\leq} & \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \big(3I\tilde{L}_1\alpha_t^2\eta^2 E_t + 3I\tilde{L}_1\alpha_t^2\gamma^2 F_t + 3I\tilde{L}_1\alpha_t^2\gamma^2 G_t + 6IL\tau^2\alpha_t^2 J_t \\
& + 6IL\tau^2\alpha_t^2 Q_t + \frac{3Ic_\nu^2\alpha_t^4}{2\tilde{L}}\frac{\sigma^2}{b_x} + \frac{6Ic_\nu^2\alpha_t^4\zeta_f^2}{\tilde{L}} + \frac{12Ic_\nu^2\alpha_t^4}{\tilde{L}}\frac{C_f^2\zeta_{g,xy}^2}{\mu^2} \\
& + 32I^2\tilde{L}_1\eta^2c_\nu^2\alpha_t^4\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 D_\ell + 32I^2\tilde{L}_1\gamma^2c_\nu^2\alpha_t^4\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 G_\ell + 6I^2L\tau^2c_\nu^2\alpha_t^4\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 J_\ell\big) \\
\overset{(b)}{\leq} & \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \big(\frac{3\eta^2}{16}\alpha_t E_t + \frac{3\gamma^2}{16}\alpha_t F_t + \frac{3\gamma^2}{16}\alpha_t G_t + \frac{3\tau^2}{8}\alpha_t J_t + \frac{3\tau^2}{8}\alpha_t Q_t \\
& + \frac{3c_\nu^2\alpha_t^3}{32\tilde{L}^2}\frac{\sigma^2}{b_x} + \frac{3c_\nu^2\alpha_t^3\zeta_f^2}{8\tilde{L}^2} + \frac{3c_\nu^2\alpha_t^3}{4\tilde{L}^2}\frac{C_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{\eta^2c_\nu^2}{8*16^3I^2\tilde{L}^4}\alpha_t D_t \\
& + \frac{\gamma^2c_\nu^2}{8*16^3I^2\tilde{L}^4}\alpha_t G_t + \frac{3\tau^2c_\nu^2}{8*16^4I^2\tilde{L}^4}\alpha_t J_t\big)
\end{aligned}
$$

In inequalities (a) and (b), we use $\alpha_t < \frac{1}{16\tilde{L}I} \leq \frac{1}{16\tilde{L}_1 I}$. Note that $\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s} \alpha_t D_t = \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t$ as $D_{\bar{t}_s} = D_{\bar{t}_{s-1}} = 0$.

Then if we choose $\eta < \frac{\tilde{L}^2}{c_\nu}$ and $\gamma < \frac{\tilde{L}^2}{c_\nu}$, $\tau < \frac{\tilde{L}^2}{c_\nu}$, we have

$$\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t \leq \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \big(\frac{\eta^2}{4}\alpha_t E_t + \frac{\gamma^2}{4}\alpha_t F_t + \frac{\gamma^2}{2}\alpha_t G_t + \tau^2 \alpha_t J_t + \frac{\tau^2}{2}\alpha_t Q_t + \frac{c_\nu^2 \alpha_t^3}{8\tilde{L}^2}\frac{\sigma^2}{b_x} + \frac{c_\nu^2 \alpha_t^3 \zeta_f^2}{2\tilde{L}^2} + \frac{c_\nu^2 \alpha_t^3}{\tilde{L}^2}\frac{C_f^2 \zeta_{g,xy}^2}{\mu^2}\big) \tag{20}$$

Based on Lemma C.9, for $t \neq \bar{t}_s$, we have:

$$G_t \leq \left(1 + \frac{33}{32I}\right) G_{t-1} + 8IL^2\eta^2\alpha_{t-1}^2 D_{t-1} + 8IL^2\eta^2\alpha_{t-1}^2 E_{t-1} + 8IL^2\gamma^2\alpha_{t-1}^2 F_{t-1} + 8Ic_\omega^2\alpha_{t-1}^4\frac{\sigma^2}{b_y} + 16Ic_\omega^2\alpha_{t-1}^4\zeta_g^2 + 16I^2L^2\eta^2c_\omega^2\alpha_{t-1}^4\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2 D_l + 16I^2L^2\gamma^2c_\omega^2\alpha_{t-1}^4\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2 G_l$$

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

$$G_t \leq \sum_{\ell=\bar{t}_{s-1}}^{t-1} \big(24IL^2\eta^2\alpha_\ell^2 D_\ell + 24IL^2\eta^2\alpha_\ell^2 E_\ell + 24IL^2\gamma^2\alpha_\ell^2 F_\ell + 24Ic_\omega^2\alpha_\ell^4\frac{\sigma^2}{b_y} + 48Ic_\omega^2\alpha_\ell^4\zeta_g^2 + 48I^2L^2\eta^2c_\omega^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 D_{\bar{\ell}} + 48I^2L^2\gamma^2c_\omega^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1}\alpha_{\bar{\ell}}^2 G_{\bar{\ell}}\big)$$

Next we multiply α_t over both sides and take sum from $\bar{t}_{s-1}+1$ to $\bar{t}_s$, use the condition that $\alpha_t < \frac{1}{16I\tilde{L}} < \frac{1}{16IL}$, $\eta < \frac{\tilde{L}^2}{c_\omega}$ and $\gamma < \frac{\tilde{L}^2}{c_\omega}$, we have:

$$\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t G_t \leq \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \big(\frac{1}{2}\eta^2\alpha_t D_t + \frac{1}{8}\eta^2\alpha_t E_t + \frac{1}{8}\gamma^2\alpha_t F_t + \frac{c_\omega^2\alpha_t^3}{8\tilde{L}^2}\frac{\sigma^2}{b_y} + \frac{c_\omega^2\alpha_t^3}{8\tilde{L}^2}\zeta_g^2\big) \tag{21}$$

Based on Lemma C.10, we have:

$$J_t \leq \left(1 + \frac{33I}{32I}\right) J_{t-1} + 16I\tilde{L}_2^2\tau^2\alpha_{t-1}^2 G_{t-1} + 16I\tilde{L}_2^2\tau^2\alpha_{t-1}^2 F_{t-1} + 16I\tilde{L}_2^2\eta^2\alpha_{t-1}^2 D_{t-1} + 16I\tilde{L}_2^2\eta^2\alpha_{t-1}^2 E_{t-1} + 16IL^2\tau^2\alpha_{t-1}^2 Q_{t-1} + 8Ic_u^2\alpha_{t-1}^4\frac{\sigma^2}{b_x} + 16Ic_u^2\alpha_{t-1}^4\zeta_f^2 + 32Ic_u^2\alpha_{t-1}^4\frac{C_f^2\zeta_{g,yy}^2}{\mu^2} + 64I^2c_u^2\eta^2\alpha_{t-1}^4\tilde{L}_2^2\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2 D_l + 64I^2c_u^2\gamma^2\alpha_{t-1}^4\tilde{L}_2^2\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2 G_l + 16I^2c_u^2\tau^2\alpha_{t-1}^4L^2\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2 J_l$$

Suppose we have $\alpha_t < \frac{1}{16\tilde{L}I}$, $\eta < \frac{\tilde{L}^2}{c_u}$, $\gamma < \frac{\tilde{L}^2}{c_u}$, $\tau < \frac{\tilde{L}^2}{c_u}$

$$\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s} \alpha_t J_t \leq \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \big(\frac{\tau^2}{2}\alpha G_t + \frac{\tau^2}{4}\alpha_t F_t + \frac{\eta^2}{2}\alpha_t D_t + \frac{\eta^2}{4}\alpha_{t-1}E_t + \frac{\tau^2}{4}\alpha_t Q_t + \frac{c_u^2\alpha_t^3}{8\tilde{L}^2}\frac{\sigma^2}{b_x} + \frac{c_u^2\alpha_t^3\zeta_f^2}{4\tilde{L}^2} + \frac{3c_u^2\alpha_t^3}{\tilde{L}^2}\frac{C_f^2\zeta_{g,yy}^2}{\mu^2}\big) \tag{22}$$

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:*

$$\mathbb{E}\big\|\nabla h(\bar{x}_t) - \bar{\mu}_t\big\|^2 \leq \frac{2\tilde{L}_1^2}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|\bar{x}_t - x_t^{(m)}\big\|^2 + 2\big\|\bar{y}_t - y_t^{(m)}\big\|^2 + 2\big\|y_{\bar{x}_t} - \bar{y}_t\big\|^2\big] + 4L^2\mathbb{E}\big\|u_{\bar{x}_t} - \bar{u}_t\big\|^2$$

where we denote $u_{\bar{x}_t} = [\nabla_{y^2} g(\bar{x}_t, y_{\bar{x}_t})]^{-1}\nabla_y f(\bar{x}_t, y_{\bar{x}_t})$ and $\tilde{L}_1^2 = \big(L^2 + \frac{2L_{xy}^2 C_f^2}{\mu^2}\big)$ is a constant.

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

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

$$\begin{aligned}\mathbb{E}\big[h(\bar{x}_{t+1})\big] \leq \mathbb{E}\big[h(\bar{x}_t)\big] - \frac{\eta\alpha_t}{4}\mathbb{E}\big[\big\|\bar{\nu}_t\big\|^2\big] - \frac{\eta\alpha_t}{2}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] + \eta\alpha_t\mathbb{E}\big[\big\|\bar{u}_t - \bar{\nu}_t\big\|^2\big] \\ + \frac{2\tilde{L}_1^2\eta\alpha_t}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|\bar{x}_t - x_t^{(m)}\big\|^2 + 2\big\|\bar{y}_t - y_t^{(m)}\big\|^2 + 2\big\|y_{\bar{x}_t} - \bar{y}_t\big\|^2\big] + 4L^2\eta\alpha_t\mathbb{E}\big\|u_{\bar{x}_t} - \bar{u}_t\big\|^2\end{aligned}$$

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

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

$$\begin{aligned}\mathbb{E}[h(\bar{x}_{t+1})] &\leq \mathbb{E}\big[h(\bar{x}_t) + \langle\nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{\bar{L}}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\big] \\ &\overset{(a)}{=} \mathbb{E}\big[h(\bar{x}_t) - \eta\alpha_t\langle\nabla h(\bar{x}_t), \bar{\nu}_t\rangle + \frac{\eta^2\alpha_t^2\bar{L}}{2}\|\bar{\nu}_t\|^2\big] \\ &\overset{(b)}{=} \mathbb{E}\big[h(\bar{x}_t) - \frac{\eta\alpha_t}{2}\big\|\bar{\nu}_t\big\|^2 - \frac{\eta\alpha_t}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta\alpha_t}{2}\big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\big\|^2 + \frac{\eta\alpha_t^2\bar{L}}{2}\big\|\bar{\nu}_t\big\|^2\big] \\ &= \mathbb{E}\big[h(\bar{x}_t) - \frac{\eta\alpha_t}{4}\big\|\bar{\nu}_t\big\|^2 - \frac{\eta\alpha_t}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta\alpha_t}{2}\underbrace{\big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\big\|^2}_{T_1}\big]\end{aligned}$$

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 $\eta\alpha_t < \frac{1}{2\bar{L}}$; For the term T_1, we have:

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

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 $\mathcal{G}(t)$:

$$\begin{aligned}\mathcal{G}_t = h(\bar{x}_t) + \frac{9bM\eta}{64\alpha_t}\big\|\bar{\nu}_t - \bar{\mu}_t\big\|^2 + \frac{18\eta\tilde{L}^2}{\mu\gamma}\big\|\bar{y}_t - y_{\bar{x}_t}\big\|^2 + \frac{9bM\eta}{64\alpha_t}\big\|\bar{q}_t - \bar{p}_t\big\|^2 \\ + \frac{9bM\eta}{64\alpha_t}\big\|\bar{\omega}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\big\|^2 + \frac{18\eta L^2}{\mu\tau}\big\|\bar{u}_t - u_{\bar{x}_t}\big\|^2\end{aligned}$$

Furthermore, we have constants $\tilde{L}_1^2 = \big(L^2 + \frac{2L_{xy}^2 C_f^2}{\mu^2}\big)$ and $\tilde{L}_2^2 = \big(L^2 + \frac{2L_{y^2}^2 C_f^2}{\mu^2}\big)$, to ease the writing, without loss of generality, we assume the second order Lipschitz constants $L_{xy} = L_{y^2}$, as a result $\tilde{L}_1^2 = \tilde{L}_2^2$, we denote it as $\tilde{L}$ in the subsequent proof.

Theorem C.14. *Suppose we choose $c_\nu = \frac{64}{9bM} + \frac{2}{3b^2M^2}$, $c_\omega = \frac{48^2}{bM\mu^2} + \frac{2}{3b^2M^2}$, $c_u = \frac{48^2}{bM\mu^2} + \frac{2}{3b^2M^2}$ $u = (bM\sigma)^2\bar{u}$, where $\bar{u} = \max\big(2, 16^2I^3\tilde{L}^2, c_\nu^{3/2}, c_\omega^{3/2}\big)$, $\delta = \frac{(bM\sigma)^{2/3}}{(16\bar{L})^{1/3}}$, $\alpha_t = \frac{\delta}{(u+t)^{1/3}}, t \in [T]$, $\gamma < \min\big(\frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, \frac{\tilde{L}^2}{c_\nu}, \frac{\tilde{L}^2}{c_\omega}, \frac{\tilde{L}^2}{c_u}, \frac{1}{2\bar{L}}, 1\big)$, $\eta < \min\big(\frac{\mu\gamma}{36\kappa\tilde{L}}, \frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, \frac{\tilde{L}^2}{c_\nu}, \frac{\tilde{L}^2}{c_\omega}, \frac{\tilde{L}^2}{c_u}, \frac{1}{2\bar{L}}, 1\big)$, $\tau <$*

$\min\big(\frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, \frac{\tilde{L}^2}{c_\nu}, \frac{\tilde{L}^2}{c_u}, \frac{1}{2L}, \frac{1}{2}\big)$ *where* C_1 *is a constant, we set the mini-batch size* $b_x = b_y = b$ *and the first batch with size* $b_1 = O(Ib)$, $r = \frac{C_f}{\mu}$, *then we have:*

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

To reach an ϵ*-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 $u \geq c_\nu^{3/2}\delta^3$, it is straightforward to verify that $c_\nu\alpha_t^2 < 1$. By Lemma C.3, we have:

$$\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \leq \big(\alpha_{t-1}^{-1} - \alpha_{t-2}^{-1} - c_\nu\alpha_{t-1}\big)A_{t-1} + \frac{2c_\nu^2\alpha_{t-1}^3\sigma^2}{bM} + \frac{16L^2\tau^2\alpha_{t-1}}{bM}(J_{t-1}+Q_{t-1})$$
$$+ \frac{8\tilde{L}^2\eta^2\alpha_{t-1}}{bM}(D_{t-1}+E_{t-1}) + \frac{8\tilde{L}^2\gamma^2\alpha_{t-1}}{bM}(F_{t-1}+G_{t-1})$$

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

$$\alpha_t^{-1} - \alpha_{t-1}^{-1} = \frac{(u+\sigma^2 t)^{1/3}}{\delta} - \frac{(u+\sigma^2(t-1))^{1/3}}{\delta} \overset{(a)}{\leq} \frac{\sigma^2}{3\delta(u+\sigma^2(t-1))^{2/3}}$$
$$\overset{(b)}{\leq} \frac{2^{2/3}\sigma^2\delta^2}{3\delta^3(u+\sigma^2 t)^{2/3}} \overset{(c)}{=} \frac{2^{2/3}\sigma^2}{3\delta^3}\alpha_t^2 \leq \frac{2}{3Ib^2M^2}\alpha_t \leq \frac{2^{2/3}\sigma^2}{3\delta^3}\alpha_t^2 \leq \frac{2}{3b^2M^2}\alpha_t$$

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 α_t, By choosing $c_\nu = \frac{64}{9bM} + \frac{2}{3b^2M^2}$, we have:

$$\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \leq -\frac{64}{9bM}\alpha_{t-1}A_{t-1} + \frac{2c_\nu^2\alpha_{t-1}^3\sigma^2}{bM} + \frac{16L^2\tau^2\alpha_{t-1}}{bM}(J_{t-1}+Q_{t-1})$$
$$+ \frac{8\tilde{L}^2\eta^2\alpha_{t-1}}{bM}(D_{t-1}+E_{t-1}) + \frac{8\tilde{L}^2\gamma^2\alpha_{t-1}}{bM}(F_{t-1}+G_{t-1})$$

Next, we telescope from $\bar{t}_{s-1}+1$ to $\bar{t}_s$:

$$\big(\frac{A_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}} - \frac{A_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}}\big) \leq -\frac{64}{9bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t A_t + \frac{2c_\nu^2\sigma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + \frac{16\tilde{L}^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t$$
$$+ \frac{8\tilde{L}^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + \frac{8\tilde{L}^2\gamma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t + \frac{16\tilde{L}^2\gamma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_t$$
$$+ \frac{32L^2\tau^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t J_t + \frac{16L^2\tau^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t Q_t \tag{23}$$

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

$$\frac{C_t}{\alpha_{t-1}} - \frac{C_{t-1}}{\alpha_{t-2}} \leq -\frac{48^2\alpha_{t-1}}{bM\mu^2}C_{t-1} + \frac{2c_\omega^2\alpha_{t-1}^3\sigma^2}{bM} + \frac{4L^2\eta^2\alpha_{t-1}}{bM}(D_{t-1}+E_{t-1}) + \frac{4L^2\gamma^2\alpha_{t-1}}{bM}(F_{t-1}+G_{t-1})$$

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

$$\frac{C_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}} - \frac{C_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}}$$
$$\leq -\frac{48^2}{bM\mu^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t + \frac{2c_\omega^2\sigma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + \frac{16L^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t$$
$$+ \frac{8L^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + \frac{8L^2\gamma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t + \frac{16L^2\gamma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_t \tag{24}$$

Next from Lemma C.2, we choose $c_u = \frac{48^2}{bM\mu^2} + \frac{2}{3b^2M^2}$, to obtain:

$$\frac{H_t}{\alpha_{t-1}} - \frac{H_{t-1}}{\alpha_{t-2}} \leq -\frac{48^2\alpha_{t-1}}{bM\mu^2}H_{t-1} + \frac{2c_u^2\alpha_{t-1}^3}{bM}\sigma^2 + \frac{8\eta^2\alpha_{t-1}\tilde{L}^2}{bM}(D_{t-1}+E_{t-1}) + \frac{8\gamma^2\alpha_{t-1}\tilde{L}^2}{bM}(F_{t-1}+G_{t-1}) + \frac{8\tau^2\alpha_{t-1}L^2}{bM}(J_{t-1}+Q_{t-1})$$

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

$$\frac{H_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}} - \frac{H_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}} \leq -\frac{48^2}{bM\mu^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t H_t + \frac{2c_u^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3\sigma^2 + \frac{8\eta^2\tilde{L}^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_{t-1}(D_t+E_t) + \frac{8\gamma^2\tilde{L}^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_{t-1}(F_t+G_t) + \frac{8\tau^2L^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t(J_t+Q_{t-1}) \tag{25}$$

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

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

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

$$\begin{aligned} B_{\bar{t}_s} - B_{\bar{t}_{s-1}} &\leq -\frac{\mu\gamma}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{\gamma^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t + \frac{9\gamma}{\mu}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t + \frac{9\kappa^2\eta^2}{2\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t \\ &\quad + \frac{9I\eta^2\gamma L^2}{\mu}\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1}\alpha_t\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 D_\ell + \frac{9I\gamma^3L^2}{\mu}\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1}\alpha_t\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 G_\ell \\ &\leq -\frac{\mu\gamma}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{\gamma^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t + \frac{9\gamma}{\mu}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t + \frac{9\kappa^2\eta^2}{2\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t \\ &\quad + \frac{9L^2\eta^2\gamma}{16^2\hat{L}^2\mu}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell D_\ell + \frac{9L^2\gamma^3}{16^2\hat{L}^2\mu}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell G_\ell \end{aligned} \tag{26}$$

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

$$I_{t+1} - I_t \leq -\frac{\mu\tau\alpha_t}{4}I_t - \frac{\tau^2\alpha_t}{4}Q_t + \frac{9\kappa^2\eta^2\alpha_t}{2\mu\tau}E_t + \frac{9\tau\alpha_t}{\mu}H_t + \frac{18I\eta^2\tau\alpha_t\tilde{L}^2}{\mu}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 D_\ell + \frac{18I\gamma^2\tau\alpha_t\tilde{L}^2}{\mu}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 G_\ell + 18I\tau^3\alpha_t L^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell^2 J_\ell$$

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

$$\begin{aligned}
I_{\bar{t}_s} - I_{\bar{t}_{s-1}} \leq & -\frac{\mu\tau}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t I_t - \frac{\tau^2}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t Q_t + \frac{9\kappa^2\eta^2}{2\mu\tau} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t E_t + \frac{9\tau}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t H_t \\
& + \frac{18I\eta^2\tau\tilde{L}^2}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 D_\ell + \frac{18I\gamma^2\tau\tilde{L}^2}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 G_\ell \\
& + 18I\tau^3 L^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_\ell^2 J_\ell \\
\leq & -\frac{\mu\tau}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t I_t - \frac{\tau^2}{4} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t Q_t + \frac{9\kappa^2\eta^2}{2\mu\tau} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t E_t + \frac{9\tau}{\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t H_t \\
& + \frac{18\eta^2\tau}{16^2\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t + \frac{18\gamma^2\tau}{16^2\mu} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t G_t + \frac{18\tau^3 L^2}{16^2\tilde{L}^2} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t J_t
\end{aligned} \tag{27}$$

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

$$\begin{aligned}
\mathbb{E}[h(\bar{x}_{t+1})] \leq & \mathbb{E}[h(\bar{x}_t)] - \frac{\eta\alpha_t}{4} E_t - \frac{\eta\alpha_t}{2} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + \eta\alpha_t A_t + 4\tilde{L}^2\eta\alpha_t B_t + 4L^2\eta\alpha_t I_t \\
& + 2\tilde{L}^2 I\eta^3\alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_l^2 D_l + 4\tilde{L}^2 I\gamma^2\eta\alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_l^2 G_l
\end{aligned}$$

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:

$$\begin{aligned}
& \mathbb{E}[h(\bar{x}_{\bar{t}_s}) - h(\bar{x}_{\bar{t}_{s-1}})] \\
& \leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{4} E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + 4L^2\eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t I_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \eta\alpha_t A_t \\
& \quad + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} 4\tilde{L}^2\eta\alpha_t B_t + 2\tilde{L}^2 I\eta^3 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_l^2 D_l + 4\tilde{L}^2 I\gamma^2\eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t \sum_{\ell=\bar{t}_{s-1}}^{t-1} \alpha_l^2 G_l \\
& \leq - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{4} E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2} \mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + 4L^2\eta \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t I_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \eta\alpha_t A_t \\
& \quad + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} 4\tilde{L}^2\eta\alpha_t B_t + \frac{\eta^3}{128} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t D_t + \frac{\gamma^2\eta}{64} \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \alpha_t G_t
\end{aligned} \tag{28}$$

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

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

$$\begin{aligned}
&\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \\
&\leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2}\mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + \big(\frac{9\eta c_\omega^2\sigma^2}{32} + \frac{9\eta c_\nu^2\sigma^2}{32} + \frac{9\eta c_u^2\sigma^2}{32}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 \\
&\quad - \frac{\tilde{L}^2\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{L^2\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t I_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\frac{9\eta\gamma\tilde{L}^2}{2\mu} - \frac{9\eta\gamma^2\tilde{L}^2}{4} - \frac{9\eta\gamma^2L^2}{8}\big)\alpha_t F_t \\
&\quad - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\frac{1}{4} - \frac{81\kappa^2\tilde{L}^2\eta^2}{\mu^2\gamma^2} - \frac{81\kappa^2L^2\eta^2}{\mu^2\tau^2} - \frac{9L^2\eta^2}{8} - \frac{9\tilde{L}^2\eta^2}{4}\big)\eta\alpha_t E_t \\
&\quad - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\frac{9\tau\eta L^2}{2\mu} - \frac{9\eta\tau^2L^2}{4}\big)\alpha_t Q_t + \big(\frac{81\kappa^2}{64} + \frac{9L^2}{4} + 9\tilde{L}^2\big)\tau^2\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t J_t \\
&\quad + \big(\frac{1}{128} + \frac{81\kappa^2}{128} + \frac{81\kappa^2}{64} + \frac{9L^2}{4} + \frac{9\tilde{L}^2}{2}\big)\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \\
&\quad + \big(\frac{1}{64} + \frac{81\kappa^2}{128} + \frac{81\kappa^2}{64} + \frac{9\tilde{L}^2}{4} + \frac{9L^2}{2}\big)\gamma^2\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_t
\end{aligned}$$

By the condition that $\eta < \frac{\mu\gamma}{36\kappa\tilde{L}}$ and $\gamma \leq \frac{1}{2L} < \frac{1}{2\mu}$. Next, we denote:

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

Then, we have:

$$\begin{aligned}
&\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \\
&\leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] + \big(\frac{9\eta c_\omega^2\sigma^2}{32} + \frac{9\eta c_\nu^2\sigma^2}{32} + \frac{9\eta c_u^2\sigma^2}{32}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 \\
&\quad - \frac{9\eta\gamma^2\tilde{L}^2}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t - \frac{\eta}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t - \frac{\tilde{L}^2\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{L^2\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t I_t \\
&\quad - \frac{9\eta\tau^2L^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t Q_t + C_1\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + C_1\gamma^2\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_t + C_1\tau^2\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t J_t
\end{aligned} \tag{29}$$

Combine Eq. (29) with Lemma C.11, and use the condition that $\eta < \min\big(\frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, 1\big)$, $\gamma < \min\big(\frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, 1\big)$ and $\tau < \min\big(\frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, 1\big)$ we have:

$$\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] + C_{\sigma,\zeta}\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3$$

For ease of notation, we denote

$$C_{\sigma,\zeta} = \big(4c_\omega^2\sigma^2 + 4c_u^2\sigma^2 + 4c_\nu^2\sigma^2 + 3c_u^2\zeta_f^2 + 3c_\nu^2\zeta_f^2 + 3c_\omega^2\zeta_g^2 + \frac{3c_\nu^2C_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{120c_u^2C_f^2\zeta_{g,yy}^2}{\mu^2}\big).$$

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

$$\mathbb{E}[\mathcal{G}_T] - \mathbb{E}[\mathcal{G}_1] \leq -\sum_{t=1}^{T-1}\frac{\eta\alpha_t}{2}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] + \eta C_{\sigma,\zeta}\sum_{t=1}^{T-1}\alpha_t^3$$

Rearranging the terms and use the fact that α_t is non-increasing, we have:

$$\begin{aligned}\frac{\eta\alpha_T}{2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] &\leq \mathbb{E}[\mathcal{G}_1] - \mathbb{E}[\mathcal{G}_T] + \eta C_{\sigma,\zeta}\sum_{t=1}^{T-1}\alpha_t^3 \\ &\leq h(x_1) - h^* + \frac{9bM\eta A_1}{64\alpha_1} + \frac{18\eta\tilde{L}^2 B_1}{\mu\gamma} \\ &\quad + \frac{9bM\eta C_1}{64\alpha_1} + \frac{9bM\eta H_1}{64\alpha_1} + \frac{18\eta L^2 I_1}{\mu\tau} + \eta C_{\sigma,\zeta}\sum_{t=1}^{T-1}\alpha_t^3\end{aligned}$$

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

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

the first inequality follows $u_t > \sigma^2$, the last inequality follows Proposition E.3.

Next, we denote the initial sub-optimality as $\Delta = h(\bar{x}_1) - h^*$, 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 $I_1 = \|u_1 - [\nabla_{y^2} g(x_1, y_{x_1})]^{-1}\nabla_y f(x_1, y_{x_1})\|^2 \leq \Delta_u$.

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 $\eta\alpha_T T/2$ to have:

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

Note that we have:

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

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

$$\frac{1}{\alpha_1} \leq \frac{u^{1/3}+\sigma^{2/3}}{\delta} = \frac{(16\tilde{L})^{1/3}((bM)^{2/3}\bar{u}^{1/3}+1)}{(bM)^{2/3}} \tag{30}$$

when $t = T$, we have:

$$\frac{1}{\alpha_T T} \leq \frac{u^{1/3}}{\delta T} + \frac{\sigma^{2/3}}{\delta T^{2/3}} = (16\tilde{L})^{1/3}\left(\frac{\bar{u}^{1/3}}{T} + \frac{1}{(bMT)^{2/3}}\right) \tag{31}$$

In summary, we have:

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

Recall that $\tilde{L} = O(\kappa)$, $\bar{L} = O(\kappa^3)$, therefore we have $c_\nu = \Theta((bM)^{-1})$, $c_\omega = \Theta(\kappa^2(bM)^{-1})$, $c_u = \Theta(\kappa^2(bM)^{-1})$ $\bar{u} = \Theta(I^3\kappa^3)$, then for η, γ, τ, we have:

$$\gamma < \min\big(\frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, \frac{\tilde{L}^2}{c_\nu}, \frac{\tilde{L}^2}{c_\omega}, \frac{\tilde{L}^2}{c_u}, \frac{1}{2L}, 1\big)$$

$$\tau < \min\big(\frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, \frac{\tilde{L}^2}{c_\nu}, \frac{\tilde{L}^2}{c_u}, \frac{1}{2L}, \frac{1}{2}\big)$$

$$\eta < \min\big(\frac{\mu\gamma}{36\kappa\tilde{L}}, \frac{1}{8C_1^{1/2}}, \frac{\tilde{L}}{4C_1^{1/2}}, \frac{\tilde{L}^2}{c_\nu}, \frac{\tilde{L}^2}{c_\omega}, \frac{\tilde{L}^2}{c_u}, \frac{1}{2\bar{L}}, 1\big)$$

where $C_1 = O(\kappa^2)$, so we have $\gamma^{-1} = O(\kappa)$, $\eta^{-1} = O(\kappa^3)$, $\tau^{-1} = O(\kappa)$, furthermore, $\alpha_1^{-1} = O(I\kappa^{4/3})$, $C_{\sigma,\zeta} = O(\kappa^6(bM)^{-2})$, assume we choose the size of the first batch to be $b_1 = Ib$.

Combine everything together, we have:

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

To reach an ϵ-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 $Gc(f,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$, $Gc(g,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$, $Jv(g,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$, $Hv(g,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$ $\square$

Algorithm 2 Federated Bilevel Optimization (**FedBiO**)

Input: Initial states x_1, y_1 and u_1; learning rates $\{\gamma_t, \eta_t, \tau_t\}_{t=1}^T$
Initialization: Set $x_1^{(m)} = x_1$, $y_1^{(m)} = y_1$, $u_1^{(m)} = u_1$;
for $t = 1$ **to** T **do**
  Randomly sample mutually independent minibatch of samples $\mathcal{B}_y$ and $\mathcal{B}_x = \{\mathcal{B}_{g,1}, \mathcal{B}_{g,2}, \mathcal{B}_{f,1}, \mathcal{B}_{f,2}\}$ of size b;
  $\omega_t^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y)$
  $\nu_t^{(m)} = \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,1}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,1}) u_t^{(m)}$;
  $\hat{y}_{t+1}^{(m)} = y_t^{(m)} - \gamma_t \omega_t^{(m)}$, $\hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta_t \nu_t^{(m)}$;
  if t mod I $= 0$ **then**
    $y_{t+1}^{(m)} = \frac{1}{M}\sum_{j=1}^M \hat{y}_{t+1}^{(j)}$; $x_{t+1}^{(m)} = \frac{1}{M}\sum_{j=1}^M \hat{x}_{t+1}^{(j)}$
  else
    $y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}$, $x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)}$
  end if
  $\hat{u}_{t+1}^{(m)} = \mathcal{P}_r(\tau_t \nabla_y f^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{f,2}) + (I - \tau_t \nabla_{y^2} g^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_{g,2})) u_t^{(m)})$;
  if t mod I $= 0$ **then**
    $u_{t+1}^{(m)} = \frac{1}{M}\sum_{j=1}^M \hat{u}_{t+1}^{(j)}$
  else
    $u_{t+1}^{(m)} = \hat{u}_{t+1}^{(m)}$
  end if
end for

C.2 Proof for the FedBiO Algorithm

Algorithm 2 follows Eq. 6, 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:*

$$\mathbb{E}\big\|\bar{y}_t - y_{\bar{x}_t}\big\|^2 \le (1 - \frac{\mu\gamma}{4})\mathbb{E}\big\|\bar{y}_{t-1} - y_{\bar{x}_{t-1}}\big\|^2 + \frac{9\kappa^2\eta^2}{2\mu\gamma}\mathbb{E}\big\|\bar{\nu}_{t-1}\big\|^2 - \frac{\gamma^2}{4}\mathbb{E}\|\bar{\omega}_{t-1}\|^2 + \frac{9L^2\gamma}{2\mu M}\sum_{m=1}^M \mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2 + \|y_{t-1}^{(m)} - \bar{y}_{t-1})\|^2\big] + \frac{4\gamma\sigma^2}{\mu b_y M}$$

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

$$\mathbb{E}\big\|\bar{u}_{t+1} - u_{\bar{x}_{t+1}}\big\|^2 \le (1 - \frac{\mu\tau}{4})\mathbb{E}\big\|\bar{u}_t - u_{\bar{x}_t}\big\|^2 + \frac{5\tau^2\sigma^2}{4b_x M} + \frac{5\eta^2\bar{L}^2}{\mu\tau}\mathbb{E}\big\|\bar{\nu}_t\big\|^2 + \frac{5}{4}\big(\frac{3\tau L^2}{\mu M} + \frac{\tau L_{y^2}^2 C_f^2}{2\mu^3 M}\big)\sum_{m=1}^M \mathbb{E}\big[\big\|\bar{x}_t - x_t^{(m)}\big\|^2 + 2\big\|\bar{y}_t - y_t^{(m)}\big\|^2 + 2\big\|y_{\bar{x}_t} - \bar{y}_t\big\|^2\big]$$

We provide the proof for Lemma C.16 here and Lemma C.15 can be derived similarly.

Proof. First, by proposition E.5 (set $\alpha = 1$) and choose $\gamma < \frac{1}{2L}$, we have:

$$\begin{aligned}
\mathbb{E}\|\bar{y}_t - y_{\bar{x}_{t-1}}\|^2 &\leq (1 - \frac{\mu\gamma}{2})\mathbb{E}\|\bar{y}_{t-1} - y_{\bar{x}_{t-1}}\|^2 - \frac{\gamma^2}{4}\mathbb{E}\|\bar{\omega}_{t-1}\|^2 \\
&\quad + \frac{4\gamma}{\mu}\mathbb{E}\|\nabla_y g(\bar{x}_{t-1}, \bar{y}_{t-1}) - \frac{1}{M}\sum_{m=1}^{M}\nabla_y g(x_{t-1}^{(m)}, y_{t-1}^{(m)})\|^2 + \frac{4\gamma\sigma^2}{\mu b_y M} \\
&\leq (1 - \frac{\mu\gamma}{2})\mathbb{E}\|\bar{y}_{t-1} - y_{\bar{x}_{t-1}}\|^2 - \frac{\gamma^2}{4}\mathbb{E}\|\bar{\omega}_{t-1}\|^2 \\
&\quad + \frac{4\gamma}{\mu M}\sum_{m=1}^{M}\mathbb{E}\|\nabla_y^{(m)} g(\bar{x}_{t-1}, \bar{y}_{t-1}) - \nabla_y g(x_{t-1}^{(m)}, y_{t-1}^{(m)})\|^2 + \frac{4\gamma\sigma^2}{\mu b_y M} \\
&\leq (1 - \frac{\mu\gamma}{2})\mathbb{E}\|\bar{y}_{t-1} - y_{\bar{x}_{t-1}}\|^2 - \frac{\gamma^2}{4}\mathbb{E}\|\bar{\omega}_{t-1}\|^2 \\
&\quad + \frac{4L^2\gamma}{\mu M}\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2 + \|y_{t-1}^{(m)} - \bar{y}_{t-1})\|^2\big] + \frac{4\gamma\sigma^2}{\mu b_y M}
\end{aligned}$$

Furthermore, by the generalized triangle inequality, we have:

$$\begin{aligned}
\mathbb{E}\big\|\bar{y}_t - y_{\bar{x}_t}\big\|^2 &\leq (1 - \frac{\mu\gamma}{4})\mathbb{E}\big\|\bar{y}_{t-1} - y_{\bar{x}_{t-1}}\big\|^2 + (1 + \frac{4}{\mu\gamma})\mathbb{E}\big\|y_{\bar{x}_t} - y_{\bar{x}_{t-1}}\big\|^2 - (1 + \frac{\mu\gamma}{4})\frac{\gamma^2}{4}\mathbb{E}\|\bar{\omega}_{t-1}\|^2 \\
&\quad + (1 + \frac{\mu\gamma}{4})\frac{4L^2\gamma}{\mu M}\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2 + \|y_{t-1}^{(m)} - \bar{y}_{t-1})\|^2\big] + \frac{4\gamma\sigma^2}{\mu b_y M} \\
&\leq (1 - \frac{\mu\gamma}{4})\mathbb{E}\big\|\bar{y}_{t-1} - y_{\bar{x}_{t-1}}\big\|^2 + \frac{9\kappa^2\eta^2}{2\mu\gamma}\mathbb{E}\big\|\bar{\nu}_{t-1}\big\|^2 - \frac{\gamma^2}{4}\mathbb{E}\|\bar{\omega}_{t-1}\|^2 \\
&\quad + \frac{9L^2\gamma}{2\mu M}\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2 + \|y_{t-1}^{(m)} - \bar{y}_{t-1})\|^2\big] + \frac{4\gamma\sigma^2}{\mu b_y M}
\end{aligned}$$

where the second inequality is due to $\gamma < 1/2L$. This completes the proof. □

C.2.2 Local Variable Drift

Lemma C.17. *For any $t \neq \bar{t}_s, s \in [S]$, we have:*

$$\|x_t^{(m)} - \bar{x}_t\|^2 \leq I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\big\|\nu_\ell^{(m)} - \bar{\nu}_\ell\big\|^2,\ \|y_t^{(m)} - \bar{y}_t\|^2 \leq I\gamma^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\big\|\omega_\ell^{(m)} - \bar{\omega}_\ell\big\|^2$$

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: $x_t^{(m)} = x_{t-1}^{(m)} - \eta\nu_{t-1}^{(m)}$, this implies that: $x_t^{(m)} = x_{\bar{t}_{s-1}}^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\nu_\ell^{(m)}$ and $\bar{x}_t = \bar{x}_{\bar{t}_{s-1}} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\bar{\nu}_\ell$. So for $t \neq \bar{t}_s$, with $s \in [S]$ we have:

$$\begin{aligned}
\|x_t^{(m)} - \bar{x}_t\|^2 &= \big\|x_{\bar{t}_{s-1}}^{(m)} - \bar{x}_{\bar{t}_{s-1}} - \big(\sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\nu_\ell^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\bar{\nu}_\ell\big)\big\|^2 = \big\|\sum_{\ell=\bar{t}_{s-1}}^{t-1}\eta\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\big\|^2 \\
&\leq I\eta^2\sum_{\ell=\bar{t}_{s-1}}^{t-1}\big\|\nu_\ell^{(m)} - \bar{\nu}_\ell\big\|^2
\end{aligned}$$

We can derive the bound for $\|y_t^{(m)} - \bar{y}_t\|^2$ similarly. This completes the proof. □

Lemma C.18. *For any $t \in [T]$, we have:*

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\big(\nu_t^{(m)}-\bar{\nu}_t\big)\big\|^2 \le \frac{4L^2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|u_t^{(m)}-\bar{u}_t\big\|^2+4\zeta_f^2+\frac{8C_f^2\zeta_{g,xy}^2}{\mu^2}+\frac{2\sigma^2}{b_x}$$
$$+\big(\frac{16L^2}{M}+\frac{32L_{xy}^2C_f^2}{\mu^2 M}\big)\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)}-\bar{x}_t\big\|^2+\big\|y_t^{(m)}-\bar{y}_t\big\|^2\big]$$

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

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\big(\omega_t^{(m)}-\bar{\omega}_t\big)\big\|^2 \le \frac{2L^2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|x_t^{(m)}-\bar{x}_t\big\|^2+\frac{2L^2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|y_t^{(m)}-\bar{y}_t\big\|^2+\frac{2\sigma^2}{b_y}+2\zeta_g^2$$

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

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\big(u_{t+1}^{(m)}-\bar{u}_{t+1}\big)\big\|^2 \le (1+\frac{1}{I})\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|u_t^{(m)}-\bar{u}_t\big\|^2+\frac{64I\tau^2C_f^2\zeta_{g,yy}^2}{\mu^2}+32I\tau^2\zeta_f^2+\frac{2\tau^2\sigma^2}{b_x}$$
$$+\big(\frac{128IL^2\tau^2}{M}+\frac{256I\tau^2L_{y^2}^2C_f^2}{\mu^2 M}\big)\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)}-\bar{x}_t\big\|^2+\big\|y_t^{(m)}-\bar{y}_t\big\|^2\big]$$

Lemma C.18-Lemma C.20 bounds the local drift of $\nu_t^{(m)}$, $\omega_t^{(m)}$ 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:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\big(\nu_t^{(m)}-\bar{\nu}_t\big)\big\|^2 &\le \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_x f^{(m)}(x_t^{(m)},y_t^{(m)})-\nabla_{xy}g^{(m)}(x_t^{(m)},y_t^{(m)})u_t^{(m)}\\ &\quad-\frac{1}{M}\sum_{j=1}^{M}\nabla_x f^{(j)}(x_t^{(j)},y_t^{(j)})-\nabla_{xy}g^{(j)}(x_t^{(j)},y_t^{(j)})u_t^{(j)}\big\|^2+\frac{2\sigma^2}{b_x}\\ &\le \underbrace{\frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_x f^{(m)}(x_t^{(m)},y_t^{(m)})-\frac{1}{M}\sum_{j=1}^{M}\nabla_x f^{(j)}(x_t^{(j)},y_t^{(j)})\big\|^2}_{T_1}\\ &\quad+\underbrace{\frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_{xy}g^{(m)}(x_t^{(m)},y_t^{(m)})u_t^{(m)}-\frac{1}{M}\sum_{j=1}^{M}\nabla_{xy}g^{(j)}(x_t^{(j)},y_t^{(j)})u_t^{(j)}\big\|^2}_{T_2}+\frac{2\sigma^2}{b_x}\end{aligned}$$

For the term T_1, we have:

$$\begin{aligned}T_1 &\le \frac{16}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_x f^{(m)}(x_t^{(m)},y_t^{(m)})-\nabla_x f^{(m)}(\bar{x}_t,\bar{y}_t)\big\|^2+\frac{4}{M}\sum_{m=1}^{M}\mathbb{E}\big\|\nabla_x f^{(m)}(\bar{x}_t,\bar{y}_t)-\nabla_x f(\bar{x}_t,\bar{y}_t)\big\|^2\\ &\le \frac{16L^2}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)}-\bar{x}_t\big\|^2+\big\|y_t^{(m)}-\bar{y}_t\big\|^2\big]+4\zeta_f^2\end{aligned}$$

Next for the term T_2, we have:

$$\begin{aligned}T_2 &\le \frac{4L^2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|u_t^{(m)}-\bar{u}_t\big\|^2+\frac{4C_f^2}{\mu^2M^2}\sum_{m=1}^{M}\sum_{j=1}^{M}\mathbb{E}\big\|\nabla_{xy}g^{(m)}(x_t^{(m)},y_t^{(m)})-\nabla_{xy}g^{(j)}(x_t^{(j)},y_t^{(j)})\big\|^2\\ &\le \frac{4L^2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|u_t^{(m)}-\bar{u}_t\big\|^2+\frac{32L_{xy}^2C_f^2}{\mu^2 M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)}-\bar{x}_t\big\|^2+\big\|y_t^{(m)}-\bar{y}_t\big\|^2\big]+\frac{8C_f^2\zeta_{g,xy}^2}{\mu^2}\end{aligned}$$

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

Lemma C.18-Lemma C.20 have recursive dependence of each other. Next, we provide an untertwined bound for each of them. For ease of notation, we denote $D_t = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$, $B_t = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\omega_t^{(m)} - \bar{\omega}_t\|^2$, $A_t = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|u_t^{(m)} - \bar{u}_t\|^2$ and $C_t = \mathbb{E}\big\|\bar{y}_t - y_{\bar{x}_t}\big\|^2$.

Lemma C.21. *For $\gamma \le \frac{1}{8I\tilde{L}_1}$ and $\eta < \frac{1}{8I\tilde{L}_2}$, $\tau < \frac{1}{128I\tilde{L}_2}$, where $\tilde{L}_1^2 = \big(L^2 + \frac{2L_{xy}^2C_f^2}{\mu^2}\big)$ and $\tilde{L}_2^2 = \big(L^2 + \frac{2L_{y^2}^2C_f^2}{\mu^2}\big)$ are constants, then we have:*

$$\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} D_t \le 96I\zeta_f^2 + 16I\zeta_g^2 + \frac{16IC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{32IC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{16I\sigma^2}{b_y} + \frac{20I\sigma^2}{b_x}$$

$$\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} B_t \le 24I\zeta_f^2 + 8I\zeta_g^2 + \frac{4IC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{8IC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{8I\sigma^2}{b_y} + \frac{5I\sigma^2}{b_x}$$

Proof. Based on Lemma C.18, and sum from $\bar{t}_{s-1}+1$ to $\bar{t}_s - 1$, we have:

$$\begin{aligned}
\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1} D_t &\le 16I\eta^2\tilde{L}_1^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1}\sum_{\ell=\bar{t}_{s-1}}^{t-1} D_\ell + 16I\gamma^2\tilde{L}_1^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1}\sum_{\ell=\bar{t}_{s-1}}^{t-1} B_\ell \\
&\quad + 4L^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1} A_t + 4(I-1)\zeta_f^2 + \frac{8(I-1)C_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{2(I-1)\sigma^2}{b_x} \\
&\le 16I^2\eta^2\tilde{L}_1^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1} D_t + 16I^2\gamma^2\tilde{L}_1^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1} B_t \\
&\quad + 4L^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1} A_t + 4(I-1)\zeta_f^2 + \frac{8(I-1)C_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{2(I-1)\sigma^2}{b_x}
\end{aligned}$$

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

$$\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} D_t \le 16I^2\tilde{L}_1^2\eta^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} D_t + 16I^2\tilde{L}_1^2\gamma^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} B_t + 4L^2 \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} A_t + 4I\zeta_f^2 + \frac{8IC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{2I\sigma^2}{b_x} \tag{32}$$

Based on Lemma C.19, and sum from $\bar{t}_{s-1}+1$ to $\bar{t}_s - 1$, we have:

$$\begin{aligned}
\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1} B_t &\le 2I\eta^2L^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1}\sum_{\ell=\bar{t}_{s-1}}^{t-1} D_\ell + 2I\gamma^2L^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1}\sum_{\ell=\bar{t}_{s-1}}^{t-1} B_\ell + 2(I-1)\sigma^2 + 2(I-1)\zeta_g^2 \\
&\le 2I^2\eta^2L^2 \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1} D_\ell + 2I^2\gamma^2L^2 \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1} B_\ell + \frac{2(I-1)\sigma^2}{b_y} + 2(I-1)\zeta_g^2
\end{aligned}$$

Combine with the case of $t = \bar{t}_s$ in Lemma C.19, we have:

$$\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} B_t \le 2I^2\eta^2L^2 \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1} D_\ell + 2I^2\gamma^2L^2 \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1} B_\ell + \frac{2I\sigma^2}{b_y} + 2I\zeta_g^2 \tag{33}$$

Apply Lemma C.20 recursively, we have:

$$\begin{aligned}
A_t &\le \sum_{\ell=\bar{t}_{s-1}}^{t-1} (1+\frac{1}{I})^{t-\ell}\big(\frac{64I\tau^2C_f^2\zeta_{g,yy}^2}{\mu^2} + 32I\tau^2\zeta_f^2 + \frac{2\tau^2\sigma^2}{b_x} + 128I^2\eta^2\tau^2\tilde{L}_2^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1} D_{\bar{\ell}} + 128I^2\gamma^2\tau^2\tilde{L}_2^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1} B_{\bar{\ell}}\big) \\
&\le \sum_{\ell=\bar{t}_{s-1}}^{t-1} \big(\frac{192I\tau^2C_f^2\zeta_{g,yy}^2}{\mu^2} + 96I\tau^2\zeta_f^2 + \frac{6\tau^2\sigma^2}{b_x} + 384I^2\eta^2\tau^2\tilde{L}_2^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1} D_{\bar{\ell}} + 384I^2\gamma^2\tau^2\tilde{L}_2^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1} B_{\bar{\ell}}\big)
\end{aligned}$$

where we denote $\tilde{L}_2^2 = (L^2 + \frac{2L_{y^2}^2 C_f^2}{\mu^2})$, 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 $(1 + a/x)^x \leq e^a$, Then we choose $a = 1$ and $x = I$. Finally, we use the fact that $e^1 \leq 3$. Next, we sum from $\bar{t}_{s-1}$ to $\bar{t}_s - 1$ to have:

$$\begin{aligned}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} A_t &\leq \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1} \big(\frac{48I^2\tau^2 C_f^2 \zeta_{g,yy}^2}{\mu^2} + 96I^2\tau^2\zeta_f^2 + \frac{6I\tau^2\sigma^2}{b_x} + 24I^3\eta^2\tau^2\tilde{L}_2^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1} D_{\bar{\ell}} + 24I^3\gamma^2\tau^2\tilde{L}_2^2 \sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell-1} B_{\bar{\ell}}\big) \\ &\leq \frac{192I^3\tau^2 C_f^2\zeta_{g,yy}^2}{\mu^2} + 96I^3\tau^2\zeta_f^2 + \frac{6I^2\tau^2\sigma^2}{b_x} + 384I^4\eta^2\tau^2\tilde{L}_2^2 \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1} D_\ell + 384I^4\gamma^2\tau^2\tilde{L}_2^2 \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1} B_\ell \end{aligned} \tag{34}$$

Combine Eq. 32, Eq. 33 and Eq. 34, and we choose η, γ and τ such that $I^2\gamma^2 L^2 < \frac{1}{4}$, $I^2\gamma^2\tilde{L}_1^2 < \frac{1}{64}$, $I^2\eta^2\tilde{L}_1^2 < \frac{1}{32}$, $I^2\eta^2 L^2 < \frac{1}{16}$, $I^2\tau^2\tilde{L}_2^2 < \frac{1}{128^2}$, $I^2\tau^2 L^2 < \frac{1}{48}$, 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.

□

C.2.3 Descent Lemma

Lemma C.22. *For all $t \in [\bar{t}_{s-1}, \bar{t}_s - 1]$, the iterates generated satisfy:*

$$\mathbb{E}\big\|\nabla h(\bar{x}_t) - \mathbb{E}_\xi[\bar{\nu}_t]\big\|^2 \leq \frac{2\tilde{L}_1^2}{M}\sum_{m=1}^M \mathbb{E}\big[\big\|\bar{x}_t - x_t^{(m)}\big\|^2 + 2\big\|\bar{y}_t - y_t^{(m)}\big\|^2 + 2\big\|y_{\bar{x}_t} - \bar{y}_t\big\|^2\big] + 4L^2\mathbb{E}\big\|u_{\bar{x}_t} - \bar{u}_t\big\|^2$$

where we denote $u_{\bar{x}_t} = [\nabla_{y^2} g(\bar{x}_t, y_{\bar{x}_t})]^{-1}\nabla_y f(\bar{x}_t, y_{\bar{x}_t})$ and $\tilde{L}_1^2 = \big(L^2 + \frac{2L_{xy}^2 C_f^2}{\mu^2}\big)$ is a constant.

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

$$\begin{aligned}\mathbb{E}\big\|\nabla h(\bar{x}_t) - \mathbb{E}_\xi[\bar{\nu}_t]\big\|^2 \leq{}& \mathbb{E}\big\|\nabla_x f(\bar{x}_t, y_{\bar{x}_t}) - \nabla_{xy} g(\bar{x}_t, y_{\bar{x}_t}) \times [\nabla_{y^2} g(\bar{x}_t, y_{\bar{x}_t})]^{-1}\nabla_y f(\bar{x}_t, y_{\bar{x}_t}) \\ &- \frac{1}{M}\sum_{m=1}^M \big(\nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)})u_t^{(m)}\big)\big\|^2 \\ \leq{}& 2\mathbb{E}\big\|\nabla_x f(\bar{x}_t, y_{\bar{x}_t}) - \frac{1}{M}\sum_{m=1}^M \nabla_x f^{(m)}(x_t^{(m)}, y_t^{(m)})\big\|^2 \\ &+ 2\mathbb{E}\big\|\nabla_{xy} g(\bar{x}_t, y_{\bar{x}_t}) \times [\nabla_{y^2} g(\bar{x}_t, y_{\bar{x}_t})]^{-1}\nabla_y f(\bar{x}_t, y_{\bar{x}_t}) \\ &- \frac{1}{M}\sum_{m=1}^M \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)})u_t^{(m)}\big\|^2\end{aligned}$$

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

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

For the second term T_2, we have:

$$\begin{aligned}T_2 \leq{}& \frac{4C_f^2}{\mu^2 M}\sum_{m=1}^M \mathbb{E}\big\|\nabla_{xy} g^{(m)}(\bar{x}_t, y_{\bar{x}_t}) - \nabla_{xy} g^{(m)}(x_t^{(m)}, y_t^{(m)})\big\|^2 \\ &+ 4L^2\mathbb{E}\big\|[\nabla_{y^2} g(\bar{x}_t, y_{\bar{x}})]^{-1}\nabla_y f(\bar{x}_t, y_{\bar{x}_t}) - \bar{u}_t\big\|^2\end{aligned}$$

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

$$T_{2,1} \leq \frac{4L_{xy}^2 C_f^2}{\mu^2 M}\sum_{m=1}^M \mathbb{E}\big[\big\|\bar{x}_t - x_t^{(m)}\big\|^2 + 2\big\|\bar{y}_t - y_t^{(m)}\big\|^2 + 2\big\|y_{\bar{x}_t} - \bar{y}_t\big\|^2\big]$$

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

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

$$\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4}\|\mathbb{E}_\xi[\nu_t^{(m)}]\|^2 + \frac{\eta^2\bar{L}\sigma^2}{2b_x M} + 2\eta L^2\mathbb{E}\big\|u_{\bar{x}_t} - \bar{u}_t\big\|^2 + \frac{\eta\tilde{L}_1^2}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|\bar{x}_t - x_t^{(m)}\big\|^2 + 2\big\|\bar{y}_t - y_t^{(m)}\big\|^2 + 2\big\|y_{\bar{x}_t} - \bar{y}_t\big\|^2\big]$$

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

Proof. Using the smoothness of f we have:

$$\begin{aligned}
\mathbb{E}[h(\bar{x}_{t+1})] &\leq \mathbb{E}[h(\bar{x}_t)] + \mathbb{E}\langle\nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{\bar{L}}{2}\mathbb{E}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&= \mathbb{E}[h(\bar{x}_t)] - \eta\mathbb{E}\langle\nabla h(\bar{x}_t), \mathbb{E}_\xi[\bar{\nu}_t]\rangle + \frac{\eta^2\bar{L}}{2}\mathbb{E}\|\mathbb{E}_\xi[\bar{\nu}_t]\|^2 + \frac{\eta^2\bar{L}\sigma^2}{2b_x M} \\
&\overset{(a)}{=} \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t) - \mathbb{E}_\xi[\bar{\nu}_t]\|^2 - \left(\frac{\eta}{2} - \frac{\eta^2\bar{L}}{2}\right)\|\mathbb{E}_\xi[\bar{\nu}_t]\|^2 + \frac{\eta^2\bar{L}\sigma^2}{2b_x M} \\
&\overset{(b)}{\leq} \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4}\|\mathbb{E}_\xi[\nu_t^{(m)}]\|^2 + \frac{\eta^2\bar{L}\sigma^2}{2b_x M} \\
&\quad + \frac{\eta\tilde{L}_1^2}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|\bar{x}_t - x_t^{(m)}\big\|^2 + 2\big\|\bar{y}_t - y_t^{(m)}\big\|^2 + 2\big\|y_{\bar{x}_t} - \bar{y}_t\big\|^2\big] + 2\eta L^2\mathbb{E}\big\|u_{\bar{x}_t} - \bar{u}_t\big\|^2
\end{aligned}$$

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.22. □

C.2.4 Proof of Convergence Theorem

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

$$\mathcal{G}_t = \mathbb{E}[h(\bar{x}_t)] + \frac{9\eta\tilde{L}_1^2}{\mu\gamma}\mathbb{E}\big\|\bar{y}_t - y_{\bar{x}_t}\big\|^2 + \frac{9\eta L^2}{\mu\tau}\mathbb{E}\big\|\bar{u}_t - u_{\bar{x}_t}\big\|^2$$

Theorem C.24. *Suppose we choose $\tau = \min(\frac{1}{128I\tilde{L}_2}, \frac{1}{144\kappa L})$, then denote $\bar{\gamma} = \min(\frac{1}{8I\tilde{L}_2}, \frac{\tau}{36\kappa\bar{L}}, \frac{1}{4\bar{L}}, \frac{1}{8I\tilde{L}_1})$, if we choose $\eta = \frac{\mu\gamma}{36\kappa\bar{L}_1}$, and $\gamma = \min\left(\bar{\gamma}, \left(\frac{\Delta'}{C_\gamma' T}\right)^{1/3}\right)$ and $r = \frac{C_f}{\mu}$ where Δ' and C_γ' 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 ϵ 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 $T = O(\kappa^6\epsilon^{-1.5})$ number of iterations.

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

potential function we have:

$$\begin{aligned}\mathcal{G}_{\bar{t}_s} - \mathcal{G}_{\bar{t}_{s-1}} \leq & -\frac{\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta L^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}F_t - \frac{\eta\tilde{L}_1^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}C_t \\ & - \frac{\eta}{4}\left(1 - \frac{162\kappa^2\eta^2\tilde{L}_1^2}{\mu^2\gamma^2} - \frac{180\kappa^2\eta^2\bar{L}^2}{\tau^2}\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}E_t + \big(\tilde{L}_1^2 + \frac{81\kappa^2\tilde{L}_1^2}{2} + \frac{45\kappa^2\tilde{L}_2^2}{16}\big)I^2\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}D_t \\ & + \big(2\tilde{L}_1^2 + \frac{81\kappa^2\tilde{L}_1^2}{2} + \frac{45\kappa^2\tilde{L}_2^2}{16}\big)I^2\gamma^2\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t + \frac{81I\kappa^2\tilde{L}_1^2\eta^3\sigma^2}{2\mu^2\gamma^2b_xM} + \frac{36I\tilde{L}_1^2\eta\sigma^2}{\mu^2b_yM} \\ & + \frac{45I\kappa^2\bar{L}^2\eta^3\sigma^2}{\tau^2b_xM} + \frac{45I\kappa L\tau\eta\sigma^2}{4b_xM} + \frac{\eta^2I\bar{L}\sigma^2}{2b_xM}\end{aligned}$$

to bound the coefficients above, we choose $\eta \leq \min\big(\frac{\mu\gamma}{36\kappa\tilde{L}_1}, \frac{\tau}{36\kappa\bar{L}}, \frac{1}{4\bar{L}}\big)$, $\tau < \frac{1}{144\kappa L}$ and we denote $C_1 = \big(2\tilde{L}_1^2 + \frac{81\kappa^2\tilde{L}_1^2}{2} + \frac{45\kappa^2\tilde{L}_2^2}{16}\big)I^2$. Then we have:

$$\begin{aligned}\mathcal{G}_{\bar{t}_s} - \mathcal{G}_{\bar{t}_{s-1}} \leq & -\frac{\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta L^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}F_t - \frac{\eta\tilde{L}_1^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}C_t - \frac{\eta}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}E_t \\ & + C_1\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}D_t + C_1\gamma^2\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t + \frac{I\eta\sigma^2}{2b_xM} + \frac{36I\tilde{L}_1^2\eta\sigma^2}{\mu^2b_yM}\end{aligned}$$

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

$$\begin{aligned}\mathcal{G}_{\bar{t}_s} - \mathcal{G}_{\bar{t}_{s-1}} \leq & -\frac{\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta L^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}F_t - \frac{\eta\tilde{L}_1^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}C_t - \frac{\eta}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}E_t + \frac{I\eta\sigma^2}{2b_xM} + \frac{36I\tilde{L}_1^2\eta\sigma^2}{\mu^2b_yM} \\ & + C_1\eta^3\big(96I\zeta_f^2 + 16I\zeta_g^2 + \frac{16IC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{32IC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{16I\sigma^2}{b_y} + \frac{20I\sigma^2}{b_x}\big) \\ & + C_1\gamma^2\eta\big(24I\zeta_f^2 + 8I\zeta_g^2 + \frac{4IC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{8IC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{8I\sigma^2}{b_y} + \frac{5I\sigma^2}{b_x}\big)\end{aligned}$$

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

$$\begin{aligned}\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq & \mathcal{G}_1 - \mathcal{G}_T + \frac{T\eta\sigma^2}{2b_xM} + \frac{36T\tilde{L}_1^2\eta\sigma^2}{\mu^2b_yM} \\ & + C_1\eta^3\big(96T\zeta_f^2 + 16T\zeta_g^2 + \frac{16TC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{32TC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{16T\sigma^2}{b_y} + \frac{20T\sigma^2}{b_x}\big) \\ & + C_1\gamma^2\eta\big(24T\zeta_f^2 + 8T\zeta_g^2 + \frac{4TC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{8TC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{8T\sigma^2}{b_y} + \frac{5T\sigma^2}{b_x}\big) \\ \leq & \Delta + \frac{9\eta\tilde{L}_1^2\Delta_y}{\mu\gamma} + \frac{9\eta L^2\Delta_u}{\mu\tau} + \frac{T\eta\sigma^2}{2b_xM} + \frac{36T\tilde{L}_1^2\eta\sigma^2}{\mu^2b_yM} \\ & + C_1\eta^3\big(96T\zeta_f^2 + 16T\zeta_g^2 + \frac{16TC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{32TC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{16T\sigma^2}{b_y} + \frac{20T\sigma^2}{b_x}\big) \\ & + C_1\gamma^2\eta\big(24T\zeta_f^2 + 8T\zeta_g^2 + \frac{4TC_f^2\zeta_{g,yy}^2}{\mu^2} + \frac{8TC_f^2\zeta_{g,xy}^2}{\mu^2} + \frac{8T\sigma^2}{b_y} + \frac{5T\sigma^2}{b_x}\big)\end{aligned}$$

we define $\Delta = h(x_1) - h^*$ as the initial sub-optimality of the function, $\Delta_y = \|y_1 - y_{x_1}\|^2$ 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:

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

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

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

Recall that, we have the condition that $\eta \leq \min\big(\frac{1}{8I\tilde{L}_2}, \frac{\mu\gamma}{36\kappa\tilde{L}_1}, \frac{\tau}{36\kappa\bar{L}}, \frac{1}{4\bar{L}}\big)$, $\gamma \leq \frac{1}{8I\tilde{L}_1}$, $\tau \leq \min(\frac{1}{128I\tilde{L}_2}, \frac{1}{144\kappa L})$. Suppose we choose $\tau = \min(\frac{1}{128I\tilde{L}_2}, \frac{1}{144\kappa L})$, then denote

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

and let $\gamma \leq \bar{\gamma}$, and $\eta = \frac{\mu\gamma}{36\kappa\tilde{L}_1}$, then we have:

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

We denote

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

then we choose γ as:

$$\gamma = \min\big(\bar{\gamma}, \left(\frac{\Delta^{'}}{C_\gamma^{'}T}\right)^{1/3}\big)$$

and obtain:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq \frac{\Delta^{'}}{\bar{\gamma}T} + \left(\frac{C_\gamma^{'}(\Delta^{'})^2}{T^2}\right)^{1/3} + \frac{18L^2\Delta_u}{\mu\tau T} + \frac{\sigma^2}{b_x M} + \frac{72\tilde{L}_1^2\sigma^2}{\mu^2 b_y M}$$

Finally, since $\tilde{L}_1 = O(\kappa)$, $\bar{L} = O(\kappa^3)$, suppose we choose $I = O(1)$, then we have $\tau^{-1} = O(\kappa)$, $\bar{\gamma}^{-1} = O(\kappa^5)$, $\Delta^{'} = O(\kappa^3)$, $C_1 = O(\kappa^4)$, $C_\eta = O(\kappa^6)$, $C_\gamma = O(\kappa^6)$, $C_\gamma^{'} = O(\kappa^6)$ 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 ϵ 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 $T = O(\kappa^6\epsilon^{-1.5})$ number of iterations. □

Algorithm 3 FedBiO- Local Lower Level Problem

Input: Initial states x_1, y_1; learning rates $\{\gamma_t, \eta_t\}_{t=1}^T$
Initialization: Set $x_1^{(m)} = x_1$, $y_1^{(m)} = y_1$;
for $t = 1$ **to** T **do**
  Randomly sample mutually independent minibatch of samples $\mathcal{B}_y$ and $\mathcal{B}_x$ of size b;
  $\omega_t^{(m)} = \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y)$ and $\nu_t^{(m)} = \Phi^{(m)}(x^{(m)}, y^{(m)}; \mathcal{B}_x)$;
  $\hat{y}_{t+1}^{(m)} = y_t^{(m)} - \gamma_t \omega_t^{(m)}$, $\hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta_t \nu_t^{(m)}$;
  if t mod I $= 0$ **then**
    $y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}$, $x_{t+1}^{(m)} = \frac{1}{M}\sum_{j=1}^M \hat{x}_{t+1}^{(j)}$
  else
    $y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}$, $x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)}$
  end if
end for

Algorithm 4 FedBiOAcc - Local Lower Level Problem

Input: Constants c_ω, c_ν, γ, η; learning rate schedule $\{\alpha_t\}$, $t \in [T]$, initial state (x_1, y_1);
Initialization: Set $y_1^{(m)} = y_1$, $x_1^{(m)} = x_1$, $\omega_1^{(m)} = \nabla_y g^{(m)}(x_1, y_1, \mathcal{B}_y)$, $\nu_1^{(m)} = \Phi^{(m)}(x_1, y_1; \mathcal{B}_x)$ for $m \in [M]$
for $t = 1$ **to** T **do**
  $\hat{y}_{t+1}^{(m)} = y_t^{(m)} - \gamma\alpha_t \omega_t^{(m)}$, $\hat{x}_{t+1}^{(m)} = x_t^{(m)} - \eta\alpha_t \nu_t^{(m)}$, $\hat{u}_{t+1}^{(m)} = u_t^{(m)} - \tau\alpha_t q_t^{(m)}$
  if t mod I $= 0$ **then**
    $y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}$, $x_{t+1}^{(m)} = \frac{1}{M}\sum_{j=1}^M \hat{x}_{t+1}^{(j)}$
  else
    $y_{t+1}^{(m)} = \hat{y}_{t+1}^{(m)}$, $x_{t+1}^{(m)} = \hat{x}_{t+1}^{(m)}$,
  end if
  Randomly sample minibatches $\mathcal{B}_y$ and $\mathcal{B}_x$
  $\hat{\omega}_{t+1}^{(m)} = \nabla_y g^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}, \mathcal{B}_y) + (1 - c_\omega\alpha_t^2)(\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)}, \mathcal{B}_y))$
  $\hat{\nu}_{t+1}^{(m)} = \Phi^{(m)}(x_{t+1}^{(m)}, y_{t+1}^{(m)}; \mathcal{B}_x) + (1 - c_\nu\alpha_t^2)(\nu_t^{(m)} - \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_x))$
  if t mod I $= 0$ **then**
    $\omega_{t+1}^{(m)} = \hat{\omega}_{t+1}^{(m)}$, $\nu_{t+1}^{(m)} = \frac{1}{M}\sum_{j=1}^M \hat{\nu}_{t+1}^{(j)}$,
  else
    $\omega_{t+1}^{(m)} = \hat{\omega}_{t+1}^{(m)}$, $\nu_{t+1}^{(m)} = \hat{\nu}_{t+1}^{(m)}$
  end if
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: $\|\nabla f^{(m)}(z) - \nabla f^{(j)}(z)\| \leq \zeta_f$, $\|\nabla_{xy} g^{(m)}(z) - \nabla_{xy} g^{(j)}(z)\| \leq \zeta_{g,xy}$, $\|\nabla_{y^2} g^{(m)}(z) - \nabla_{y^2} g^{(j)}(z)\| \leq \zeta_{g,yy}$, $\|y_x^{(m)} - y_x^{(j)}\| \leq \zeta_{g^*}$, where ζ_f, $\zeta_{g,xy}$, $\zeta_{g,yy}$, ζ_{g^*} are constants.

Note that we remove the requirement of gradient dissimilarity ζ_g in Assumption 3.5 and add the dissimilarity bound ζ_{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 ζ.

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

a) $\|\mathbb{E}[\Phi^{(m)}(x, y; \mathcal{B}_x)] - \Phi^{(m)}(x, y)\| \leq G_1$, *where* $G_1 = \kappa(1 - \tau\mu)^{Q+1} C_f$

b) $\mathbb{E}\|\Phi^{(m)}(x,y;\mathcal{B}_x)-\mathbb{E}[\Phi^{(m)}(x,y;\mathcal{B}_x)]\|^2 \le G_2^2$, *where* $G_2^2 = (2C_f^2+12C_f^2L^2\tau^2(Q+1)^2+4C_f^2L^2(Q+2)(Q+1)^2\tau^4\sigma^2)/b_x$

Proposition D.3. *Suppose Assumptions 3.2 and 3.3 hold, the following statements hold:*

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

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

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

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

Proposition D.4. *In Eq. 10, suppose Assumption 3.1, 3.2, 3.3, D.1 hold, we have:*

$$\|\nabla h^{(m)}(x)-\nabla h^{(j)}(x)\| \le (1+\kappa)\zeta_f+\frac{C_f}{\mu}\zeta_{g,xy}+\frac{\kappa C_f}{\mu}\zeta_{g,yy}+\big((1+\kappa)L+\frac{C_fL_{xy}}{\mu}+\frac{\kappa C_fL_{y^2}}{\mu}\big)\zeta_{g^*} \coloneqq \zeta$$

Proof. For $h^{(m)}(x)=f^{(m)}(x,y_x^{(m)}), m\in[M]$ in Eq. 10, we have:

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

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

$$\begin{aligned}
\big\|\nabla_x f^{(m)}(x,y_x^{(m)})-\nabla_x f^{(j)}(x,y_x^{(j)})\big\| &\le \big\|\nabla_x f^{(m)}(x,y_x^{(m)})-\nabla_x f^{(j)}(x,y_x^{(m)})\big\|\\
&\quad+\big\|\nabla_x f^{(j)}(x,y_x^{(m)})-\nabla_x f^{(j)}(x,y_x^{(j)})\big\|\\
&\le \zeta_f+L\big\|y_x^{(m)}-y_x^{(j)}\big\| \le \zeta_f+L\zeta_{g^*} \qquad (37)
\end{aligned}$$

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:

$$\begin{aligned}
&\big\|\nabla_{xy}g^{(m)}(x,y_x^{(m)})-\nabla_{xy}g^{(j)}(x,y_x^{(j)})\big\|\big\|\big(\nabla_{yy}g^{(m)}(x,y_x^{(m)})\big)^{-1}\nabla_y f^{(m)}(x,y_x^{(m)})\big\|\\
&\le \frac{C_f}{\mu}\big\|\nabla_{xy}g^{(m)}(x,y_x^{(m)})-\nabla_{xy}g^{(j)}(x,y_x^{(j)})\big\|\\
&\le \frac{C_f}{\mu}\big\|\nabla_{xy}g^{(m)}(x,y_x^{(m)})-\nabla_{xy}g^{(j)}(x,y_x^{(m)})\big\|+\frac{C_f}{\mu}\big\|\nabla_{xy}g^{(j)}(x,y_x^{(m)})-\nabla_{xy}g^{(j)}(x,y_x^{(j)})\big\|\\
&\le \frac{C_f\zeta_{g,xy}}{\mu}+\frac{C_fL_{xy}}{\mu}\big\|y_x^{(m)}-y_x^{(j)})\big\| \le \frac{C_f\zeta_{g,xy}}{\mu}+\frac{C_fL_{xy}\zeta_{g^*}}{\mu}
\end{aligned}$$

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:

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

where the first inequality is by Assumption 3.3; the third inequality is by Assumption 3.3, 3.2; the fourth inequality is by Cauchy Schwartz inequality; the last inequality is by Assumption 3.1, 3.3 and the result in Eq. 37. Combine everything together, we have:

$$
\begin{aligned}
\big\|\nabla_x f^{(m)}(x, y_x^{(m)}) - \nabla_x f^{(j)}(x, y_x^{(j)})\big\| \le \zeta_f + L\zeta_{g^*} + \frac{C_f\zeta_{g,xy}}{\mu} + \frac{C_f L_{xy}\zeta_{g^*}}{\mu} + \frac{L(\zeta_f + L\zeta_{g^*})}{\mu} \\
+ \frac{C_f L(\zeta_{g,yy} + L_{y^2}\zeta_{g^*})}{\mu^2}
\end{aligned}
$$

which completes the proof. □

D.1 Proof for the FedBiOAcc-Local Algorithm

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

Lemma D.5. *Suppose we have $c_\nu\alpha_t^2 < 1$, then:*

$$
\begin{aligned}
\mathbb{E}\big[\big\|\bar{\nu}_t - \mathbb{E}_\xi[\bar{\mu}_{t,\mathcal{B}_x}]\big\|^2\big] \le (1 - c_\nu\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{\nu}_{t-1} - \mathbb{E}_\xi[\bar{\mu}_{t-1,\mathcal{B}_x}]\big\|^2\big] + \frac{2c_\nu^2\alpha_{t-1}^4}{b_x M}G_2^2 \\
+ \frac{2\hat{L}^2}{b_x M^2}\sum_{m=1}^{M}\mathbb{E}\big[\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \big\|y_t^{(m)} - y_{t-1}^{(m)}\big\|^2\big]
\end{aligned}
$$

where $\mu_{t,\xi}^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}; \mathcal{B}_x)$ and the expectation outside is w.r.t *all the stochasity of the algorithm.*

Proof. For ease of notation, we denote $\mu_{t,\xi}^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)}; \xi_x)$, and $\mu_t^{(m)} = \Phi^{(m)}(x_t^{(m)}, y_t^{(m)})$, then by the definition of $\bar{\nu}_t$ we have:

$$
\begin{aligned}
&\mathbb{E}\big[\big\|\bar{\nu}_t - \mathbb{E}_\xi[\bar{\mu}_{t,\mathcal{B}_x}]\big\|^2\big] = \mathbb{E}\big[\big\|\frac{1}{M}\sum_{m=1}^{M}\big(\hat{\nu}_t^{(m)} - \mathbb{E}_\xi[\mu_{t,\mathcal{B}_x}^{(m)}]\big)\big\|^2\big] \\
&= \mathbb{E}\big[\big\|\frac{1}{M}\sum_{m=1}^{M}\big(\mu_{t,\mathcal{B}_x}^{(m)} + (1 - c_\nu\alpha_{t-1}^2)(\nu_{t-1}^{(m)} - \mu_{t-1,\mathcal{B}_x}^{(m)}) - \mathbb{E}_\xi[\mu_{t,\mathcal{B}_x}^{(m)}]\big)\big\|^2\big] \\
&= \mathbb{E}\big[\big\|(1 - c_\nu\alpha_{t-1}^2)\big(\bar{\nu}_{t-1} - \mathbb{E}_\xi[\bar{\mu}_{t-1,\mathcal{B}_x}]\big) + \big(\bar{\mu}_{t,\mathcal{B}_x} - \mathbb{E}_\xi[\bar{\mu}_{t,\mathcal{B}_x}] + (1 - c_\nu\alpha_{t-1}^2)(\mathbb{E}_\xi[\bar{\mu}_{t-1,\mathcal{B}_x}] - \bar{\mu}_{t-1,\mathcal{B}_x})\big)\big\|^2\big] \\
&\le (1 - c_\nu\alpha_{t-1}^2)\mathbb{E}\big[\big\|\bar{\nu}_{t-1} - \mathbb{E}_\xi[\bar{\mu}_{t-1,\mathcal{B}_x}]\big\|^2\big] \\
&\quad + \frac{1}{b_x^2 M^2}\sum_{m=1}^{M}\sum_{\xi_x\in\mathcal{B}_x}\mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mathbb{E}_\xi[\mu_{t,\xi_x}^{(m)}] + (1 - c_\nu\alpha_{t-1}^2)(\mathbb{E}_\xi[\mu_{t-1,\xi_x}^{(m)}] - \mu_{t-1,\xi_x}^{(m)})\big\|^2\big]
\end{aligned}
$$

where inequality (a) uses the fact that the cross product term is zero in expectation, the condition that $c_\nu \alpha_t^2 < 1$ and the fact that clients independently choose samples.

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

$$
\begin{aligned}
T_1 &\overset{(a)}{\leq} 2(c_\nu \alpha_{t-1}^2)^2 \mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mathbb{E}_\xi[\mu_{t,\xi_x}^{(m)}]\big\|^2\big] + 2(1-c_\nu\alpha_{t-1}^2)^2 \mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mu_{t-1,\xi_x}^{(m)} - (\mathbb{E}_\xi[\mu_t^{(m)}] - \mathbb{E}_\xi[\mu_{t-1}^{(m)}])\big\|^2\big] \\
&\overset{(b)}{\leq} 2(c_\nu \alpha_{t-1}^2)^2 \mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mathbb{E}_\xi[\mu_t^{(m)}]\big\|^2\big] + 2\mathbb{E}\big[\big\|\mu_{t,\xi_x}^{(m)} - \mu_{t-1,\xi_x}^{(m)}\big\|^2\big] \\
&\overset{(c)}{\leq} 2(c_\nu \alpha_{t-1}^2)^2 G_2^2 + 2\hat{L}^2 \mathbb{E}\big[\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \big\|y_t^{(m)} - y_{t-1}^{(m)}\big\|^2\big]
\end{aligned}
$$

where inequality (a) follows the generalized triangle inequality; (b) follows Proposition E.2 due to the definition of $\mu_t^{(m)}$; (c) follows the smoothness property of $\hat{L}$ and the bounded variance assumption; This completes the proof. □

Lemma D.6. *Suppose we have $c_\omega \alpha_{t-1}^2 < 1$, then we have:*

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

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

$$
\begin{aligned}
\mathbb{E}\big[\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2\big] &\leq (1 - \frac{\mu\gamma\alpha_{t-1}}{4})\mathbb{E}\big[\big\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2\big] - \frac{\gamma^2\alpha_{t-1}}{4}\mathbb{E}\big[\big\|\omega_{t-1}^{(m)}\big\|^2\big] \\
&\quad + \frac{9\gamma\alpha_{t-1}}{2\mu}\mathbb{E}\big[\big\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\big\|^2\big] + \frac{9\kappa^2\eta^2\alpha_{t-1}}{2\mu\gamma}\mathbb{E}\big[\|\nu_{t-1}^{(m)}\|^2\big]
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}\big[\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2\big] &\leq (1 - \frac{\mu\gamma\alpha_{t-1}}{4})\mathbb{E}\big[\big\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2\big] - \frac{\gamma^2\alpha_{t-1}}{4}\mathbb{E}\big[\big\|\omega_{t-1}^{(m)}\big\|^2\big] \\
&\quad + \frac{9\gamma\alpha_{t-1}}{2\mu}\mathbb{E}\big[\big\|\omega_{t-1}^{(m)} - \nabla_y g^{(m)}(x_{t-1}^{(m)}, y_{t-1}^{(m)})\big\|^2\big] \\
&\quad + \frac{9\kappa^2\eta^2\alpha_{t-1}}{\mu\gamma}\mathbb{E}\big[\|\nu_{t-1}^{(m)}\|^2\big] + \frac{9\kappa^2}{\mu\gamma\alpha_{t-1}}\mathbb{E}\big[\|\hat{x}_t^{(m)}) - \bar{x}_t\|^2\big]
\end{aligned}
$$

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

$$
\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 \leq (1 - \frac{\mu\gamma\alpha_t}{2})\|y_{x_t^{(m)}}^{(m)} - y_t^{(m)}\|^2 - \frac{\gamma^2\alpha_t}{4}\|\omega_t^{(m)}\|^2 + \frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2. \tag{38}
$$

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

$$
\begin{aligned}
\|y_{t+1}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)}\|^2 &\leq (1 + \frac{\mu\gamma\alpha_t}{4})\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\|y_{x_t^{(m)}}^{(m)} - y_{x_{t+1}^{(m)}}^{(m)})\|^2 \\
&\leq (1 + \frac{\mu\gamma\alpha_t}{4})\|y_{t+1}^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\kappa^2\|x_t^{(m)} - x_{t+1}^{(m)}\|^2
\end{aligned} \tag{39}
$$

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

$$\|y_{t+1}^{(m)} - y_{\hat{x}_{t+1}^{(m)}}^{(m)})\|^2 \leq (1 + \frac{\mu\gamma\alpha_t}{4})(1 - \frac{\mu\gamma\alpha_t}{2})\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)})\|^2 - (1 + \frac{\mu\gamma\alpha_t}{4})\frac{\gamma^2\alpha_t}{4}\|\omega_t^{(m)}\|^2$$
$$+ (1 + \frac{\mu\gamma\alpha_t}{4})\frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t^{(m)}\|^2 + (1 + \frac{4}{\mu\gamma\alpha_t})\kappa^2\|x_t^{(m)} - x_{t+1}^{(m)}\|^2$$

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

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

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

$$\|y_{t+1}^{(m)} - y_{\hat{x}_{t+1}^{(m)}}^{(m)})\|^2 \leq \big(1 - \frac{\mu\gamma\alpha_t}{4}\big)\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)})\|^2 - \frac{\gamma^2\alpha_t}{4}\|\omega_t^{(m)}\|^2$$
$$+ \frac{9\gamma\alpha_t}{2\mu}\|\nabla_y g(x_t^{(m)}, y_t^{(m)}) - w_t\|^2 + \frac{9\kappa^2}{2\mu\gamma\alpha_t}\underbrace{\|x_t^{(m)} - x_{t+1}^{(m)}\|^2}_{T_1}$$

Note for the term T_1 we have $T_1 = \big\|\eta\alpha_t\nu_t^{(m)}\big\|^2$ for $t+1 \neq \bar{t}_s$ and $T_1 = \big\|\bar{x}_{t+1} - x_t^{(m)}\big\|^2 \leq 2\big\|\hat{x}_{t+1}^{(m)} - \bar{x}_{t+1}\big\|^2 + 2\big\|\eta\alpha_t\nu_t^{(m)}\big\|^2$ for $t+1 = \bar{t}_s$. This completes the proof. □

D.1.3 Upper Variable Drift

Lemma D.8. *For $t \in [\bar{t}_{s-1}+1, \bar{t}_s]$, with $s \in [S]$ we have:*

$$\|\hat{x}_t^{(m)} - \bar{x}_t\|^2 \leq \sum_{\ell=\bar{t}_{s-1}}^{t-1} I\eta^2\alpha_l^2\big\|\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\big\|^2$$

Proof. Since we have $\hat{x}_t^{(m)} = x_{t-1}^{(m)} - \eta\alpha_{t-1}\nu_{t-1}^{(m)}$, this implies that:

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

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

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

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

Lemma D.9. *Suppose $\alpha_t < \frac{1}{16I\hat{L}}$, $\eta < 1$, then for $t \neq \bar{t}_s, s \in [S]$, we have:*

$$\sum_{m=1}^M \mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2 \leq \big(1 + \frac{33}{32I}\big)\sum_{m=1}^M \mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 + 4I\hat{L}^2\alpha_{t-1}^2\sum_{m=1}^M \mathbb{E}\big[2\|\eta\bar{\nu}_{t-1}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\big]$$
$$+ 8IM(c_\nu\alpha_{t-1}^2)^2G_1^2 + \frac{8IM(c_\nu\alpha_{t-1}^2)^2G_2^2}{b_x} + 16IM(c_\nu\alpha_{t-1}^2)^2\zeta^2$$
$$+ 128I\hat{L}^2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^M \mathbb{E}\big[\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\|^2\big]$$
$$+ 128I\bar{L}^2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^M\sum_{\ell=\bar{t}_{s-1}}^{t-2} I\eta^2\alpha_l^2\mathbb{E}\big\|\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\big\|^2$$

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

Proof. By the update step in Line 7 of Algorithm 1, for $t \neq \bar{t}_s$, we have:

$$\begin{aligned}
\mathbb{E}\|\hat{\nu}_t^{(m)} - \bar{\nu}_t\|^2 &= \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} + (1-c_\nu\alpha_{t-1}^2)\big(\nu_{t-1}^{(m)} - \mu_{t-1,\mathcal{B}_x}^{(m)}\big) - \big(\bar{\mu}_{t,\mathcal{B}_x} + (1-c_\nu\alpha_{t-1}^2)\big(\bar{\nu}_{t-1} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big)\big\|^2 \\
&= \mathbb{E}\big\|(1-c_\nu\alpha_{t-1}^2)\big(\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\big) + \mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 \\
&\overset{(a)}{\leq} (1+\frac{1}{I})(1-c_\nu\alpha_{t-1}^2)^2\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 \\
&\quad + \big(1+I\big)\mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 \\
&\leq \left(1+\frac{1}{I}\right)\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 + \big(1+I\big)\mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2
\end{aligned} \tag{40}$$

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

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

$$\begin{aligned}
&\sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - (1-c_\nu\alpha_{t-1}^2)\big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 \\
&\leq 2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t,\mathcal{B}_x} - \big(\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big)\big\|^2 + 2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big\|^2 \\
&\leq 2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \mu_{t-1,\mathcal{B}_x}^{(m)}\big\|^2 + 2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big\|^2
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{m=1}^M \mathbb{E}\big\|\mu_{t,\mathcal{B}_x}^{(m)} - \mu_{t-1,\mathcal{B}_x}^{(m)}\big\|^2 \leq \hat{L}^2\sum_{m=1}^M \mathbb{E}\big[\|x_t^{(m)} - x_{t-1}^{(m)}\|^2 + \|y_t^{(m)} - y_{t-1}^{(m)}\|^2\big] \\
&\leq \hat{L}^2\alpha_{t-1}^2\sum_{m=1}^M \mathbb{E}\big[\|\eta\nu_{t-1}^{(m)}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\big]
\end{aligned} \tag{41}$$

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

Next for the second term, we have:

$$\begin{aligned}
&\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \bar{\mu}_{t-1,\mathcal{B}_x}\big\|^2 = \sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \mu_{t-1}^{(m)} - \big(\bar{\mu}_{t-1,\mathcal{B}_x} - \bar{\mu}_{t-1}\big) + \mu_{t-1}^{(m)} - \bar{\mu}_{t-1}\big\|^2 \\
&\overset{(a)}{\leq} 2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \mu_{t-1}^{(m)} - \big(\bar{\mu}_{t-1,\mathcal{B}_x} - \bar{\mu}_{t-1}\big)\big\|^2 + 2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1}^{(m)} - \bar{\mu}_{t-1}\big\|^2 \\
&\overset{(b)}{\leq} \underbrace{2\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \mu_{t-1}^{(m)}\big\|^2}_{T_1} + \underbrace{4\sum_{m=1}^M \mathbb{E}\big\|\nabla h^{(m)}(\bar{x}_{t-1}) - \nabla h(\bar{x}_{t-1})\big\|^2}_{T_2} \\
&\quad + \underbrace{4\sum_{m=1}^M \mathbb{E}\big\|\mu_{t-1}^{(m)} - \nabla h^{(m)}(\bar{x}_{t-1}) + \nabla h(\bar{x}_{t-1}) - \bar{\mu}_{t-1}\big\|^2}_{T_3}
\end{aligned} \tag{42}$$

Note for the term T_1 of Eq. 42, we have $\mathbb{E}\big\|\mu_{t-1,\mathcal{B}_x}^{(m)} - \mu_{t-1}^{(m)}\big\|^2 \leq G_1^2 + \frac{G_2^2}{b_x}$; For the term T_2 of Eq. 42, by the bounded intra-node heterogeneity assumption we have:

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

Finally, For the term T_3 of Eq. 42

$$T_3 \le 8\sum_{m=1}^{M}\mathbb{E}\big\|\mu_{t-1}^{(m)} - \nabla h^{(m)}(\bar{x}_{t-1})\big\|^2 + 8\sum_{m=1}^{M}\mathbb{E}\big\|\nabla h(\bar{x}_{t-1}) - \bar{\mu}_{t-1}\big\|^2 \le 16\sum_{m=1}^{M}\mathbb{E}\big\|\mu_{t-1}^{(m)} - \nabla h^{(m)}(\bar{x}_{t-1})\big\|^2$$
$$\le 32\sum_{m=1}^{M}\mathbb{E}\big[\big\|\mu_{t-1}^{(m)} - \nabla h^{(m)}(x_{t-1}^{(m)})\big\|^2 + \big\|\nabla h^{(m)}(x_{t-1}^{(m)}) - \nabla h^{(m)}(\bar{x}_{t-1})\big\|^2\big]$$
$$\overset{(a)}{\le} 32\bar{L}^2\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2\big] + 32\hat{L}^2\sum_{m=1}^{M}\mathbb{E}\big[\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\|^2\big]$$

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

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

$$\sum_{m=1}^{M}\mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$$
$$\le \big(1+\frac{1}{I}\big)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 + 4I\hat{L}^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[\underbrace{\|\eta\nu_{t-1}^{(m)}\|^2}_{T_1} + \|\gamma\omega_{t-1}^{(m)}\|^2\big]$$
$$+ 8IM(c_\nu\alpha_{t-1}^2)^2G_1^2 + \frac{8IM(c_\nu\alpha_{t-1}^2)^2G_2^2}{b_x} + 16IM(c_\nu\alpha_{t-1}^2)^2\zeta^2$$
$$+ 128I\bar{L}^2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2\big] + 128I\hat{L}^2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\|^2\big]$$

We separate the term T_1 with triangle inequality to get:

$$\sum_{m=1}^{M}\mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$$
$$\le \left(1+\frac{1}{I}+8I\hat{L}^2\eta^2\alpha_{t-1}^2\right)\sum_{m=1}^{M}\mathbb{E}\|\nu_{t-1}^{(m)} - \bar{\nu}_{t-1}\|^2 + 4I\hat{L}^2\alpha_{t-1}^2\sum_{m=1}^{M}\mathbb{E}\big[2\|\eta\bar{\nu}_{t-1}\|^2 + \|\gamma\omega_{t-1}^{(m)}\|^2\big]$$
$$+ 8IM(c_\nu\alpha_{t-1}^2)^2G_1^2 + \frac{8IM(c_\nu\alpha_{t-1}^2)^2G_2^2}{b_x} + 16IM(c_\nu\alpha_{t-1}^2)^2\zeta^2$$
$$+ \underbrace{128I\bar{L}^2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|x_{t-1}^{(m)} - \bar{x}_{t-1}\|^2\big]}_{T_1} + 128I\hat{L}^2(c_\nu\alpha_{t-1}^2)^2\sum_{m=1}^{M}\mathbb{E}\big[\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\|^2\big]$$

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

Next, to simply the notation, we denote $A_t = \mathbb{E}\|\bar{\nu}_t - \mathbb{E}_\xi[\bar{\mu}_{t,\mathcal{B}_x}]\|^2$, $B_t = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2$, $C_t = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\|^2$, $D_t = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\nu_t^{(m)} - \bar{\nu}_t\|^2$, $E_t = \mathbb{E}\|\bar{\nu}_t\|^2$, $F_t = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\omega_t^{(m)}\|^2$.

Lemma D.10. *For $\alpha_t < \frac{1}{16\hat{L}I}$, we have:*

$$\big(1 - \frac{3\kappa^2\eta^2c_\nu^2}{4*16^3I^5\hat{L}^4}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \le \frac{3c_\nu^2}{2*16^2I^4\hat{L}^2}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell B_\ell + \frac{3\eta^2}{32I}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell E_\ell + \frac{3\gamma^2}{64I}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell F_\ell$$
$$+ \left(\frac{3c_\nu^2G_1^2}{32I\hat{L}^2} + \frac{3c_\nu^2G_2^2}{32Ib_x\hat{L}^2} + \frac{3c_\nu^2\zeta^2}{16I\hat{L}^2}\right)\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell^3$$

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:

$$D_t \leq \big(1+\frac{33}{32I}\big)D_{t-1} + 128I\hat{L}^2c_\nu^2\alpha_{t-1}^4B_{t-1} + 8I\hat{L}^2\alpha_{t-1}^2\eta^2E_{t-1} + 4I\hat{L}^2\alpha_{t-1}^2\gamma^2F_{t-1}$$
$$+ 8Ic_\nu^2\alpha_{t-1}^4G_1^2 + \frac{8Ic_\nu^2\alpha_{t-1}^4G_2^2}{b_x} + 16Ic_\nu^2\alpha_{t-1}^4\zeta^2 + 128I^2\bar{L}^2\eta^2c_\nu^2\alpha_{t-1}^4\sum_{\ell=\bar{t}_{s-1}}^{t-2}\alpha_l^2D_l$$

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

$$D_t \leq \sum_{\ell=\bar{t}_{s-1}}^{t-1}\big(1+\frac{33}{32I}\big)^{t-\ell}\big(128I\hat{L}^2c_\nu^2\alpha_\ell^4B_\ell + 8I\hat{L}^2\eta^2\alpha_\ell^2E_\ell + 4I\hat{L}^2\gamma^2\alpha_\ell^2F_\ell + 8Ic_\nu^2G_1^2\alpha_\ell^4 + \frac{8Ic_\nu^2G_2^2\alpha_\ell^4}{b_x}$$
$$+ 16Ic_\nu^2\zeta^2\alpha_\ell^4 + 128I^2\bar{L}^2\eta^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell}\alpha_{\bar{\ell}}^2D_{\bar{\ell}}\big)$$
$$\leq \sum_{\ell=\bar{t}_{s-1}}^{t-1}\big(384I\hat{L}^2c_\nu^2\alpha_\ell^4B_\ell + 24I\hat{L}^2\eta^2\alpha_\ell^2E_\ell + 12I\hat{L}^2\gamma^2\alpha_\ell^2F_\ell + 24Ic_\nu^2G_1^2\alpha_\ell^4 + \frac{24Ic_\nu^2G_2^2\alpha_\ell^4}{b_x}$$
$$+ 16Ic_\nu^2\zeta^2\alpha_\ell^4 + 384I^2\bar{L}^2\eta^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell}\alpha_{\bar{\ell}}^2D_{\bar{\ell}}\big)$$

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

$$\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s}\alpha_tD_t \leq \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t\sum_{\ell=\bar{t}_{s-1}}^{t-1}\big(384I\hat{L}^2c_\nu^2\alpha_\ell^4B_\ell + 24I\hat{L}^2\eta^2\alpha_\ell^2E_\ell + 12I\hat{L}^2\gamma^2\alpha_\ell^2F_\ell$$
$$+\left(24Ic_\nu^2G_1^2 + \frac{24Ic_\nu^2G_1^2}{b_x} + 48Ic_\nu^2\zeta^2\right)\alpha_\ell^4 + 384I^2\bar{L}^2\eta^2c_\nu^2\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell}\alpha_{\bar{\ell}}^2D_{\bar{\ell}}\big)$$
$$\overset{(a)}{\leq}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(24I^{1/2}\hat{L}c_\nu^2\alpha_\ell^4B_\ell + \frac{3I^{1/2}\hat{L}\eta^2}{2}\alpha_\ell^2E_\ell + \frac{3I^{1/2}\hat{L}\gamma^2}{4}\alpha_\ell^2F_\ell$$
$$+\left(\frac{3I^{1/2}c_\nu^2G_1^2}{2\hat{L}} + \frac{3I^{1/2}c_\nu^2G_2^2}{2b_x\hat{L}} + \frac{3I^{1/2}c_\nu^2\zeta^2}{\hat{L}}\right)\alpha_\ell^4 + \frac{24I^{3/2}\bar{L}^2\eta^2c_\nu^2}{\hat{L}}\alpha_\ell^4\sum_{\bar{\ell}=\bar{t}_{s-1}}^{\ell}\alpha_{\bar{\ell}}^2D_{\bar{\ell}}\big)$$
$$\overset{(b)}{\leq}\frac{3c_\nu^2}{2*16^2I^4\hat{L}^2}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell B_\ell + \frac{3\eta^2}{32I}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell E_\ell + \frac{3\gamma^2}{64I}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell F_\ell$$
$$+\left(\frac{3c_\nu^2G_1^2}{32I\hat{L}^2} + \frac{3c_\nu^2G_2^2}{32Ib_x\hat{L}^2} + \frac{3c_\nu^2\zeta^2}{16I\hat{L}^2}\right)\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell^3 + \frac{3\kappa^2\eta^2c_\nu^2}{4*16^3I^5\hat{L}^4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_tD_t$$

In inequalities (a) and (b), we use $\alpha_t < \frac{1}{16\hat{L}I^{3/2}}$. Note that $\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s}\alpha_tD_t = \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_tD_t$ as $D_{\bar{t}_s} = D_{\bar{t}_{s-1}} = 0$, so we have:

$$\big(1-\frac{3\kappa^2\eta^2c_\nu^2}{4*16^3I^5\hat{L}^4}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_tD_t \leq \frac{3c_\nu^2}{2*16^2I^4\hat{L}^2}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell B_\ell + \frac{3\eta^2}{32I}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell E_\ell + \frac{3\gamma^2}{64I}\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell F_\ell$$
$$+\left(\frac{3c_\nu^2G_1^2}{32I\hat{L}^2} + \frac{3c_\nu^2G_2^2}{32Ib_x\hat{L}^2} + \frac{3c_\nu^2\zeta^2}{16I\hat{L}^2}\right)\sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_\ell^3$$

This completes the proof. □

D.1.4 Descent Lemma

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

$$\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta\alpha_t}{4}\mathbb{E}[\|\bar{\nu}_t\|^2] - \frac{\eta\alpha_t}{2}\mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + 2\eta\alpha_t\mathbb{E}[\|\mathbb{E}_\xi[\bar{\mu}_{t,\mathcal{B}_x}] - \bar{\nu}_t\|^2] + 4\eta\alpha_t G_1^2$$
$$+ \frac{\bar{L}^2 I\eta^3\alpha_t}{M}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2\sum_{m=1}^{M}\mathbb{E}\|\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\|^2 + \frac{4\hat{L}^2\eta\alpha_t}{M}\sum_{m=1}^{M}\mathbb{E}\|y^{(m)}_{x_t^{(m)}} - y_t^{(m)}\|^2$$

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

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

$$\begin{aligned}
\mathbb{E}[h(\bar{x}_{t+1})] &\leq \mathbb{E}\big[h(\bar{x}_t) + \langle\nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{\bar{L}}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2\big] \\
&\overset{(a)}{=} \mathbb{E}\big[h(\bar{x}_t) - \eta\alpha_t\langle\nabla h(\bar{x}_t), \bar{\nu}_t\rangle + \frac{\eta^2\alpha_t^2\bar{L}}{2}\|\bar{\nu}_t\|^2\big] \\
&\overset{(b)}{=} \mathbb{E}\big[h(\bar{x}_t) - \frac{\eta\alpha_t}{2}\big\|\bar{\nu}_t\big\|^2 - \frac{\eta\alpha_t}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta\alpha_t}{2}\big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\big\|^2 + \frac{\eta\alpha_t^2\bar{L}}{2}\big\|\bar{\nu}_t\big\|^2\big] \\
&= \mathbb{E}\big[h(\bar{x}_t) - \frac{\eta\alpha_t}{4}\big\|\bar{\nu}_t\big\|^2 - \frac{\eta\alpha_t}{2}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta\alpha_t}{2}\underbrace{\big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\big\|^2}_{T_1}\big]
\end{aligned}$$

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 $\eta\alpha_t < \frac{1}{2\bar{L}}$; For the term T_1, we have:

$$\mathbb{E}\big[\big\|\nabla h(\bar{x}_t) - \bar{\nu}_t\big\|^2\big] \leq 2\mathbb{E}\big[\big\|\nabla h(\bar{x}_t) - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\big\|^2\big] + 4\mathbb{E}\big[\big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \mathbb{E}_\xi[\bar{\mu}_{t,\mathcal{B}_x}]\big\|^2\big]$$
$$+ 4\mathbb{E}\big[\big\|\mathbb{E}_\xi[\bar{\mu}_{t,\mathcal{B}_x}] - \bar{\nu}_t\big\|^2\big]$$

For the first term, we have:

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

where the last inequality uses Lemma D.8. For the second term, we have:

$$4\mathbb{E}\big[\big\|\frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)}) - \mathbb{E}_\xi[\bar{\mu}_{t,\xi}]\big\|^2\big] \leq \frac{8}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|\nabla h(x_t^{(m)}) - \mu_t^{(m)}\big\|^2\big] + \frac{8}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|\mu_t^{(m)} - \mathbb{E}_\xi[\mu_{t,\mathcal{B}_x}^{(m)}]\big\|^2\big]$$
$$\leq \frac{8\hat{L}^2}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|y^{(m)}_{x_t^{(m)}} - y_t^{(m)}\big\|^2\big] + 8G_1^2$$

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 $\mathcal{G}(t)$:

$$\mathcal{G}_t = h(\bar{x}_t) + \frac{9bM\eta}{16\alpha_t}\big\|\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla h(x_t^{(m)})\big\|^2 + \frac{18\eta\hat{L}^2}{\mu\gamma}\times\frac{1}{M}\sum_{m=1}^{M}\big\|y_t^{(m)} - y^{(m)}_{x_t^{(m)}}\big\|^2$$
$$+ \frac{9bM\hat{L}^2\eta}{16L^2\alpha_t}\times\frac{1}{M}\sum_{m=1}^{M}\big\|\omega_t^{(m)} - \nabla_y g^{(m)}(x_t^{(m)}, y_t^{(m)})\big\|^2$$

Theorem D.12. *Suppose* $\gamma \le \frac{1}{2L}$, $\eta < \min\big(\frac{\mu\gamma}{144\kappa\hat{L}}, \frac{\hat{L}^2}{C_1^{1/2}c_\nu}, \frac{1}{(C_1 I)^{1/2}}, \frac{\hat{L}}{(C_1 I)^{1/2}}, \frac{I\hat{L}^2}{\kappa c_\nu}, 1\big)$, $c_\nu = \frac{32}{9bM} + \frac{\hat{L}}{24Ib^2M^2}$, $c_\omega = \frac{144L^2}{bM\mu^2} + \frac{\hat{L}}{24Ib^2M^2}$, $u = (bM\sigma)^2\bar{u}$, *where* $\bar{u} = \max\big(2, 16^3 I^{9/2}\hat{L}, c_\nu^{3/2}, c_\omega^{3/2}\big)$, $\delta = \frac{(bM\sigma)^{2/3}}{\hat{L}^{2/3}}$, *then we have:*

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

To reach an ϵ-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 $Q = O(\kappa\log(\frac{\kappa}{bM\epsilon}))$.

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

$$\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \le \big(\alpha_{t-1}^{-1} - \alpha_{t-2}^{-1} - c_\nu\alpha_{t-1}\big)A_{t-1} + \frac{2c_\nu^2\alpha_{t-1}^3 G_2^2}{bM} + \frac{4\hat{L}^2\eta^2\alpha_{t-1}}{bM}(D_{t-1}+E_{t-1}) + \frac{2\hat{L}^2\gamma^2\alpha_{t-1}F_{t-1}}{bM}$$

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

$$\begin{aligned}\alpha_t^{-1} - \alpha_{t-1}^{-1} &= \frac{(u+\sigma^2 t)^{1/3}}{\delta} - \frac{(u+\sigma^2(t-1))^{1/3}}{\delta} \overset{(a)}{\le} \frac{\sigma^2}{3\delta(u+\sigma^2(t-1))^{2/3}} \\ &\overset{(b)}{\le} \frac{2^{2/3}\sigma^2\delta^2}{3\delta^3(u+\sigma^2 t)^{2/3}} \overset{(c)}{=} \frac{2^{2/3}\sigma^2}{3\delta^3}\alpha_t^2 \le \frac{2\hat{L}^2}{3M^2}\alpha_t^2 \le \frac{\hat{L}}{24Ib^2M^2}\alpha_t\end{aligned}$$

where inequality (a) results from the concavity of $x^{1/3}$ as: $(x+y)^{1/3} - x^{1/3} \le y/3x^{2/3}$, inequality (b) used the fact that $u_t \ge 2\sigma^2$, inequality (c) uses the definition of α_t. By choosing $c_\nu = \frac{32}{9bM} + \frac{\hat{L}}{24Ib^2M^2}$, we have:

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

When $t = \bar{t}_s$, by Lemma D.5 and Lemma D.8, we have:

$$\begin{aligned}\frac{A_t}{\alpha_{t-1}} - \frac{A_{t-1}}{\alpha_{t-2}} \le -\frac{32}{9bM}\alpha_{t-1}A_{t-1} + \frac{2c_\nu^2\alpha_{t-1}^3G_2^2}{bM} + \frac{8\hat{L}^2\eta^2\alpha_{t-1}}{bM}(D_{t-1}+E_{t-1}) \\ + \frac{2\hat{L}^2\gamma^2\alpha_{t-1}}{bM}F_{t-1} + \frac{8\hat{L}^2}{bM}\sum_{\ell=\bar{t}_{s-1}}^{t-1} I\eta^2\alpha_\ell D_\ell\end{aligned}$$

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

$$\begin{aligned}\frac{\alpha_t}{\alpha_{\bar{t}_s-1}} &= \frac{(u_{\bar{t}_s-1}+\sigma^2(\bar{t}_s-1))^{1/3}}{(u_t+\sigma^2 t)^{1/3}} = \big(1 + \frac{u_{\bar{t}_s-1} - u_t + \sigma^2(\bar{t}_s-1-t)}{u_t+\sigma^2 t}\big)^{1/3} \\ &\le \big(1+\frac{(I-1)\sigma^2}{u_t+\sigma^2 t}\big)^{1/3} \le 1 + \frac{(I-1)}{3(t+I+1)} \le 2\end{aligned}$$

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

$$\begin{aligned}\big(\frac{A_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}} - \frac{A_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}}\big) \le -\frac{32}{9bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t A_t + \frac{2c_\nu^2G_2^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + \frac{16I\hat{L}^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \\ + \frac{8\hat{L}^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + \frac{2\hat{L}^2\gamma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t \qquad (43)\end{aligned}$$

Next, we follow similar derivation as $A_t/\alpha_{t-1} - A_{t-1}/\alpha_{t-2}$. By Lemma D.6, For $t \ne \bar{t}_s$, we choose $c_\omega = \frac{144L^2}{bM\mu^2} + \frac{\hat{L}}{24Ib^2M^2}$, to obtain:

$$\frac{C_t}{\alpha_{t-1}} - \frac{C_{t-1}}{\alpha_{t-2}} \le -\frac{144L^2\alpha_{t-1}}{bM\mu^2}C_{t-1} + \frac{2c_\omega^2\alpha_{t-1}^3\sigma^2}{bM} + \frac{4L^2\eta^2\alpha_{t-1}}{bM}(D_{t-1}+E_{t-1}) + \frac{2L^2\gamma^2}{bM}\alpha_{t-1}F_{t-1}$$

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

$$\frac{C_t}{\alpha_{t-1}} - \frac{C_{t-1}}{\alpha_{t-2}} \leq -\frac{144L^2\alpha_{t-1}}{bM\mu^2}C_{t-1} + \frac{2c_\omega^2\alpha_{t-1}^3\sigma^2}{bM} + \frac{8L^2\eta^2\alpha_{t-1}}{bM}(D_{t-1}+E_{t-1}) + \frac{2L^2\gamma^2\alpha_{t-1}}{bM}F_{t-1} + \frac{4L^2}{bM}\sum_{\ell=\bar{t}_{s-1}}^{t-1} I\eta^2\alpha_\ell D_\ell$$

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

$$\left(\frac{C_{\bar{t}_s}}{\alpha_{\bar{t}_s-1}} - \frac{C_{\bar{t}_{s-1}}}{\alpha_{\bar{t}_{s-1}-1}}\right) \leq -\frac{144L^2}{2bM\mu^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t + \frac{2c_\omega^2\sigma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + \frac{16IL^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + \frac{8L^2\eta^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + \frac{2L^2\gamma^2}{bM}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t. \tag{44}$$

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

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

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

$$B_t - B_{t-1} \leq -\frac{\mu\gamma\alpha_{t-1}B_{t-1}}{4} - \frac{\gamma^2\alpha_{t-1}F_{t-1}}{4} + \frac{9\gamma\alpha_{t-1}C_{t-1}}{2\mu} + \frac{18\kappa^2\eta^2\alpha_{t-1}D_{t-1}}{\mu\gamma} + \frac{18\kappa^2\eta^2\alpha_{t-1}E_{t-1}}{\mu\gamma} + \frac{9\kappa^2 I\eta^2\alpha_{t-1}}{\mu\gamma}\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_\ell D_\ell$$

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

$$B_{\bar{t}_s} - B_{\bar{t}_{s-1}} \leq -\frac{\mu\gamma}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t - \frac{\gamma^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t + \frac{9\gamma}{2\mu}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t C_t + \frac{36I\kappa^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + \frac{18\kappa^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t \tag{45}$$

Next, by Lemma D.11, we have:

$$\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta\alpha_t}{4}E_t - \frac{\eta\alpha_t}{2}\mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + \bar{L}^2 I\eta^3\alpha_t\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2 D_l + 2\eta\alpha_t A_t + 4\hat{L}^2\eta\alpha_t B_t + 4\eta\alpha_t G_1^2$$

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

$$\begin{aligned}\mathbb{E}[h(\bar{x}_{\bar{t}_s}) - h(\bar{x}_{\bar{t}_{s-1}})] &\leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{4}E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + 4\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_1^2 \\ &\quad + \bar{L}^2 I\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t\sum_{\ell=\bar{t}_{s-1}}^{t-1}\alpha_l^2 D_l + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}2\eta\alpha_t A_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}4\hat{L}^2\eta\alpha_t B_t \\ &\leq -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{4}E_t - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + 4\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_1^2 \\ &\quad + \frac{\kappa^2\eta^3}{64}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}2\eta\alpha_t A_t + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}4\hat{L}^2\eta\alpha_t B_t\end{aligned} \tag{46}$$

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

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

$$\begin{aligned}
\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2}\mathbb{E}[\|\nabla h(\bar{x}_t)\|^2] + \big(\frac{9\hat{L}^2\eta c_\omega^2\sigma^2}{8L^2} + \frac{9\eta c_\nu^2 G_2^2}{8}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 - \frac{\hat{L}^2\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t \\
& - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\frac{9\eta\gamma^2\hat{L}^2}{2} - \frac{9\eta\gamma^2\hat{L}^2}{8} - \frac{9\eta\gamma\hat{L}^2}{8\mu}\big)\alpha_t F_t \\
& - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\big(\frac{1}{4} - \frac{324\kappa^2\hat{L}^2\eta^2}{\mu^2\gamma^2} - \frac{9\hat{L}^2\eta^2}{2} - \frac{9\hat{L}^2\eta^2}{2}\big)\eta\alpha_t E_t \\
& + \big(\frac{\kappa^2}{64} + \frac{648I\kappa^2\hat{L}^2}{\mu^2\gamma^2} + 9I\hat{L}^2 + 9I\hat{L}^2\big)\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t + 4\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_1^2
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1} \frac{\eta\alpha_t}{2}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] + \big(\frac{9\eta c_\omega^2\sigma^2}{8\mu\gamma} + \frac{9\eta c_\nu^2 G_2^2}{8\mu\gamma}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3 + 4\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_1^2 \\
& - \frac{9\eta\gamma^2\hat{L}^2}{4}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t - \frac{3\eta}{16}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t - \frac{\hat{L}^2\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t + C_1 I\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t
\end{aligned} \tag{47}$$

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

$$\begin{aligned}
\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t D_t \leq & \frac{c_\nu^2}{128I^4\hat{L}^2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t B_t + \frac{\eta^2}{8I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t E_t + \frac{\gamma^2}{16I}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t F_t \\
& + \big(\frac{c_\nu^2 G_1^2}{8I\hat{L}^2} + \frac{c_\nu^2 G_2^2}{8Ib\hat{L}^2} + \frac{c_\nu^2\zeta^2}{4I\hat{L}^2}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3
\end{aligned} \tag{48}$$

Combine Eq. (47) and Eq. (48), and use the condition that $\eta < \min\big(\frac{\hat{L}^2}{C_1^{1/2}c_\nu}, \frac{1}{C_1^{1/2}}, \frac{\hat{L}}{C_1^{1/2}}, 1\big)$, the fact that $I \geq 1$, we have:

$$\begin{aligned}
\mathbb{E}[\mathcal{G}_{\bar{t}_s}] - \mathbb{E}[\mathcal{G}_{\bar{t}_{s-1}}] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta\alpha_t}{2}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] + 4\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t G_1^2 \\
& + \eta\big(\frac{9\hat{L}^2 c_\omega^2\sigma^2}{8L^2} + \frac{9c_\nu^2 G_2^2}{8} + \frac{c_\nu^2 G_1^2}{8} + \frac{c_\nu^2 G_2^2}{8b} + \frac{c_\nu^2\zeta^2}{4}\big)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\alpha_t^3
\end{aligned}$$

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

$$\mathbb{E}[\mathcal{G}_T] - \mathbb{E}[\mathcal{G}_1] \leq -\sum_{t=1}^{T-1}\frac{\eta\alpha_t}{2}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] + \big(\eta C_{\sigma,\zeta} + \frac{4\eta G_1^2}{\alpha_T^2}\big)\sum_{t=1}^{T-1}\alpha_t^3$$

For ease of notation, we denote $C_{\sigma,\zeta} = \big(\frac{9\hat{L}^2 c_\omega^2\sigma^2}{8L^2} + \frac{9c_\nu^2 G_2^2}{8} + \frac{c_\nu^2 G_1^2}{8} + \frac{c_\nu^2 G_2^2}{8b} + \frac{c_\nu^2\zeta^2}{4}\big)$. Rearranging the terms and use the fact that α_t is non-increasing, we have:

$$\begin{aligned}
\frac{\eta\alpha_T}{2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] \leq & \mathbb{E}[\mathcal{G}_1] - \mathbb{E}[\mathcal{G}_T] + \big(\eta C_{\sigma,\zeta} + \frac{4\eta G_1^2}{\alpha_T^2}\big)\sum_{t=1}^{T-1}\alpha_t^3 \\
\leq & h(x_1) - h^* + \frac{9bM\eta A_1}{16\alpha_1} + \frac{18\eta\hat{L}^2 B_1}{\mu\gamma} + \frac{9bM\eta C_1}{16\alpha_1} + \big(\eta C_{\sigma,\zeta} + \frac{4\eta G_1^2}{\alpha_T^2}\big)\sum_{t=1}^{T-1}\alpha_t^3
\end{aligned}$$

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

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

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 initial inner variable estimation error *i.e.* $B_1 = \frac{1}{M} \sum_{m=1}^{M} \|y_1^{(m)} - y_{x_1^{(m)}}^{(m)}\|^2 \leq \Delta_y$, and we assume $\Delta_y = O(\kappa^{-1})$, furthermore, we have:

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

and

$$C_1 = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\|\omega_1^{(m)} - \nabla_y g^{(m)}(x_1^{(m)}, y_1^{(m)})\|^2 \leq \frac{\sigma^2}{b_1}$$

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 $\eta \alpha_T T/2$ to have:

$$\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] \leq \frac{2\Delta}{\eta \alpha_T T} + \frac{9b\sigma^2}{8b_1 T \alpha_1 \alpha_T} + \frac{36\hat{L}^2 \Delta_y}{\kappa \mu \gamma T \alpha_T} + \frac{9b\sigma^2}{8b_1 T \alpha_1 \alpha_T} + \frac{2b^2 M^2 C_{\sigma,\zeta} \ln(T)}{\hat{L}^2 T \alpha_T} + \frac{8b^2 M^2 G_1^2 \ln(T)}{\hat{L}^2 T \alpha_T^3}$$

Note that we have:

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

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

$$\frac{1}{\alpha_1} \leq \frac{u^{1/3} + \sigma^{2/3}}{\delta} = \frac{\hat{L}^{2/3}(\bar{u}^{1/3}(bM)^{2/3} + 1)}{(bM)^{2/3}}$$

when $t = T$, we have:

$$\frac{1}{\alpha_T T} \leq \frac{u^{1/3}}{\delta T} + \frac{\sigma^{2/3}}{\delta T^{2/3}} = \frac{\hat{L}^{2/3} \bar{u}^{1/3}}{T} + \frac{\hat{L}^{2/3}}{(bMT)^{2/3}}$$

In summary, we have:

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}\big[\|\nabla h(\bar{x}_t)\|^2\big] \leq \big(&\frac{2\Delta}{\eta} + \frac{9b\sigma^2}{8b_1\alpha_1} + \frac{36\hat{L}^2\Delta_y}{\kappa\mu\gamma} + \frac{9b\sigma^2}{8b_1\alpha_1} \\ &+ 2\ln(T)\big(\frac{b^2M^2C_{\sigma,\zeta}}{\hat{L}^2} + \frac{4b^2M^2G_1^2}{\hat{L}^2\alpha_T^2}\big)\big)\big(\frac{\hat{L}^{2/3}\bar{u}^{1/3}}{T} + \frac{\hat{L}^{2/3}}{(bMT)^{2/3}}\big)\end{aligned}$$

Note that $\hat{L} = O(\kappa^2)$, $\bar{L} = O(\kappa^3)$, $c_\nu = O((bM)^{-1}\kappa^2)$ and $c_\omega = O((bM)^{-1}\kappa^2)$, $\bar{u} = O(I^{9/2}\kappa^2 + (bM)^{-3/2}\kappa^3)$, $\alpha_1^{-1} = O(I^{3/2}\kappa^2 + (bM)^{-1/2}\kappa^{7/3})$, then for η, we have:

$$\eta \leq \min\big(\frac{\mu\gamma}{144\kappa\hat{L}}, \frac{\hat{L}^2}{\kappa c_\nu}, \frac{\hat{L}^2}{C_1^{1/2}c_\nu}, \frac{1}{C_1^{1/2}}, \frac{\hat{L}}{C_1^{1/2}}, \frac{1}{2\bar{L}}, 1\big)$$

Recall that $C_1 = \big(\frac{\kappa^2}{64} + \frac{648\kappa^2\hat{L}^2}{\mu^2\gamma^2} + 9\hat{L}^2 + 9\hat{L}^2\big)$, suppose we choose $\gamma = \frac{1}{2\hat{L}}$, then $C_1 = O(\kappa^8)$ and $\eta^{-1} = O(\kappa^4)$, $\mu\gamma = O(\kappa^{-1})$. recall that $C_{\sigma,\zeta} = \big(\frac{9\hat{L}^2c_\omega^2\sigma^2}{8\hat{L}^2} + \frac{9c_\nu^2G_2^2}{8} + \frac{c_\nu^2G_1^2}{8} + \frac{c_\nu^2G_2^2}{8b} + \frac{c_\nu^2\zeta^2}{4}\big)$, 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:

$$\frac{8b^2M^2\ln(T)}{\hat{L}^{1/3}\alpha_T^2}\big(\frac{\bar{u}^{1/3}}{T} + \frac{1}{(bMT)^{2/3}}\big) \leq \frac{16b^2M^2\bar{u}}{T} + 16\left(\frac{b^2M^2\bar{u}}{T}\right)^{2/3} + 16\left(\frac{b^2M^2\bar{u}}{T}\right)^{1/3} + 16 = O(\kappa^3 b^2 M^2 I^{9/2})$$

Then, we have:

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

To reach an ϵ-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 $Q = O(\kappa\log(\frac{\kappa}{bM\epsilon}))$. The communication cost is $E = T/I \geq \kappa^{62/9}(bM)^{-1/3}\epsilon^{-7/6}$, the sample complexity is $Gc(f,\epsilon) = O(M^{-1}\kappa^8\epsilon^{-1.5})$, $Gc(g,\epsilon) = O(\kappa^8\epsilon^{-1.5})$, $Jv(g,\epsilon) = O(\kappa^8\epsilon^{-1.5})$, $Hv(g,\epsilon) = O(\kappa^9\epsilon^{-1.5})$

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

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 $\gamma < \frac{1}{L}$, when $t \neq \bar{t}_s$, we have:*

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

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

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2 \leq (1-\frac{\mu\gamma}{2})\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\big\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2 + \frac{10\kappa^2\eta^2}{\mu\gamma M}\sum_{m=1}^{M}\mathbb{E}\big\|\nu_{t-1}^{(m)}\big\|^2 + \frac{10\kappa^2}{\mu\gamma M}\sum_{m=1}^{M}\mathbb{E}\big\|x_t^{(m)} - \bar{x}_t\big\|^2 + \frac{3\gamma^2\sigma^2}{b_y}$$

Proof. First, we have:

$$\begin{aligned}
\mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2 &\leq (1+\frac{\mu\gamma}{2})\mathbb{E}\big\|y_t^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2 + (1+\frac{2}{\mu\gamma})\mathbb{E}\big\|y_{x_t^{(m)}}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2 \\
&\leq (1+\frac{\mu\gamma}{2})(1-\mu\gamma)\mathbb{E}\big\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2 + (1+\frac{2}{\mu\gamma})\mathbb{E}\big\|y_{x_t^{(m)}}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2 + (1+\frac{\mu\gamma}{2})\frac{2\gamma^2\sigma^2}{b_y} \\
&\leq (1-\frac{\mu\gamma}{2})\mathbb{E}\big\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2 + \frac{3\kappa^2}{\mu\gamma}\mathbb{E}\big\|x_t^{(m)} - x_{t-1}^{(m)}\big\|^2 + \frac{3\gamma^2\sigma^2}{b_y}
\end{aligned}$$

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

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

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

$$\mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2 \leq (1-\frac{\mu\gamma}{2})\mathbb{E}\big\|y_{t-1}^{(m)} - y_{x_{t-1}^{(m)}}^{(m)}\big\|^2 + \frac{10\kappa^2\eta^2}{\mu\gamma}\mathbb{E}\big\|v_{t-1}^{(m)}\big\|^2 + \frac{10\kappa^2}{\mu\gamma}\mathbb{E}\|x_t^{(m)} - \bar{x}_t\|^2 + \frac{3\gamma^2\sigma^2}{b_y}$$

Average over all clients, we get the claim in the lemma. □

D.2.2 Upper Variable Drift

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

$$\|x_t^{(m)} - \bar{x}_t\|^2 \leq I\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \left\|\nu_\ell^{(m)} - \bar{\nu}_\ell\right\|^2$$

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: $x_t^{(m)} = x_{t-1}^{(m)} - \eta\nu_{t-1}^{(m)}$, this implies that: $x_t^{(m)} = x_{\bar{t}_{s-1}}^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\nu_\ell^{(m)}$ and $\bar{x}_t = \bar{x}_{\bar{t}_{s-1}} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\bar{\nu}_\ell$. So for $t \neq \bar{t}_s$, with $s \in [S]$ we have:

$$\|x_t^{(m)} - \bar{x}_t\|^2 = \left\|x_{\bar{t}_{s-1}}^{(m)} - \bar{x}_{\bar{t}_{s-1}} - \Big(\sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\nu_\ell^{(m)} - \sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\bar{\nu}_\ell\Big)\right\|^2 \overset{(a)}{=} \left\|\sum_{\ell=\bar{t}_{s-1}}^{t-1} \eta\big(\nu_\ell^{(m)} - \bar{\nu}_\ell\big)\right\|^2$$

$$\leq I\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1} \left\|\nu_\ell^{(m)} - \bar{\nu}_\ell\right\|^2$$

This completes the proof. □

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

$$\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left\|\big(\nu_t^{(m)} - \bar{\nu}_t\big)\right\|^2 \leq \frac{4\hat{L}^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2 + \frac{12\bar{L}^2 I\eta^2}{M}\sum_{m=1}^{M}\sum_{\ell=\bar{t}_{s-1}}^{t-1} \mathbb{E}\left\|\nu_\ell^{(m)} - \bar{\nu}_\ell\right\|^2$$

$$+ 8\zeta^2 + 4G_1^2 + \frac{4G_2^2}{b_x}$$

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

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

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

$$\begin{aligned}
\mathbb{E}\left\|\big(\nu_t^{(m)} - \bar{\nu}_t\big)\right\|^2 &\overset{(a)}{\leq} 2\mathbb{E}\left\|\big(\nu_t^{(m)} - \nabla h^{(m)}(x_t^{(m)})\big) - \big(\bar{\nu}_t - \frac{1}{M}\sum_{m=1}^{M}\nabla h^{(j)}(x_t^{(j)})\big)\right\|^2 \\
&\quad + 2\mathbb{E}\left\|\big(\nabla h^{(m)}(x_t^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_t^{(j)})\big)\right\|^2 \\
&\overset{(b)}{\leq} 2\mathbb{E}\left\|\nu_t^{(m)} - \nabla h^{(m)}(x_t^{(m)})\right\|^2 + 2\mathbb{E}\left\|\nabla h^{(m)}(x_t^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_t^{(j)})\right\|^2 \\
&\leq 4\hat{L}^2\mathbb{E}\left\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\right\|^2 + 4G_1^2 + \frac{4G_2^2}{b_x} + 2\mathbb{E}\left\|\nabla h^{(m)}(x_t^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_t^{(j)})\right\|^2
\end{aligned} \tag{49}$$

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:

$$\begin{aligned}
&\sum_{m=1}^{M}\left\|\nabla h^{(m)}(x_t^{(m)}) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_t^{(j)})\right\|^2 \\
&\overset{(a)}{\leq} 2\sum_{m=1}^{M}\left\|\nabla h^{(m)}(x_t^{(m)}) - \nabla h^{(m)}(\bar{x}_t)\right\|^2 + 4M\left\|\nabla h(\bar{x}_t) - \frac{1}{M}\sum_{j=1}^{M}\nabla h^{(j)}(x_t^{(j)})\right\|^2 \\
&\quad + 4\sum_{m=1}^{M}\left\|\nabla h^{(m)}(\bar{x}_t) - \nabla h(\bar{x}_t)\right\|^2 \overset{(b)}{\leq} 6\bar{L}^2\sum_{m=1}^{M}\left\|x_t^{(m)} - \bar{x}_t\right\|^2 + 4M\zeta^2
\end{aligned} \tag{50}$$

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 50 back to 49, we get the results in the lemma. □

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

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

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 $B_t = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2$. Based on Lemma D.15, we have:

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

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

$$\begin{aligned}
\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s - 1} D_t &\leq 4\hat{L}^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s - 1} B_t + 12\bar{L}^2 I\eta^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s - 1} \sum_{\ell=\bar{t}_{s-1}}^{t-1} D_\ell + 8(I-1)\zeta^2 + 4(I-1)G^2 \\
&\leq 4\hat{L}^2 \sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s - 1} B_t + 12\bar{L}^2 I^2\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{\bar{t}_s - 1} D_\ell + 8(I-1)\zeta^2 + 4(I-1)G_1^2 + \frac{4(I-1)G_2^2}{b_x}
\end{aligned}$$

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:

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

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:*

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

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

$$\begin{aligned}
\mathbb{E}\big\|\nabla h(\bar{x}_t) - \mathbb{E}_\xi[\bar{\nu}_t]\big\|^2 &\overset{(a)}{\leq} \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\big\|\mathbb{E}_\xi[\nu_t^{(m)}] - \nabla h^{(m)}(\bar{x}_t)\big\|^2 \\
&\leq \frac{2}{M}\sum_{m=1}^{M} \mathbb{E}\big[\big\|\mathbb{E}_\xi[\nu_t^{(m)}] - \mu_t^{(m)}\big\|^2 + \big\|\mu_t^{(m)} - \nabla h^{(m)}(\bar{x}_t)\big\|^2\big] \\
&\overset{(b)}{\leq} \frac{\hat{L}^2}{M}\sum_{m=1}^{M} \big(\mathbb{E}\big\|x_t^{(m)} - \bar{x}_t\big\|^2 + \mathbb{E}\big\|y_t^{(m)} - y_{\bar{x}_t}^{(m)}\big\|^2\big) + 2G_1^2 \\
&\leq \frac{2\hat{L}^2}{M}\sum_{m=1}^{M} \big(\mathbb{E}\big\|x_t^{(m)} - \bar{x}_t\big\|^2 + \mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)} + y_{x_t^{(m)}}^{(m)} - y_{\bar{x}_t}^{(m)}\big\|^2\big) + 2G_1^2 \\
&\leq \frac{2\hat{L}^2}{M}\sum_{m=1}^{M} \big(\big(1 + 2\kappa^2\big)\mathbb{E}\big\|x_t^{(m)} - \bar{x}_t\big\|^2 + 2\mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2\big) + +2G_1^2
\end{aligned}$$

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:*

$$\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4}\mathbb{E}\big\|\mathbb{E}_\xi[\bar{\nu}_t]\big\|^2 + \frac{\eta^2\bar{L}G_2^2}{2b_xM} + \eta G_1^2 + \frac{\eta\hat{L}^2}{M}\sum_{m=1}^{M}\Big(4\kappa^2 I\eta^2 \sum_{\ell=\bar{t}_{s-1}}^{t-1}\mathbb{E}\big\|\nu_\ell^{(m)} - \bar{\nu}_\ell\big\|^2 + 2\mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2\Big)$$

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

$$\mathbb{E}[h(\bar{x}_{t+1})] \leq \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4}\mathbb{E}\big\|\mathbb{E}_\xi[\bar{\nu}_t]\big\|^2 + \frac{\eta^2\bar{L}G_2^2}{2b_xM} + \eta G_1^2 + \frac{2\eta\hat{L}^2}{M}\sum_{m=1}^{M}\mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2$$

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

Proof. Using the smoothness of f we have:

$$\begin{aligned}
\mathbb{E}[h(\bar{x}_{t+1})] &\leq \mathbb{E}[h(\bar{x}_t)] + \mathbb{E}\langle\nabla h(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{\bar{L}}{2}\mathbb{E}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&\overset{(a)}{=} \mathbb{E}[h(\bar{x}_t)] - \eta\mathbb{E}\langle\nabla h(\bar{x}_t), \mathbb{E}_\xi[\bar{\nu}_t]\rangle + \frac{\eta^2\bar{L}}{2}\mathbb{E}\|\mathbb{E}_\xi[\bar{\nu}_t]\|^2 + \frac{\eta^2\bar{L}G_2^2}{2b_xM} \\
&\overset{(b)}{=} \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 + \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t) - \mathbb{E}_\xi[\bar{\nu}_t]\|^2 - \left(\frac{\eta}{2} - \frac{\eta^2\bar{L}}{2}\right)\mathbb{E}\|\mathbb{E}_\xi[\bar{\nu}_t]\|^2 + \frac{\eta^2\bar{L}G_2^2}{2b_xM} \\
&\overset{(c)}{\leq} \mathbb{E}[h(\bar{x}_t)] - \frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4}\mathbb{E}\big\|\mathbb{E}_\xi[\nu_t^{(m)}]\big\|^2 + \frac{\eta^2\bar{L}G_2^2}{2b_xM} + \eta G_1^2 \\
&\quad + \frac{\eta\hat{L}^2}{M}\sum_{m=1}^{M}\Big(\underbrace{4\kappa^2\mathbb{E}\big\|x_t^{(m)} - \bar{x}_t\big\|^2}_{T_1} + 2\mathbb{E}\big\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\big\|^2\Big)
\end{aligned}$$

where equality (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 $\eta < 1/2\bar{L}$. 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 $\mathcal{G}(t)$:

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

Theorem D.19. *Suppose we have constant $\bar{\eta} = \min\big(\frac{1}{2C_1^{1/2}}, \frac{\mu\gamma}{12\kappa\hat{L}}, \frac{1}{2\bar{L}}, \frac{1}{6I\hat{L}}\big)$, if we choose $\eta = \min\big(\bar{\eta}, \big(\frac{2\Delta}{C_\eta T}\big)^{1/3}\big)$ and $\gamma = \frac{1}{2L}$, we have:*

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

To reach an ϵ 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 $Q = O(\kappa\log(\frac{\kappa}{\epsilon}))$ in Eq. 11, and $T = O(\kappa^8\epsilon^{-1.5})$ number of iterations.

Proof. Similar to Lemma D.16, we denote $D_t = \frac{1}{M}\sum_{m=1}^{M}\|\left(\nu_t^{(m)} - \bar{\nu}_t\right)\|^2$, $B_t = \frac{1}{M}\sum_{m=1}^{M}\|y_t^{(m)} - y_{x_t^{(m)}}^{(m)}\|^2$, additionally, we denote $E_t = \|\mathbb{E}_\xi[\bar{\nu}_t]\|^2$. First, by Lemma D.13, when $t \neq \bar{t}_s$, by the triangle inequality, we have:

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

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

$$B_t - B_{t-1} \leq -\frac{\mu\gamma}{2}B_{t-1} + \frac{20\kappa^2\eta^2}{\mu\gamma}D_{t-1} + \frac{20\kappa^2\eta^2}{\mu\gamma}E_{t-1} + \frac{10\kappa^2 I\eta^2}{\mu\gamma}\sum_{\ell=\bar{t}_{s-1}}^{t-1}D_\ell + \frac{10\kappa^2\eta^2G_2^2}{\mu\gamma b_x M} + \frac{3\gamma^2\sigma^2}{b_y}$$

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

$$B_{\bar{t}_s} - B_{\bar{t}_{s-1}} \leq -\frac{\mu\gamma}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t + \frac{40\kappa^2 I\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}D_t + \frac{20\kappa^2\eta^2}{\mu\gamma}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}E_t + \frac{10I\kappa^2\eta^2G_2^2}{\mu\gamma b_x M} + \frac{3I\gamma^2\sigma^2}{b_y} \tag{51}$$

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

$$\mathbb{E}[h(\bar{x}_{t+1})] - \mathbb{E}[h(\bar{x}_t)] \leq -\frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{\eta}{4}E_t + 4\kappa^2\hat{L}^2 I\eta^3\sum_{\ell=\bar{t}_{s-1}}^{t-1}D_l + 2\eta\hat{L}^2B_t + \frac{\eta^2\bar{L}G_2^2}{2b_x M} + \eta G_1^2$$

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

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

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

$$\begin{aligned}
\mathbb{E}[h(\bar{x}_{\bar{t}_s})] - \mathbb{E}[h(\bar{x}_{\bar{t}_{s-1}})] \leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta}{4}E_t + 4\kappa^2\hat{L}^2 I\eta^3\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s-1}\sum_{\ell=\bar{t}_{s-1}}^{t-1}D_l \\
& + \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}2\hat{L}^2\eta B_t + \frac{I\eta^2\bar{L}G_2^2}{2b_x M} + I\eta G_1^2 \\
\leq & -\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta}{2}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\frac{\eta}{4}E_t + 4\kappa^2\hat{L}^2 I^2\eta^3\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}D_l \\
& + 2\hat{L}^2\eta\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t + \frac{I\eta^2\bar{L}G_2^2}{2b_x M} + I\eta G_1^2
\end{aligned} \tag{52}$$

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:

$$\begin{aligned}
\mathcal{G}_{\bar{t}_s} - \mathcal{G}_{\bar{t}_{s-1}} \leq & -\frac{\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{5\eta\hat{L}^2}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t - \frac{\eta}{4}\left(1 - \frac{720\kappa^2\eta^2\hat{L}^2}{\mu^2\gamma^2}\right)\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}E_t \\
& + \left(\frac{360}{\mu^2\gamma^2} + 4I\right)\eta^3\kappa^2\hat{L}^2 I\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}D_t + \frac{27I\hat{L}^2\gamma\eta\sigma^2}{b_y\mu} \\
& + \frac{90I\kappa^2\hat{L}^2\eta^3G_2^2}{\mu^2\gamma^2 b_x M} + \frac{I\eta^2\bar{L}G_2^2}{2b_x M} + I\eta G_1^2
\end{aligned}$$

to bound the coefficients above, we choose $\eta \leq \frac{\mu\gamma}{48\kappa\hat{L}}$. Then we have:

$$\begin{aligned}\mathcal{G}_{\bar{t}_s} - \mathcal{G}_{\bar{t}_{s-1}} \leq & -\frac{\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{5\eta\hat{L}^2}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t - \frac{\eta}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}E_t \\ & + \left(\frac{360}{\mu^2\gamma^2} + 4I\right)\eta^3\kappa^2\hat{L}^2 I\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}D_t + \frac{27I\hat{L}^2\gamma\eta\sigma^2}{b_y\mu} \\ & + \frac{90I\kappa^2\hat{L}^2\eta^3G_2^2}{\mu^2\gamma^2 b_x M} + \frac{I\eta^2\bar{L}G_2^2}{2b_x M} + I\eta G_1^2\end{aligned}$$

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

$$\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}D_t \leq 6\hat{L}^2\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t + 18I\zeta^2 + 6IG_1^2 + \frac{6IG_2^2}{b_x}$$

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

$$\begin{aligned}\mathcal{G}_{\bar{t}_s} - \mathcal{G}_{\bar{t}_{s-1}} \leq & -\frac{\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 - \frac{5\eta\hat{L}^2}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}B_t - \frac{\eta}{8}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}E_t \\ & + C_1 I\eta^3\left(6\hat{L}^2\sum_{t=\bar{t}_{s-1}+1}^{\bar{t}_s}B_t + 18I\zeta^2 + 6IG_1^2 + \frac{6IG_2^2}{b_x}\right) \\ & + \frac{27I\hat{L}^2\gamma\eta\sigma^2}{b_y\mu} + \frac{5I\eta G_2^2}{8b_x M} + \frac{I\eta^2\bar{L}G_2^2}{2b_x} + I\eta G_1^2 \\ \leq & -\frac{\eta}{2}\sum_{t=\bar{t}_{s-1}}^{\bar{t}_s-1}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 + 18C_1 I\hat{L}^2\eta^3\zeta^2 + 6C_1 I\hat{L}^2\eta^3G_1^2 + \frac{6C_1 I\hat{L}^2\eta^3G_2^2}{b_x} + \frac{5I\eta G_2^2}{8b_x M} \\ & + \frac{27I\hat{L}^2\gamma\eta\sigma^2}{b_y\mu} + \frac{I\eta^2\bar{L}G_2^2}{2b_x M} + I\eta G_1^2\end{aligned}$$

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

$$\begin{aligned}\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq & \mathcal{G}_1 - \mathcal{G}_T + 18C_1 T\hat{L}^2\eta^3\zeta^2 + 6C_1 T\hat{L}^2\eta^3G_1^2 + \frac{6C_1 T\hat{L}^2\eta^3G_2^2}{b_x} \\ & + \frac{5T\eta G_2^2}{8b_x M} + \frac{27T\hat{L}^2\gamma\eta\sigma^2}{b_y\mu} + \frac{T\eta^2\bar{L}G_2^2}{2b_x M} + T\eta G_1^2 \\ \leq & \Delta + \frac{9\eta\hat{L}^2\Delta_y}{\mu\gamma} + 18C_1 T\hat{L}^2\eta^3\zeta^2 + 6C_1 T\hat{L}^2\eta^3G_1^2 + \frac{6C_1 T\hat{L}^2\eta^3G_2^2}{b_x} \\ & + \frac{5T\eta G_2^2}{8b_x M} + \frac{27T\hat{L}^2\gamma\eta\sigma^2}{b_y\mu} + \frac{T\eta^2\bar{L}G_2^2}{2b_x M} + T\eta G_1^2\end{aligned}$$

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

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq & \underbrace{\frac{2\Delta}{\eta T} + \frac{\eta\bar{L}G_2^2}{2b_x M} + \left(36C_1\hat{L}^2\zeta^2 + 12C_1\hat{L}^2G_1^2 + \frac{12C_1\hat{L}^2G_2^2}{b_x}\right)\eta^2}_{T_1} \\ & + \underbrace{\frac{18\hat{L}^2\Delta_y}{\mu\gamma T} + \frac{54\hat{L}^2\gamma\sigma^2}{b_y\mu}}_{T_2} + \underbrace{\frac{5G_2^2}{4b_x M} + G_1^2}_{T_3}\end{aligned}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq \frac{2\Delta}{\eta T} + \left(36C_1\hat{L}^2\zeta^2 + 12C_1\hat{L}^2G_1^2 + \frac{12C_1\hat{L}^2G_2^2}{b_x}\right)\eta^2 + \frac{18\hat{L}^2\Delta_y}{\mu\gamma T} + \frac{54\hat{L}^2\gamma\sigma^2}{b_y\mu} + \frac{G_2^2}{4b_xM} + \frac{5G_2^2}{4b_xM} + G_1^2 \tag{53}$$

Next, we denote constant $\bar{\eta} = \min\left(\frac{1}{2C_1^{1/2}}, \frac{\mu\gamma}{12\kappa\hat{L}}, \frac{1}{2\hat{L}}, \frac{1}{6I\hat{L}}\right)$ and $C_\eta = \left(36C_1\hat{L}^2\zeta^2 + 12C_1\hat{L}^2G_1^2 + \frac{12C_1\hat{L}^2G_2^2}{b_x}\right)$ we choose

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla h(\bar{x}_t)\|^2 \leq \frac{2\Delta}{\bar{\eta}T} + \frac{36\kappa\hat{L}^2\Delta_y}{T} + \left(\frac{4C_\eta\Delta^2}{T^2}\right)^{1/3} + \frac{27\hat{L}^2\sigma^2}{\mu b_y L} + \frac{3G_2^2}{2b_xM} + 2G_1^2$$

Finally, since $\hat{L} = O(\kappa^2)$, $\bar{L} = O(\kappa^3)$ and and $\mu\gamma = O(\kappa^{-1})$. Suppose we choose $I = O(1)$, then $\bar{\eta} = O(\kappa^{-4})$ and $C_1 = O(\kappa^8)$, $\zeta = O(\kappa^2)$, $C_\eta = O(\kappa^{16})$, thus, we have

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

and to reach an ϵ 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 $Q = O(\kappa\log(\frac{\kappa}{\epsilon}))$ 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. $||x_i + x_j||^2 \leq (1+a)||x_i||^2 + (1+\frac{1}{a})||x_j||^2$ *for any $a > 0$, and*
2. $||\sum_{k=1}^{K} x_k||^2 \leq K\sum_{k=1}^{K}||x_k||^2$

Proposition 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*

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

Proposition E.4. *Suppose we have function $g(y)$, which is L-smooth and μ-strongly-convex, then suppose $\gamma < \frac{1}{L}$, 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 σ^2.

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

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

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

$$\begin{aligned} g(y^*) &\geq \mathbb{E} g(y_{t+1}) + \mathbb{E}\langle \nabla_y g(y_t), y^* - y_{t+1} \rangle + \frac{\mu}{2}\|y^* - y_t\|^2 - \frac{L}{2}\mathbb{E}\|y_{t+1} - y_t\|^2 \\ &\geq \mathbb{E} g(y_{t+1}) + \gamma \|\nabla_y g(y_t)\|^2 + \langle \nabla_y g(y_t), y^* - y_t \rangle + \frac{\mu}{2}\|y^* - y_t\|^2 - \frac{L\gamma^2}{2}\mathbb{E}\|\nabla_y g(y_t, \xi)\|^2 \\ &\geq \mathbb{E} g(y_{t+1}) + \gamma \|\nabla_y g(y_t)\|^2 + \langle \nabla_y g(y_t), y^* - y_t \rangle + \frac{\mu}{2}\|y^* - y_t\|^2 - \frac{L\gamma^2}{2}\mathbb{E}\|\nabla_y g(y_t)\|^2 - \frac{L\gamma^2\sigma^2}{2} \\ &\geq \mathbb{E} g(y_{t+1}) + \langle \nabla_y g(y_t), y^* - y_t \rangle + \frac{\mu}{2}\|y^* - y_t\|^2 + \left(\gamma - \frac{L\gamma^2}{2}\right)\|\nabla_y g(y_t)\|^2 - \frac{L\gamma^2\sigma^2}{2} \end{aligned}$$

By definition of y^*, we have $g(y^*) \geq g(y_{t+1})$. Thus, we obtain

$$0 \geq \langle \nabla_y g(y_t), y^* - y_t \rangle + \frac{\mu}{2}\|y^* - y_t\|^2 + \left(\gamma - \frac{L\gamma^2}{2}\right)\|\nabla_y g(y_t)\|^2 - \frac{L\gamma^2\sigma^2}{2}$$

By $y_{t+1} = y_t - \gamma \nabla_y g(y_t, \xi)$, we have:

$$\begin{aligned} \mathbb{E}\|y_{t+1} - y^*\|^2 &= \mathbb{E}\|y_t - \gamma \nabla_y g(y_t, \xi) - y^*\|^2 = \|y_t - y^*\|^2 - 2\gamma\langle \nabla_y g(y_t), y_t - y^* \rangle + \gamma^2 \mathbb{E}\|\nabla_y g(y_t, \xi)\|^2 \\ &\leq \big(1 - \mu\gamma\big)\|y_t - y^*\|^2 - 2\gamma\big(\gamma - \frac{L\gamma^2}{2} - \frac{\gamma}{2}\big)\|\nabla_y g(y_t)\|^2 + \big(L\gamma^3 + \gamma^2\big)\sigma^2 \end{aligned}$$

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

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

This completes the proof. □

Proposition E.5. *Suppose we have function $g(y)$, which is L-smooth and μ-strongly-convex, then suppose $\gamma < \frac{1}{2L}$ and $\alpha_t < 1$, the progress made by one step of gradient descent is:*

$$\begin{aligned} \|y_{t+1} - y^*\|^2 \leq (1 - \frac{\mu\gamma\alpha_t}{2})\|y_t - y^*\|^2 - \frac{\gamma^2\alpha_t}{4}\|\omega_t\|^2 \\ + \frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(x_t, y_t) - \mathbb{E}[w_t]\|^2 + \frac{3\gamma^2\alpha_t}{2} Var[\omega_t]. \end{aligned}$$

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 $g(y)$, we have:

$$\begin{aligned} g(y^*) &\geq g(y_t) + \langle \nabla_y g(y_t), y^* - y_t \rangle + \frac{\mu}{2}\|y^* - y_t\|^2 \\ &= g(y_t) + \mathbb{E}\langle \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1} \rangle \\ &\quad - \gamma\langle \nabla_y g(y_t), \mathbb{E}[w_t] \rangle + \frac{\mu}{2}\|y^* - y_t\|^2 \end{aligned} \tag{54}$$

where the expectation is *w.r.t* the stochasity of ω_t. Then by L-smoothness, we have:

$$\frac{L}{2}\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 \geq \mathbb{E} g(\tilde{y}_{t+1}) - g(y_t) + \gamma\langle \nabla_y g(y_t), \mathbb{E}[\omega_t] \rangle \tag{55}$$

Combining the 54 with 55, we have

$$\begin{aligned}
g(y^*) &\geq \mathbb{E}g(\tilde{y}_{t+1}) + \mathbb{E}\langle \mathbb{E}[w_t], y^* - \tilde{y}_{t+1}\rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1}\rangle \\
&\quad + \frac{\mu}{2}\|y^* - y_t\|^2 - \frac{L}{2}\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 \\
&\geq \mathbb{E}g(\tilde{y}_{t+1}) + \gamma\|\mathbb{E}[w_t]\|^2 + \langle \mathbb{E}[w_t], y^* - y_t\rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1}\rangle \\
&\quad + \frac{\mu}{2}\|y^* - y_t\|^2 - \frac{L\gamma^2}{2}\mathbb{E}\|\omega_t\|^2 \\
&\geq \mathbb{E}g(\tilde{y}_{t+1}) + \langle \mathbb{E}[w_t], y^* - y_t\rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1}\rangle \\
&\quad + \frac{\mu}{2}\|y^* - y_t\|^2 + \left(\gamma - \frac{L\gamma^2}{2}\right)\|\mathbb{E}[\omega_t]\|^2 - \frac{L\gamma^2}{2}Var[\omega_t]
\end{aligned}$$

where Var denotes the variance. By definition of y^*, we have $g(y^*) \geq g(\tilde{y}_{t+1})$. Thus, we obtain

$$\begin{aligned}
0 &\geq \langle \mathbb{E}[w_t], y^* - y_t\rangle + \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1}\rangle \\
&\quad + \frac{\mu}{2}\|y^* - y_t\|^2 + \left(\gamma - \frac{L\gamma^2}{2}\right)\|\mathbb{E}[\omega_t]\|^2 - \frac{L\gamma^2}{2}Var[\omega_t]
\end{aligned} \tag{56}$$

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

$$\begin{aligned}
&- \mathbb{E}\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - \tilde{y}_{t+1}\rangle \\
&= -\langle \nabla_y g(y_t) - \mathbb{E}[w_t], y^* - y_t\rangle + \langle \nabla_y g(y_t) - \mathbb{E}[w_t], \mathbb{E}[\omega_t]\rangle \\
&\leq \frac{1}{\mu}\|\nabla_y g(y_t) - \mathbb{E}[w_t]\|^2 + \frac{\mu}{4}\|y^* - y_t\|^2] + \frac{1}{\mu}\|\nabla_y g(y_t) - \mathbb{E}[w_t]\|^2 + \frac{\mu\gamma^2}{4}\|\mathbb{E}[\omega_t]\|^2 \\
&= \frac{2}{\mu}\|\nabla_y g(y_t) - \mathbb{E}[w_t]\|^2 + \frac{\mu}{4}\|y^* - y_t\|^2 + \frac{\mu\gamma^2}{4}\|\mathbb{E}[\omega_t]\|^2.
\end{aligned}$$

Combining with Eq. 56:

$$0 \geq \langle \mathbb{E}[w_t], y^* - y_t\rangle - \frac{2}{\mu}\|\nabla_y g(y_t) - \mathbb{E}[w_t]\|^2 + \big(\gamma - \frac{3L\gamma^2}{4}\big)\|\mathbb{E}[w_t]\|^2 + \frac{\mu}{4}\|y^* - y_t\|^2 - \frac{L\gamma^2}{2}Var[\omega_t]$$

By $y_{t+1} = y_t - \gamma\alpha_t\omega_t$, we have:

$$\begin{aligned}
\mathbb{E}\|y_{t+1} - y^*\|^2 = \mathbb{E}\|y_t - \gamma\alpha_t\omega_t - y^*\|^2 &= \|y_t - y^*\|^2 - 2\gamma\alpha_t\langle \mathbb{E}[\omega_t], y_t - y^*\rangle + \gamma^2\alpha_t^2\mathbb{E}[\|\omega_t\|^2] \\
&\leq \big(1 - \frac{\mu\gamma\alpha_t}{2}\big)\|y_t - y^*\|^2 - 2\gamma\alpha_t\big(\gamma - \frac{\gamma\alpha_t}{2} - \frac{3L\gamma^2}{4}\big)\|\mathbb{E}[\omega_t]\|^2 \\
&\quad + \frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(y_t) - \mathbb{E}[w_t]\|^2 + (L\gamma^3\alpha_t + \gamma^2\alpha_t^2)Var[\omega_t]
\end{aligned}$$

Then since we choose $\gamma < \frac{1}{2L}$, $\alpha_t < 1$, we obtain:

$$\begin{aligned}
\mathbb{E}\|y_{t+1} - y^*\|^2 \leq (1 - \frac{\mu\gamma\alpha_t}{2})\|y^* - y_t\|^2 &- \frac{\gamma^2\alpha_t}{4}\|\mathbb{E}[\omega_t]\|^2 \\
&+ \frac{4\gamma\alpha_t}{\mu}\|\nabla_y g(x_t, y_t) - \mathbb{E}[w_t]\|^2 + \frac{3\gamma^2\alpha_t}{2}Var[\omega_t].
\end{aligned}$$

This completes the proof. □